https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Open Access Earth System Science Data Discussions

Median bed-material sediment particle size across rivers in the contiguous U.S.

Guta Wakbulcho Abeshu[1], Hong-Yi Li[1*], Zhenduo Zhu[2*], Zeli Tan[3], L. Ruby Leung[3]

[1]Department of Civil & Environmental Engineering, University of Houston, Texas, 77204, USA

[2]Department of Civil, Structural and Environmental Engineering, University at Buffalo, the State University of New York, New York, 14260, USA

[3]Pacific Northwest National Laboratory, Washington, 99352, USA

Correspondence to: Hong-Yi Li (hongyili.jadison@gmail.com) and Zhenduo Zhu (zhenduoz@buffalo.edu)

Abstract

Bed-material sediment particle size data, particularly for the median sediment particle size (D50), are critical for understanding and modeling riverine sediment transport. However, sediment particle size observations are primarily available at individual sites. Large-scale modeling and assessment of riverine sediment transport are limited by the lack of continuous regional maps of bed-material sediment particle size. We hence present a map of D50 over the contiguous U.S. in a vector format that corresponds to millions of river segments (i.e., flowlines) in the National Hydrography Dataset Plus (NHDplus) dataset. We develop the map in four steps: 1) collect and process the observed D50 data from 2577 U.S. Geological Survey stations or U.S. Army Corps of Engineers sampling locations; 2) collocate these data with the NHDplus flowlines based on their geographic locations, resulting in 1691 flowlines with collocated D50 values; 3) develop a predictive model using the eXtreme Gradient Boosting (XGBoost) machine learning method based on the observed D50 data and the corresponding climate, hydrology, geology and other attributes retrieved from the NHDplus dataset; 4) estimate the D50 values for flowlines without observations using the XGBoost predictive model. We expect this map to be useful for various purposes such as research in large-scale river sediment transport using model- and data-driven approaches, teaching of environmental and earth system sciences, planning and managing floodplain zones, etc. The map is available at http://doi.org/10.5281/zenodo.4921987 (Li et al., 2021).

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Open Access Earth System Science Data Discussions

1 Introduction

Bed-material sediment particle size information is critical for understanding and modeling riverine sediment processes, including sediment erosion, entrainment, deposition, and transportation. Various sedimentology formulas have been proposed to quantify the sediment processes, with sediment particle size being a critical parameter in those formulas (Meyer-Peter and Müller, 1948; Einstein, 1950; Engelund and Hansen, 1967; Ackers and White, 1973; Van Rijn, 1984; Parker 1990; Garcia and Parker, 1991; Wu et al., 2000; An et al., 2021). Moreover, sediment particle size is a critical factor in riverine dynamics of heavy metal (Unda-Calvo et al., 2019; Zhang et al., 2020), nutrients (Xia et al., 2017; Glaser et al., 2020), microplastic (Corcoran et al., 2019; He et al., 2020), and fish habitats and benthic lives (Dalu et al., 2020; Rieck and Sullivian, 2020).

The sediment transport modes can be classified into bed-material load and wash load (Garcia, 2008). The bed-material load consists of all sizes of particles existing in a river bed regardless of whether they are being transported along the bed (bedload) or in suspension (suspended load). Wash load consists of very fine particles (diameter less than 0.062 mm) that are always in suspension in the water and rarely reside on the bed (Garcia, 1975). Wash load is usually controlled by only land surface processes (soil erosion in hillslopes and transport from hillslopes into rivers), but not much by riverine hydraulic conditions (Garcia, 1975). In this study, we focus on the bed-material sediment particle size data that are critical in applying sediment transport formulas to estimate bed-material load. For example, the Engelund-Hansen equation estimates bed-material load, and median bed-material sediment particle size (denoted as D50, i.e., the size larger than 50% of sediment particles) is one of the most important parameters (Engelund and Hansen, 1967).

Despite the importance of bed-material sediment particle size, such data has limited availability due to the expensive costs of measuring and analyzing such data. As one of the most data-rich countries in the world, the United States (U.S.) collects and disseminates the sediment particle size data mainly through two federal agencies: The U.S. Geological Survey (USGS) and the U.S. Army Corps of Engineers (USACE). USGS manages the most gauges and distributes the river-related measurements on the U.S. rivers. As of April 2021, there are 424948 stations with field/laboratory samples in the USGS water quality portal, among which 1.2% (4991) include bed-material sediment particle data for rivers over the contiguous U.S., and 0.5% (2277) have complete percentiles of bed-material sediment particle data to compute D50.

Spatial approximation, i.e., interpolation or extrapolation, is a typical method to overcome data sparsity when there is no universal relationship between the variable of interest (e.g., D50) and other extensively available information. In the case of sediment particle size, a simple spatial approximation should be conducted within the same river system, assuming similar geological and hydrological settings. Here we denote a river system as the whole river network discharging to the ocean (or inland lakes) via the same outlet. Such a simple spatial approximation is nevertheless not feasible in many river systems, where there are few or no measurement data to support meaningful interpolation and extrapolation. Several studies have reported empirical relationships between bed-material sediment particle size with river channel characteristics (e.g., channel slope) and flow regimes (Niño, 2002; Zhang et al., 2017). Such relations are nonetheless site-specific and not universal enough to apply over various river systems.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

An alternative approach is to establish complex correlations between sediment particle size and other data that are extensively available over the contiguous U.S. Such correlations can then be applied across the U.S. for predicting sediment particle size. Conventional linear or nonlinear regression methods usually require good prior knowledge of the mechanisms controlling sediment size distribution, and thus are not suitable for use to establish complex correlations when understanding of factors that control sediment size is somewhat limited. Machine learning offers an effective way forward because of its ability to establish nonlinear, complex predictive models without the prerequisite of sufficient process-based knowledge (Afan et al., 2016).

Therefore, our objective is to develop a spatial map of D50 over the contiguous U.S. rivers by establishing a predictive model between D50 and other extensively available hydroclimatological and geological data using state-of-the-art machine learning techniques. In the following, we describe the data in Section 2, introduce the machine learning model development in Section 3, and present our results in Section 4. We also explain the limitations of our method in Section 5, potential usage of the D50 map in Section 6, and data availability in Section 7. We finally conclude with Section 8.

2 Data

2.1 Bed-material sediment particle size observations

The USGS sediment data are available to the public through the National Water Information System (NWIS) water quality data portal. There are 4991 USGS stations with at least one sample of bed-material sediment particle size, but only 2277 stations have complete data to allow meaningful computation of D50, as shown in Figure 1a. The USACE sediment particle size data are available in a technical report by Gaines and Priestas (2016). Gaines and Priestas (2016) include the bed-material sediment particle size samples taken at 442 locations along the Mississippi River main stem between Head of Passes, Louisiana and Grafton, Illinois. We exclude locations without exact geographic coordinates and eventually yield 300 locations, as shown in Figure 1a. In total, we have 2577 locations with complete bed-material sediment particle size percentiles to allow for the D50 calculation. At each location, the sediment particle size might have been sampled more than once in different years, although almost half of the locations are sampled only once (see Figure 1b). Figure 1c shows the years that the latest samples were taken. About 94% of these locations have been sampled after 1970.

We compute the D50 values from the measured sediment particle size distributions in three steps: 1) the cumulative sediment size distribution curve is drawn with log-2-transformed sediment size following the concept of the Krumbein phi scale (Krumbein, 1934). 2) A linear interpolation is performed between the percentiles smaller and larger than the 50^{th} percentile to obtain the D50 value. 3) For the stations with multiple sampling times, a representative D50 value is computed as the mean D50 value from all the sampling times. The D50 values calculated following this procedure are denoted as "observed D50 values" to differentiate them from the predicted D50 values using machine learning techniques described later. Figure 1d shows the histogram of the computed D50 values in the Krumbein phi scale. About 75% of these D50 values are between 0.0625 mm and 2.0 mm. Garcia (2008) suggested that a river can be a sand-bed or gravel-bed river if the D50 value is below

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

or above 2.0 mm. The D50 values computed from the observed sediment particle size distributions thus dominantly reflect sand-bed river conditions, while only approximately 25% are gravel-bed rivers.

2.2 Predictive variables

The predictive variables are retrieved from the NHDplus database (McKay et al., 2012) and additional attributes for the NHDPlus catchments from the ScienceBase dataset (Wieczorek et al., 2018). ScienceBase is a comprehensive scientific data and information management platform hosted by USGS (sciencebase.gov). In the medium resolution NHDplus, there are about 2.7 million stream segments (average length of 1.93 km, denoted as flowlines from now on). NHDplus directly provides 138 attributes of flowlines, most of which are descriptive instead of quantitative. We select eight quantitative attributes relevant to the channel geometry and hydrology, such as upstream drainage area, channel bed slope, mean annual flow velocity, sinuosity, etc. ScienceBase provides additional attributes related to the NHDplus watersheds (local drainage area corresponding to a single flowline) and associated upstream drainage areas in thirteen themes (Wieczorek et al., 2018). We select 68 hydroclimatological and geological attributes from ScienceBase, such as climate, hydrologic, topographic, soil, and geologic conditions. In total, 76 attributes are selected as potential predictive variables for input to the machine learning algorithm. We provide a detailed list of these predictive variables in Supplimentary Table S1 and four illustrative maps in Supplementary Figure S1.

We then establish the spatial correspondence between the observed D50 values and the 76 predictive variables. In NHDplus, there are ~26000 USGS stations associated with a portion of the flowlines through the common identifiers. This common identifier is unique for every flowline, but several USGS stations may be located on the same flowline and have the same common identifier. We match the 2277 USGS stations (with observed D50 values) with stations in NHDplus. Some of the 2277 USGS stations are not included in NHDplus, so we obtain 1530 matching stations. The 300 USACE sampling locations are collocated with the flowlines via their geographic coordinates. Several USGS stations or USACE sampling locations may be on the same flowline. In that case, we assign the average of the D50 values of these USGS stations to the flowline. We further exclude a few flowlines with missing attribute values. Finally, we have a total of 1691 flowlines corresponding to the observed D50 values. In other words, in each of these 1691 flowlines, we have established a good correspondence between the observed D50 values and the 76 predictive attributes.

3 Model Development

Among various machine learning methods, eXtreme Gradient Boosting (XGBoost) is a version of the gradient tree boosting algorithm known for its high efficiency and superior performance in recent years (Chen and Guestrin, 2016; Zheng et al., 2019; Fan et al., 2021). Therefore, we adopt XGBoost to develop a predictive model with the Optuna optimization framework (Akiba et al., 2019) for tuning hyperparameters and the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2016) for feature importance analysis and thus feature selection. More details are explained as follows.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

3.1 XGBoost: eXtreme Gradient Boosting

Tree boosting is a machine learning framework that combines weak learners to develop a strong learner, where the base learners are decision trees that are trained sequentially, with the latter focusing on mistakes made by the preceding one. Gradient boosting machines are a family of tree boosting techniques where errors are minimized by gradient descent algorithms. One of the most recent offspring of gradient boosting techniques is the XGBoost, a scalable end-to-end tree boosting system (Chen & Guestrin, 2016). It has been successfully utilized across a wide array of applications, such as snowpack estimation (Zheng et al., 2019) and water storage change in a large lake (Fan et al., 2021). XGBoost dataset is represented as $D = \{(X_i, Y_i), i = 1, 2, \ldots, N\}$, where $X_i = [X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip}]$ is a row vector with input features with real value elements and $Y_i \epsilon R$. The tree ensemble model employs M additive functions to predict the output of interest as

$$\hat{Y}_i = \phi(X_i) = \sum_{m=1}^{M} f_m(X_i) , \qquad f_m \epsilon \mathrm{F} \qquad 1$$

where F is the space of regression trees. The model is trained in an additive manner by minimizing a regularized objective to learn the set of functions employed in the model. At each iteration, a differentiable convex loss function that measures the difference between the prediction $\hat{Y}_i$ and the target Y_i is computed, and the model is also penalized for the complexity of the regression tree functions.

3.2 Tuning Hyperparameters

Tuning hyperparameters is a cumbersome task and is often performed by reducing the parameter search space through randomized search and applying a grid search on the reduced space. Alternatively, hyperparameter optimization frameworks like Hyperopt (Bergstra et al., 2013) and Optuna (Akiba et al., 2019) are commonly preferred since they can continually narrow down the bulky hyperparameter search space to an optimal space based on the preceding results. This study implements Optuna with a Tree-structured Parzen Estimator (TPE) parameter sampling framework to obtain the optimal hyperparameter sets.

The procedure for tuning hyperparameters relies on two major components: cross-validation and evaluation metrics. Cross-validation measures the model's predictive power with a given hyperparameter set by dividing a dataset into folds. In k-fold cross-validation, the dataset is randomly split into k mutually exclusive subsets of approximately equal size as, $D = \{D_1, D_2, D_3, \ldots, D_k\}$. In each iteration $k - 1$ folds of D are used for training, and the remaining one is used for validation. The predictions resulting from a given set of hyperparameters are made by iterating through the folds, so the model is trained and validated k times. Hence, k model performance values and the mean value is reported as the model performance for this set of hyperparameters. Optuna allows the use of user-defined metrics for model evaluation during the k-fold cross-validation. Taking advantage of this structure, we use the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) as the model performance metric.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021


Open Access Earth System Science Data Discussions

$$KGE = 1 - \sqrt{(1-r)^2 + \left(1 - \frac{\sigma_{sim}}{\sigma_{obs}}\right)^2 + \left(1 - \frac{\mu_{sim}}{\mu_{obs}}\right)^2} \quad (2)$$

where σ is the standard deviation, μ is the mean, and r is the linear correlation between the observed and simulated series. A perfect agreement between observation and simulation gives the theoretical maximum KGE value at 1.0. The higher the KGE value, the closer the match between the observed and simulated series. KGE offers some advantages over commonly used metrics like root mean squared errors (RMSE) or coefficient of determination (R^2) because: 1) it is not dominated by relatively large values; and 2) it simultaneously captures both the magnitude and phase differences between the observed and simulated series (Gupta et al., 2009).

3.3 Feature Selection

Feature selection is also an essential step in developing a simpler model that is still capable of reasonably predicting the target variable with fewer attributes. Feature importance is a technique of computing each predictive variable's degree of contribution towards the optimal prediction model, which can be used for determining feature selection. The approaches of computing feature importance scores include correlation coefficient, the coefficients calculated as part of decision trees, or advanced approaches like SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2016). In this study, we use the mean absolute SHAP values as feature importance measures.

Initially, we begin with 76 predictive variables. For feature selection purposes, we add a new "predictor" of randomly generated real number values. We train the model and compare the feature importance scores (i.e., the mean absolute SHAP values) of all predictors. Then, all attributes with scores less than the random number attribute are dropped out. The procedure is repeated using the new set of predictors until the random number attribute is the least important feature. Lastly, the remaining features are utilized for tuning the final optimal set of hyperparameter values.

3.4 General Steps

The general steps of the model development procedure are as follows.

1. The predictors are scaled using the Minimum-Maximum scaler method, i.e., all features will be transformed into a range of [0,1]. The main advantage of having this bounded range normalization is that it can supress the effect of outliers.
2. The dataset is randomly split into training (70%) and testing (30%) sets, respectively. Only the training data are used in steps 3 and 4, while the testing data are reserved for step 5.
3. Optuna and k-fold (k=5) cross-validation are used for tuning hyperparameters, with a maximum tree of 5000 and an early stopping value of 50. The objective function for the hyperparameter optimization procedure is to maximize the mean Kling-Gupta Efficiency (KGE) value returned from the k-fold cross-validation.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021


Open Access Earth System Science Data Discussions

4. Feature selection is performed as described in section 3.3, so step 3 is repeated with the new and smaller set of predictors. Steps 3 and 4 are repeated until no more predictor can be excluded.
5. The final model is developed by fitting on the whole training data using the optimal hyperparameters, and evaluated using the testing data reserved in step 2.
6. The model from step 5 is used to predict the D50 values for the contiguous U.S. river flowlines.

4 Results

We discuss our results in three steps: the subset of flowlines as the basis to formulate our predictive model, the development and validation of our predictive model, and the national D50 map derived based on the predictive model.

4.1 Measured D50

Figure 2 shows the 1691 flowlines with the associated observed D50 values. The Mississippi River has relatively denser measurements attributed to the USACE database, while the southwest (e.g., the Rio Grande) and the Great Basin have fewer measurements. Overall, the 1691 flowlines are distributed throughout the contiguous United States, providing a good spatial representation of the NHDplus flowlines. Similar to all observed D50 values in Figure 1d, most of the D50 values associated with the flowlines represent sand-bed rivers (D50 < 2.0 mm). Larger-D50 (> 2.0 mm) flowlines are mainly located in the basins of California, Upper Colorado, Missouri, Ohio and Upper Mississippi.

4.2 Predictive Model

4.2.1 Feature Selection

After iterations of feature selection (procedure described in sections 3.3 and 3.4), 13 out of 76 predictive variables are eventually selected and shown in Table 1. These variables are identified as more significant than a random-number input vector based on the mean absolute SHAP value. 2 out of 8 channel characteristics and 11 out of 68 basin characteristics remain as the significant predictors (see Table 1 for description). The most important predictor is found to be the soil erodibility factor (TOT_KFACT), followed by average annual wet day (TOT_WDANN) and mean annual snow as a percent of total precipitation (TOT_PRSNOW).

These three basin-related predictors rank higher than the two channel-related characteristics. Channel slope (Slope) and distance between flowline and the river mouth (Pathlength) are found to be the most important channel characteristics for predicting D50, which agrees with the downstream fining phenomena and sediment transport mechanisms (Nino, 2002). It is somewhat surprising that some hydraulic channel characteristics such as mean annual flow velocity are not included in the final feature selection. Studies on river hydraulics show relations between channel flow (i.e., velocity and water depth) and channel bed characteristics (i.e., slope and roughness), such as the Manning's equation, Chezy's law, etc., and channel bed roughness can be related to bed sediment size (Garcia, 2008). However, the feature selection with the XGBoost model and

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

SHAP value indicates that mean annual flow velocity may not be a good predictor for D50 in this case. A possible reason is that mean annual flow velocity is dependent on some of the selected parameters such as TOT_WDANN, Slope, etc., so excluding this variable avoids overfitting the data.

It should be noted that the ranking of feature importance according to the mean absolute SHAP values is quite different from the correlation coefficients between D50 and predictors, as shown in Figure 3. TOT_KFACT and TOT_WDANN, the two most important features in Table 1, have correlation coefficients of only 0.08 and 0.06, respectively. TOT_PRSNOW has the strongest correlation with D50, with a correlation coefficient of 0.29. The scatter plots between D50 and all the selected features do not show any apparent relationship between D50 and any single feature (see Supplementary Figure S2). This indicates that the XGBoost model can reveal higher-order interactions among the predictors for better predictions.

Although feature selection sheds light on the contribution of input variables to model outputs, a drawback of the machine learning technique is that it cannot explain mechanistically why selected features are more important than unselected ones. Therefore, the goal of feature selection is to find the best (i.e., most robust) input variables to feed the best model for D50 predictions. If a different machine learning algorithm from XGBoost is used, the selected features, especially their rankings, can be different. Feature selection is dependent on the selection of the algorithm, so the selected features in this study should not be directly used in other models or studies.

4.2.2 Model Hyperparameters and Performance

Table 2 shows the tuned hyperparameters of the best XGBoost model that is trained using the 13 selected predictors and 70% of the training data set. For a detailed explanation of the hyperparameters please refer to Chen and Guestrin (2016). Figure 4 shows the performance of the optimal XGBoost model on the training and testing data sets, respectively. Here we consider an optimal model based on two criteria: 1) the model performance is satisfactory in both the training and testing phases, ad indicated by good metrics values (e.g., KGE in this study), and 2) the model performance is relatively consistent between the training and testing phases. Here the optimal XGBoost model gives the KGE value 0.794 for training and 0.513 for testing. The testing value is above 0.5, suggesting satisfactory model performance (Gupta et al., 2009; Knoben et al., 2019). The performance on the testing data is noticeably worse than that on the training data, as expected. This difference is nevertheless acceptable given the complexity of the prediction problem. The relatively consistent model performance between the training and testing phase suggests that the model validation (via the testing phase) is successful.

4.2.3 Model Uncertainty

We carry out further analysis to shed light on how the modelling results may be sensitive to some of the key steps as outlined in Section 3.4. We focus on Steps 2 to 4 only because Steps 1 and 5 are standard practice and Step 6 is to utilize the modelling results.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

For Step 2, the 2/3 (train) and 1/3(test) split is typical in machine learning for splitting training and testing data. This can be readjusted up to 4/5(train) and 1/5(test) if the total sample size is sufficiently large, which is nonetheless not the case here. For Step 3, the choice of k value is usually 5 or 10 depending on the training sample size. We use 5 since using 10 significantly reduces the number of samples per fold and the left-out sample will be too small for validation during cross-validation. Increasing k-fold to 6 or decreasing it to 4 still gives a similar satisfactory performance in both the training and testing phases, with training/testing KGE of 0.759/0.505 and 0.795/0.512 respectively. For Step 4, we evaluate the model sensitivity to each selected feature by dropping one of the 13 variables at a time and repeating the same modelling procedure for the remaining 12 variables. Figure 5 shows that dropping the variables leads to the model performance dropping below KGE = 0.475 in the training phase for most features except for TOT_SATOF and TOT_WBM_TAV. Even with those two, the KGE difference between the training and testing phases increases from 0.28 to 0.36 by including them as predictors. Thus, all the variables remaining after feature selections play a significant role in the final model.

4.3 National Map

Using the developed machine learning model and NHDplus channel/basin characteristics data, we are able to produce a national map of bed sediment D50 values (Figure 6). To our best knowledge, it is the first of its kind D50 data for the whole contiguous U.S. The spatial pattern of D50 in Figure 6 are generally consistent with the observed D50 in Figure 2. High D50 values are mostly distributed in the west coast, upper Missouri and Ohio regions, and low D50 values are concentrated in the east coast. The consistency between Figures 2 and 6 suggests that the observed D50 data are reasonably representative of the whole contiguous U.S., despite the sparse distribution. Given that the testing data set is independent of the training dataset, we expect that the error statistics derived for the testing data should be relatively consistent with the error statistics in applying the model to derive the national map of D50.

5 Limitations of the method

The predicted D50 values may be subject to several limitations despite using state-of-the-art machine learning techniques to develop the predictive model. These limitations include: 1) Limited data availability. Although the 1691 observed D50 values are adequately representative of the contiguous U.S. (i.e., consistent spatial patterns between Figures 2 and 6), limited data availability prevents us from establishing a separate predictive model for each river basin. For example, there is little observed D50 data in the Rio Grande and Great Basin, so the predicted D50 values over these basins should be used cautiously. 2) Our methodology is statistical in nature and lacks explicit process-based understanding. For example, Figure 4 shows the model tends to overestimate D50 for smaller D50 values (particularly < 0.25 mm) and underestimate D50 for larger D50 values (particularly > 1 mm). However, in various trials we have performed, the current result is closest to the 1:1 relationship based on both the KGE metric and visual check. Further process-based understanding of this systematic bias is beyond the scope of this study and left for a future work. 3) We have not explicitly incorporated the effects of lakes and reservoirs but rather

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

assumed these effects have been indirectly reflected in the NHDplus hydrologic attributes adopted in the predictive model. As such, our predictive results are certainly not free from uncertainties. Therefore, we recommend using our D50 map for sediment modeling and assessment at the regional or national scales instead of local studies at the individual river segment.

6 Potential usage

The D50 map might be used for large-scale sediment transport modeling over the whole contiguous U.S., or a major river basin such as the Mississippi River basin. There is inevitably some uncertainty embedded in these maps sourced from the original D50 observations and NHDplus attributes, the XGBoost modeling, and the spatial extrapolation process. This uncertainty should be taken into account while evaluating the uncertainties in the model simulations.

The D50 map may also be used to derive other important parameters for large-scale hydrological or hydraulic modeling. For instance, Manning's roughness coefficient is an important parameter for river routing modeling. It is, however, largely empirical and hard to directly measure, hence not readily available at a regional or national scale. Previous studies have established some empirical relationships between Manning's roughness coefficient and D50 for river channels (Coon, 1998; Gillen, 1996; Julien, 2002; Meyer-Peter & Müller, 1948). Thus, one might derive a map of Manning's roughness coefficient over the contiguous U.S. based on the empirical relationship and the D50 maps.

7 Data availability

The national D50 map is freely available at http://doi.org/10.5281/zenodo.4921987 (Li et al., 2021). The input data are obtained from the USGS water quality portal (https://nwis.waterdata.usgs.gov/usa/nwis/qwdata), NHDplus (https://www.epa.gov/waterdata/nhdplus-national-data) and ScienceBase (https://doi.org/10.5066/F7765D7V).

8 Conclusions

We develop a new national map of the median bed sediment particle size by combining the USGS sediment observations, the channel and watershed characteristics from NHDplus and ScienceBase, and state-of-the-art machine learning techniques. Despite the limitations, the map is highly valuable for sediment modeling and assessment at the regional and larger scales, which has not been feasible previously.

Author contributions

GA conducted the analysis. HL and ZZ designed the study. HL and RL conceived the idea. All the authors contributed to the writing.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Competing interests

The authors declare that there is no conflict of interest.

Acknowledgments

This research is supported by the Office of Science of the U.S. Department of Energy as part of the Earth System Model Development program area through the Energy Exascale Earth System Model (E3SM) project. The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RLO1830.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

References

Ackers, P., & White, W. R. (1973). Sediment transport: new approach and analysis. Journal of the Hydraulics Division, 99(11), 2041-2060.

Afan, H. A., El-shafie, A., Mohtar, W. H. M. W., & Yaseen, Z. M. (2016). Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. Journal of Hydrology, 541, 902–913.

An, C., Gong, Z., Naito, K., Parker, G., Hassan, M. A., Ma, H., & Fu, X. (2021). Grain size-specific Engelund-Hansen type relation for bed material load in sand-bed rivers, with application to the Mississippi River. Water Resources Research, 57, e2020WR027517. https://doi.org/10.1029/2020WR027517.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 163, 785–794.

Corcoran PL, Belontz SL, Ryan K, Walzak MJ (2019) Factors controlling the distribution of microplastic particles in benthic sediment of the Thames River, Canada. Environ Sci Technol 54(2):818–825.

Coon, W. (1998). *Estimation of Roughness Coefficients for Natural Stream Channels with Vegetated Banks*. *U.S. Geological Survey water-supply paper*

Einstein, H. A. (1950). The Bedload Function for Bedload Transportation in Open Channel Flows. Technical Bulletin No. 1026, USDA Soil Conservation Service, 1–71.

Engelund F, & Hansen E. (1967). A monograph on sediment transport in alluvial streams: Teknish Vorlag, Copenhagen, Denmark.

Fan, C., Song, C., Liu, K., Ke, L., Xue, B., Chen, T., ... & Cheng, J. (2021). Century-Scale Reconstruction of Water Storage Changes of the Largest Lake in the Inner Mongolia Plateau Using a Machine Learning Approach. Water Resources Research, 57(2), e2020WR028831.

Gaines, R.A.; Priestas, A.M. Particle Size Distribution of Bed Sediments along the Mississippi River, Grafton, Illinois, to Head of Passes, Louisiana, November 2013; Technical Report 7; US Army Corps of Engineers: Vicksburg, MS, USA, 2016.

M. H. García (Ed.) (2008), *Sedimentation Engineering: Processes, Measurements, Modeling, and Practice*, 1150 pp., Am. Soc. of Civ. Eng., Reston, Va.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Garcia, M., & Parker, G. (1991). Entrainment of bed sediment into suspension. Journal of Hydraulic Engineering, 117(4), 414-435.

Gillen, D. F. (1996). Determination of Roughness Coefficients for Streams in West-Central Florida. *Journal of Hydrology*, 203–250.

345 Glaser, C.; Zarfl, C.; Rügner, H.; Lewis, A.; Schwientek, M. Analyzing Particle-Associated Pollutant Transport to Identify In-Stream Sediment Processes during a High Flow Event. Water 2020, 12, 1794.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

350 Julien, P. Y. (2002). *River Mechanics*. Cambridge University Press. https://doi.org/10.1017/9781316107072

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Krumbein, W. C. (1934). Size frequency distributions of sediments. Journal of Sedimentary Research, 4(2), 65-77.

355 Li, H.-Y.; Wigmosta, M. S.; Wu, H.; Huang, M. Y.; Ke, Y. H.; Coleman, A. M.; Leung, L. R., A Physically Based Runoff Routing Model for Land Surface and Earth System Models. J Hydrometeorol 2013, 14 (3), 808-828.

Li, H.-Y.; Leung, L. R.; Tesfa, T.; Voisin, N.; Hejazi, M.; Liu, L.; Liu, Y.; Rice, J.; Wu, H.; Yang, X. F., Modeling stream temperature in the Anthropocene: An earth system modeling approach. J Adv Model Earth Sy 2015, 7 (4), 1661-1679.

Li, H.-Y.; Abeshu, G.; Zhu, Z.; Tan, Z.; Leung, L. R. (2021). A national map of riverine median bed-material particle size
360 over CONUS (Version 1.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4921987.

He, B; Wijesiri, B.; Ayoko, G.A.; Egodawatta, P.; Rintoul, L.; Goonetilleke, A. (2020) Influential factors on microplastics occurrence in river sediments. Sci Total Environ 738:139901.

McKay, L.; Bondelid, T.; Dewald, T.; Johnston, J.; Moore, R.; Rea, A., NHDPlus Version 2: User Guide. 2012.

Meyer-Peter, E., & Müller, R. (1948). Formulas for bedload transport. In IAHSR 2nd meeting, Stockholm, appendix 2. IAHR.

365 Niño, Y. (2002). Simple Model for Downstream Variation of Median Sediment Size in Chilean Rivers. Journal of Hydraulic Engineering, 128(10), 934–941.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Parker, G. (1990), Surface-based bedload transport relation for gravel rivers, J. Hydraul. Res., 28(4), 417–436.

Rieck, L. O., Sullivan, S. M. P. (2020) Coupled fish-hydrogeomorphic responses to urbanization in streams of Columbus, Ohio, USA. PLoS ONE 15(6): e0234303. https://doi.org/10.1371/journal.pone.0234303.

Zhang, W., Wang, H., Li, Y., Lin, L., Hui, C., et al. (2020), Bend-induced sediment redistribution regulates deterministic processes and stimulates microbial nitrogen removal in coarse sediment regions of river, Water Res., 170 (2020), p. 115315

Unda-Calvo, J.; Ruiz-Romera, E.; Fdez-Ortiz de Vallejuelo, S.; Martínez-Santos, M.; Gredilla, A. Evaluating the role of particle size on urban environmental geochemistry of metals in surface sediments. Sci. Total Environ. 2019, 646, 121–133.

Van Rijn, L. C. (1984). Sediment transport, part II: suspended load transport. Journal of Hydraulic Engineering, 110(11), 1613-1641.

Wieczorek, M.E., Jackson, S.E., and Schwarz, G.E., 2018, Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States (ver. 3.0, January 2021): U.S. Geological Survey data release, https://doi.org/10.5066/F7765D7V.

Wu, W., Wang, S. S. Y., Jia, Y. (2000). Nonuniform sediment transport in alluvial rivers. J. Hydraul. Res., 38(6), 427-434.

Xia, X., Z. Jia, T. Liu, S. Zhang, and L. Zhang (2017), Coupled nitrification-denitrificaton caused by suspended sediment (SPS) in rivers: Importance of SPS size and composition, Environ. Sci. Technol., 51(1), 212– 221, doi:10.1021/acs.est.6b03886.

Zhang, L., Zhang, H., Tang, H., & Zhao, C. (2017). Particle size distribution of bed materials in the sandy river bed of alluvial rivers. International Journal of Sediment Research, 32(3), 331–339.

Zheng, Z., Ma, Q., Jin, S., Su, Y., Guo, Q., & Bales, R. C. (2019). Canopy and terrain interactions affecting snowpack spatial patterns in the Sierra Nevada of California. Water Resources Research, 55(11), 8721-8739.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Table 1. Most important predictors according to the feature selection

Predictor	Description	Mean Absolute SHAP Value
TOT_KFACT	Soil erodibility factor of Universal Soil Loss Equation	0.51
TOT_WDANN	Average annual number of wet days	0.46
TOT_PRSNOW	Mean annual snow as a percent of total precipitation	0.37
Slope	Channel bed slope for each flowline	0.36
Pathlength	Distance from the downstream end of a flowline to the end of the network (river mouth)	0.34
TOT_RFACT	R factor of Universal Soil Loss Equation	0.34
TOT_SATOF	Annual saturation overland flow as a percent of total runoff	0.33
TOT_CONTACT	Time it takes for water to drain along subsurface flow paths to the stream	0.31
TOT_BASIN_SLOPE	Average topographic slope within the upstream drainage area	0.30
TOT_ELEV_MEAN	Average surface elevation within the upstream drainage area	0.29
AI	Aridity index defined as the ratio of annual mean potential evaporation to annual mean precipitation	0.29
TOT_WBM_TAV	Average mean annual temperature within the upstream drainage area	0.29
TOT_RUN	Average annual runoff within the upstream drainage area	0.23

Note: here we use the same names as those in the NHDplus attribute tables, but moderately revise the description using terminologies that can be understood by a broader audience

395

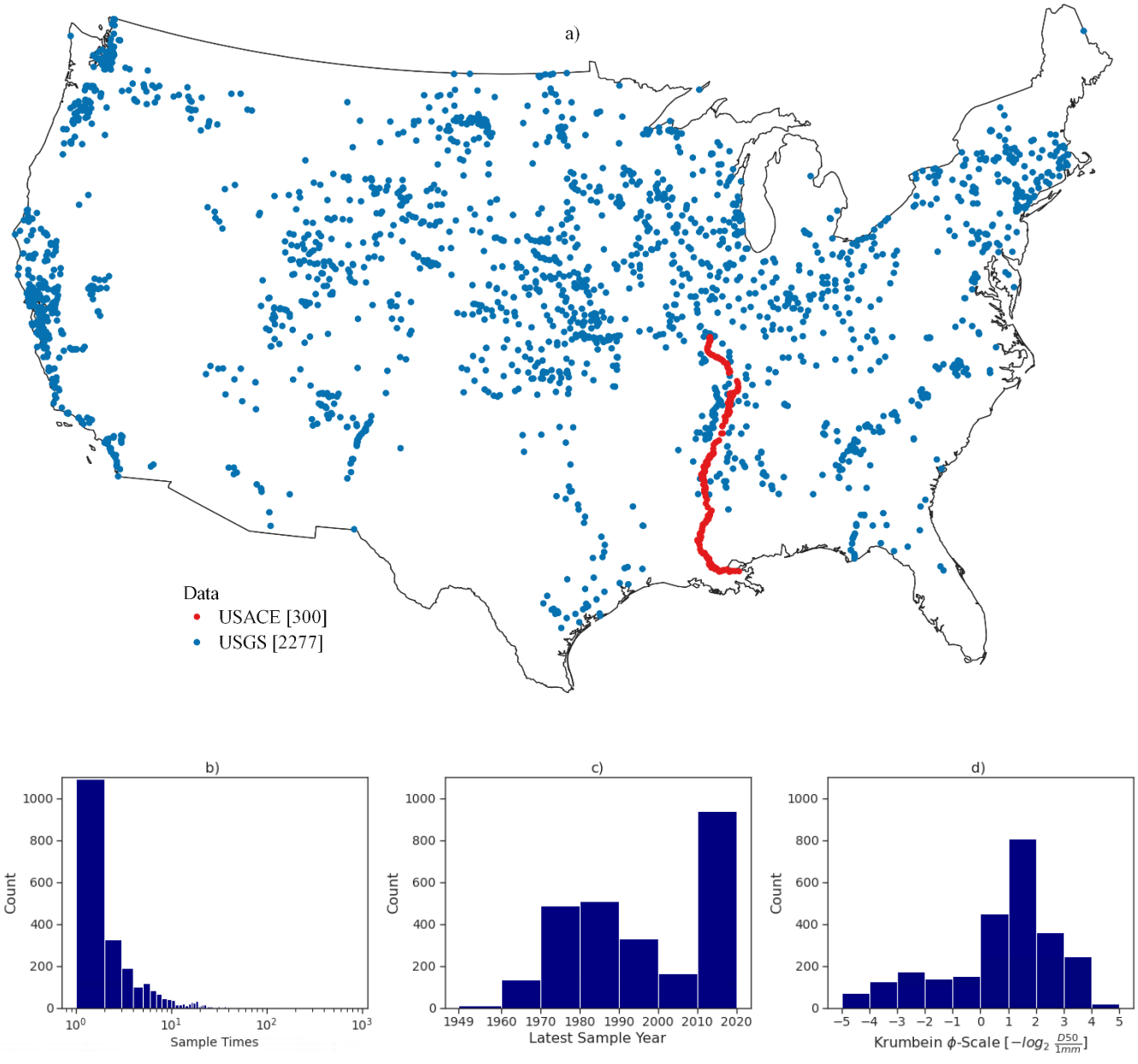
https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Table 2. Optimal value of the XGBoost model hyperparameters

Hyperparameter	Optimal Value	Tuning Range
learning_rate	0.825	[0,1]
min_split_loss	1	[0,∞]
max_depth	6	[0,∞]
min_child_weight	58	[0,∞]
max_delta_step	22	[0,∞]
subsample	0.695	[0,1]
colsample_bytree	0.712	[0,1]
reg_lambda	26.821	[0,∞]
reg_alpha	2.561	[0,∞]
n_estimators	155	[1,∞]

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

Open Access Earth System Science Data Discussions



405

Figure 1: Sediment sample stations. a. Location of USGS stations and USACE sampling locations; b. Number of samples at each station/location; c. sample year at each station/location; d. D50 values.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

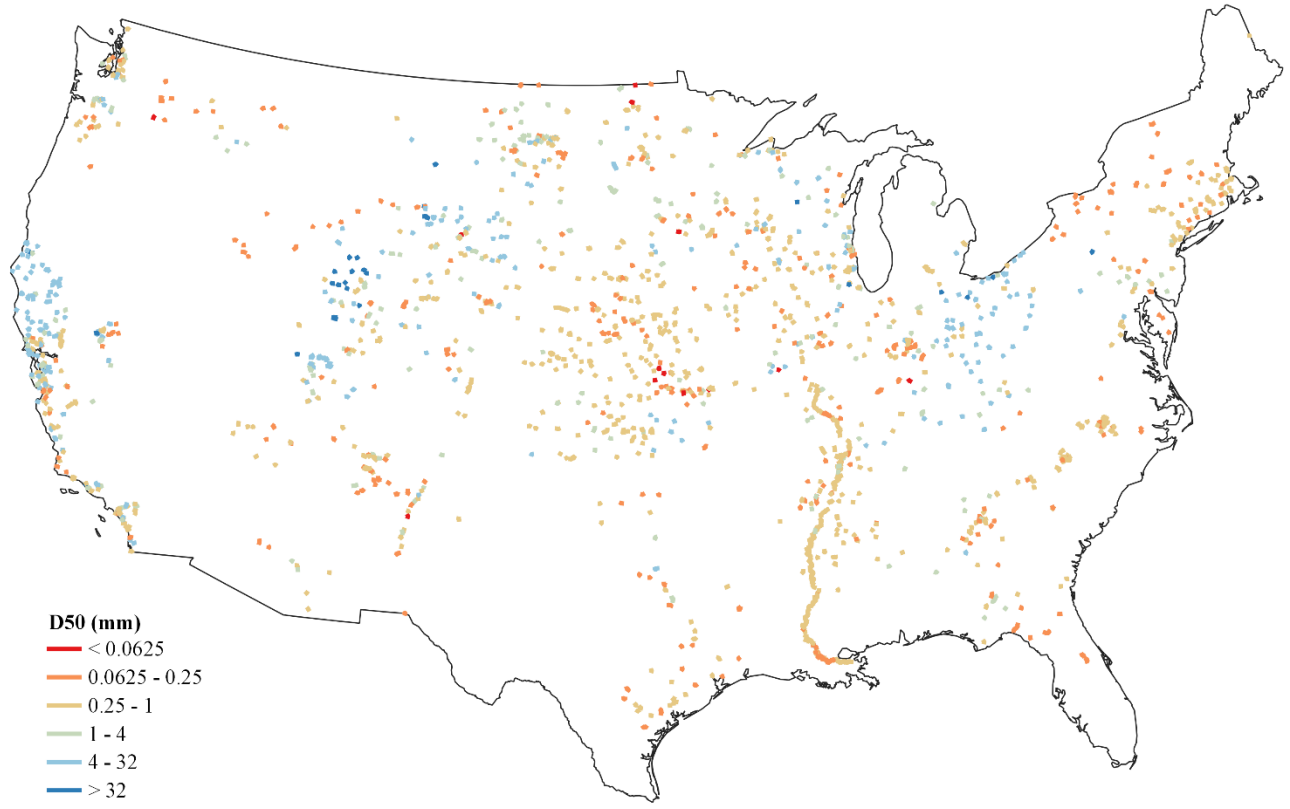


Figure 2. 1,691 flowlines with measured D50.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

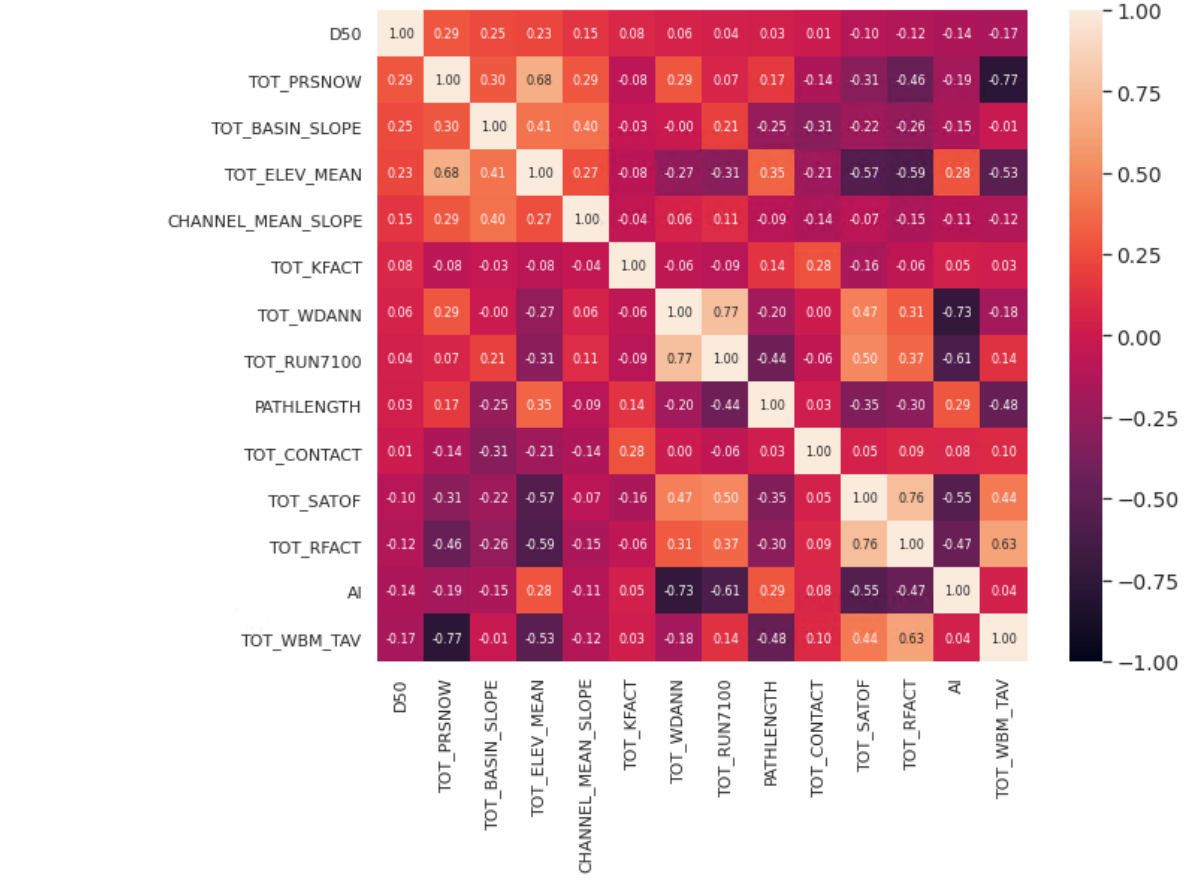


Figure 3. Correlation coefficients among D50 and the 13 selected predictors.

415

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

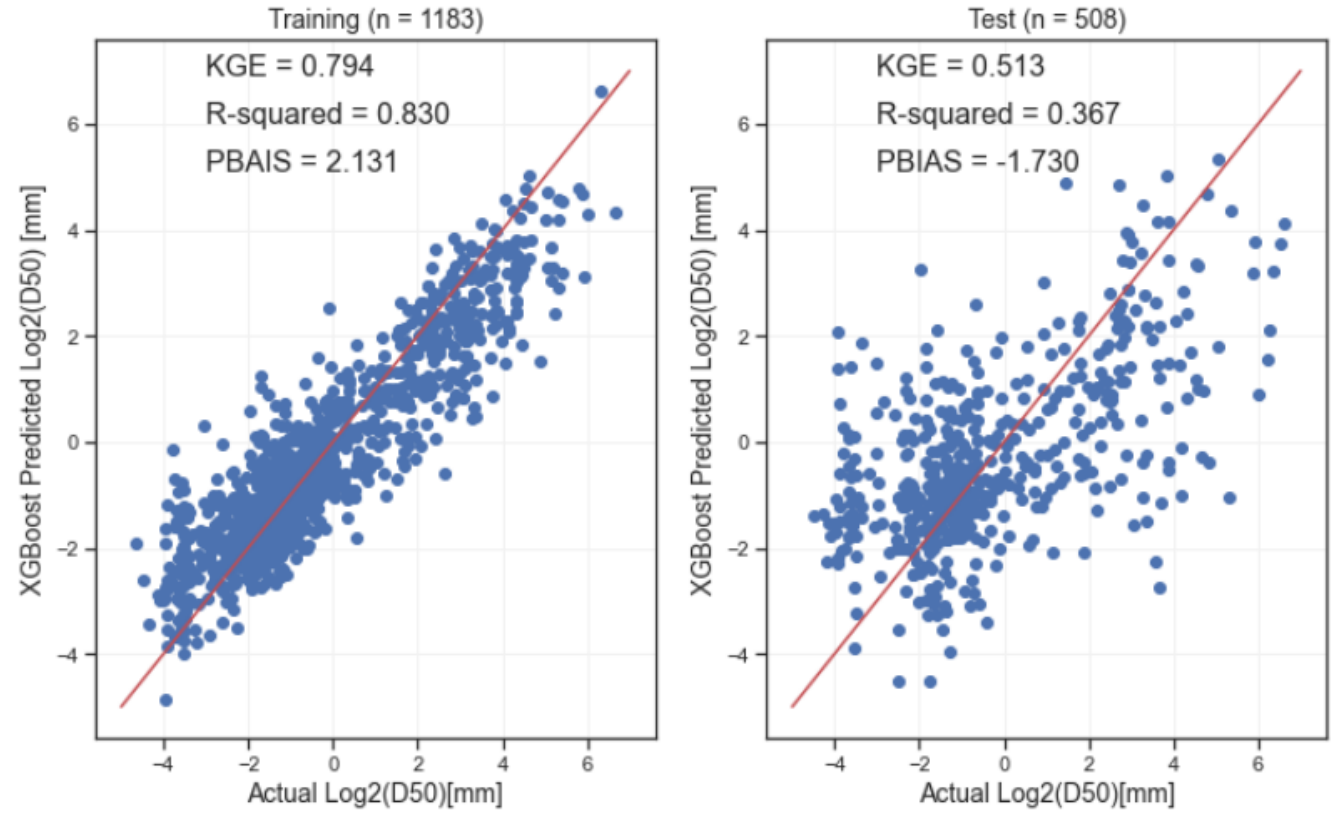


Figure 4: XGBoost model performance with the training (left) and testing (right) data sets.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

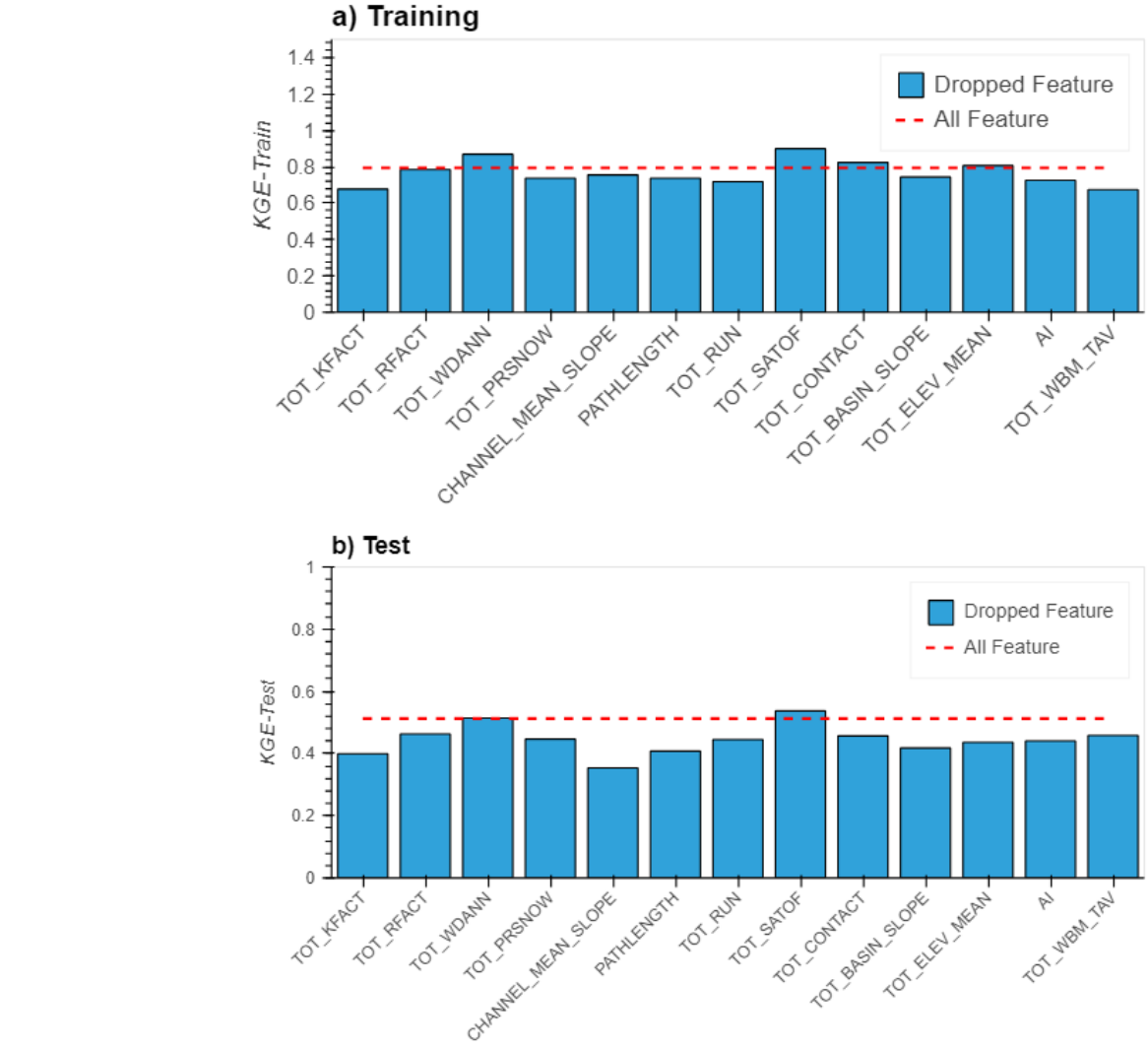


Figure 5. Sensitivity of the XGBoost model to the selected features. The result shown in blue bars are obtained by dropping the corresponding labelled feature from the 13 selected features. The dashed red line represents the model performance with all variables.

https://doi.org/10.5194/essd-2021-201
Preprint. Discussion started: 8 July 2021

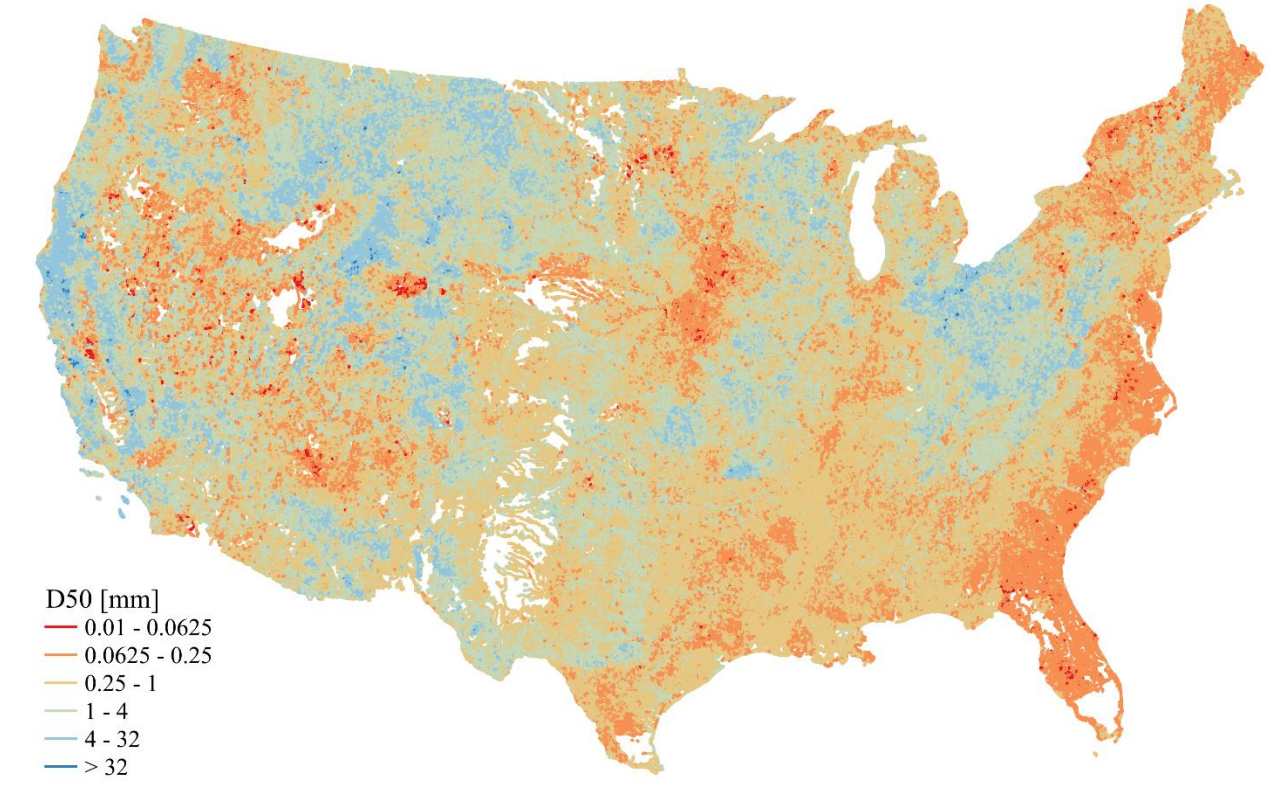


Figure 6. Predicted D50 in ~2.7 million flowlines across the contiguous U.S. using the XGBoost model.