Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

Depth-to-Bedrock Map of China at a Spatial Resolution of 100 Meters

Fapeng Yan[1], Wei Shangguan[2]*, Jing Zhang[1] and Bifeng Hu[3,4,5]

[1] College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

[2] Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China.

[3] Unité de Recherche en Science du Sol, INRA, Orléans 45075, France

[4] InfoSol, INRA, US 1106, Orléans F-4075, France

[5] Sciences de la Terre et de l'Univers, Orléans University, 45067 Orleans, France

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

Abstract. Depth to bedrock serves as the lower boundary of soil, which influences or controls many of the Earth's physical and chemical processes. It plays important roles in geology, hydrology, land surface processes, civil engineering, and other related fields. This paper describes the materials and methods to produce a high-resolution (100 m) depth-to-bedrock map of China. Observations were interpreted from borehole log data (ca. 6,382 locations) sampled from the Chinese National Important Geological Borehole Database. To fill in large sampling gaps, additional pseudo-observations generated based on expert knowledge were added. Then, we overlaid the training points on a stack of 133 covariates including climatic images, DEM-derived parameters, land-cover and land-use maps, MODIS surface reflectance bands, vegetation index images, and the Harmonized World Soil Database. Spatial prediction models were developed using the random forests and gradient boosting tree, and ensemble prediction results were then obtained by these two independently fitted models. Finally, uncertainty estimation was generated by the quantile regression forest model. The 10-fold cross-validation showed that the ensemble models explain 57% of the variation in depth to bedrock. Based on comparison with depth-to-bedrock maps of China extracted from previous global predictions, our predictions showed higher accuracy. More observations, especially those in data-sparse areas, should be added to training data, and more covariates with high precision should be used to further improve the accuracy of spatial predictions. The resulting maps of this study are available on Figshare at the following DOI: https://doi.org/10.6084/m9.figshare.7011524.v1. And they are also available for download at http://globalchange.bnu.edu.cn/.

1 Introduction

Soil is the loose layer on the surface of the geosphere. It is the foundation of the whole terrestrial ecosystem (van Breemen and Buurman, 2002). The International Union of Soil Sciences (IUSS) divides the soil profile into six main genetic horizons: O (organic horizon), A (humus horizon), E (eluvial horizon), B (illuvial horizon), C (parent rock horizon), and R (hard rock). Of these, the bedrock (i.e., the R horizon) is the consolidated solid rock underlying unconsolidated surface materials, such as soil or other regolith (Jain, 2014). Depth to bedrock (DTB) is the depth to the R horizon, which is equivalent to the total thickness of the solum and weathered rocks; DTB controls or influences many physical and chemical processes of the Earth (Jain, 2014).

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

DTB information plays an important role in many fields of Earth system science. In geology, DBT has been used for applications such as mineral exploration, earthquake modeling, and landslide risk assessment (Schenk and Jackson, 2005; Fan et al., 2013). In land surface modeling, DTB is an important input parameter that affects the energy, water, and carbon cycles. However, in most land surface models, DTB has been set as a constant value because of a lack of data, which limits the performance of land surface modeling (Gochis et al., 2010). DTB Information is also indispensable to civil engineering in building homes, roads, railways, and bridges (Price, 2009). Furthermore, DTB is of great importance to the study and applications of hydrology, ecology, agriculture, and other relevant fields (Tromp-van Meerveld et al., 2007; Fu et al., 2011).

Although DTB is often considered equal to the thickness of the soil, there are great differences between different measurement results. Soil thickness is mostly determined based on soil profiles from soil surveys and borehole profiles from geological surveys. The observed depth of a soil profile is generally less than 2 meters, and the thickness of the soil is therefore recorded as a value lower than 2 meters (Shangguan et al., 2017). However, in reality, the DTB (the depth to the R horizon) ranges from 0 meters to more than 1 kilometer, which is much greater than the average depth of soil profiles. Limited by external factors such as equipment and technological constraints, traditional soil surveys cannot reach bedrock in most cases. However, in contrast to traditional soil surveys, geological borehole drillings usually reach depths of hundreds of meters or even deeper, and most boreholes reach bedrock. Thus, borehole drilling logs are the most effective sources of DTB data.

Ground observations of DTB, which include soil profiles from soil surveying and borehole drilling log data such as water well records and other measurements, have been widely used as training data to produce spatial predictions of DTB (Tesfa et al., 2009; Shafuque et al., 2011; Miller and White, 1998; Hengl et al., 2014; Shangguan et al., 2017). Various mapping methods, which include physically based models, interpolation from samples, and empirical-statistical models (Kuriakose et al., 2009), have been employed for this purpose. Pelletier and Rasmussen (2009) proposed a geomorphically based model that uses digital elevation model data to predict soil thicknesses based on a hypothesis that there is a long-term balance between soil production and erosion. Karlsson et al. (2013) developed a simplified regolith model modified from a trigonometric approach to estimate regolith thickness based on slopes, outcrops, and distance to outcrops in eight directions, and compared the results with those of linear regression and inverse distance weighting interpolation.

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

Shafique et al. (2011) proposed a multivariate linear model based on elevation, landform, and distance to stream information to predict regolith thickness in a data-sparse environment. Hengl et al. (2014) used zero-inflated models to predict global depth to bedrock based on a compilation of major international soil profile databases. Dahlke et al. (2009) used a soil landscape model to predict soil depth based on class means of merged spatial explanatory variables. Tesfa et al. (2009) applied generalized additive and random forest models based on topographic and land-cover attributes to predict soil depth at the watershed scale. Shangguan et al. (2017) predicted global depth to bedrock using the random forest and gradient boosting tree models. Based on previous studies, machine learning methods, especially random forest (RF) and gradient boosting tree (GBT) methods, showed better performance than traditional interpolation methods under normal circumstances, and are available in the "*randomForest*" (Breiman, 2001) and "*xgboost*" (Chen et al., 2016) packages in the R software.

Although information about DTB is very important, to date, information about DTB in China is very deficient, and there is no independent map of depth to bedrock in China. However, researchers have advanced toward this target. Globally, there are several existing maps of DTB covering the area of China (FAO, 1996; Hengl et al., 2014; Pelletier et al., 2016, Shangguan et al., 2017). The earliest global distribution of DTB was produced by the FAO (Food and Agriculture Organization) (1996); the depth was limited to the uppermost 2 meters and mapped using expert rules, and was primarily based on soil unit classification, soil phase, and slope class. Hengl et al. (2014) developed a global depth-to-bedrock map at 1-km resolution based on zero-inflated models using a compilation of major international soil profile databases and 75 global environmental covariates representing soil-forming factors. Pelletier et al. (2016) produced a global data set of the average thicknesses of soil, intact regolith, and sedimentary deposits by representing uplands using soil data and lowlands using water well data, with topographic, climatic, and geological data used as input. In China, Shangguan et al. (2013) developed a comprehensive 30 × 30 arc-second resolution gridded data set of soil characteristics that included soil depth derived from soil profiles and the Soil Map of China (1:1,000,000), but the soil-depth data quality was relatively low because there were fewer observations of deep soil. In addition, Shangguan et al. (2017) produced another global map of depth to bedrock based on machine learning, using soil profile data, borehole data, and pseudo-observations.

Among above-mentioned maps of DTB, most have relatively coarse resolutions (1 km or coarser), except the map produced by Shangguan et al. (2017) (250 m resolution). In addition, observations of DTB (FAO, 1996; Shangguan et al., 2013; Hengl et al., 2014) have been based solely on soil data; thus, the predictions are often limited to soil surfaces with depths limited to several meters. This depth is not consistent with the actual distribution of DTB. In addition, most samples (Pelletier et al., 2016; Shangguan et al., 2017) were located in North America, whereas no samples or only a small number of samples were located in China, which resulted in high uncertainty for predictions in China. However, a large number of borehole logs produced by geologists in China provide DTB information and are now available. In addition, several environmental covariates with high resolution have been produced, which can be used to produce a high-resolution DTB map of China. These data sources provide the cornerstone for producing a new map of DTB with higher accuracy and resolution.

In this study, we aim to estimate DTB in China using machine learning methods. Observations interpreted from geological borehole profiles and pseudo-observations of DTB are used as training points. An extensive list of remote-sensing-based covariates, including DEM-derived parameters, climatic images, MODIS products, land cover/land use, and the latest lithological/soil maps of China are used as covariates. The objective of this paper is to (1) produce a DTB map of China at a high spatial resolution of 100 meters; (2) compare and evaluate this map with observations and existing DTB maps; and (3) estimate the uncertainty of the DTB map and discuss the outlook for generating more accurate DTB maps in the future.

2 Materials and methods

2.1 Borehole data

A total of 6,382 borehole logs sampled from the Chinese National Important Geological Borehole Database (NIGBD http://zkinfo.cgsi.cn) were used in our study. The NIGBD comprises about 80 million boreholes from across China (except Taiwan province). In every borehole log, geographic coordinates and detailed lithological records are provided in the form of scanned images. Therefore, the DTB of each borehole can be interpreted by finding the boundary between the regolith and fresh bedrock.

2.1.1 Observations sampled from the NIGBD

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

The DTB of every borehole must be interpreted manually, and interpreting more than 80 million boreholes logs therefore demands an immense amount of work and has high costs. However, many boreholes that are located close to each other have similar DTB and environmental factors. Therefore, we developed a sampling scheme to take a fraction of borehole drillings from the NIGBD as the observation data sets in this study. Mapping methods, regardless of methods based on spatial autocorrelation or soil environmental correlation, have requirements based on the number, distribution, and typicality of the samples, which ensure global representation of the samples (Zhang et al., 2012). To obtain representative samples from these boreholes, we used a sampling scheme similar to stratified sampling to acquire our training points from the NIGBD.

The stratified sampling scheme includes designation of grid shape (such as a square grid, triangular grid, or hexagonal grid) and grid size. A square grid is the easiest and most effective, and is most widely used in sampling (Zhang et al., 2012). In general, smaller grid size leads to more accurate predictions, but with greater sampling costs. Here, we used square grid sampling with a 0.2 × 0.2 arc-degree grid, in consideration of the balance between representativeness and cost. Usually, one observation or a number of observations are sampled at random locations from each grid. However, the locations of boreholes in this study were determined in a previous geological survey. Thus, we have taken one borehole randomly from each grid instead of one borehole from a random location.

The depths of the boreholes range from 0 meters to more than 1 kilometer. Among these boreholes, we were unable to determine the DTB from a few boreholes because of the limitations of the records (see details in Sect. 2.1.2). This constraint resulted in vacancies of many grid cells after the interpretation of all boreholes from the first sampling. To resolve this problem, we used an additive sampling method; that is, additional samplings were taken multiple times until no new observations could be added to the observation data sets. Thus, the latter samplings were aimed at grids without DTB data based on the previous samplings. After a finite number of additive samplings, the borehole logs of the NIGBD were considered efficiently used, and samples from all the samplings were used in our study. The distribution of DTB observations interpreted from boreholes is shown in Fig. 1.

2.1.2 Interpretation of borehole records

Interpreting DTB from borehole profiles sampled from the NIGBD was one of the crucial aspects of this study. Borehole profiles, which were previously recorded by geologists, have longitudinal verbal descriptions of soil layers and lithological layers with corresponding depths from the land surface to the top and bottom of each layer. A typical simplified borehole profile diagram is shown in Fig. 2.

Each borehole profile has several layers. Generally, the top layer of a borehole profile is pedolith, where pedological processes have destroyed the original bedrock structure, principally through the weathering of primary bedrock minerals and the formation and re-distribution of secondary materials (National Committee on Soil and Terrain, 2009). Below is saprolite, referring to the zone where the bedrock fabric is largely isovolumetrically weathered but primary bedrock structures are still recognized. At the bottom is the unweathered bedrock. Because different boreholes were drilled by different geological teams at different times, the details of stratification in the profiles often differ, and the lithological description of each layer may be detailed or vague. These differences result in inconsistencies or uncertainties in the borehole database, which were propagated into our DTB observations.

To interpret the DTB from a borehole profile in the form of a scanned picture, we must manually determine the boundary between the regolith and fresh bedrock based on lithological descriptions and the dip angle of the borehole. The dip angles of a minority of boreholes whose dip angle were not given were about 90°. Then, the DTB was calculated as the product of boundary depth and sine of the dip angle. DTB can be interpreted from most sampled boreholes. However, some boreholes are too shallow (several meters or less than 1 m) to reach the bedrock, and some have lithological records that are unclear, which can make it is very difficult to determine the DTB (as described in Sect. 2.1.1). Therefore, we used additive samplings. Because a number of boreholes went to depths of more than 100 meters but still did not reach the bedrock, we could not obtain accurate DTB data from these borehole profiles either. In this case, we regarded the depths of those boreholes as approximations of the real DTB value. In addition, most research and applications focus on relatively shallow depths.

2.2 Pseudo-observations

As shown in Fig. 1, DTB observations interpreted from borehole logs cover an extensive area across China, except for the Qinghai-Tibet Plateau where boreholes are difficult to drill. Any purely data-driven model fitted with large gaps in the covariate space is most likely to result in considerable omissions, especially for areas that are often inaccessible or not of interest to soil surveys or geological exploration. Therefore, we used pseudo-observations added to training data to fill such gaps, which will avoid extrapolation for these areas (e.g., deserts and steep mountainous areas). Deserts consist mainly of sand, and the DTB of such areas could be found in some publications. Steep-slope areas without vegetation typically have very shallow or zero DTB; that is, rock outcrop. Therefore, we used the following data sources to generate pseudo-observations to add to the training points:

(1) The distribution map of deserts in China from the Data Center of Environmental and Ecological Science in Western China (http://zgsm.westgis.ac.cn).

(2) Steep, bare surface areas generated using a slope map of China and remote-sensing-based data.

(3) Previously published detailed geological maps reporting DTB or bedrock outcrops.

We generated a certain number of points in random positions within deserts based on the distribution map of China's deserts. The DTB values of these points were obtained from existing material and previous studies of the sand thickness of the deserts. We must note that the number of points was limited to less than 10% of the whole number of observations to prevent adding too many soft observations, and we only used points whose values had high credibility. In addition, several points located in high-slope areas (> 60°) were added to the observations with DTB values that varied between 0 and 0.1 m.

2.3 Environmental covariates

In our study, a total of 133 related environmental layers, which cover five types of factors (climate, topography, living organisms, water dynamics, and parent material) and represent the factors of soil formation according to Jenny (1994), were selected to generate a DTB map of China. These predictors were generalized into seven predictive "*scorpan*" factors (McBratney et al., 2003). The 133 covariates classified as "*scorpan*" factors included:

(1) Harmonized soil database images: percent coverage of Andosols, Histosols, and dozens of other soil types.

(2) Climatic images: images indicating the values of 8-day MODIS day-time and night-time local standard time (LST), long-term and monthly precipitation data, etc.

(3) Land use and land cover images: including vegetation maps, land cover and land use

classifications, biomass and yield maps, etc.

(4) Relief data, mainly derived from digital elevation models: slope maps, the topographic wetness index, the topographic openness index, physiographic landform units, elevation and secondary terrain attributes, etc.

(5) Geological and parent material maps: geological ages based on surface geology.

The complete list of the 133 environmental covariates is given in Supplement File A.

2.4 Spatial prediction model

The framework of our research is shown in Fig. 3. This framework consists of four main processes:

1. Overlaying observations of DTB and covariates to generate a regression matrix for modeling;
2. Obtaining the best parameters for modeling using cross-validation;
3. Fitting the prediction models based on the whole regression matrix;
4. Applying spatial prediction models using covariates and comparing the prediction with existing maps.

2.4.1 Model fitting

In this study, we overlaid observations of DTB and covariates under the same coordinate reference to generate a matrix including DTB and covariate columns. The matrix was used as input data for machine learning. Then, we separately used RF and GBT to fit the prediction models. Finally, the spatial predictions were generated using an ensemble model based on the two models. RF and GBT are decision-tree-based ensemble methods. The RF model uses fully grown decision trees and reduces error by reducing variance (Breiman, 2001). The GBT model uses shallow trees and reduces error mainly by reducing bias, and to some extent by reducing variance by aggregating the outputs from many models (Chen and Guestrin, 2016). RF and GBT were implemented respectively in the "*randomForest*" and "*xgboost*" packages in the R environment. Parallel computing was employed to improve data processing efficiency.

2.4.2 Model validation and evaluation

Ten-fold cross-validation was used to evaluate prediction accuracy. Comparison with previously existing DTB maps was then employed to evaluate our results.

In cross validation, samples were divided into a training set (5,740 samples) and validation set (642 samples). The training set was used to fit models, and the validation set was used to validate model performance. Some widely used indicators such as the coefficient of determination (R^2 or the

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

amount of variation explained by the model), mean error (ME), and root mean square error (RMSE) were used to evaluate model performance. Of these indicators, the coefficient of determination is calculated by:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

where SSR is the regression sum of squares, SST is the total variation sum of squares, and SSE is the residual sum of squares, which is the difference of SST and SSR. The variable y_i is the measured target value, $\hat{y}$ is the prediction of each point, $\bar{y}$ is the average of the measurements, and n is number of validation points. The value of R^2 is usually between 0 and 1; a value close to 1 indicates a perfect model, and values around 0 indicate a failed model. The RMSE, which is also called standard error, is calculated by:

$$\mathrm{RMSE} = \sqrt{MSE} = \sqrt{SSE / n}, \tag{2}$$

where MSE is the mean squared error. RMSE estimates the deviation between predictions and observed values. A smaller RMSE indicates a better prediction.

Different covariates have different importance to DTB. Covariates with no or weak relations with DTB may produce noise in fitted models. This noise results in higher error of predictions. Our results based on modeling with different covariates showed that the noise has a certain degree of influence on the accuracy of the models, especially for the gradient boosting tree model. Therefore, we removed some covariates with low importance based on the random forests model to reduce prediction errors. The covariates we ultimately used are marked in Supplement File A.

In addition, to verify whether our predictions are more accurate than existing DTB maps of China, we compared our predictions with existing DTB maps using the validation set.

2.4.3 Model prediction and uncertainty estimation

The final model was fitted based on all samples with parameters selected by cross-validation. The final spatial predictions were generated using an ensemble model based on random forests and the gradient boosting tree method, which can avoid the overshooting effect (Sollich and Krogh, 1996). To predict DTB in China at 100 m resolution, we used the available environmental covariates at 100 m resolution.

Because any model for digital soil mapping inevitably suffers from different sources of error, it is important to quantify the uncertainty associated with the produced maps (Poggio et al., 2016). Analyzing and evaluating help data users to understand its existence and also can help to improve decision quality (Liang et al., 2018). In this study, we used quantile regression forests to estimate the uncertainty of estimations. Quantile regression forests are a tree-based ensemble algorithm for estimation of conditional quantiles. This method is particularly suitable for high-dimensional data. Quantile regression forests were implemented via the R environment in the "*quantregForest*" package (Meinshausen, 2014). To estimate the uncertainty of predictions at every location, we generated the uncertainty map of predictions by:

$$uncertainty = \frac{qp_{0.9} - qp_{0.1}}{qp_{0.5}}$$

where $qp_{0.9}$ is the 0.9 quantile prediction of DTB, $qp_{0.1}$ is the 0.1 quantile prediction of DTB, and $qp_{0.5}$ is the 0.5 quantile prediction of DTB. The uncertainty map is the reference when using the DTB map of China.

All code used to generate predictions is available from the Github channels (https://github.com/yanfp/DTB100China).

3 Results

3.1 Model input statistical summary

A summary of the DTB statistics is provided in Table 1. The DTB ranged from 0 to 1,106.91 m, with a mean DTB of 36.62 m and a median value of 8.24 m. Fig. 4(a) shows the histogram of DTB within 100 m. The DTB after logarithmic transformation had a distribution similar to a normal distribution but with many zero values (i.e., outcrops) (Fig. 4(b)).

3.2 Model accuracy and variable importance

As is shown in Table 2, the GBT model had good ability to estimate DTB and yielded relatively higher R^2 (0.81) and lower RMSE than the RF model (Table 2) based on the training set.

The importance of covariates measured based on the residual sum of squares of the random forests model is shown in Fig. 5. The four most important covariates for DTB in this study were the topographic wetness index, physiographic landform units, the topographic openness index, and slope. In contrast, the most important covariate for the DTB according to Pelletier et al. (2016) and Shangguan et al. (2017) was precipitation. The relationships between DTB and four important covariates are shown in Fig. 6. This figure shows that DTB had a positive correlation with the topographic wetness index. The topographic wetness index is a secondary terrain attribute related to the geomorphometry of the surface or landform classification. In addition, DTB showed a positive correlation with the topographic openness index and elevation, and a negative correlation with the slope. These relations are consistent with our knowledge about DTB.

3.3 Estimation accuracy

The cross-validation summary statistics of interpolation for models based on RF and GBT are shown in Table 3 and Fig. 7. These statistics show that RF produced more accurate estimations than GBT. Because the GBT model showed relatively higher R^2 and lower RMSE than the RF model based on the training set (Sect. 3.2), this result means that the GBT model had a large degree of overfitting. Our results showed significant overestimation in lower values of DTB, which is a common problem in regression, especially when the model is not able to explain > 50% of variability in the target variable (Shangguan et al., 2017).

3.4 Prediction results

Output estimations of DTB by the ensemble model based on RF and GBT at 100 meters resolution are shown in Fig. 8. Our estimated results reveal that the predicted mean DTB was 54.42 m. High values of DTB were mainly distributed in desert areas, the North China Plain (including areas in Hebei province, Henan province, and Jiangsu province) and the Northeast China Plain (including areas in Heilongjiang province, Jilin province, and Liaoning province). Relatively lower values of DTB were mainly located in hilly and mountainous areas, such as Sichuan province, Chongqing city, Guangxi province, and the mountainous areas of Northeast China. The spatial pattern of the DTB map of this study is similar to those of the maps produced by Pelletier et al. (2016) and Shangguan et al. (2017).

In addition, estimations of three percentiles (0.1 (Fig. 9(a)), 0.50 (Fig. 9(b), and 0.9 (Fig. 9(c)) were produced by the quantile regression forests model. The mean values of the estimated DTB for the three percentiles were 4.95 m, 31.22 m, and 99.56 m, respectively. The maps show that the spatial pattern of DTB predicted by the quantile regression forests model was similar to that of the ensemble model based on the random forest and gradient boosting tree methods.

The uncertainty map of the prediction of DTB is shown in Fig. 10. The uncertainty in the predictions in part depends on the density of sampling (Zhou et al., 2018). In our study it was low in deserts, sandy areas, the North China Plain, and the Northeast China Plain, where the topography is relatively simple and sampling was relatively dense. In the Tibetan Plateau and western Inner Mongolia, where sampling was sparse and DTB is low, the uncertainty was high. The uncertainty was also relatively high in the Yun-Gui Plateau where the topography is complex with widespread karst landforms.

3.5 Comparison with existing study results

We compared our results with existing maps produced by Pelletier et al. (2016) and Shangguan et al (2017). Our results show similar spatial patterns with these maps. Of course, DTB values in deserts, sandy areas, and the North China Plain were relatively high, and values in hilly and mountainous areas, such as Chongqing City and Yunnan province, were relatively low in the map of this study and in maps from global predictions. The estimated mean DTB was 54.42 m in our study, whereas the mean values predicted by Pelletier et al. (2016) (Fig. 11 (a)) and Shangguan et al. (2017) (Fig. 11 (b)) were 11.81 m and 26.64 m. The correlation coefficient between DTB observations and predictions in our study is 0.75, which is significantly higher than the estimation results of Pelletier et al. (2016) and Shangguan et al. (2017) (Table 4). In addition, compared with the prediction results of Pelletier et al. (2016) and Shangguan et al. (2017), our estimation results had obviously lower RMSE (47.57) and ME (1.82).

In addition, our prediction result shows similar spatial patterns to the maps produced by Pelletier et al. (2016) and Shangguan et al. (2017), but revealed more detailed information than previous predictions. There are more jumping points in the map of Shangguan et al. (2017) than others, and the map predicted by Pelletier et al. (2016) shows low continuity in space with high values and low values in a wide range. From the comparison in a typical region in the North China Plain (Fig. 12), our map revealed more spatial details, especially in high DTB areas, than the maps by Shangguan et al. (2017) and Pelletier et al. (2016) (Fig. 12(a)). In contrast, the map estimated by Pelletier et al. (2016) shows abrupt change between highland and lowland areas (Fig. 12(c)).

4 Data availability

The resulting maps are available on Figshare at the following DOI: https://doi.org/10.6084/m9.figshare.7011524.v1. And they are also available for download at http://globalchange.bnu.edu.cn/.

5 Discussion

5.1 Success and limitations of the data set

Our training observations were selected by using square grid sampling with a 0.2 × 0.2 arc-degree grid. We sampled at least one observation within each grid cell. Under this condition, the training data are most representative under the current sampling method, which will produce the most accurate predictions. However, boreholes have uneven spatial distribution. Very few boreholes were located in inhospitable areas such as deserts and mountainous areas (Fig. 13). In addition, we were unable to interpret the DTB from some borehole profiles. These limitations resulted in vacancies of observations in many grid cells. Lack of observations will increase the uncertainty of predictions in these areas.

The reliability of training data and covariates together determines the accuracy of predictions. Although observations in this study were less heavily distributed in western China, which may limit the accuracy of our predictions, the number of observations in China is far greater than that in other studies. In addition, the DTB values interpreted from borehole profiles were more accurate than those from soil profiles. Therefore, the DTB maps produced from borehole profiles were also more accurate than maps solely based on soil profiles, especially for deep-DTB areas. In addition, the predictions show a higher correlation coefficient with observations than did previous DTB maps based on the validation set. The amount of variation explained by models for the DTB is about 57%, which means that more than half of the variation is explained. We produced the DTB maps of China at a resolution of 100 m. Although only a few covariates had spatial resolution of 100 m because of the lack of data, the spatial resolution of most covariates was about 1 kilometer. Thus, spatial variation at 100-meter scale may not be fully explained. However, covariates with high correlation with DTB, such as DEM-derived parameters and land cover, have high resolutions (Fig. 5 and 6). More observations and more covariates with high precision should be used in the future to improve prediction accuracy.

5.2 Error from interpretation of borehole records

As described above, the DTB observations were visually interpreted from every borehole profile. Because different borehole profiles were mapped by different organizations, the basis of layer stratification differed slightly for different profiles. This issue contributes to the disunity of DTB observations. In addition, the level of detail for different borehole profile stratifications is discrepant because of their original uses. Furthermore, lithological records of some borehole profiles that give vague information about soil and lithology were not distinct enough for us to interpret the DTB accurately. All these factors contributed to errors in our DTB observations.

5.3 Models built from different topographic partitions

The DTB was determined based on many covariates including factors of topography, climate, geology, vegetation, age, and human activity. Soils at the surface of the Earth are formed under the combined effects of those factors (Zhou et al., 2016). However, the mechanisms of soil formation and the importance of each covariate still are not completely clear (Li et al., 2004). The most important covariates related to the DTB may be different in different geographic partitions. Therefore, a model based on observations over the whole area of China may not be able to capture the major factors in some regional areas. Models built from regional partitions may produce more accurate predictions than global models within the partitions. Pelletier et al. (2016) distinguished global land surfaces into three landform components, upland hillslope, upland valley bottom, and lowland, and used different models for each component to estimate the DTB. Peng et al. (2018) divided training data into subsets according to the similarity of the predicted variables and attain the independent prediction model, which improved the prediction accuracy. In the future, different models should be built and spatial predictions should be applied separately in different topographic partitions.

6 Conclusions

In this study, we demonstrated the use of an ensemble model to produce a DTB map of China at a resolution of 100 meters using the most reliable ground observations of DTB interpreted from borehole profiles. This study provides the final prediction map of DTB as well as an uncertainty estimation map for China. The cross-validation showed that the R^2 of the ensemble model was 0.57, and the comparison showed that our DTB map is more accurate than existing DTB maps. Even though the shortage of data used in this study, including DTB observations and environmental covariates, limited the precision of the DTB map at a scale of 100 meters, this data set provides more accurate information for Earth system researches compared with previous maps of DTB. Based on the spatial prediction framework, data processing, model fitting, and spatial prediction are

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

fully automated and can be updated easily. By adding more DTB observations and using more accurate covariates, we will be able to produce more accurate DTB maps of China in the future.

Author contributions.

Wei Shangguan, Jing Zhang, and Fapeng Yan designed the experiment and control the planning. Fapeng Yan collected and compiled DTB observation data, prepared a part of environmental covariate data, built models and implemented the spatial prediction, and wrote the paper. Wei Shangguan prepared the other part of environmental covariate data. Bifeng Hu contributed to the process on the data compilation and data validation. Jing Zhang initiated and coordinated the work. All authors contributed to the scientific discussion of the results, the editing, and revision of the paper.

Competing interests.

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of China under grants 41575072 and the National Key Research and Development Program of China under grants 2017YFA0604303.

References

Boer, M., Barrio, G. D., and Puigdefiibregas, J.: Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model, Geoderma, 72, 99-118, 1996.

Breemen;, N. v., and Buurman, P.: Soil formation, Springer Science & Business Media, The Netherlands, 2002.

Breiman, L.: Random forests, Machine learning, 45, 5-32, 2001.

Chen, J., Li, M., and Wang, W.: Statistical Uncertainty Estimation Using Random Forests and Its Application to Drought Forecast, Mathematical Problems in Engineering, 2012, 1-12, 10.1155/2012/915053, 2012.

Chen, T., and Guestrin, C.: Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794, 10.1145/2939672.2939785, 2016.

Dahlke, H. E., Behrens, T., Seibert, J., and Andersson, L.: Test of statistical means for the extrapolation of soil depth point information using overlays of spatial environmental data and bootstrapping techniques, Hydrological Processes, 23, 3017-3029, 10.1002/hyp.7413, 2009a.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G., Walsh, M. G., and Gonzalez, M. R.: SoilGrids1km--global soil information based on automated mapping, PLoS One, 9, e105992, 10.1371/journal.pone.0105992, 2014.

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

Jain, S.: Fundamentals of Physical Geology, Springer, New Delhi, 2014.

Jie, P., Asim, B., Qingsong, J., Ruiying, Z., Jie, H., Bifeng, H., Zhou, S., Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China, https://doi.org/10.1016/j.geoderma.2018.08.006, 2018.

Karlsson, C. S. J., Jamali, I. A., Earon, R., Olofsson, B., and Mörtberg, U.: Comparison of methods for predicting regolith thickness in previously glaciated terrain, Stockholm, Sweden, Geoderma, 226-227, 116-129, 10.1016/j.geoderma.2014.03.003, 2014.

Kuriakose, S. L., Devkota, S., Rossiter, D. G., and Jetten, V. G.: Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the Western Ghats of Kerala, India, Catena, 79, 27-38, 10.1016/j.catena.2009.05.005, 2009.

Li, T., Zhao, Y., Zhang, K., Zheng, Y., and Wang, Y.: Soil geography, China, 2004.

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3-52, 10.1016/s0016-7061(03)00223-4, 2003.

Meinshausen, N.: Quantregforest: quantile regression forests, R package, 2014.

Miller, D. A., and White, R. A.: A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling, American Meteorological Society, 2, 1-26, 1998.

Osborne, J. W.: Improving your data transformations: Applying the Box-Cox transformation, Practical Assessment, Research & Evaluation, 15, 2, 2010.

Pelletier, J. D., and Rasmussen, C.: Geomorphically based predictive mapping of soil thickness in upland watersheds, Water Resources Research, 45, 10.1029/2008wr007319, 2009.

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., Williams, Z., Brunke, M. A., and Gochis, D.: A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, Journal of Advances in Modeling Earth Systems, 8, 41-65, 10.1002/2015ms000526, 2016.

Piskin, K., and Bergstorm, R. E.: Glacial Drift in Illinois Thickness and Character, 1975.

Poggio, L., Gimona, A., Spezia, L., and Brewer, M. J.: Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA, Geoderma, 277, 69-82, 10.1016/j.geoderma.2016.04.026, 2016.

Richard, S. M., Shipman, T. C., Greene, L. C., and Harris, R. C.: Estimated Depth to Bedrock in Arizona, Arizona Geological Survey, Digital Geologic Map DGM-52, layout scale, 1, 2007.

Roesset, J. M., Stokoe II, K. H., and Seng, C.-R.: Determination of Depth to Bedrock from Falling Weight Deflectometer Test Data, Transportation Research Record, 1995.

Rue, H., and Martino, S.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Royal Statistical Society, 71, 319-392, 2009.

Shafique, M., der Meijde, M. v., and Rossiter, D. G.: Geophysical and remote sensing-based approach to model regolith thickness in a data-sparse environment, Catena, 87, 11-19, 10.1016/j.catena.2011.04.004, 2011a.

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

Shangguan, W., Hengl, T., Jesus, J. S. M. d., and Dai, Y.: Mapping the global depth to bedrock for land surface modeling, Advances in Modeling Earth Systems, 9, 65-88, 2016.

Sollich, P., and Krogh', A.: Learning with ensembles: How over-fitting can be useful, Advances in Neural Information Processing Systems, 190-196, 1995.

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., and McNamara, J. P.: Modeling soil depth from topographic and land cover attributes, Water Resources Research, 45, 10.1029/2008wr007474, 2009a.

Wilford, J.: A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis, Geoderma, 183, 124-142, 10.1016/j.geoderma.2010.12.022, 2012.

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

Y. Zhou, R. Webster, R.A. Viscarra Rossel, Z. Shi, and Chen, S.: Baseline map of soil organic carbon in Tibet and its uncertainty in the 1980s, Geoderma, 334, 124-133, 2018.

Yin Z., Asim B., Zhiqiang M., Yanli L., Qiuxiao C., and Shi, Z.: Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain, Geoderma, 271, 71-79, 2016.

Zhang, S., Zhu, A., Liu, j., and Yang, L.: Summarization of Digital Soil Attribute Mapping Methods and Sample Design Based on Samples, Soils, 44, 917-923, 2012.

Zongzheng, L., Songchao C., Yuanyuan Y., Ruiying Z., Zhou S., and Rossel, R. A. V.: Baseline map of soil organic matter in China and its associated uncertainty, Geoderma, 335, 47-56, 2018.

Table 1: Summary statistics of depth to bedrock in meters

DTB	Number
=0	1026
0~2.00	585
2.00~10.00	1833
10.00~50.00	1768
50.00~100.00	427
100.00~300.00	630
>300.00	113

Table 2: Model fitting results for the depth to bedrock.

Model	Unit	R^2	RMSE	ME
Random forests	M	0.575	47.48	1.75
Gradient boosting tree	M	0.811	31.43	2.13

Table 3: Mapping performance for the depth to bedrock.

	Unit	R^2	RMSE	ME
Random forests	M	0.573	47.57	1.82
Gradient boosting tree	M	0.547	49.53	2.18
Ensemble	M	0.566	48.57	2.50

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

Table 4: Correlation index between observations and predictions

Study	Unit	R	RMSE	ME
This study	m	0.752	47.57	1.82
Pellertier et al. (2016)	m	0.486	81.98	36.52
Shangguan et al. (2017)	m	0.475	67.32	14.71

R denotes the correlation coefficient

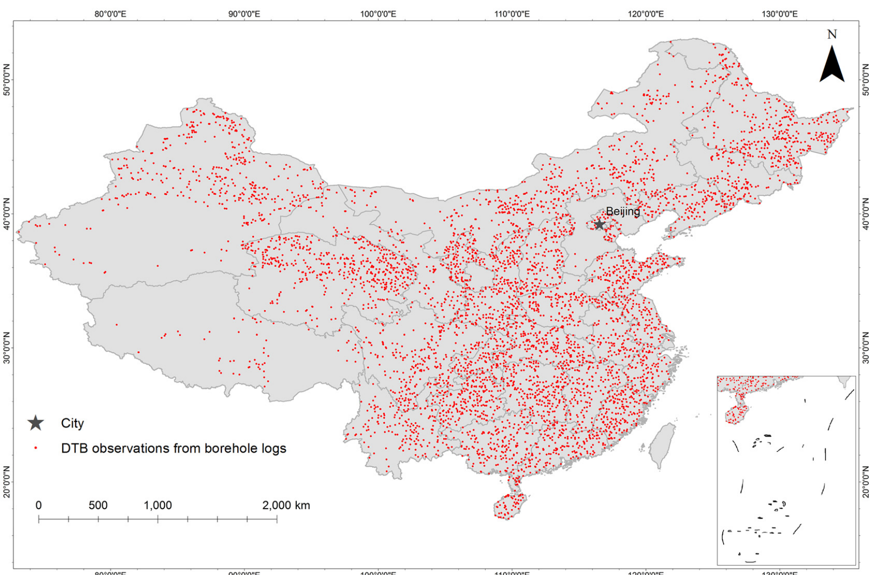

Figure 1: Distribution of DTB observations interpreted from boreholes.

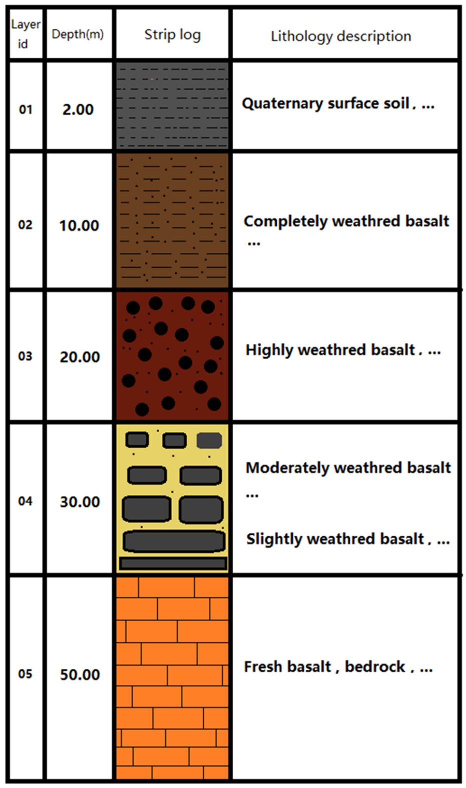

Figure 2: A typical borehole log sketch column. A borehole log describes the materials, color, and composition of each layer, and provides the depth, dip, and other relevant information. The original logs are in Chinese.

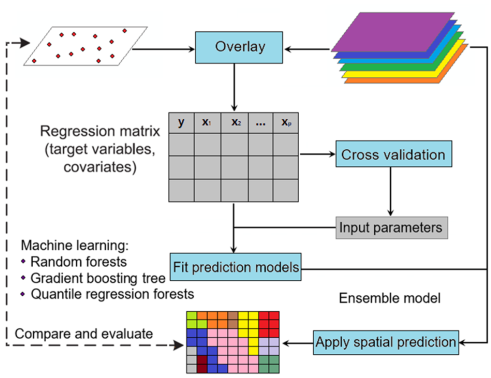

Figure 3: The spatial prediction framework used to fit models and apply spatial prediction of DTB in China at 100 m resolution.

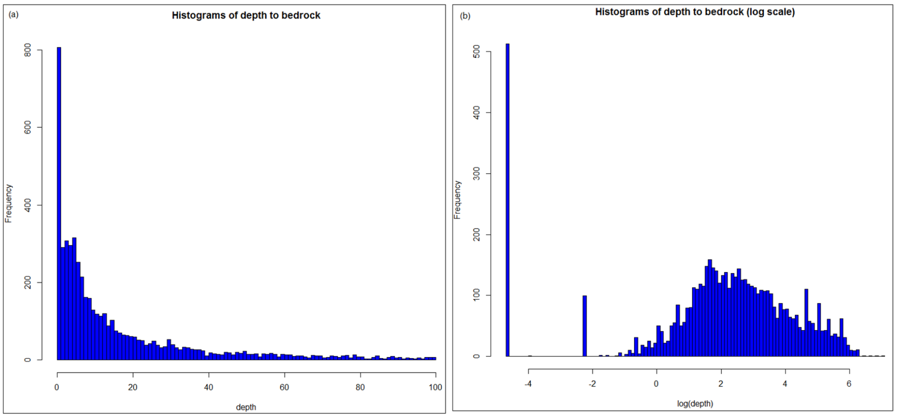

Figure 4: Histogram of depth to bedrock (a) and (b) after logarithmic transformation (values large than 100 m are not shown).

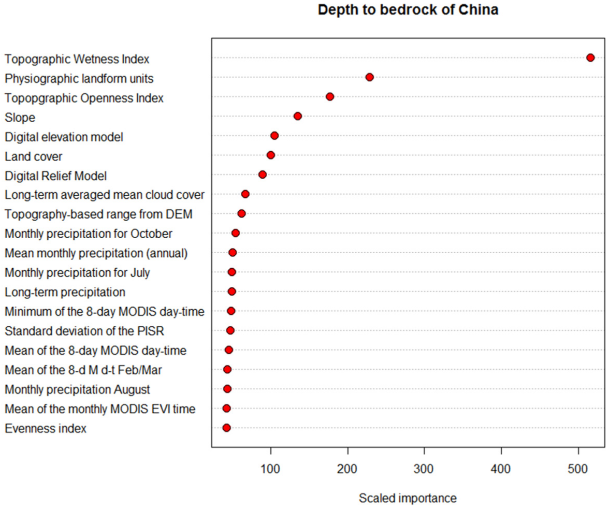

Figure 5: Importance of covariates for the depth to bedrock based on the random forest model.

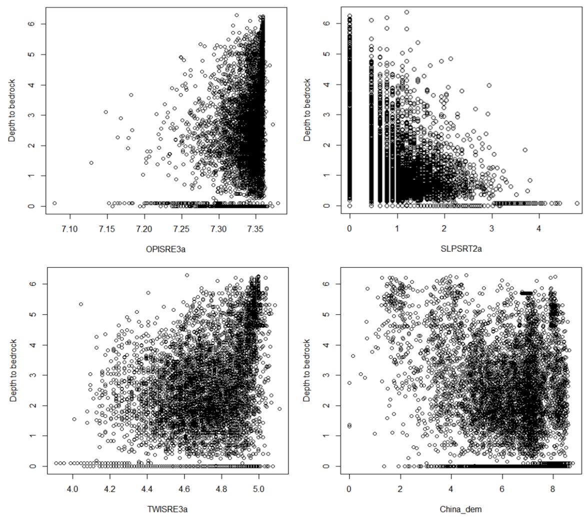

Figure 6: Relationships for target variables and the most important covariates (logarithmic scale). (*TWISRE3a* is the SAGA Topographic Wetness Index; *SLPSRT2a* is a slope map in percent; *OPISRE3a* is the SAGA Topographic Openness Index; *China_dem* is a digital elevation model of China.)

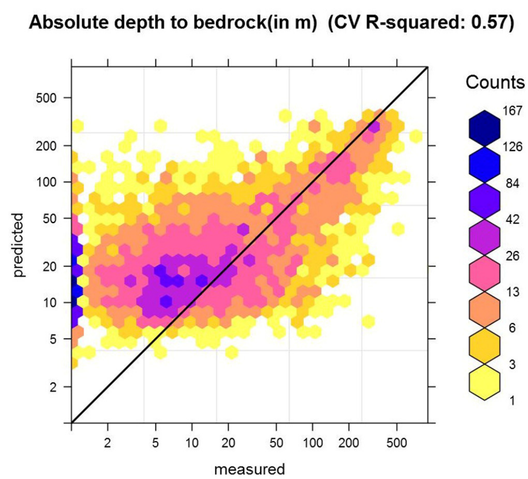

Figure 7: Plot showing cross-validation results for depth to bedrock on a logarithmic scale; R^2 is calculated using Eq. (1).

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

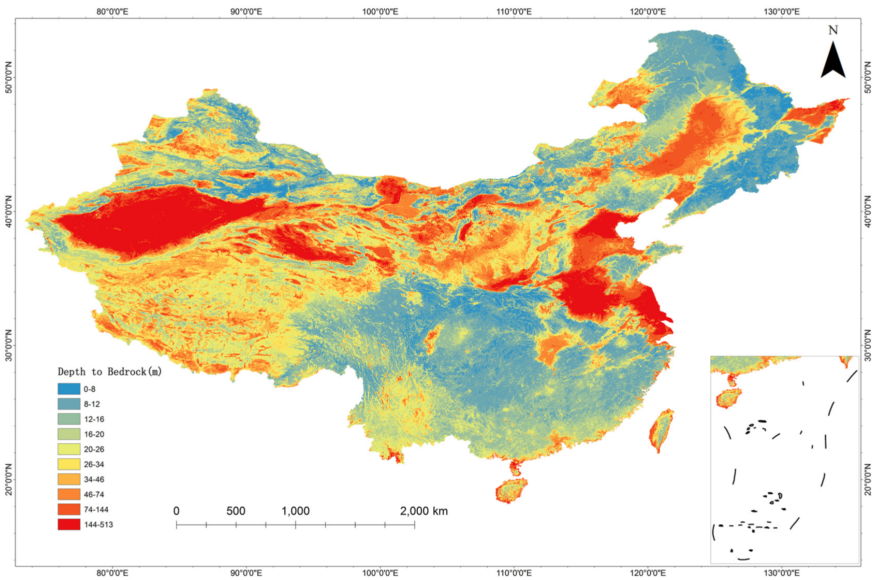

Figure 8: Final prediction of the depth to bedrock based on the ensemble model

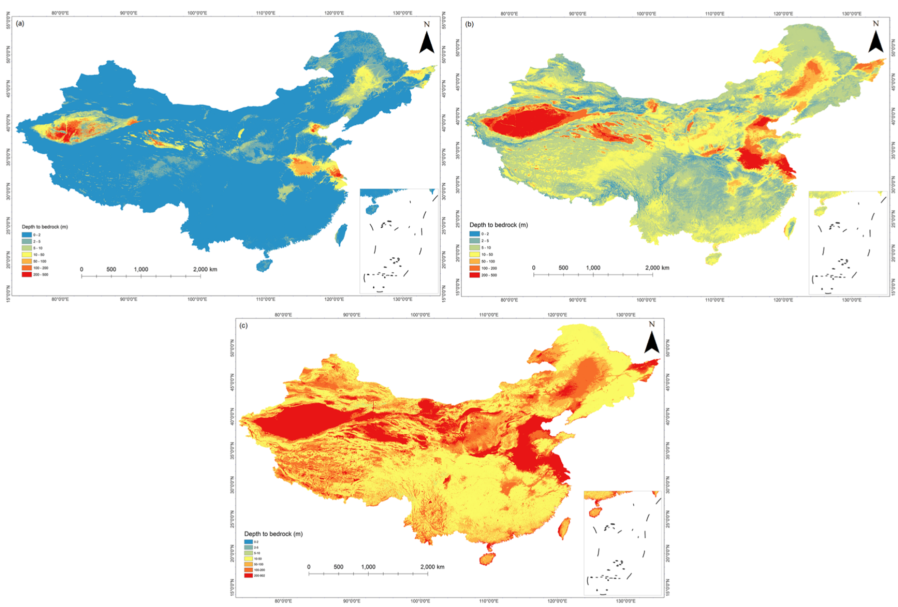

Figure 9: Depth to bedrock maps produced by the quantile regression forests model at the percentiles of 0.1 (a), 0.50 (b), and 0.9 (c).

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

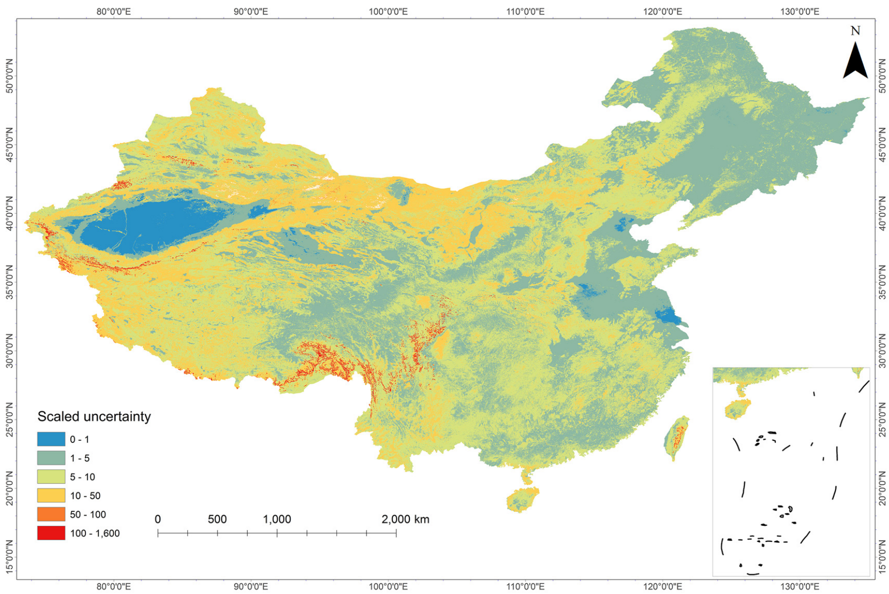

Figure 10: Uncertainty map of prediction of the depth to bedrock

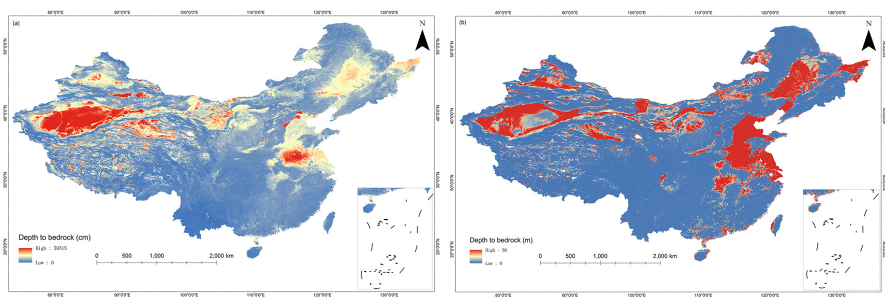

Figure 11: Extracted maps from global predictions of (a) Shangguan et al. (2017) and (b) Pelletier et al. (2016)

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

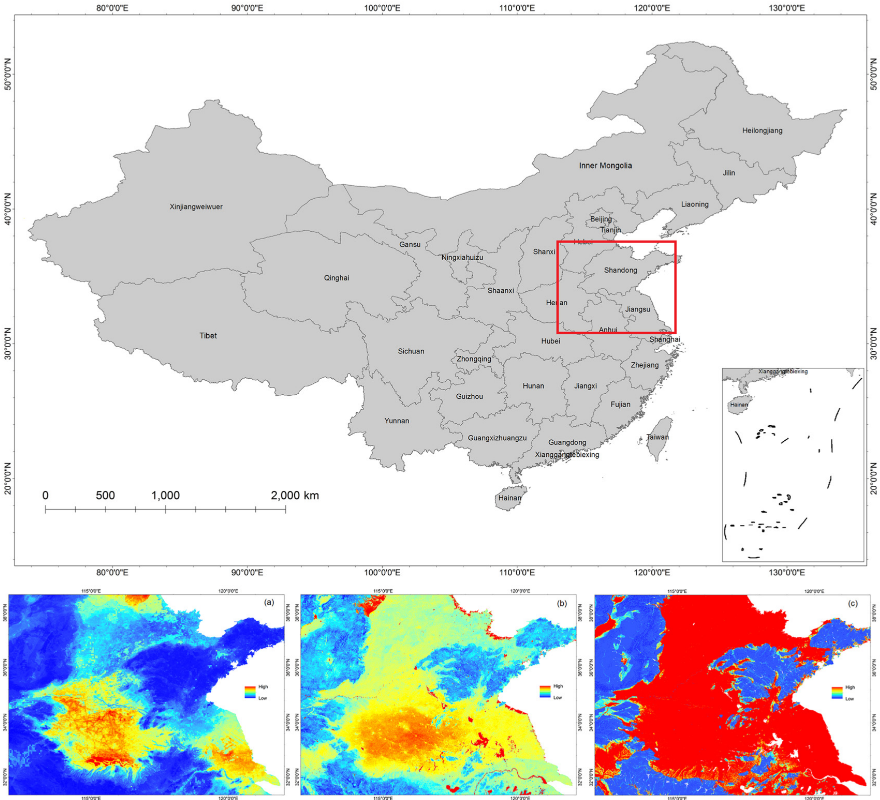

Figure 12: Regional maps of (a) this study, (b) Shangguan et al. (2017), and (c) Pelletier et al. (2016).

Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 13 September 2018

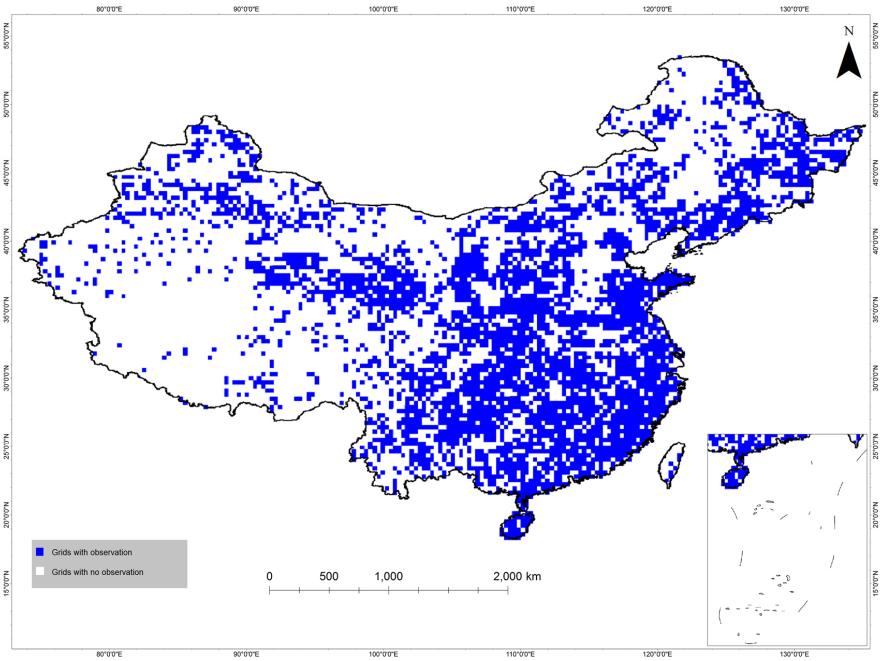

Figure 13: The distribution of 0.2 × 02 arc-degree grid with observation (blue color).