Geoderma 361 (2020) 114061

Contents lists available at ScienceDirect

Geoderma

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

High-resolution and three-dimensional mapping of soil texture of China

Feng Liu[a], Gan-Lin Zhang[a,b,*], Xiaodong Song[a], Decheng Li[a], Yuguo Zhao[a,b], Jinling Yang[a], Huayong Wu[a], Fei Yang[a]

[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

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Digital soil mapping
Machine learning
Large extent
Environmental factors
Uncertainty

ABSTRACT

The lack of detailed three-dimensional soil texture information largely restricts many applications in agriculture, hydrology, climate, ecology and environment. This study predicted 90 m resolution spatial variations of sand, silt and clay contents at a national extent across China and at multiple depths 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm. We used 4579 soil profiles collected from a national soil series inventory conducted recently and currently available environmental covariates. The covariates characterized environmental factors including climate, parent materials, terrain, vegetation and soil conditions. We constructed random forest models and employed a parallel computing strategy for the predictions of soil texture fractions based on its relationship with the environmental factors. Quantile regression forest was used to estimate the uncertainty of the predictions. Results showed that the predicted maps were much more accurate and detailed than the conventional linkage maps and the SoilGrids250m product, and could well represent spatial variation of soil texture across China. The relative accuracy improvement was around 245–370% relative to the linkage maps and 83–112% relative to the SoilGrids250m product with regard to the R^2, and it was around 24–26% and 14–19% respectively with regard to the RMSE. The wide range between 5% lower and 95% upper prediction limits may suggest that there was a substantial room to improve current predictions. Besides, we found that climate and terrain factors are major controllers for spatial patterns of soil texture in China. The heat and water-driven physical and chemical weathering and wind-driven erosion processes primarily shape the pattern of clay content. The terrain, wind and water-driven deposition, erosion and transportation sorting processes of soil particles primarily shape the pattern of silt. The findings provide clues for modeling future soil evolution and for national soil security management under the background of global and regional environmental changes.

1. Introduction

Soil texture is an important soil property that controls most physical, chemical and biological processes in soils. It can influence soil thermal capacity, permeability, water holding capacity and solute movement which are closely associated with applications of climate, ecological, hydrological modelling, smart agricultural management and soil pollution control. It often has high spatial heterogeneity over regions and landscapes in both lateral and vertical dimensions. Currently, there is an increasing demand for detailed three-dimensional soil texture information in dealing with global and national issues such as climate change, soil degradation, water resource shortage, environmental pollution, agricultural and ecosystem sustainability (Sanchez et al., 2009; Montanarella and Vargas, 2012; McBratney et al., 2014).

The GlobalSoilMap project was proposed in 2006 and officially launched in 2009. Its aim is to make a new digital soil map of the world using state-of-the-art technologies for soil mapping and predicting soil properties at a 90 m resolution and six standard depth layers 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm (Arrouays et al., 2014). The soil attributes to be mapped include sand, silt and clay contents, coarse fragments, bulk density, organic carbon, pH, cation exchange capacity, available water capacity, electrical conductivity and soil depth. Currently, only several countries have attempted to create three-dimensional national soil information using digital soil mapping methods at a fine resolution (Arrouays et al., 2017). Adhikari et al. (2013) predicted spatial distribution of soil texture fractions of Denmark at the six depth layers at 30 m resolution using the Cubist

Abbreviations: DEM, digital elevation model; TWI, topographic wetness index; ETM+, Enhanced Thematic Mapper Plus; MODIS, Moderate Resolution Imaging Spectrometer; MAT, annual mean temperature; MAP, annual precipitation; NDVI, normalized difference vegetation index; NDWI, mean nomalized difference water index; LST, land surface temperature; R^2, coefficient of determination; RMSE, root mean square error; ME, mean error; CCC, concordance correlation coefficient; RI, relative improvement; CV, coefficient of variation; SD, standard deviation

* Corresponding author at: State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China.
E-mail address: glzhang@issas.ac.cn (G.-L. Zhang).

https://doi.org/10.1016/j.geoderma.2019.114061
Received 27 June 2019; Received in revised form 24 October 2019; Accepted 29 October 2019
Available online 28 November 2019
0016-7061/

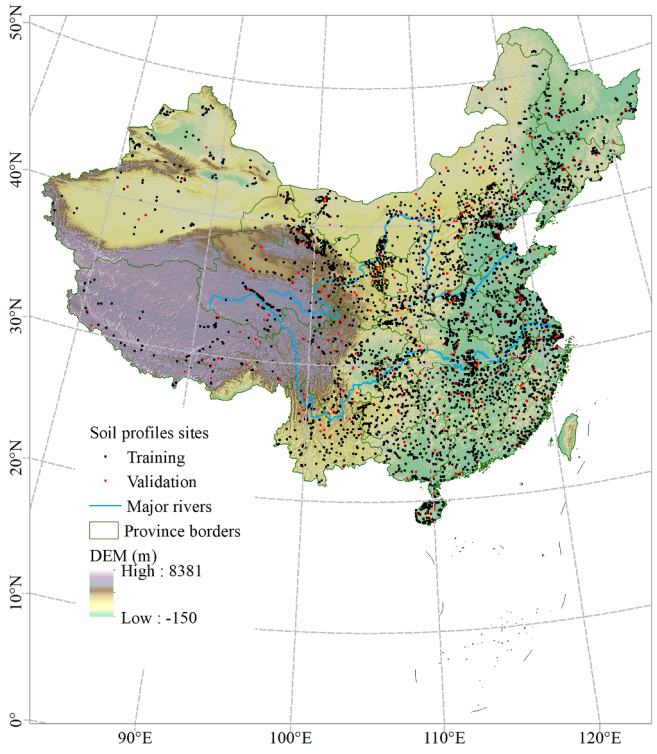

Fig. 1. Spatial distribution of soil profiles sites used in the study.

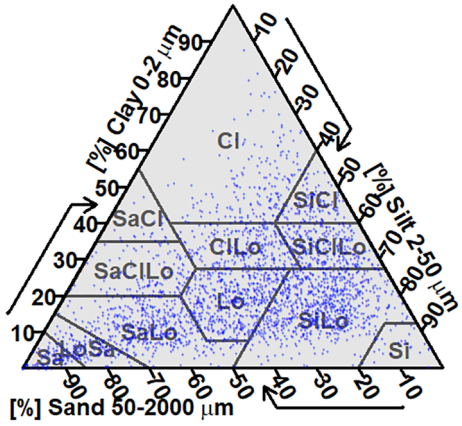

Fig. 2. Ternary plot of the sand, silt and clay contents for horizon observations.

decision tree algorithm. With the same algorithm, Viscarra Rossel et al. (2015) generated a set of 90 m resolution and Australia-wide maps of clay, silt and sand contents at the six layers. Mulder et al. (2016a) made 90 m resolution maps of the soil texture fractions across France. Padarian et al. (2017) modeled 100 m resolution national maps of the fractions of Chile using the classification and regression tree algorithm. Kempen et al. (2014) mapped clay content of the Netherlands using regression-kriging method. Ramcharan et al. (2018) generated complete coverage gridded predictions at 100 m spatial resolution of sand and clay contents for the conterminous United States. Besides, at global extent, Hengl et al. (2014) developed 1 km resolution soil grids for the soil texture fractions using a three-dimensional regression-kriging method. And they recently updated the maps to 250 m resolution using an ensemble of random forest and gradient boosting methods (Hengl et al., 2017a).

China has a large span in geographical extent, covering an area of 9.6 million km^2. Its soil landscapes are very diverse and complex. It is a challenge to accurately predict soil spatial variations with a limited number of soil survey sites across the country. So far, high-resolution and three dimensional predictive soil mapping at a national extent has not yet been attempted although there were some studies in watershed and field scales (Chen et al., 2013; Li et al., 2013; Liu et al., 2013, 2016; Yang et al., 2017). Shangguan et al. (2012) developed 1 km resolution maps of

Table 1
Environmental covariates for characterizing soil formative factors in the soil texture prediction (s, soil; c, climate; o, organisms; r, relief; p, parent material).

Variable	Description	Factors	Resolution
MAT	Annual mean temperature (°C)	c	1 km
diurnalRange	Mean diurnal range (°C)	c	1 km
tempSeason	Temperature seasonality (°C)	c	1 km
tempMax	Maximum temperature of warmest month (°C)	c	1 km
tempMin	Minimum temperature of coldest month (°C)	c	1 km
annualRange	Temperature annual range (°C)	c	1 km
MAP	Annual precipitation (mm)	c	1 km
precipSeason	Precipitation standard deviation (mm)	c	1 km
precipSummer	Precipitation of warmest quarter (mm)	c	1 km
solarR	Mean annual solar radiation (Jm^{-2}yr^{-1})	c	1 km
vaporPressure	Water vapor pressor (kpa)	c	1 km
windS	Wind speed (m/s)	c	1 km
elevation	Elevation above sea level (m)	r	90 m
slope	Slope gradient (%)	r	90 m
asp2n	Aspect angle distance from north (°)	r	90 m
curpln	Plan curvature	r	90 m
curprf	Profile curvature	r	90 m
TWI	topographic wetness index	r	90 m
posOpen	Positive terrain openness	r	90 m
negOpen	Negative terrain openness	r	90 m
topoExp	Topographic exposure to wind	r, p	90 m
regolithick	Regolith thickness	s	90 m
Band5	Surface reflectance at shortwave infrared (1.55–1.75um)	s	30 m
Band7	Surface reflectance at shortwave infrared (2.08–2.35um)	s, p	30 m
clayi	Clay mineral index	s, p	30 m
NDWI	Annual mean nomalized difference water index	s, c	30 m
NDVI	Mean NDVI during the growing season	o, c	30 m
ndviSeason	Standard deviation of NDVI over a year	o, c	250 m
LSTfm	Mean daytime LST of Feb & Mar (°C)	s, c	1 km
LSTam	Mean daytime LST of Apr & May (°C)	s, c	1 km
LSTjj	Mean daytime LST of Jun & Jul (°C)	s, c	1 km
LSTas	Mean daytime LST of Aug & Sep (°C)	s, c	1 km
LSTon	Mean daytime LST of Oct & Nov (°C)	s, c	1 km

Table 2
Statistical description of the splines-fitted sand, silt and clay percentages at different depths based on the 4121 training soil profiles.

Depth (cm)	Mean (%)	SD (%)	CV	Skewness	Kurtosis
Clay:					
0–5	19.97	12.38	0.62	0.94	1.17
5–15	20.10	12.32	0.61	0.90	1.05
15–30	20.66	12.82	0.62	0.86	0.82
30–60	21.31	13.80	0.65	0.87	0.77
60–100	21.79	14.56	0.67	0.87	0.69
100–200	21.71	15.28	0.70	0.79	0.41
Silt:					
0–5	41.98	19.50	0.46	−0.22	−0.66
5–15	41.92	19.37	0.46	−0.21	−0.69
15–30	41.81	19.61	0.47	−0.18	−0.76
30–60	41.28	20.04	0.49	−0.10	−0.80
60–100	40.73	20.46	0.50	−0.05	−0.86
100–200	40.47	21.34	0.53	−0.06	−0.92
Sand:					
0–5	38.02	24.65	0.65	0.68	−0.36
5–15	37.92	24.56	0.65	0.69	−0.35
15–30	37.40	24.92	0.67	0.71	−0.39
30–60	37.33	25.58	0.69	0.70	−0.46
60–100	37.37	26.21	0.70	0.67	−0.58
100–200	37.79	27.69	0.73	0.68	−0.69

Table 3
Mean and standard deviation of prediction performance of soil texture fractions based on 30 repeats of 10-fold cross validation with the 4121 training soil profiles.

Depth (cm)	R^2	CCC	RMSE	ME
Clay:				
0–5	0.45 (0.003)	0.60 (0.002)	9.23 (0.022)	0.22 (0.017)
5–15	0.46 (0.003)	0.61 (0.002)	9.06 (0.023)	0.21 (0.014)
15–30	0.46 (0.003)	0.61 (0.002)	9.44 (0.023)	0.23 (0.021)
30–60	0.46 (0.002)	0.61 (0.002)	10.18 (0.018)	0.25 (0.016)
60–100	0.43 (0.003)	0.59 (0.002)	10.97 (0.027)	0.27 (0.018)
100–200	0.43 (0.003)	0.59 (0.002)	11.53 (0.026)	0.25 (0.024)
Silt:				
0–5	0.48 (0.002)	0.63 (0.002)	14.09 (0.030)	−0.09 (0.025)
5–15	0.49 (0.003)	0.64 (0.002)	13.86 (0.031)	−0.09 (0.026)
15–30	0.48 (0.003)	0.63 (0.002)	14.20 (0.032)	−0.09 (0.020)
30–60	0.45 (0.003)	0.60 (0.002)	14.90 (0.035)	−0.09 (0.020)
60–100	0.44 (0.003)	0.59 (0.002)	15.32 (0.033)	−0.06 (0.028)
100–200	0.44 (0.003)	0.59 (0.002)	16.01 (0.033)	−0.11 (0.031)
Sand:				
0–5	0.49 (0.002)	0.64 (0.002)	17.54 (0.032)	0.28 (0.022)
5–15	0.50 (0.002)	0.65 (0.002)	17.35 (0.033)	0.29 (0.031)
15–30	0.48 (0.002)	0.63 (0.002)	17.91 (0.036)	0.34 (0.031)
30–60	0.46 (0.002)	0.61 (0.001)	18.80 (0.033)	0.38 (0.027)
60–100	0.45 (0.003)	0.60 (0.002)	19.52 (0.044)	0.41 (0.034)
100–200	0.45 (0.003)	0.61 (0.002)	20.46 (0.048)	0.54 (0.043)

R^2: coefficient of determination; CCC: concordance correlation coefficient; RMSE: root mean square error; ME: mean error.

Table 4
Prediction interval coverage probability (PICP) for clay, silt and sand contents and different depths based on the 458 random-held back soil profiles.

Depth (cm)	Clay PICP %	Silt PICP %	Sand PICP %
0–5	90.8	89.4	91.3
5–15	90.2	91.3	90.8
15–30	90.3	90.3	88.8
30–60	88.9	88.2	90.6
60–100	89.1	90.3	90.8
100–200	89.8	90.7	91.9

sand, silt and clay contents of China at multiple depths (different from those specified by the GlobalSoilMap project) through a conventional linkage method which linked legacy soil profiles with polygons of 1:1,000,000 scale soil type map. The soil profiles were originated from the Second National Soil Survey conducted three decades ago, which lacked of accurate georeference due to unavailable GPS system at that time. With the same legacy soil data, Chen et al. (2019) made a 90 resolution map of topsoil soil pH of China using random forest and gradient boosting methods, and Liang et al. (2019) made a map of topsoil soil organic matter content of China using the Cubist method.

Despite the efforts above, the gap between detailed and accurate soil data demands and availability is still large. Our understanding on soil variation over landscapes at horizontal and vertical dimensions is still limited (Heuvelink and Webster, 2001; Stockmann et al., 2013; Phillips, 2016). Meanwhile, current digital soil mapping methods are not perfect when applying them in large areas (Minasny and McBratney, 2010; Mulder et al., 2016b; Hengl et al., 2017a; Tifafi et al., 2018).

Therefore, the objective of the study was to conduct high-resolution and three-dimensional predictive mapping of soil texture across China, and to reveal the controlling environmental factors and processes for spatial pattern of soil texture at a national extent.

2. Material and methods

2.1. Soil data source

A total of 4579 soil profiles were used in this study (Fig. 1). They were obtained by a recent project of National Soil Series Survey and Compilation of Soil Series of China conducted from 2009 to 2019 (Zhang et al., 2013). The soil reconnaissance survey sampled all typical soil types according to Chinese Soil Taxonomy (Cooperative Research Group on Chinese Soil Taxonomy, 2001), which are so-called central concepts of all soil types down to soil series level. Typical soil profiles representing main soil-landscapes were collected. The geographical coordinates of each soil profile location were recorded with a GPS receiver. The soil pits were generally dug to a depth of 1.5–2 m or until a lithic or paralithic contact. All soil profiles were divided vertically into different pedological horizons according to specific profile morphology, and soil samples were collected from each horizon. In laboratory, samples were air-dried at room temperature and then passed through a 2 mm sieve. Basic soil properties including soil texture, soil organic carbon, bulk density, pH, cation exchange capacity, total nitrogen, total phosphorus, total potassium, available phosphorus and available potassium were measured. Among them, soil texture fractions, namely sand (2–0.05 mm), silt (0.05–0.002 mm) and clay (< 0.002 mm) percentages, were determined using the pipette method (Zhang and Gong, 2012; USDA-NRCS, 2004). Fig. 2 shows horizon observations of sand, silt and clay content in a soil texture triangle.

Fig. 1 shows that the northwestern and southwestern parts had obviously less soil profiles than other parts. The reasons had two aspects: one was poor accessibility in the west and limited funding for the soil survey, and another was relatively small soil spatial variation in the northwest. The northwestern part is an arid region, where there are widespread gobi, semi-deserts and deserts. The southwestern part is an alpine region (mainly the Qinghai-Tibet Plateau), where there are many high-relief mountains and extensive depopulated zones. The harsh environments and underdeveloped road networks lead to poor accessibility in the two parts. Out of all soil profiles, 458 (10%) were randomly selected as random-held back evaluation samples. They were used for evaluating the results of this study through comparisons with existing traditional soil texture maps and the SoilGrids250 maps, and also evaluating the uncertainty estimation of soil texture predictions. The remaining 4121 (90%) were used for model training and 10-fold cross validation (Fig. 1).

2.2. Environmental covariates

The SCORPAN (soil, climate, organisms, topography, parent material, age and space) concept (McBratney et al., 2003) provides a framework for the choice of environmental covariates. We selected the covariates associated with the formation, accumulation and transportation processes of soil particles and also those can reflect spatial difference of soil texture. Redundant covariates with high Pearson correlation coefficient values (i.e., 0.85) were removed in the selection. Table 1 lists the covariates used for spatial prediction of the soil texture fractions in this study. The climatic variables over 1970–2000 were obtained from the WorldClim database at 1 km resolution (Hijmans et al., 2005). Solar radiation reflects the intensity of solar heating on land surface. Wind speed mainly reflects soil erosion and evapotranspiration. MAT reflects mean status of air temperature while tempMAX and tempMin express two extreme status. DiurnalRange, tempSeason and annualRange represent changes of air temperature at diural, seasonal and annual scales. Primary and secondary terrain variables were derived from a 90 m digital elevation model (DEM) of the Shuttle Radar Topographic Mission (http://srtm.csi.cgiar.org/srtmdata/) using the SAGA GIS tool (http://www.saga-gis.org). Surficial geology was partly represented by Landsat8 ETM + Band7 (shortwave infrared at 2.08–2.35um) and clay mineral ratio (Band5/Band7, Drury, 1987). The Band7 designed for geologists has the potential to detect surficial lithology and minerals. Vegetation conditions were represented by a mean normalized difference vegetation index (NDVI) derived from ETM + observations during the growing season of 2017 and a NDVI standard deviation from MODIS (Moderate Resolution Imaging Spectrometer) observations over the same year. Land surface

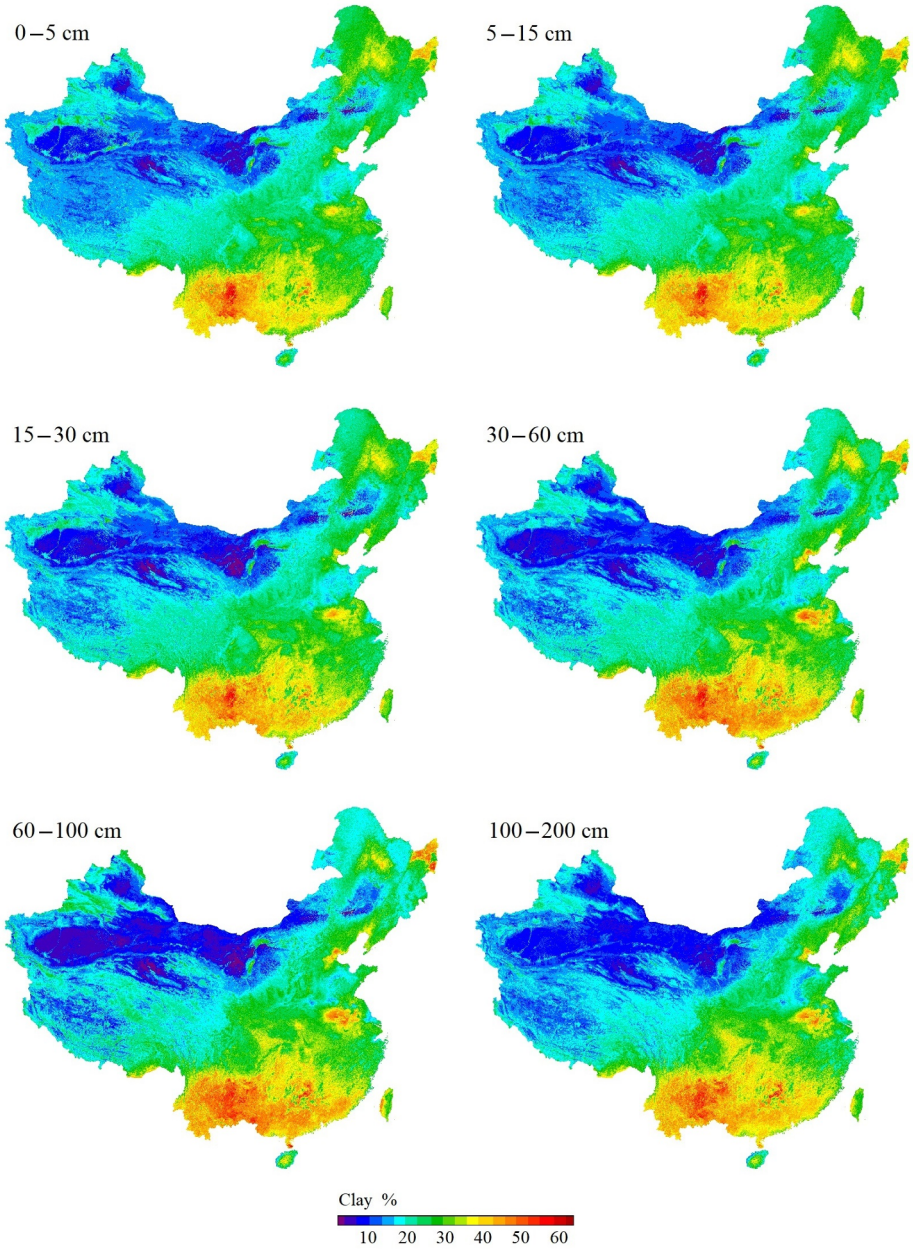

Fig. 3. The predicted maps of clay content at six depth intervals.

moisture conditions were represented by ETM+ shortwave infrared Band5 and Band7 and a normalized difference water index (NDWI) of the growing season (Gao, 1996). Land surface temperature conditions were represented by a set of double-month averages (Feb & Mar, Apr & May, Jun & Jul, Aug & Sep, Oct & Nov) of a time series of 8-day composite MODIS land surface temperature (LST) observations during the year of 2017 (http://modis.gsfc.nasa.gov). Regolith thickness (Shangguan et al., 2017) was used as a covariate because it can largely indicate the surface conditions of weathering and accumulation and erosion or removal. All data layers of the covariates were resampled to a raster cell size of 90 m.

2.3. Deriving sample data at predefined depths

For each profile of a soil texture fraction (sand, silt or clay percentages), we used equal-area quadratic splines to fit a continuous depth function to original horizon sample data. The splines are a set of local quadratic polynomials tied together with 'knots' located at horizon boundaries. It goes through each horizon, maintaining the average value of the soil attribute, and is linear between the horizons and quadratic within the horizons giving a linear-quadratic smoothing spline. The areas above and below the fitted curve in a horizon are equal (Ponce-Hernandez et al., 1986; Malone et al., 2009). A spline-smoothing parameter lamda (λ) controls the trade-off between fidelity and roughness penalty. In this study, its default value of 0.1 was adopted for the fittings of all soil texture fractions. From the fitted spline, we derived the mean values of a soil texture fraction within the predefined six standard depth layers 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm. The mean values were taken as the standardized sample data for the following spatial prediction of soil texture. The fittings were performed using the Spline Tool version 2.0 developed by the team of Australian Collaborative Land Evaluation Program (Jacquier and Seaton, 2012). The mathematical description of the equal-area quadratic splines can be found in Bishop et al. (1999).

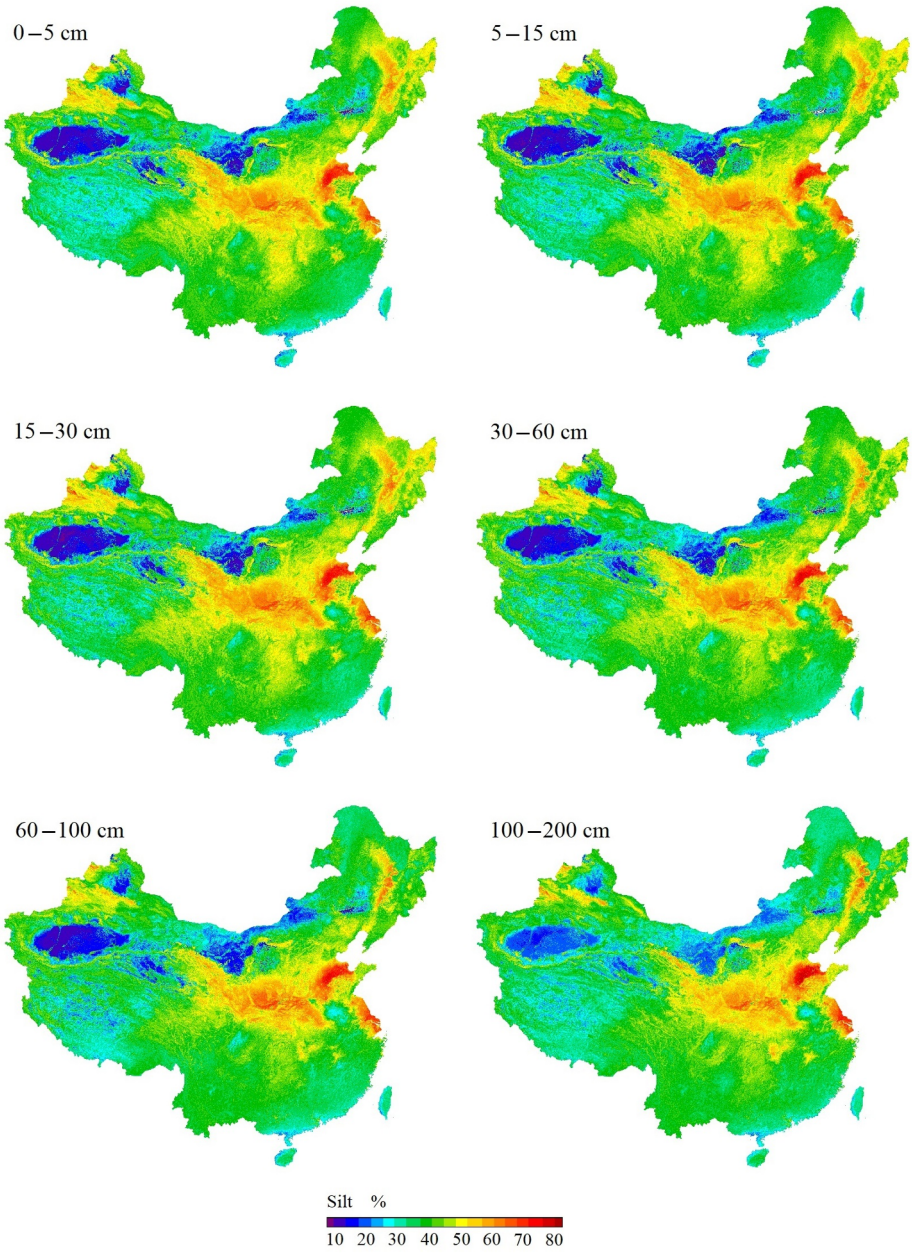

Fig. 4. The predicted maps of silt content at six depth intervals.

2.4. Predicting spatial variation of soil texture and estimating uncertainty

Based on the 4121 training soil profile sites, an ensemble machine learning algorithm, random forest (Breiman, 2001), was used to model the relationships between each soil texture fraction (sand, silt or clay percentage) and the environmental covariates at each depth interval. The model training was performed separately for sand, silt and clay percentages and depth by depth. This algorithm was chosen because it can deal with complex soil-environmental relationships and reduces overfitting problem of a single decision or regression tree model. It has been proven to be an excellent machine learning method currently available (Brungard et al., 2015; Hengl et al., 2015; Nussbaum et al., 2018). It generates multiple classification and regression trees, and all trees are grown to maximum size without pruning. Each tree is trained based on a random subset of the sample data (with replacement). A random subset of covariates is also chosen for the tree training. The use of bootstrap sampling in model training allows the remaining (out-of-bag) samples to be used for error estimation. Final predictions of random forest are an average of the predictions of individual trees. The algorithm also provides an estimation of relative importance of covariates based on the increase in mean square error (i.e., %IncMSE) when a covariate is randomly permuted. The bigger the %IncMSE value the more importance of the covariate is. There are four important parameters: number of variables used to grow each tree (mtry), number of trees to be grown in the forest (ntree), minimum number of terminal nodes (nodesize) and proportion of samples taken in a single tree. The mtry is related with the strength of each tree and correlations between trees, and increasing it can increase the strength of each tree and correlations between trees. The default values of the parameters are empirical values which are chosen based on a number of (if not many) data experiments with different datasets when developing the model (Liaw and Wiener, 2002; Svetnik et al., 2003). Svetnik et al. (2003), Diaz-Uriarte and de Andres (2006) and Grimm et al. (2008) suggested that the default of mtry is often a good choice. In this study, we made trials with three values of mtry: one third (default), one half and two third of the total number of covariates, but found that they had no obvious

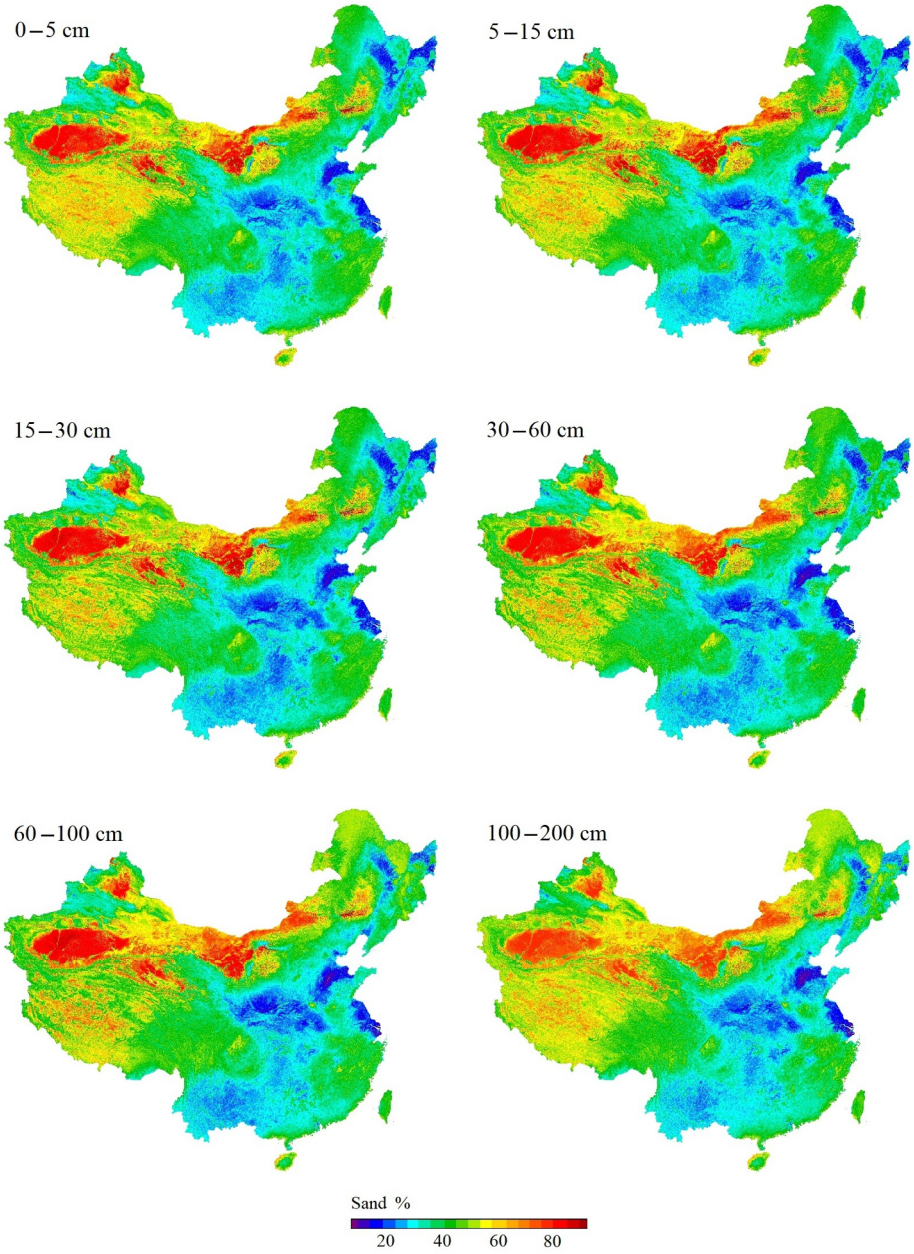

Fig. 5. The predicted maps of sand content at six depth intervals.

differences in the prediction accuracy. The default values of mtry and nodesize (i.e., 5) thus were used for all soil texture fractions and depths. The parameter ntree does not really need to be fine-tuned, its default value of 500 was used in this study because we tried different values and found the default value is sufficient to yield stable results (see supplement material). Then, with the spatially exhaustive environmental covariates, the trained random forest model of each soil texture fraction and depth interval was applied over space. We consequently generated 90 resolution national map of sand, silt and clay content at the depth layers 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm.

Uncertainty representation is a crucial aspect of digital soil mapping (Arrouays et al., 2014). Digital soil mapping models are not only expected to deliver accurate soil predictions at a given location but their suitability to deliver maps should encompass ability to predict how uncertain these predictions are (Vaysse and Lagacherie, 2017). The uncertainties of the predicted maps of clay, silt and sand contents were estimated using quantile regression forest (Vaysse and Lagacherie, 2017). The estimation produced maps of 0.05 and 0.95 quantiles for each fraction and depth. That is to say, for every pixel location of the study area and every depth, there were values of uncertainty for the predictions of clay, silt and sand percentages respectively.

The random forest modeling and mapping were implemented in the open source R environment (R Core Team, 2016) with the packages 'randomForest' (Liaw and Wiener, 2002), 'quantregForest' (Meinshausen and Schiesser, 2015), 'rgdal' (Keitt et al., 2009), 'raster' (Hijmans and van Etten, 2013), 'ggplot2' (Wickham et al., 2019), and 'dismo' (Hijmans et al., 2017). In addition, due to large mapping area and fine resolution, the amount of data and computation are quite large. For each fraction at each depth, prediction was performed on almost 1.2 billion pixels. To improve computation efficiency, a parallel computing strategy was employed in data processing, which separated the whole area into many tiles and run multiple-thread parallel computation.

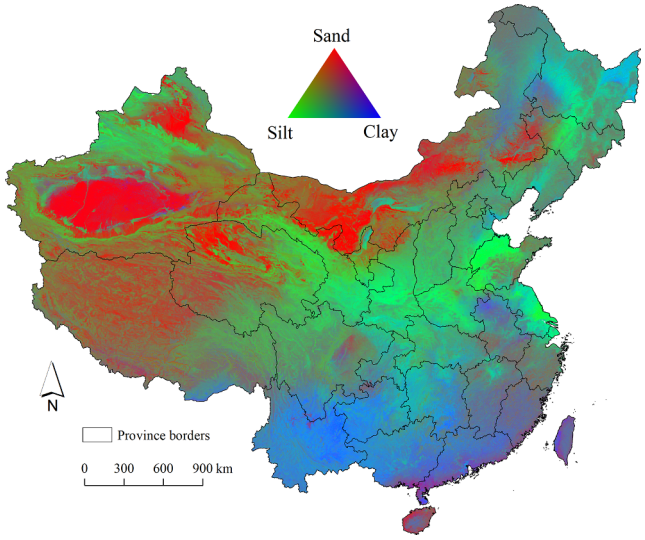

Fig. 6. Colour composite map of the three soil texture fractions at 0–5 cm depth.

2.5. Evaluation criteria

2.5.1. Criteria for evaluating the soil texture predictions

Based on the 4121 training soil profiles, 10-fold cross validation was used to evaluate the performance of the random forest method for each soil texture fraction and depth interval. We used four measurements: coefficient of determination (R^2), Concordance Correlation Coefficient (CCC; Lin, 1989), root mean square error (RMSE) and mean error (ME). They were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (1)$$

$$CCC = \frac{2r\sigma_o\sigma_p}{\sigma_o^2 + \sigma_p^2 + (\bar{O} - \bar{P})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (3)$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i) \quad (4)$$

where P_i and O_i are respectively the predicted and observed values of a soil texture fraction at sample point i; n is the total number of sample points; $\bar{P}$ and $\bar{O}$ are respectively the averages of the predicted and observed values; σ_p and σ_o are the corresponding standard deviations; and r is the correlation coefficient value between the predicted and observed values. We performed 30 repeats of 10-fold cross validation and calculated their values of mean and standard deviation of the measurements.

On the other hand, based on the 458 random-held back evaluation soil profiles, we compared our soil texture predictions with existing soil texture maps. One is the 1 km resolution linkage maps made by Shangguan et al. (2012) using the conventional linkage method. The method linked texture values of legacy soil profiles with polygons of 1:1,000,000 scale soil type map, which was not predictive soil mapping. Another is the 250 m resolution SoilGrids250m maps made by Hengl et al. (2017a) using an ensemble of random forest and gradient boosting methods and the same set of soil profiles in the extent of China. The improvement of our predictions relative to a reference work was calculated based on the R^2 and RMSE using the Eqs. (5) and (6) respectively:

$$RI_{R^2} = \frac{R^2_{new} - R^2_{ref}}{R^2_{ref}} \quad (5)$$

$$RI_{RMSE} = \frac{RMSE_{ref} - RMSE_{new}}{RMSE_{ref}} \quad (6)$$

where $RI_{R^2}$and RI_{RMSE} are relative improvement of our predictions with regard to R^2 and RMSE respectively, R^2_{new} and $RMSE_{new}$ are accuracy measurements for our predictions, and R^2_{ref} and $RMSE_{ref}$ are accuracy measurements for a reference work.

2.5.2. Criteria for evaluating the uncertainty estimation of the predictions

Lagacherie et al. (2019) demonstrated that uncertainty estimation could be itself highly uncertain, especially when using sparse soil datasets. It is thus important to evaluate the performance of uncertainty estimation. Based on the 458 random-held back soil profiles, prediction interval coverage probability (PICP) was used for the evaluation. The PICP is simply the proportion of observations at each depth that are encapsulated by the corresponding prediction interval (Solomatine and Shrestha, 2009; Malone et al., 2011). In this study, the predication interval was estimated by the quantile regression forest models mentioned above. If the uncertainty estimates have been reasonably defined, the PICP should result in an estimate of 90% for a 90% prediction interval. In addition, the standard deviations of the accuracy measurements (R^2, RMSE and ME) derived from the 30 repeats of 10-fold cross validation were also used to reflect, to some extent, stability of uncertainty of the predictions.

3. Results and analysis

3.1. Statistical summary of soil texture samples

Table 2 lists statistical description of the splines-fitted sand, silt and clay percentages at different depths based on the soil profiles. Overall, mean silt and sand contents were remarkably higher than mean clay content at every depth layer. Mean clay content slightly increased with the increase of depth while both mean silt and sand contents slightly decreased with the increase of depth. No matter what situation for the vertical change of mean content, standard deviation (SD) of all three fractions exhibited an increasing trend with the increase of depth. Sand and clay contents had higher variability for all depth layers with the coefficient of variation (CV) between 0.61 and 0.73, whereas silt content had lower variability with the CVs between 0.46 and 0.53.

3.2. Performance of model prediction and uncertainty estimation

Table 3 lists the mean and standard deviation of model prediction accuracy indicators of the soil texture fractions based on 30 repeats of 10-fold cross validation with the 4121 training soil profiles. Overall, the mean CCC values ranged from 0.59 to 0.65, indicating good agreement between the predicted and observed values. The mean ME values were very close to zero, suggesting overall unbiased predictions. The mean R^2 values of the predictions of soil texture fractions at different depth intervals were between 0.43 and 0.50. This indicates the models explained around 43–50% of soil texture variation present.

Specifically, sand and silt contents had slightly higher mean R^2 values than clay content, suggesting that sand and silt were slightly more predictable than clay. For every fraction, the R^2 values slightly decreased downward from the depth of 5 cm while the RMSE values increased, exhibiting a vertical decline of predictability of soil texture. The 5–15 cm depth interval had a better model performance than the 0–5 cm depth interval, with the higher R^2 and lower RMSE values.

Table 4 lists the values of prediction interval coverage probability (PICP) for clay, silt and sand contents and different depths, which were calculated based on the 458 random-held back soil profiles. For a 90% prediction interval, we would expect 90% of observations to fall within

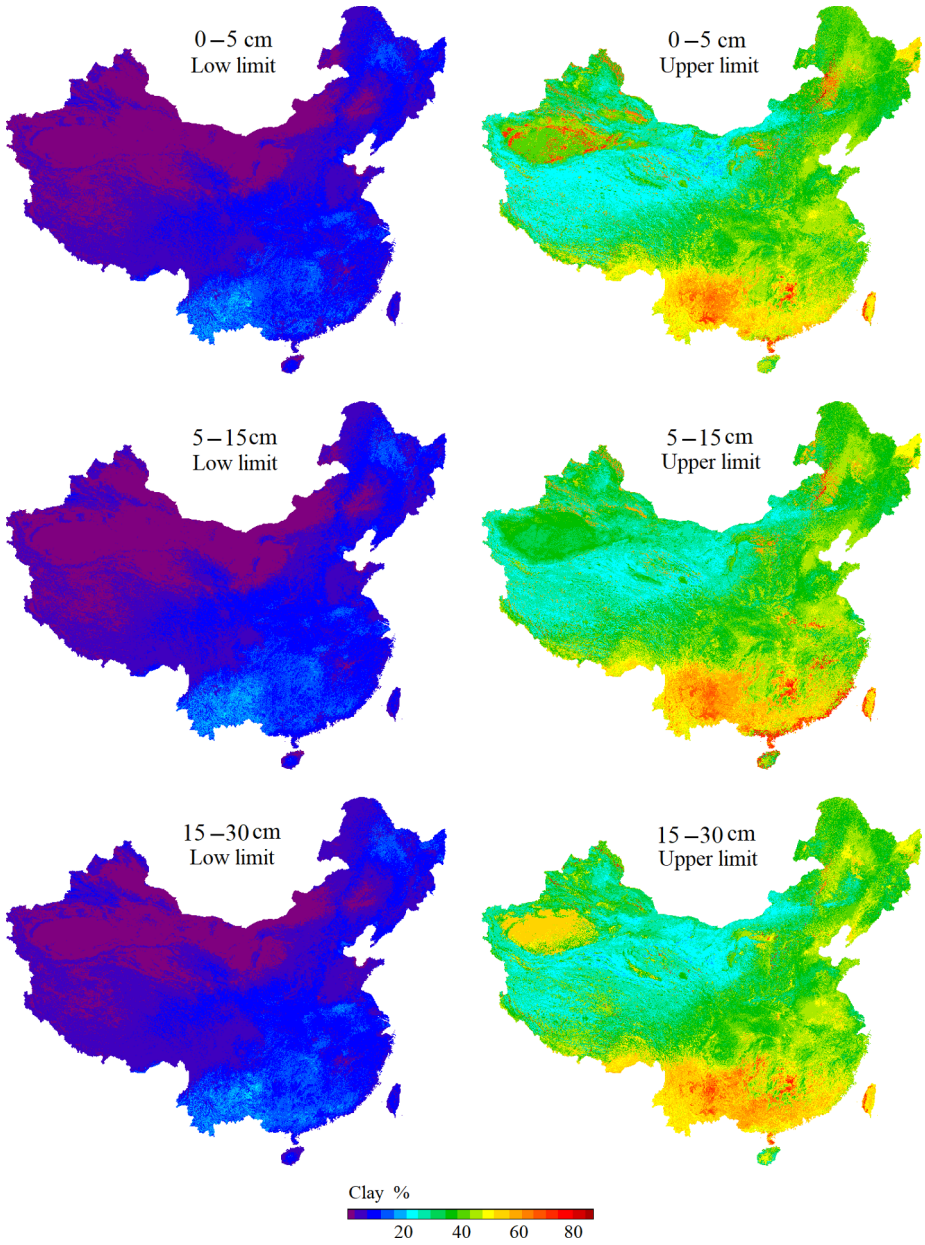

Fig. 7. Maps of lower and upper limits of 90% prediction interval of clay content for each depth.

the lower and upper prediction limits. It can be seen that all the fractions achieved PICP values very close to 90%, suggesting that these lower and upper prediction limits estimated by the quantile random forest method were of an appropriate magnitude. That is to say, the uncertainty estimations, to a large extent, were reliable. This can also be indicated by the small values of standard deviation of overall prediction accuracy indicators (R^2, CCC, RMSE and ME) in Table 3.

3.3. Spatial patterns of the predictions and their uncertainty

Figs. 3, 4 and 5 show the predicted maps of clay, silt and sand contents at different depths across China. Overall, clay content was predicted to be low in the north and northwest but high in the south. The lowest occurred in deserts of the northwest while the highest in the Yunnan-Guizhou Plateau covered by the Quaternary red clay which is believed to be formed in a more humid and warmer geological period but elevated with uplift of the Qinghai-Tibet Plateau (Huang and Lu, 2019). The relative high clay content occurred in some provinces (Hunan, Guangdong and Guangxi) in the south and the low-lying lacustrine deposits areas mainly including the Huaibei Plain of Anhui province and northern Songnen Plain of Heilongjiang province. Silt content was predicted to be high in the Loess Plateau, Yellow River alluvial plain of Shandong province, Jiangsu Plain, Yili valley of Xinjiang province, and eastern Songnen Plain. It was low in the desert areas in the north and northwest. Loess deposition is widely distributed in the country and thus there are considerable silt content in soils. Sand content generally exhibited an opposite pattern to silt content, i.e., high sand content corresponded to low silt content and vice versa. The obvious exception was the southwestern part with relatively low sand content and middle silt content due to high clay content. Sand content was predicted to be high in the north and northwest of China but low in the south. Within the south, most mountainous areas tended to have relatively higher sand content than low relief areas. The highest content occurred in desert areas in the northwest, followed by Gobi areas and western Qinghai-Tibet Plateau, whereas the lowest occurred in the Loess Plateau, Yellow River alluvial plain of Shandong province, Jiangsu Plain, and northeastern Songnen Plain. Deserts occupy large areas in Xinjiang, Inner Mongolia and Qinghai provinces, which are rich in sand content and poor in silt and clay contents. The predicted soil texture patterns conformed

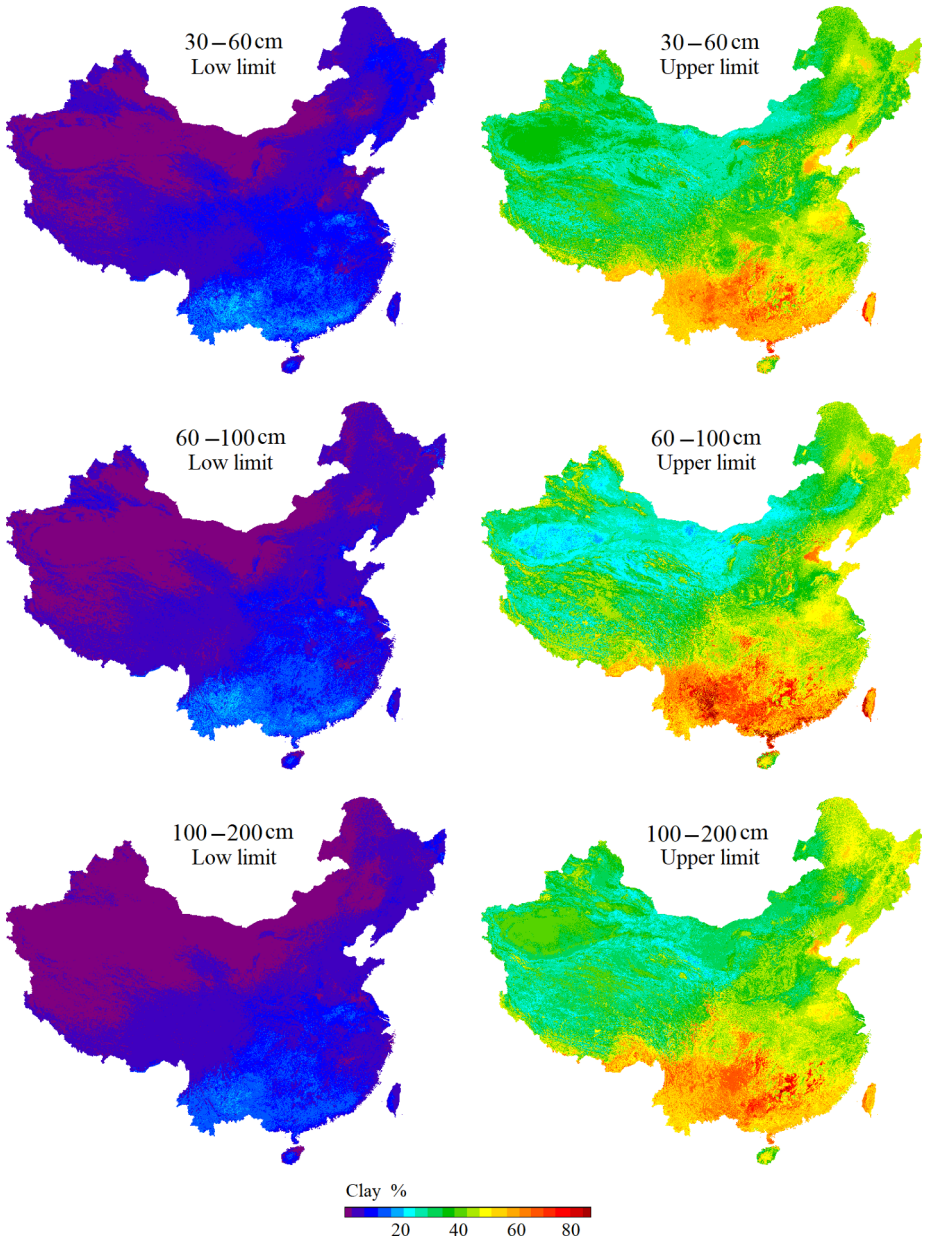

Fig. 7. (*continued*)

well with what is known about their general characteristics and distribution of soils in China (Gong et al., 2014).

Fig. 6 shows a colour composite map of the three soil texture fractions for the 0–5 cm depth layer. The northwest and north China is dominated by sandy textured soils (red colours). The middle part, i.e., Yellow River watershed mainly including Loess Plateau and lower reaches alluvial plains, is dominated by silty textured soils (green colours). The south and some low-lying lacustrine deposits areas is dominated by clay textured soils (blue colours). These can also be evidenced by the soil texture triangle plot in Fig. 2.

Figs. 7, 8 and 9 show maps of the uncertainties in predictions of clay, silt and sand contents for each depth. The uncertainty was expressed as lower and upper prediction limits at a 90% confidence interval. The 5% lower and 95% upper prediction limits had a similar spatial patterns with the mean predictions shown in Figs. 3–5, i.e., higher mean prediction values responded to higher values of lower and upper limits. The range between 5% lower and 95% upper prediction limits appeared rather wide for all three soil texture fractions, suggesting that there is room to improve the current spatial predictions.

3.4. Comparison with the existing maps

Table 5 shows accuracy assessments of our soil texture predictions and the conventional linkage maps from Shangguan et al. (2012) and the SoilGrids250m from Hengl et al. (2017a) at five depth intervals based on the 458 random-held back evaluation samples. Our predictions had much higher R^2 values and lower RMSE values than these existing soil texture maps for all the depths. The lowest accuracy was found for the linkage maps, showing the advantage of digital soil mapping over conventional linkage method. The remarkably lower accuracy of SoilGrids250m than our predictions is not out of the expectation that global model building could be less accurate than national model building when focusing on a national extent. The differences in the R^2 values indicate that the improvement for clay content was around 248% relative to the linkage maps and 92% relative to the SoilGrids250m maps respectively. The improvement for silt content was around 370% and 112% respectively, and that for sand content was around 245% and 83% respectively. As far as RMSE is concerned, the improvement for clay content was around 24% relative to the linkage

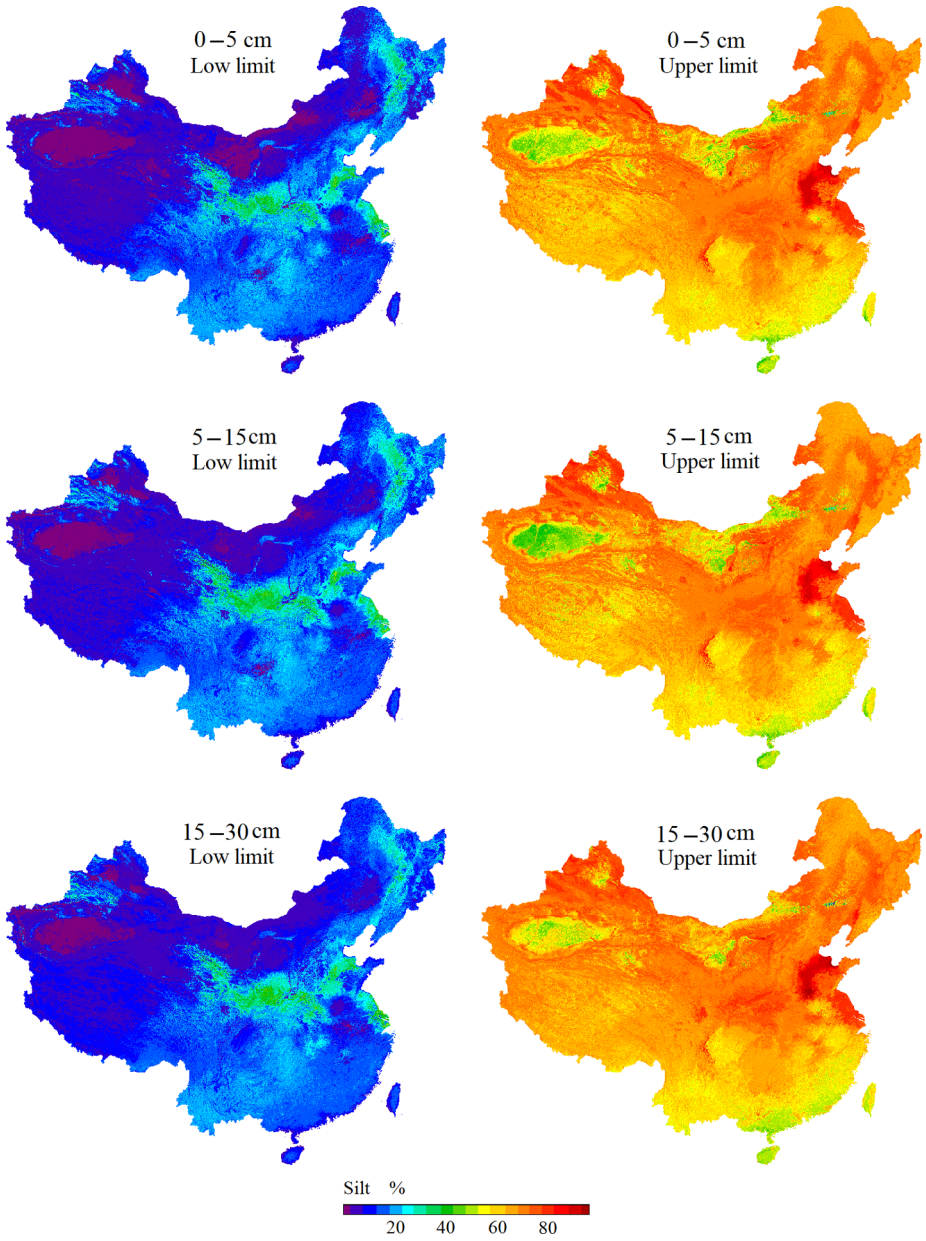

Fig. 8. Maps of lower and upper limits of 90% prediction interval of silt content for each depth.

maps and 14% relative to the maps of the SoilGrids250m maps respectively, that for silt content was around 26% and 19% respectively, and that for sand content was around 26% and 17% respectively. Thus, the predictions were much more accurate than the existing maps of soil texture fractions.

There were obvious differences between our predicted maps and the existing maps although they had a similar trend of spatial distribution. Take silt content at 5–15 cm depth as an example (Fig. 10). The linkage silt map looks fragmented and stepped. There are many abrupt changes between patches on the map. The abrupt changes inherited the limitations of original polygon-based soil type map and assumed that there is no spatial variation of soil texture within a soil polygon and the variation appears only at the boundaries of polygons. This was obviously reflected in the upper right scatter plot in Fig. 6, where the mapped silt content values were the same (within a polygon) but the observed silt content values were very different. The assumption often does not hold in reality. Although abrupt changes of soils over landscapes do exist, changes in soil properties generally occur in a gradual and continuous way. The silt map of the SoilGrids250m did not well represent the desert areas in Xinjiang and Inner Mongolia provinces, where the lowest silt content should occur. It also did not well represent the Loess Plateau and the loess deposit areas in the northeast, where relatively high silt content should occur. Another obvious problem of the SoilGrids250m map is the condensed range of the predicted silt content, i.e., overestimation for low silt content and underestimation for high silt content. This can be seen from the SoilGrids250m silt map and the corresponding scatter plot in the middle of Fig. 10, which may be a result of smoothing effect from the averaging operations in ensemble algorithms. Both random forest and gradient boosting both are already ensemble models. The two models were ensembled again to produce the SoilGrids250m map. The averaging operation thus was actually performed three times in the mapping, leading to a relatively serious smoothing effect on the resulting map. Hengl et al. (2017a) also found the smoothing effect in the predictions of Tasmania and California. Besides, there are remarkable difference in the level of spatial detail between our predictions and the existing maps. Take clay content at 5–15 cm depth as an example and focus on a 67 km × 46 km local area situated in the Huangling county of Shanxi province, China

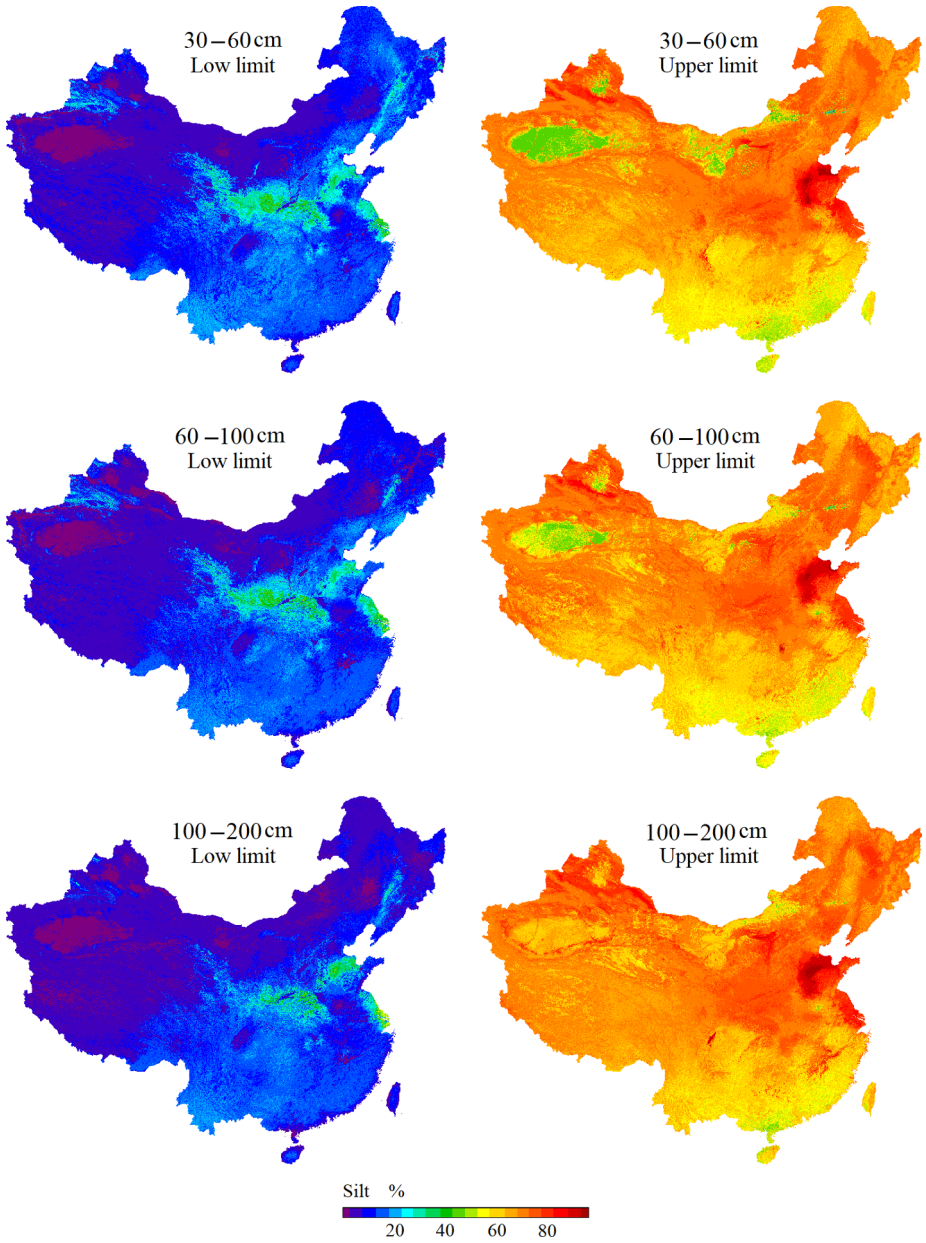

Fig. 8. (*continued*)

(Fig. 11). From the map excerpt it can be seen that our clay map were much more detailed than the existing clay maps over space. Thus, our predictions better represented spatial variation of the soil texture fractions across China than the existing maps.

3.5. Controlling factors of China soil texture patterns

Figs. 12, 13 and 14 show relative importance of the first 15 important covariates used in the predictions of soil texture fractions at multiple depths. In clay predictions, the importance of the covariates did not have obvious changes with depth. Solar radiation, wind speed and Band5 were the most important covariates for all depth. Temperature-related variables and regolith thickness were the second important covariates. It appeared that changes of air temperature at diurnal and seasonal and even annual scales became more important than its mean status with the increase of depth. Terrain variables such as slope gradient, TWI and elevation had relatively low importance, which play roles through controlling local moisture and thermal conditions and mass redistribution over landscapes. Relatively high solar radiation and temperature changes and low moisture conditions in the north and northwest of China lead to relatively strong physical weathering and weak chemical weathering in soil forming process. This affects the production of secondary minerals, plus severe wind erosion caused by high wind speed of this area, resulting in relatively low clay content and high sand content. On the contrary, relatively high mean temperature and moisture and low temperature changes in the south lead to strong chemical weathering of parent materials and consequently massive production of secondary clay minerals, plus high vegetation cover of this area, resulting in relatively high clay content. Thus, heat and water drive physical and chemical weathering and wind drive erosion processes which primarily have shaped the pattern of clay content of China.

In silt predictions, elevation, solar radiation and air temperature seasonality were the most important covariates for almost all depths. Other covariates except terrain variables and annual precipitation all had some changes of relative importance with depth. Although these changes, Band7, slope gradient, wind speed and air temperature annual range were the second important covariates. The following important covariates include the air temperature max value, Band5, mean daytime

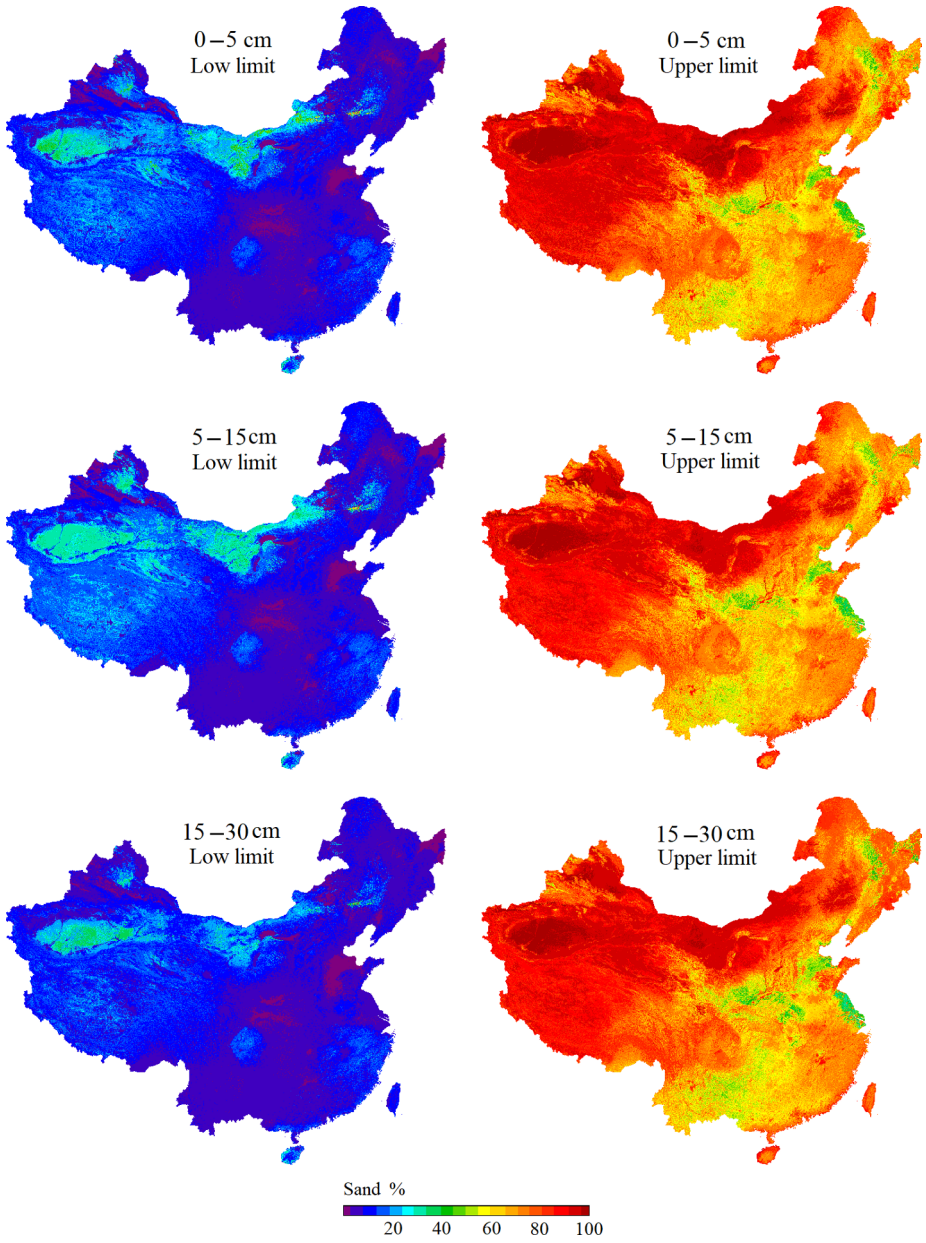

Fig. 9. Maps of lower and upper limits of 90% prediction interval of sand content for each depth.

LST, annual precipitation, TWI and terrain wind effect. Obviously, terrain was a major control for the silt distribution. The northwest wind prevailing in winter weakens when encountering the obstruction of the Qilian and Qinling Mountains. The mountain chain consists of a high and rugged barrier extending from Gansu to Henan province. The weakening results in massive deposits of dust carried by the wind in the north-central China, contributing to the formation the Loess Plateau. The dust is rich in silt content. In addition to the large-scale terrain arrangements, local terrain features represented by elevation, slope gradient and TWI to a large extent determine gravity and hydraulic power conditions and thus the intensity of erosion, redistribution and sorting processes of soil particles. Wind is an important control for the silt distribution. Relative high wind speed in the north and northwest lead to strong soil wind erosion and loss of fine soil particles. And the variables of terrain wind effects to the winter wind from the dominant northwest played a significant role in the silt predictions. Besides, the importance of Band7, Band5, MAP and TWI indicate that water is also an important control for the silt distribution. Relatively high moisture and annual precipitation lead to strong water erosion, transportation and sorting of soil particles especially for the loess deposition areas. The Yellow River runs through the Loess Plateau and carries large amount of fine soil particles to plain areas of the lower reaches, forming a zone of high silt content in the northwestern Shandong province. Thus, the terrain, wind and water have driven deposition, erosion and transportation sorting processes of soil particles which have primarily shaped the pattern of silt of China.

In sand predictions, solar radiation, wind speed, elevation, Band5 and Band7 were the most important covariates for almost all depths. The changes and extremely high value of air temperature appeared to be more important than its mean status, plus the importance of solar radiation, indicating that physical weathering was an important process for affecting the formation of sand distribution. Wind speed was as important as in clay predictions. Annual mean precipitation became much more important than that in both silt and clay predictions, plus the importance of Band5 and Band7, indicating the importance of water erosions. Although elevation was still important, slope gradient contributed little below the depth of 30 cm. The elevation may primarily exert its influence in gravity and water erosion which takes away fine

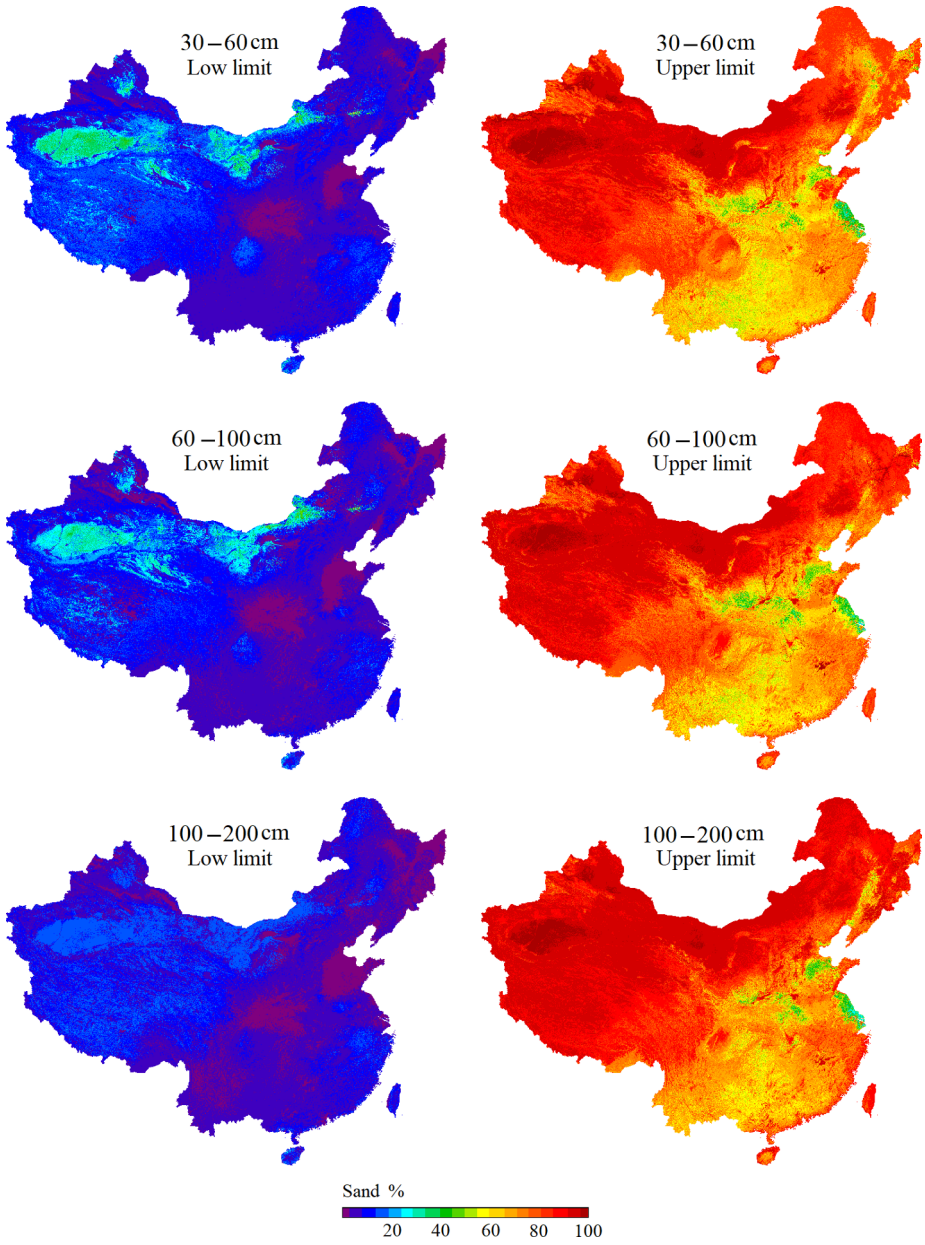

Fig. 9. (*continued*)

particles and leaves coarse particles. This can be seen in the south where most mountainous areas tended to have relatively higher sand content than neighbouring low relief areas. Thus, heat has driven physical weathering and wind, water and terrain have driven erosion processes which have primarily shaped the pattern of sand distribution in China.

4. Discussion

4.1. Comparison of contemporary digital soil mapping assessments

The R^2 values of the predictions of soil texture fractions at depths in this study (0.43–0.50) was at the same level as the national three-dimensional soil texture mapping studies of Australia (0.39–0.53; Viscarra Rossel et al., 2015), Denmark (0.26–0.55; Adhikari et al., 2013) and United States (0.46–0.57; Ramcharan et al., 2018), but was relatively better than that of France (0.19–0.44; Mulder et al., 2016a), and much better than that of Chile (0–0.09; Padarian et al., 2017) although their differences in soil landscapes, prediction methods and data conditions. Moreover, the R^2 values had a smaller range than other studies, indicating a more steady predictive performance among different depths. The substantial unexplained variation can be attributed to the limited number of sparse soil profiles sites, i.e., nearly one soil profile site per 2000 km^2 on average in the study area. This may be not enough to capture short-range spatial variation of soil texture as noted by Arrouays et al. (2017).

The decline of prediction accuracy with the increase of depth was consistent with quite a few studies of three-dimensional prediction of soil properties (Minasny et al., 2006; Liu et al., 2013; Kempen et al., 2014; Viscarra Rossel et al., 2015). This may be associated with the increase of soil texture variability towards deep layers, which can also be seen from the vertical changes of SD and CV values in Table 2. Moreover, the covariates used in this study mainly characterize surface environmental conditions and processes and have relatively weak relationships with soil properties of deep layers. The 0–5 cm depth interval often had a slightly worse model performance than the 5–15 cm depth interval. One important reason may be the bias resulted from the splines fitting on original soil horizons data. It is well known that if the

Table 5
Accuracy assessments of the soil texture predictions with the existing linkage maps from Shangguan et al. (2012) and the SoilGrids250m from Hengl et al. (2017a,b) based on the 458 random-held back soil profiles.

	Our predictions			The linkage maps			SoilGrids250m		
Depth (cm)	R^2	RMSE (%)	ME (%)	R^2	RMSE (%)	ME (%)	R^2	RMSE (%)	ME (%)
Clay									
0–5	0.43	9.17	0.38	0.13	12.05	−0.31	0.23	10.71	1.81
5–15	0.44	9.21	0.21	0.13	12.09	−0.37	0.22	10.72	1.62
15–30	0.41	9.73	0.25	0.10	12.74	−0.38	0.22	11.21	2.26
30–60	0.43	10.10	0.77	0.12	13.52	1.42	0.22	11.93	3.38
60–100	0.42	10.74	0.91	0.14	14.04	1.88	0.22	12.67	3.89
Silt									
0–5	0.53	14.25	−0.68	0.10	20.01	−1.02	0.25	17.94	−0.76
5–15	0.53	14.05	−0.55	0.11	19.58	−1.14	0.25	17.65	−0.39
15–30	0.50	14.38	−0.62	0.10	19.77	−1.25	0.23	17.83	−0.95
30–60	0.46	15.37	−0.20	0.11	19.98	−1.45	0.22	18.46	−1.05
60–100	0.46	15.73	0.29	0.11	20.42	−1.25	0.22	19.06	−0.85
Sand									
0–5	0.52	17.72	0.67	0.15	24.56	1.31	0.27	21.87	−1.07
5–15	0.53	17.46	0.67	0.15	24.30	1.53	0.27	21.58	−1.21
15–30	0.49	18.38	0.77	0.14	24.97	1.75	0.27	22.06	−1.23
30–60	0.45	19.78	−0.29	0.13	25.83	0.11	0.26	23.05	−2.26
60–100	0.46	19.89	−0.75	0.14	26.02	−0.55	0.27	23.52	−2.96

initial data are for horizons of topsoil e.g. 0–12 or 0–20 cm and if there is a change in soil texture below, then the fitted spline curve of 0–5 cm may be biased. The magnitude of the biases depends on the constrast between topsoil horizons and subsoil ones (Arrouays et al., 2014; Odgers et al., 2012). Another possible reason may be the exposure of the very surface layer to complex external environments, which makes it easier to be influenced by some random or disturbing factors such as human activities.

4.2. Conventional linkage method and digital soil mapping

The linkage method is relatively simple and easy to operate. It only needs a polygon-based soil type map and a dataset of soil samples, and does not need intensive computation. But, the method has several drawbacks. First, it is constrained by the scale of the soil type map. Due to the lack of detailed soil type map in large extents, high resolution soil properties maps often cannot be made. Second, it assumes that soil property value is the same within a polygon and its variation only occurs at polygon boundaries. This may not be consistent with the reality of soil spatial variation over landscapes. The spatial misrepresentation of soil property would be severe when the scale of soil type map is small. Third, the process of selecting the linkage between polygons and soil samples was to some extent arbitrary as mentioned in Shangguan et al. (2012). Although these drawbacks, under the situations of lacking soil property information of large extents such as national, continental or global, the conventional linkage method would be useful for quick production of usable maps. For example, Reynolds et al. (2000) made the 10 km resolution global distributions of sand and clay fractions for 0–30 and 30–100 cm depth intervals through linking the Food Agriculture Organization (FAO) soil map of the world with global pedon databases. And FAO et al. (2009) developed the widely used Harmonized World Soil Database.

In comparison, digital soil mapping method involves relatively complicated operations including characterizing soil formative environments, modeling soil and environmental relationships, and predicting soil variation over space. It usually needs soil samples, environmental covariates and an appropriate predictive model. It can utilize rich environmental data such as digital elevation model based terrain variables and remote sensing images to assist its mapping process. With high spatial resolution of environmental datasets available currently, we can conduct spatially detailed digital soil mapping to represent the details of soil property variation. High resolution and large extent soil mapping usually needs the support of high performance computing facility. Besides, digital soil mapping method can often achieve relatively high mapping accuracy compared to the conventional techniques. In addition, it is easy to evaluate the uncertianty of soil property predictions. Thus, the digital soil mapping is a promising alternative to the conventional linkage method.

4.3. Some issues of large extent digital soil mapping

Previous digital soil mapping studies mainly focused on relatively small extent with the purpose of developing and testing methods (Moore et al., 1993; Robinson and Metternicht, 2006; Malone et al., 2009; Arrouays et al., 2017). Studies in recent years have been extending to large extent such as national (Odgers et al., 2012; Grundy et al., 2015; Viscarra Rossel et al., 2015; Mulder et al., 2016a; Chaney et al., 2016; Chen et al., 2019), continental (Hengl et al., 2015, 2017b; Ballabio et al., 2016) and global extents (Hengl et al., 2014, 2017a). With the increase of spatial extent, more landscapes are usually included. This leads to two aspects of changes. One aspect is that the difference in field accessibility over the mapping extent of interest may become obvious. For example, plains are often more accessible than hilly or mountainous areas. Another aspect is that soil-landscape relationship to be considered may become diverse and complex, with relatively strong nonlinearity, spatial nonstationarity, and the involvement of multiple factors. Arrouays et al. (2012) noted such complexity in designing soil monitoring networks in large areas.

The first change makes it difficult to achieve a spatially even distribution of soil samples. Of course, most large extent mapping used legacy soil samples to avoid new sampling but such legacy samples are usually not distributed evenly over space. The spatially uneven distribution of samples is an issue for soil prediction modelling and its cross validation (Richer-de-Forges et al., 2017, Hengl et al., 2017b). In this study, the national soil survey did not use a statistical design, but purposive or typical one. Although the coverage of almost all soil landscapes across China, the samples were unevenly distributed over space. The 10-fold cross validation for evaluating model performance was based on the uneven distribution of samples. It used a way of random selection to form a subset of samples for validation. The random selection gives more importance to the local areas where samples are dense, leading to relatively small number of validation samples in the areas with sparse samples. This may result in a bias in the evaluation and obtain a more optimistic accuracy. Brus et al. (2011)

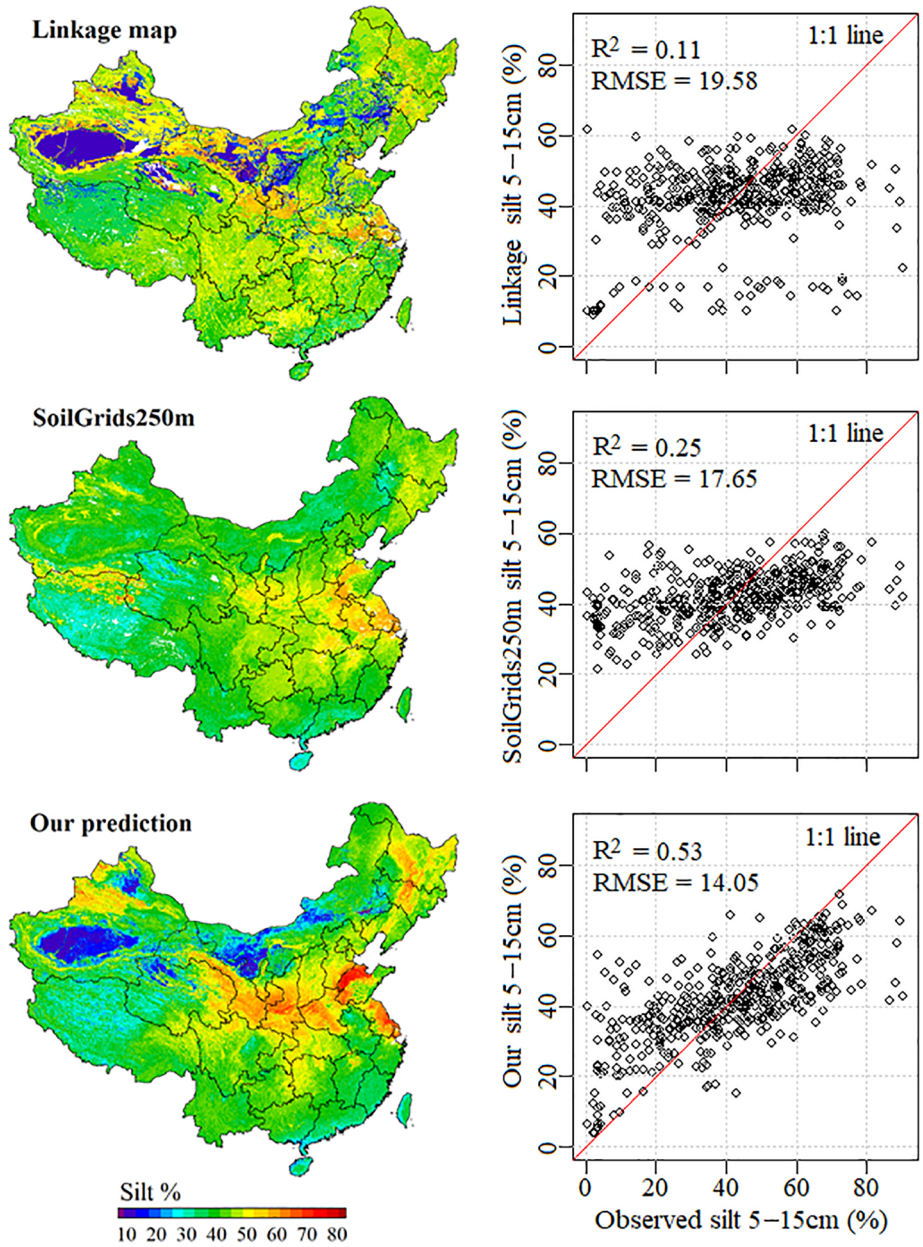

Fig. 10. Comparison of the silt prediction with the linkage maps from Shangguan et al. (2012) and the SoilGrids250m from Hengl et al. (2017a,b) at 5–15 cm depth based on the 458 random-held back evaluation samples.

noted that the sample subset randomly selected from the samples of non-probability sampling design will be biased, i.e., not a true representation of total population, and consequently the validation based on the subset will provide biased estimates of model quality. Thus, the results of the cross validation in this study only give an indication of the true accuracy of the predictions. Richer-de-Forges et al. (2017) developed an evaluation procedure to take into account spatial configuration of samples in evaluation. They provides a promising idea to reduce the evaluation bias.

The second change poses a challenge for current digital soil mapping methodologies. Lagacherie and Voltz (2000) speculated that, especially over large areas, predictive capabilities are limited because relationships between soil properties and landscape attributes are nonlinear or unknown. Minasny and McBratney (2010) noted that knowledge and techniques for regional soil mapping may not be applicable at a global extent due to spatial variability of soil-landscape relationships. Thus large extent soil mapping needs models that can deal with the complex soil-landscape relationships. Such models themselves are often complex in model structure and algorithm. Mitran et al. (2018) found that geographically weighted regression kriging model had better predictive performance of surface total carbon stocks than linear regression kriging model in two adjacent states in southern India. The former model was more complex than the latter model because it took into account the spatial non-stationarity of the relationships between soil carbon and environmental covariates. Keskin et al. (2019) showed that random forest as ensemble method outperformed other relatively simple methods such as classification and regression trees for digital mapping of soil carbon fractions in Florida. It would be helpful to explore more complicated (maybe also better) models but the computational challenge arising therefrom needs to be considered.

In addition, although the model built in a large extent may be unbiased for the whole extent, its prediction is often biased in local areas. Mulder et al. (2016b) found that the global SoilGrids1km product developed by Hengl et al. (2014) overestimated soil organic carbon content of France. Liang et al. (2019) also observed the overestimation for soil organic matter content of China. Vitharana et al. (2019) observed

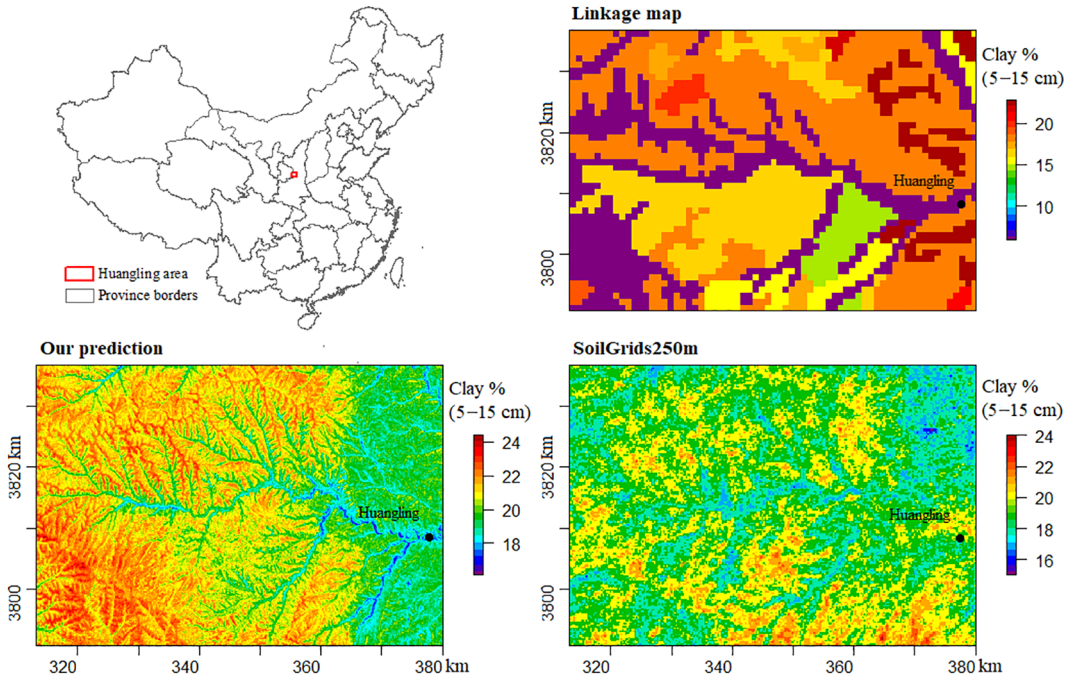

Fig. 11. Comparison of spatial details of the predictions with the existing maps: clay content at 5–15 cm depth in the map excerpt of Huangling area of Shanxi province, China.

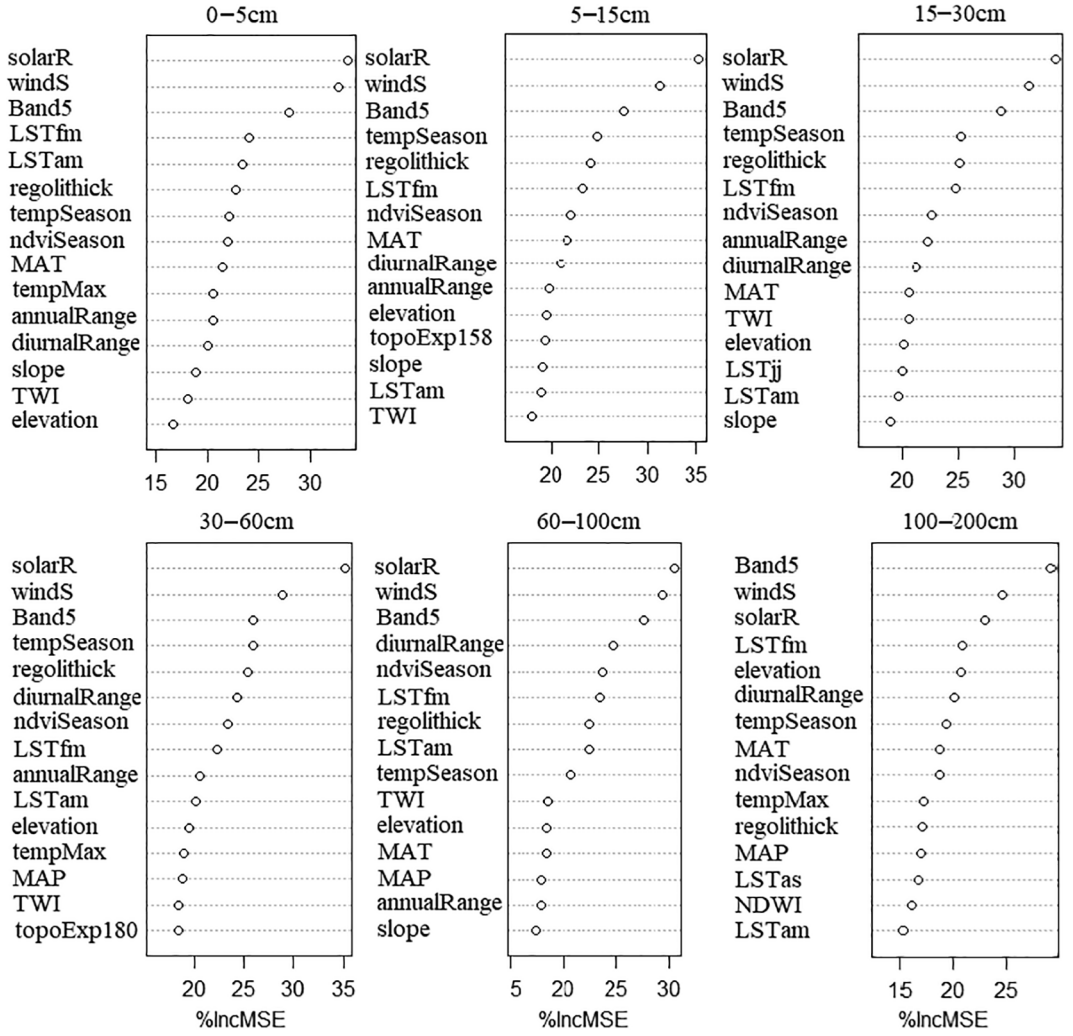

Fig. 12. Relative importance (%IncMSE) of the covariates used in clay prediction.

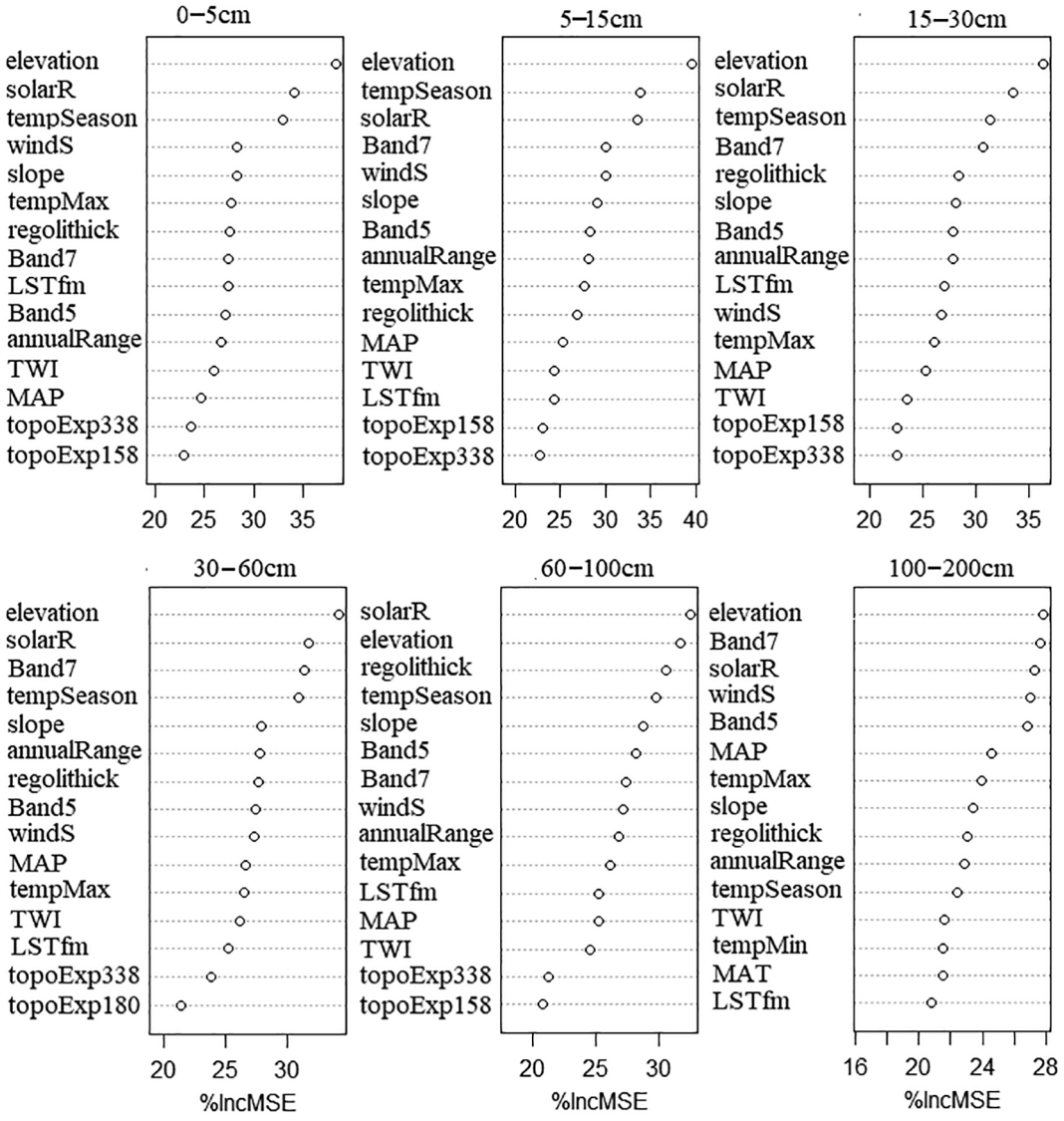

Fig. 13. Relative importance (%IncMSE) of the covariates used in silt prediction.

the overestimation of total soil organic carbon stocks by SoilGrids250m was 122% for the 0–30 cm layer and 209% for the 30–100 cm layer in Sri Lanka. This is not surprising but indeed a problem that needs to be addressed by digital soil mapping communities. This problem may indicate that current soil prediction models are still not flexible enough to deal with the complexity of soil-landscape relationships in large extents. Arrouays et al. (2017) highlighted that both top-down and bottom-up approaches are necessary to enhance the quality of digital soil maps and map the entire world. Hengl et al. (2017a) recommended a strategy of merging local and global predictions. Padarian et al. (2019) proposed a transfer learning method to localise a continental soil vis-NIR calibration model for accurate local soil predictions, which can be borrowed to digital soil mapping. Thus, new methodologies for large extent digital soil mapping still need to be developed. From that, the GlobalSoilMap project would greatly benefit.

4.4. Potential applications of high resolution national soil texture maps

The high resolution soil texture maps produced in this study have many potential applications such as climate, ecological, hydrological modelling, water resource management and soil pollution control. First, the maps can be used for estimating important soil hydrological parameters including soil available water capacity, permanent wilting point and saturated hydraulic conductivity through pedotransfer functions (Shiri et al., 2017). As is well known, these parameters are highly heterogeneous over space and their measurements at field or laboratory are laborious and time-consuming, leading to a dearth of their spatial information. High resolution maps of soil texture and hydrological parameters are useful to model ecological and hydrological processes and make scientific agricultural irrigation planning in consonance with local soil conditions. Second, the maps can be used for estimating soil erodibility K-value through empirical models (Liang et al., 2013). Soil erosion is a severe problem of ecological environments in China. According to the Bulletin of First National Census for Water, the total area of land affected by water and wind was estimated to be almost one third of the territory of China (Ministry of Water Resources and National Bureau of Statistics of China, 2013). High resolution map of the K-value is key information for assessing the amount and risk of soil erosion and guiding the construction of ecological environments. Third, soil pollution is currently a big issue of concern in China. The maps provide critical soil spatial information for assessing the environmental risks caused by pollutant leaching and developing solutions for soil pollution remediation. Last, spatial uncertainty distribution of the maps can be used for guiding further soil sampling design to improve the map

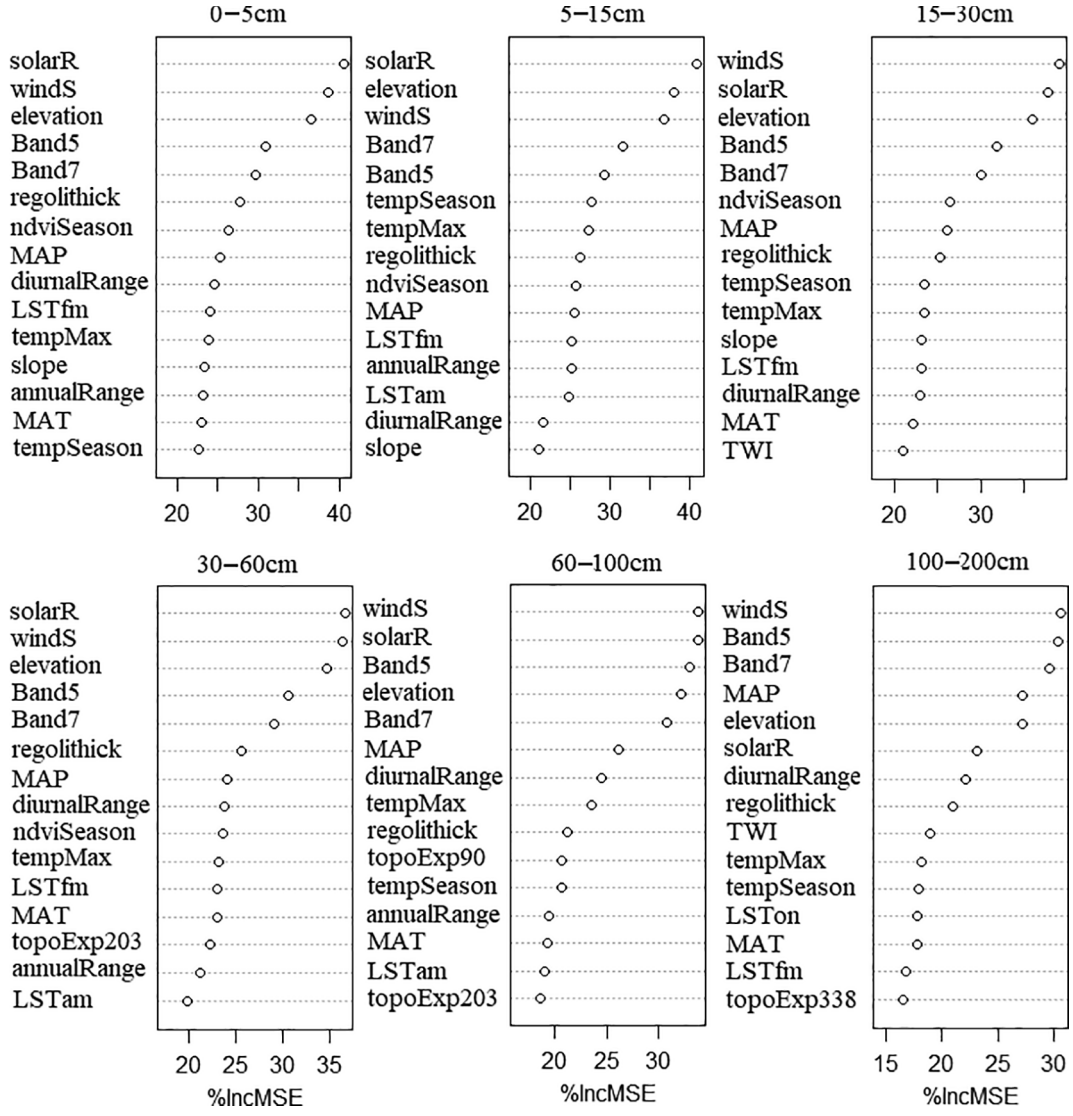

Fig. 14. Relative importance (%IncMSE) of the covariates used in sand prediction.

quality because areas with high uncertainty are usually under represented and need to collect new samples.

5. Conclusions

The study shows that the combination of machine learning techniques with currently available high-resolution soil formative environmental covariates can effectively predict spatial variation of soil texture at a national extent and a detailed level. We provided the first version of 90 m resolution maps of soil texture fractions and their uncertainty across China. It was much more accurate and detailed than the existing soil texture maps and can well represent spatial variation of soil texture. The predicted maps represent a contribution of China to the GlobalSoilMap project, and provide critical soil information for water-related applications. Besides, new methodologies still need to be explored for large extent digital soil mapping, from which many global initiatives will greatly benefit.

In addition, we found that heat, water, wind and terrain were major controlling factors for the spatial patterns of soil texture in China. The heat and water have driven physical and chemical weathering and wind have driven erosion processes which have primarily shaped the pattern of clay content. The terrain, wind and water have driven deposition, erosion and transportation sorting processes of soil particles which have primarily shaped the pattern of silt. Heat-driven physical weathering and wind, water and terrain-driven erosion processes have primarily shaped the patter of sand. The findings provide clues for developing mechanistic soil evolution models to simulate spatiotemporal evolution of soil texture at a national extent. The simulation can support national soil management to ensure the soil is secured in the future. This is particularly important under the background of increasing climate changes and intensified human activities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Key Basic Research Special Foundation of China (2008FY110600 and 2014FY110200), National Natural Science Foundation of China (41571212), and National Key Research and Development Program (2018YFE0107000). We thank the colleagues of the project of Chinese Soil Series Survey and Compilation of Chinese Soil Series and all soil surveyors and technical assistants

involved in the survey and laboratory work. We also thank the anonymous reviewers and the Editor Prof. Alex McBratney for their comments and suggestions which have greatly improved this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2019.114061.

References

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-resolution 3-D mapping of soil texture in Denmark. Soil Sci. Soc. Am. J. 77, 860–876.

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., Zhang, G.L., 2014. Globalsoilmap: toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134.

Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Regional 9, 1–4.

Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Martin, M.P., Bellamy, P.H., Lark, R.M., Kibblewhite, M., 2012. Generic issues on broad-scale soil monitoring schemes: a review. Pedosphere 22 (4), 456–469.

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.

Bishop, T., McBratney, A.B., Laslett, G., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45.

Breiman, L., 2001. Random forests. Mach. Learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, Jr.T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67.

Chen, C., Hu, K.L., He, R., 2013. 3D stochastic simulation and uncertainty assessment of soil texture at field scale. Soils 45 (2), 319–325.

Chen, S.C., Liang, Z.Z., Webster, R., Zhang, G.L., Zhou, Y., Teng, H.F., Hu, B.F., Arrouays, D., Shi, Z., 2019. 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.

Cooperative Research Group on Chinese Soil Taxonomy. 2001. Keys to Chinese Soil Taxonomy (third ed.). Hefei: Press of University of Science and Technology of China.

Diaz-Uriarte, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinf. 7. https://doi.org/10.1186/1471-2105-7-3.

Drury, S., 1987. Image Interpretation in Geology. London: Allen and Unwin, pp. 243.

FAO, IIASA, ISRIC, ISS-CAS, JRC, 2009. Harmonized World Soil Database (version1.1). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Gao, B., 1996. NDWI–a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266.

Gong, Z.T., Huang, J.R., Zhang, G.L., 2014. Soil geography of China. Beijing: Science Press. ISBN 978-7-03-038905-3. pp. 636. (In Chinese).

Grimm, R., Behrens, T., Marker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-digital soil mapping using random forests analysis. Geoderma 146, 102–113.

Grundy, M.J., Viscarra Rossel, R.A., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L.J., 2015. Soil and landscape grid of Australia. Soil Res. 53, 835–844.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X.Y., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017a. Soilgrids250m: global gridded soil information based on machine learning. PLoS ONE 122, e0169748.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km – global soil information based on automated mapping. PLoS ONE 9 (8), e105992.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS ONE 10, e0125814.

Hengl, T., Leenaars, J.G.B., Shepherd, K.D., Walsh, M.G., Heuvelink, G.B.M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., Kwabena, N.A., 2017b. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250m spatial resolution using machine learning. Nutr. Cycl. Agroecosyst. 109 (1), 77–102.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100, 269–301.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., 2017. Package 'dismo': species distribution modeling. Version 1.1-4. http://rspatial.org/sdm/.

Hijmans, R.J., van Etten, J., 2013. raster: raster: Geographic data analysis and modeling. R package version 2.1-25. http://CRAN.R-project.org/package=raster.

Huang B., Lu S.G., 2019. Soil series in Yunnan province. Beijing: Science Press. (In Chinese).

Jacquier, D.W., Seaton, S., 2012. Spline Tool for Estimating Soil Attributes at Standard Depths. CSIRO Land and Water, Australia. http://www.asris.csiro.au/methods.html#Method_Downloads.

Keitt, T., Bivand, R., Pebesma, E., Rowlingson, B., 2009. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.6-21. http://CRAN.R-project.org/package=rgdal.

Kempen, B., Heuvelink, G.B.M., Brus, D., Walvoort, D., 2014. Towards globalsoilmap.net products for the Netherlands. In: Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A.B. (Eds.), GlobalSoilMap–Basis of the Global Spatial Soil Information System. CRC Press, pp. 85–90.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N.P.A., 2019. 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 337, 1320–1328.

Lagacherie, P., Voltz, M., 2000. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma 97, 187–208.

Li, H.Y., Shi, Z., Webster, R., Triantafilis, J., 2013. Mapping the three-dimensional variation of soil salinity in a rice-paddy soil. Geoderma 195–196, 31–41.

Liang, Z.Z., Chen, S.C., Yang, Y., Zhao, R., Shi, Z., Viscarra Rossel, R.A., 2019. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. Geoderma 335, 47–56.

Liang, Y., Liu, X.C., Cao, L.X., Zheng, F.L., Zhang, P.C., Shi, M.C., Cao, Q.Y., Yuan, J.Q., 2013. K-value calculation of soil erodibility of China water erosion areas and its macro-distribution. Soil Water Conserv. China 10, 35–40 (In Chinese).

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.

Liu, F., Rossiter, D.G., Song, X.D., Zhang, G.L., Yang, R.M., Zhao, Y.G., Li, D.C., Ju, B., 2016. A similarity-based method for three-dimensional prediction of soil organic matter concentration. Geoderma 263, 254–263.

Liu, F., Zhang, G.L., Sun, Y.J., Zhao, Y.G., Li, D.C., 2013. Mapping the three-dimensional distribution of soil organic matter across a subtropical hilly landscape. Soil Sci. Soc. Am. J. 77 (4), 1241–1253.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.

Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626.

McBratney, A.B., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.

McBratney, A.B., Mendonca-Santos, L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Meinshausen, N., Schiesser, L., 2015. quantregForest: quantile regression forests. R Package.

Minasny, B., McBratney, A.B., 2010. Chapter 34: Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, And Operation. Springer, Dordrecht, pp. 429–436.

Minasny, B., McBratney, A.B., Mendonca-Santos, M.L., Odeh, I.O.A., Guyon, B., 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. Aust. J. Soil Res. 44, 233–244.

Ministry of Water Resources and National Bureau of Statistics of China, 2013. Bulletin of First National Census for Water. China Water & Power Press, Beijing. (in Chinese).

Mitran, T., Mishra, U., Lal, R., Ravisankar, T., Screenivas, K., 2018. Spatial distribution of soil carbon stocks in a semi-arid region of India. Geoderma Regional 15, e00192.

Montanarella, L., Vargas, R., 2012. Global governance of soil resources as a necessary condition for sustainable development. Curr. Opin. Env Sust. 4 (5), 559–564.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57 (2), 443–452.

Mulder, V.L., Lacoste, M., de Forges, A.R., Arrouays, D., 2016a. Globalsoilmap France: high-resolution spatial modelling the soils of France up to two meter depth. Sci. Total Environ. 573, 1352–1369.

Mulder, V.L., Lacoste, M., de Forges, A.R., Martin, M., Arrouays, D., 2016b. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16–34.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4 (1), 1–22.

Odgers, N.P., Libohova, Z., Thompson, J.A., 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190, 153–163.

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma 340, 270–288.

Phillips, J.D., 2016. Identifying sources of soil landscape complexity with spatial

adjacency graphs. Geoderma 267, 58–64.
Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil profile from analysis of a small number of samples. J. Soil Sci. 37, 455–467.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., Thompson, J., 2018. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. Soil Sci. Soc. Am. J. 82, 186–201.
Reynolds, C.A., Jackson, T.J., Rawls, W.J., 2000. Estimating soil water-holding capacities by linking the Food Agriculture Organization soil map of the world with global pedon databases, continuous pedotransfer functions. Water Resour. Res. 36, 3653–3662.
Richer-de-Forges, A.C., Saby, N.P.A., Mulder, V.L., Laroche, B., Arrouays, D., 2017. Probability mapping of iron pan presence in sandy podzols in South-West France, using digital soil mapping. Geoderma Regional 9, 39–46.
Robinson, T.P., Metternicht, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Comput. Electron. Agric. 50, 97–108.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny Budiman, Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital soil map of the world. Science 325, 680–681.
Shangguan, W., Dai, Y.J., Liu, B.Y., Ye, A., Yuan, H., 2012. A soil particle-size distribution dataset for regional land and climate modelling in China. Geoderma 171–172, 85–91.
Shangguan, W., Hengl, T., de Jesus, J.M., Yuan, H., Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. J. Adv. Model. Earth Syst. 9. https://doi.org/10.1002/2016MS000686.
Shiri, J., Keshavarzi, A., Kisi, O., Karimi, S., 2017. Using soil easily measured parameters for estimating soil water capacity: soft computing approaches. Comput. Electron. Agric. 141, 327–339.
Solomatine, D.P., Shrestha, D.L., 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resour. Res. 45, W00B11. https://doi.org/10.1029/2008WR006839.
Stockmann, U., Adams, M.A., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A.B., de Courcelles, V.D.R., Singh, K., Wheeler, I., Abbott, L., Angers, D.A., Baldock, J., Bird, M., Brookes, P.C., Chenu, C., Jastrow, J.D., Lal, R., Lehmann, J., O'Donnellk, A.G., Parton, W.J., Whitehead, D., Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric. Ecosyst. Environ. 164, 80–99.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.
Tifafi, M., Guenet, B., Hatte, C., 2018. Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: intercomparison and evaluation based on field data from USA, England, Wales, and France. Global Biogeochem. Cy. 32, 42–56.
USDA-NRCS, 2004. Soil survey laboratory methods manual. Soil Survey Investigations Report No. 42 (Version 4.0).
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.
Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53, 845–864.
Vitharana, U.W.A., Mishra, U., Mapa, R.B., 2019. National soil organic carbon estimates can improve global estimates. Geoderma 337, 55–64.
Wickham, H., Chang, W., Henry, L., Pedersen, L.T., Takahashi, K., Wilke, C., Woo, K., Yutani, H., 2019. ggplot2: create elegant data visualisations using the grammar of graphics. R package version 3.2.1. https://cran.r-project.org/web/packages/ggplot2.
Yang, R.M., Yang, F., Yang, F., Huang, L.M., Liu, F., Yang, J.L., Zhao, Y.G., Li, D.C., Zhang, G.L., 2017. Pedogenic knowledge-aided modelling of soil inorganic carbon stocks in an alpine environment. Sci. Total Environ. 599–600, 1445–1453.
Zhang, G.L., Gong, Z.T., 2012. Soi survey laboratory methods. Beijing: Science Press. pp. 8–23. (In Chinese).
Zhang, G.L., Wang, Q.B., Zhang, F.R., Wu, K.N., Cai, C.F., Zhang, M.K., Li, D.C., Zhao, Y.G., Yang, J.L., 2013. Criteria for establishment of soil family and soil series in Chinese Soil Taxonomy. Acta Pedol. Sin. 50 (4), 826–834 (In Chinese).