Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2018-103-AC1, 2018

Interactive comment on “Depth-to-Bedrock Map of China at a Spatial Resolution of 100 Meters” *by* Fapeng Yan et al.

Fapeng Yan et al.

yanfap@hotmail.com

Received and published: 31 December 2018

Thanks for your valuable advices:

Comment: This study provides an important DTB data of China. However, there are many contents which were not clearly illustrated, such as the sampling way of the observations, the generation of pseudo points, etc. The introduction is not straightforward. According to the demonstration in the introduction, the innovation is the resolution of the DTB data?

Reply: Thanks for the reviewer's comments. For the contents which were not clearly illustrated, we have addressed in the following specific comments of the reviewer. To present the innovation in the introduction more clearly, we modified the corresponding

paragraph as follows:

Among above-mentioned maps of DTB, there were still some deficiencies, including coarse resolution, limited observations, and limited accuracy. Most of them have relatively coarse resolutions (1 km or coarser), except the map produced by Shangguan et al. (2017) (250 m resolution). However, several environmental covariates (mainly remote sensing data) with high resolution have been produced, which can be used to produce a high-resolution DTB map of China. 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. Both the site observations of boreholes and environmental covariates provide the cornerstone for producing a new map of DTB with higher accuracy and resolution.

Comment: What is the basis of the ensemble method of random forest and gradient boosting tree?

Reply: The use of ensemble model is to avoid the overshooting effect (Sollich, P. & Krogh, A. Learning with ensembles: How over-fitting can be useful. In Proceedings of the 1995 Conference, vol. 8, 190 (1996)). Ensemble prediction has been widely used in areas such as climate modeling. Ensemble method is a common and successful practice in machine learning. Please refer to the following website to see more about ensemble learning: https://en.wikipedia.org/wiki/Ensemble_learning and https://en.wikipedia.org/wiki/Ensemble_averaging_(machine_learning). One of the most relevant statement for tree-based model is “Using a variety of strong learning algorithms, however, has been shown to be more effective than using techniques that attempt to dumb-down the models in order to promote diversity." We will add the following contents to Line 241: We use the ensemble of the two models because ensemble predictions of strong learning algorithms are proved to be more effective in producing better results (Gashler et al., 2008).

Comment: The authors need to try to explain the relationships of DTB and those important covariates, that is why those covariates play an important role in the prediction.

Reply: We agree that it may be useful to understand the relationships of DTB and those important covariates. We plan to add a section in the discussion:

In this section, we try to explain relationships of DTB and those important covariates, which is why those covariates play an important role in the prediction. Topographic wetness index is a function of both the slope and the upstream contributing area per unit width orthogonal to the flow direction. It is commonly used to quantify topographic control on hydrological processes. The greater the wetness is, the more serious the water erosion is. This results in deeper DTB. Slop is an important topographic factor that influence flow speed, soil erosion. High-slope areas often have higher water loss and more serious soil erosion. After a long-term development, these areas are more likely to have relatively low DTB. Especially for the sharp-slope areas, the DTB is close to 0. Topographic openness index is a topographic character expressing the degree of dominance or enclosure of a location on an irregular surface. It reflects the Earth's seafloor, planetary landforms, and features on any irregular surface (Ryuzo Yokoyama et al., 2002). These are the influence factor of the process of transformation of soil and bedrock. Precipitation covariates have high impact on DTB because precipitation determines the humidity of earth surface directly and influence the water erosion seriously. DEM is the basic topographic index, it influences the variation of radiation, water and soil material, which will further influence the soil formation.

Comment: Line 146-147, does this mean that one location has multiple observations?

Reply: It does not mean that one location has multiple observations. Each observation

has unique coordinates. It means that there may be no observation, one observation or multiple observations in each $0.2^{\circ}\times0.2^{\circ}$grid. If there is no observation in that grid, this results in vacancy of observation of the grid. If there is one observation, we use it. If multiple observations, we take one observation randomly among them. We will revise Line 146-147 to the following content: In each $0.2^{\circ}\times0.2^{\circ}$grid, there may be no observation, one observation or multiple observations at different locations. If there is no observation in a grid, it results in vacancy of observation of this grid. If there is one observation, we use it. If there are multiple observations, we take one observation randomly among them.

Comment: Line 149-154, this additive sampling is not clearly illustrated.

Reply: We added some explanation about the additive sampling here: After the first sampling, we get one borehole profile for each $0.2^{\circ}\times0.2^{\circ}$grid as far as possible. Then we try to interpret all the borehole profiles to get the DTB value, but we may be unable to interpret the DTB sometimes. This results in that some grids still have no DTB observation though these grids have more boreholes other than the borehole in the first sampling. In the second sampling, we try to sample another borehole for each of these grids without DTB interpretation. Then we try to interpret all the new borehole profiles in the second sampling to get the DTB value, but we may be still unable to interpret the DTB sometimes. And a third sampling is taken and so on.

Comment: Line 172, so the authors demonstrated the uncertainties, but is there a way to deal with these uncertainties.

Reply: The uncertainties are sourced from the limitation of borehole profiles. To reduce the error of DTB in the process of interpreting borehole profiles as far as possible, we did not use borehole profiles whose lithological descriptions are too vague (such as a layer that is composed of bedrock mixed with weathered rocks) to get the accurate DTB.

Comment: Line 174-185, this part is not clear.

Reply: We modified this part as follows and drew a figure (see attachment Fig1) to show it: There are four kind of cases in the interpretation of DTB. In most cases, DTB are reached for most boreholes and the dip angle is 90°. Then, the DTB is taken as the boundary depth. The dip angles of a minority of boreholes whose dip angle were not 90°. Then, the DTB was calculated as the product of boundary depth and sine of the dip angle. In the third kind of cases, 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. In the fourth kind of cases, because a number of boreholes were drilled 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 as most researches and applications focus on relatively shallow depths.

Comment: Line 206, how do you know the credibility of those "Points"? And I'm not sure how did the authors get those points from existing materials and previous studies.

Reply: We can get some information about the thickness of desert and sand dunes from websites such as Baidu Encyclopedia. Some literatures give information about the profiles of China's desert. Boreholes logs of Pishan, Moyu, Yutian etc. in Taklimakan Desert show that it is mainly medium fine sand and silt below 200 meters (Wu Zheng, 1981). Shahan zone in the South Rim of Taklimakan Desert has a thickness of lower 80 meters (Li Baosheng, Jin Jiong., 1988).

Comment: 224: what is the data source for parent material data?

Reply: The Rock type is based on the global lithology map from USGS (RTMUSG15 in the supplement file) and Geological ages is based on the surface geology (GEAISG3 in the supplement file). We obtained these data from http://www.worldgrids.org/. However, this website is not available now.

Comment: Section 2.4, what do you mean by fitting the prediction models based on the regression matrix?

Reply: The regression matrix is generated by overlaying points and covariates (step 1) and it is used as the input data of machine learning models.

Comment: Line 267, which covariates were removed and what is the removing rule?

Reply: We use the random forest model to estimate the importance of each covariate. Covariates with the lowest correlation will be removed under a given condition. A covariate was removed only when it does not make the model without this covariate significantly worse, i.e., when the R2 of the model decreased less than 0.01 or increased. In this way, we kept the balance between the model complexity (i.e. number of covariates) and model accuracy. The covariates we used are listed in Supplement File A. According to the second reviewer's comment, we revised this rule and will update the supplement file.

Comment: Line 332, the uncertainty map was calculated based on quantile regression forests, therefore is not the uncertainty of the DTB map generated by the resemble method.

Reply: This comment is also proposed in RC2 (reviewer 2). We were aware of the inconsistency between the prediction by RF and GBT models and the uncertainty derived by quantile regression forest. Due to the reviewer's comment, we will offer two sets of data in the next revision. One is the prediction by the ensemble of RF and GBT models, which avoids the overshooting effect (Sollich and Krogh, 1996) and provides a more robust estimation. The other is the prediction and the uncertainty by quantile regression forest. Because most users do not need an uncertainty map in their applications, it is recommended to use the ensemble prediction and take the uncertainty map as a reference. In cases where a consistent prediction and uncertainty are needed, it is recommended to use the estimation by quantile regression forest. Although the mean prediction (quantile is 0.5) is somewhat different from the prediction of ensemble model, it was also validated and will be a better choice for users who need both final prediction and uncertainty estimation. We will also show the map and its accuracy by quantile regression in the revised manuscript. We added the following sentenced in section 2.4: Two sets of data are provided for users. One is the prediction by the ensemble of RF and GBT models and the other is the prediction and the uncertainty by quantile regression forest. Because most users do not need an uncertainty map in their applications, it is recommended to use the ensemble prediction and take the uncertainty map by quantile regression forest as a reference. In cases where a consistent prediction and uncertainty are needed, it is recommended to use the estimation by quantile regression forest.

Comment: Line 352, how did authors obtain the RMSE and ME results for the results of Pelletier et al., and Shangguan et al.? I mean what were the validation points.

Reply: We added the following sentences: Table 4 shows statistics between the 6,328 observations used in this study and predictions of three studies. For our study, we used the cross validation to calculate the statistics (Note that RMSE and ME of our study is the same as Table 3). For the maps by Pelletier et al. (2016) and Shangguan et al. (2017), we calculated the statistics between the 6,328 observations and the prediction directly.

Comment: Line 367-371, these contents were already illustrated in the previous sections.

Reply: We will revise Line 367-374 to the following contents to make it more concise: Although we were trying to get the most representative samples under the current sampling method and add a number of pseudo-observations to our training data. Very few boreholes were located in inhospitable areas such as deserts and mountainous areas (Fig. 13). Lack of observations will increase the uncertainty of predictions in these areas.

Comment: Section 5.2, all these errors make the study unreliable, do the authors have

any suggestions for reducing or dealing with these errors?

Reply: All kinds of observations have some errors. We can choose not to use those with too high errors. However, we disagree that this will make the study unreliable, though we should be aware of the uncertainty. We do not have confident suggestions for reducing or dealing with these errors. Maybe organize a project to harmonize and quality control the borehole DTB observations, which will need joint efforts of geologists of China or even the world.

Comment: The publication year of “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.” is wrong.

Reply: We will revise it to 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, 2017.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-103, 2018.

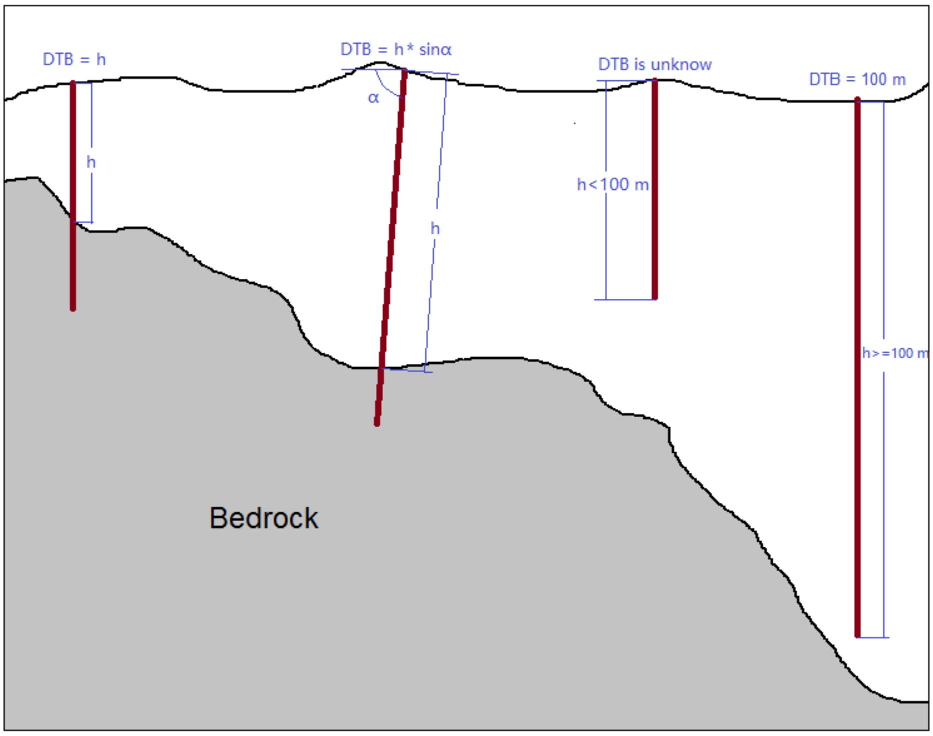
DTB = h
h
DTB = h* sinα
α
h
DTB is unknow
h<100 m
DTB = 100 m
h>=100 m
Bedrock

Fig. 1. Fig1