Earth Syst. Sci. Data, 14, 929–942, 2022
https://doi.org/10.5194/essd-14-929-2022

Open Access Earth System Science Data

Median bed-material sediment particle size across rivers in the contiguous US

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

[1]Department of Civil and 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, Buffalo, New York 14260, USA
[3]Pacific Northwest National Laboratory, Richland, Washington 99352, USA

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

Received: 15 June 2021 – Discussion started: 8 July 2021
Revised: 21 December 2021 – Accepted: 23 December 2021 – Published: 24 February 2022

Abstract. Bed-material sediment particle size data, particularly 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 US in a vector format that corresponds to approximately 2.7 million 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; and (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 environmental and earth system sciences, planning and managing floodplain zones, etc. The map is available at https://doi.org/10.5281/zenodo.4921987 (Li et al., 2021a).

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 (Ackers and White, 1973; An et al., 2021; Einstein, 1950; Engelund and Hansen, 1967; Garcia and Parker, 1991; Meyer-Peter and Müller, 1948; Parker, 1990; Parker and Andrews, 1985; Van Rijn, 1985; Wu et al., 2000). Moreover, sediment particle size is a critical factor in riverine dynamics of heavy metal (Unda-Calvo et al., 2019; Zhang et al., 2020), nutrients (Glaser et al., 2020; Xia et al., 2017), microplastic (Corcoran et al., 2020; He et al., 2020), and fish habitats and benthic lives (Dalu et al., 2020; Riecki et al., 2020).

The sediment transport modes can be classified into bed-material load and wash load (García, 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 (bed load) or in suspension (suspended load). Wash load consists of very fine particles (diameter less than 0.0625 mm) that are always in suspension in the water and rarely reside on the bed (García, 2008). Wash load is usually controlled only by land surface processes (soil erosion in hillslopes and transport from hillslopes into rivers) but not much by riverine hydraulic conditions (García, 2008). In this study, we fo-

cus on the bed-material sediment particle size data that are critical in applying sediment transport formulas to estimate bed-material load. For example, 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 in the Engelund–Hansen equation (Engelund and Hansen, 1967).

Despite the importance of bed-material sediment particle size, such data have 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 (US) 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 US rivers. As of April 2021, there are 424 948 stations with field and/or laboratory samples in the USGS water quality portal, among which 1.2 % (3644) include bed-material sediment particle data for rivers over the contiguous US, and 0.6 % (2277) have complete percentiles of bed-material sediment particle data for computing 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.

An alternative approach is to establish complex correlations between sediment particle size and other data that are extensively available over the contiguous US. Such correlations can then be applied across the US 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 they 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 US 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 Sect. 2, introduce the machine learning model development in Sect. 3, and present our results in Sect. 4. We also explain the limitations of our method in Sect. 5, potential usage of the D50 map in Sect. 6, and data availability in Sect. 7. We finally conclude with Sect. 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 3644 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 Fig. 1a. There are 1367 USGS stations with incomplete percentiles of bed-material sediment particle data, which can be divided into three groups: (1) 1183 stations have no effective percentiles provided, (2) 147 stations have only percentiles above the 50th percentile, and (3) 37 stations have only percentiles below the 50th percentile. Therefore, we neglect these data in further analysis.

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 the locations without exact geographic coordinates and eventually obtain 300 locations, as shown in Fig. 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 at different times, although almost half of the locations are sampled only once (see Fig. 1b for the histogram and Fig. S1a in the Supplement for the spatial map). For about 94 % of these stations, the latest samples were taken after the 1970s (see Fig. 1c for the histogram and Fig. S1b in the Supplement for the spatial map). We calculated the coefficient of variation (CV) for the 760 stations that have at least five samples over time. For the rest of the stations, the number of samples is too small for meaningful calculation of CV. For most of these 760 stations, the CV values range between 0.3 and 1.2, with a median of approximately 0.6 (see Fig. S2 in the Supplement). The small CV values indicate the good stability of D50 (at the same location) over time.

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 (in mm) following the concept of the Ψ scale (Parker and Andrews, 1985), (2) a linear inter-

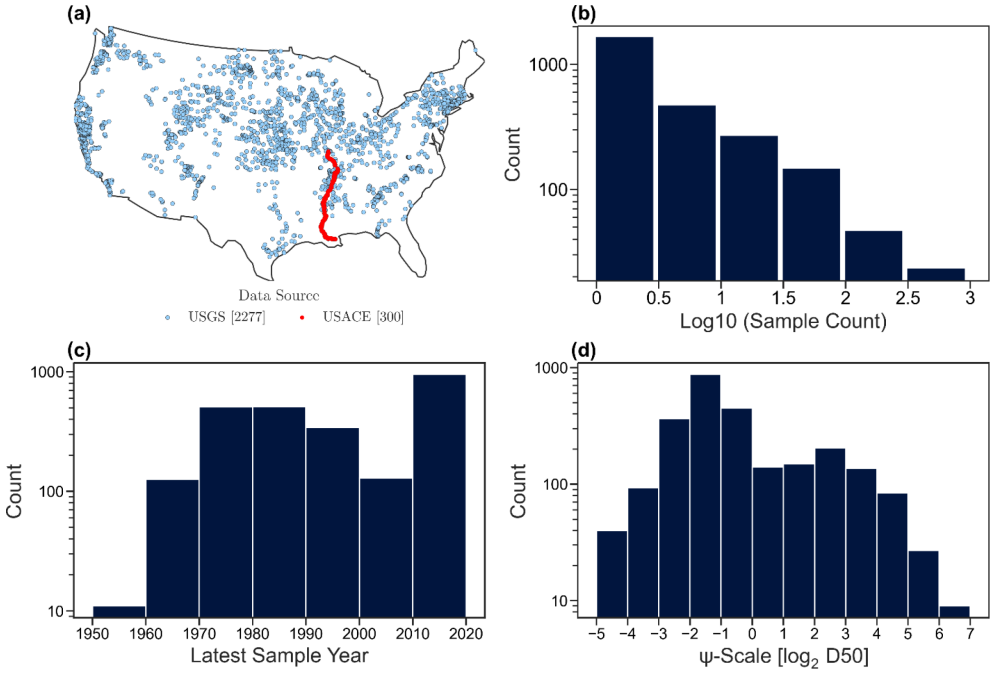

Figure 1. Sediment sample stations. **(a)** Locations of 2277 USGS stations and 300 USACE sampling locations. **(b)** Number of samples at each station/location. **(c)** Histogram of latest sample years. **(d)** Histogram of D50 values on the Ψ scale ($\log_2$D50, D50 in mm).

polation is performed between the percentiles smaller and larger than the 50th percentile to obtain the D50 value, and (3) for the stations with multiple sampling times, a representative D50 value is computed as the mean D50 value from all the sampling times. We take the mean as a representative D50 to simply account for possible uncertainties in sampling and measurement. Although the sampling and measurement procedures are carefully designed (Edwards and Glysson, 1999), it is practically impossible to avoid uncertainties in such sampling and measurement procedures. Thus, we believe a representative D50 can be better estimated by taking a mean. 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 on the Ψ scale. About 75 % of these D50 values are between 0.0625 and 2.0 mm. It is suggested that a river can be a sand-bed or gravel-bed river if the D50 value is below or above 2.0 mm (García, 2008). The D50 values computed from the observed sediment particle size distributions thus mostly reflect sand-bed river conditions, while only approximately 25 % are gravel-bed rivers.

One might wonder how the sites with observed D50 values are distributed between small and large streams (e.g., whether or not smaller streams have more observed D50 data than larger streams). We use stream order (Fig. S3a and b in the Supplement) and upstream drainage area (Fig. S3c and d) as the indicators of stream size and examine the distributions of flowline lengths (Fig. S3a, c) and D50 samples (Fig. S3b and d), respectively. The total flowline length increases with the stream size (i.e., stream order or drainage area), which is expected since overall larger rivers have longer lengths. Interestingly, the number of D50 stations follows a bell distribution except for the largest stream order or drainage area, which is primarily due to the USACE measurements on the lower Mississippi River (198 sample locations). Therefore, there is no clear indication that larger or smaller streams dominate the D50 data points.

2.2 Predictive variables

The predictive variables are retrieved from the National Hydrography Dataset Plus (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 (https://www.sciencebase.gov, last access: 6 February 2022). 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 up-

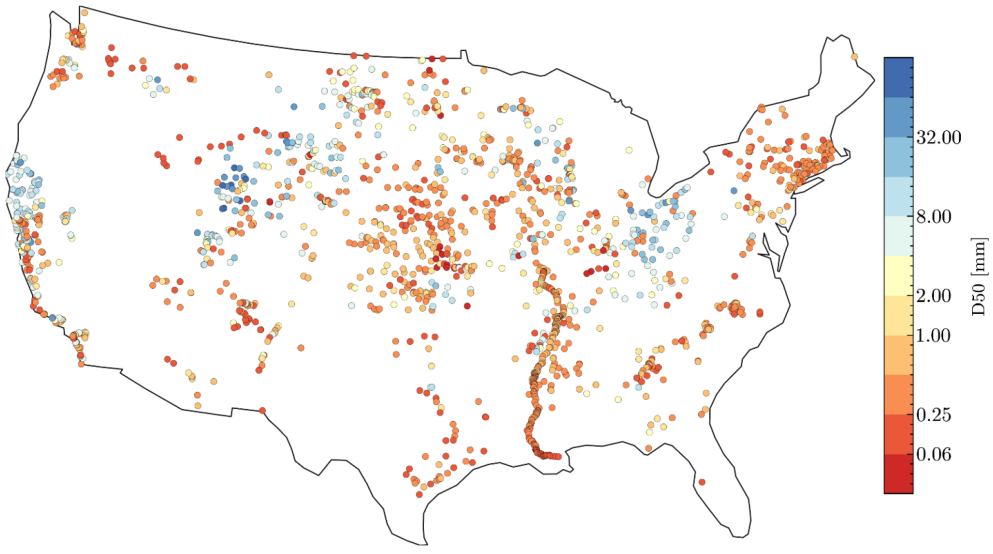

Figure 2. The 1691 flowlines with measured D50.

stream drainage areas in 13 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 Table S1 in the Supplement and four illustrative maps in Fig. S4 in the Supplement.

We then establish the spatial correspondence between the observed D50 values and the 76 predictive variables. In NHDPlus, there are ∼ 26 000 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 that have 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. There are 12 flowlines with 2 sampling locations and 2 flowlines with 3 sampling locations. In those cases, we assign the average of the D50 values of these USGS stations to the flowline. The mean length of the 14 flowlines is 6.63 km. In such a length, only two or three sampling locations cannot capture the spatial variability in a meaningful way. Therefore, we simply calculate the average without making further assumptions. We further exclude a few flowlines with missing attribute values. Finally, we have a total of 1691 flowlines corresponding to the observed D50 values, as shown in Fig. 2. As such, 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; Fan et al., 2021; Zheng et al., 2019). The relations between the input predictors (e.g., watershed characteristics) and D50 are too complex to be established with traditional linear regression or dimensionless analysis methods. 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, 2017) for feature importance analysis and thus feature selection. We also consider the representativeness of input predictors in the feature selection. More details are explained as follows.

3.1 XGBoost: eXtreme Gradient Boosting

Tree boosting is a machine learning framework that combines weak learners to develop a strong learner, in which 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. One of the most recent offspring of gradient boosting techniques is the XGBoost, a scalable end-to-end tree boosting system (Chen and 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 $\mathbf{D} = \{(X_i, Y_i), i = 1, 2, \ldots, N\}$, where $\mathbf{X}_i = [X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip}]$ is a row vector with input features with real value elements and $Y_i \in R$. The tree ensemble

model employs M additive functions to predict the output of interest as

$$\hat{Y}_i = \phi(X_i) = \sum_{m=1}^{M} f_m(X_i), \quad f_m \in F, \tag{1}$$

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., 2015) 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 $\mathbf{D} = \{D_1, D_2, D_3, \ldots, D_k\}$. 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 are 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.

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

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 is, the closer the match between the observed and simulated series is. KGE offers some advantages over commonly used metrics like root mean squared errors (RMSEs) or the 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 (Lundberg and Lee, 2017). 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.

Then, we further examine the representativeness of the data by comparing the ranges of the selected features between the D50-available data and the nationwide data. We use percentiles 2.5 and 97.5 to represent the lower and higher ends of ranges in the available data. We do not directly use the absolute min and max values to avoid the impacts of outliers. We then calculate the percentage of the nationwide data below the percentile 2.5 of the available data. A percentage value of no more than 10 % indicates a good match of lower ends between the available and nationwide data. Similarly, we calculate the percentage of the nationwide data above the percentile 97.5 of the available data. A percentage value of no more than 10 % indicates a good match of upper ends between the available and nationwide data. Taken together, for any feature, if more than 80 % of the nationwide data are located within the percentiles 2.5 and 97.5 of the available data, we consider that the available data are sufficiently representative of the nationwide data for this specific feature. Otherwise, a feature is considered non-representative and thus is removed from model development. 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 and illustrated in Fig. 3:

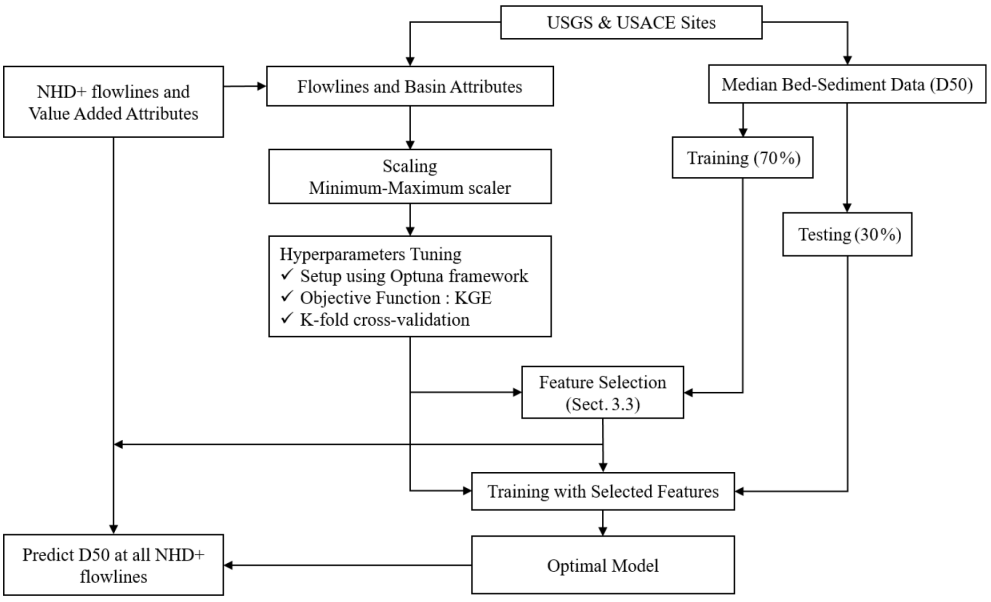

Figure 3. Flowchart for XGBoost training and prediction.

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 suppress the effect of outliers.

2. The dataset is randomly split into training (70 %) and testing (30 %) sets. 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.

4. Feature selection is performed as described in Sect. 3.3, so Step 3 is repeated with the new and smaller set of predictors. Steps 3 and 4 are repeated until no more predictors 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 US 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 Fig. 1b, 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 Sect. 3.3 and 3.4), 12 out of 76 predictive variables, or predictors, are eventually selected. Firstly, 13 variables are identified as more significant than a random-number input vector based on the mean absolute SHAP value, as shown in Table 1. A total of 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 (Soil_erod_factor), followed by average annual wet day

Table 1. Most important predictors according to the feature selection.

Predictor		Description	Mean absolute SHAP value (with slope)	Mean absolute SHAP value (without slope)
Name used in this study	Name in NHDPlus			
Basin_slope	TOT_Basin_slope	Average topographic slope within the upstream drainage area	0.30	0.42
Ann_runoff	TOT_RUN	Average annual runoff within the upstream drainage area	0.23	0.41
Chan_length	Pathlength	Distance from the downstream end of a flowline to the end of the network (river mouth)	0.34	0.39
Ann_snow_perc	TOT_PRSNOW	Mean annual snow as a percent of total precipitation	0.37	0.37
Aridity_index	AI	Aridity index defined as the ratio of annual mean potential evaporation to annual mean precipitation	0.29	0.37
Ann_wet_days	TOT_WDANN	Average annual number of wet days	0.46	0.36
Mean_temp	TOT_WBM_TAV	Average mean annual temperature within the upstream drainage area	0.29	0.35
R_factor	TOT_RFACT	*R* factor of Universal Soil Loss Equation	0.34	0.35
T_Qsub	TOT_CONTACT	Time it takes for water to drain along subsurface flow paths to the stream	0.31	0.34
Mean_elev	TOT_ELEV_MEAN	Average surface elevation within the upstream drainage area	0.29	0.31
Qsat_to_Qtotal	TOT_SATOF	Annual saturation overland flow as a percent of total runoff	0.33	0.27
Soil_erod_factor	TOT_KFACT	Soil erodibility factor of Universal Soil Loss Equation	0.51	0.22
Channel_mean_slope	Slope	Channel slope	0.36	

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.

(Ann_wet_days) and mean annual snow as a percent of total precipitation (Ann_snow_perc).

These three basin-related predictors rank higher than the two channel-related characteristics. Channel slope (Slope) and distance between flowline and the river mouth (Chan_length) are found to be the most important channel characteristics for predicting D50, which agrees with the downstream fining phenomena and sediment transport mechanisms (Niño, 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 (García, 2008). However, the feature selection with the XGBoost model and 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 features such as Ann_wet_days, 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 Fig. 4. Soil_erod_factor and Ann_wet_days, the two most important features in Table 1, have correlation coefficients of only 0.08 and 0.06, respectively. Ann_snow_perc has the strongest correlation with D50, with a correlation coefficient of 0.29. The individual scatter plots between D50 and each of the selected features do not show any apparent relationship between D50 and any single feature (see Fig. S5 in the Supplement), indicating that there might exist some higher-order interactions among the predictors which the traditional regression analysis cannot reveal.

We further examine the representativeness of the 1691 flowlines with observed D50 values of the rest of the flowlines included in the NHDPlus database in terms of the ranges of the 13 features. For convenience, we denote the subset of NHDPlus data associated with the 1691 flowlines

Table 2. Comparison of the ranges and percentiles of 13 input features between the D50-available and nationwide datasets.

Attributes	Percent of nationwide data that is		Relative difference in percentiles between the D50-available and nationwide data		
	below percentile 2.5 of the D50-available data	above percentile 97.5 of the D50-available data	25th	50th	75th
Soil_erod_factor	5.1	3.4	0.07	0.03	0.03
R_factor	6.9	10.2	0.22	0.09	0.44
Ann_wet_days	2.8	3.1	0.01	0.07	0.03
Ann_snow_perc	0.0	3.9	0.71	0.20	0.12
Channel_mean_slope	0.0	19.9	1.72	1.44	1.43
Chan_length	2.3	5.2	0.07	0.21	0.04
Ann_runoff	4.3	1.6	0.00	0.28	0.23
Qsat_to_Qtotal	0.0	7.1	2.00	0.13	0.38
T_Qsub	6.2	3.1	0.68	0.59	0.37
Basin_slope	9.3	4.8	0.17	0.04	0.10
Mean_elev	9.9	2.5	0.35	0.27	0.22
Aridity_index	2.6	6.5	0.07	0.14	0.03
Mean_temp	4.5	8.6	0.02	0.12	0.18

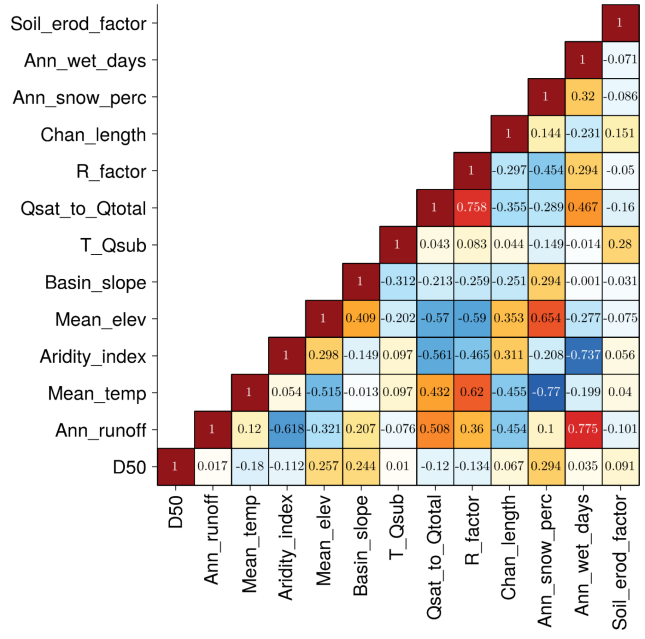

Figure 4. Correlation coefficients among D50 and the 12 selected predictors.

as *D50-available* and the whole NHDPlus database as *nationwide*. For most of the 13 features, the percentages of the nationwide data that are beyond the lower or higher ends of the D50-available are no more than 10 %, except for channel slope, i.e., "Channel_mean_slope". Table 2 also lists the relative difference in the 25th, 50th, and 75th percentiles between the D50-available and nationwide data. For instance, for the 25th percentile, we calculated the relative difference as the ratio of the difference between the 25th percentile of the D50-available data and that of the nationwide data over the average between the 25th percentile of the D50-available data and that of the nationwide data. This relative difference is less than 0.5 for most of the features, again except for "Channel_mean_slope". For a better visual illustration, Fig. 5 shows the cumulative distribution functions (CDFs) and corresponding 5th, 25th, 50th, 75th, and 95th percentiles. The CDFs are close between the D50-available and nationwide data, except for "Channel_mean_slope". A similar message can be seen from the box plots in Fig. S6 in the Supplement. In addition, we would like to point out that the 1691 sampling stations we use to train and test our model are located across the whole contiguous US and are hence geographically representative. Therefore, we conclude that the 1691 flowlines with observed D50 values are representative enough of all the flowlines nationwide in terms of the 12 input predictors, except for "Channel_mean_slope".

We remove "Channel_mean_slope" and use the remaining 12 predictors to develop the final model, following the same model training and testing procedures as before. Figure 6 shows the comparison of model performances between the previous and new models. The model performance metrics are similar. Actually, R^2 became slightly better in both training (0.834 vs. 0.830) and testing (0.405 vs. 0.367), while KGE became slightly worse in training (0.775 vs. 0.794) but better in testing (0.527 vs. 0.513). The slight decrease in KGE in training data is reasonable since the model hyperparameter tuning was based on the objective of maximizing KGE, and losing one predictor will slightly reduce the space of parameter tuning. Nevertheless, now the KGE value in the testing phase is closer to that in the training phase.

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 mechanis-

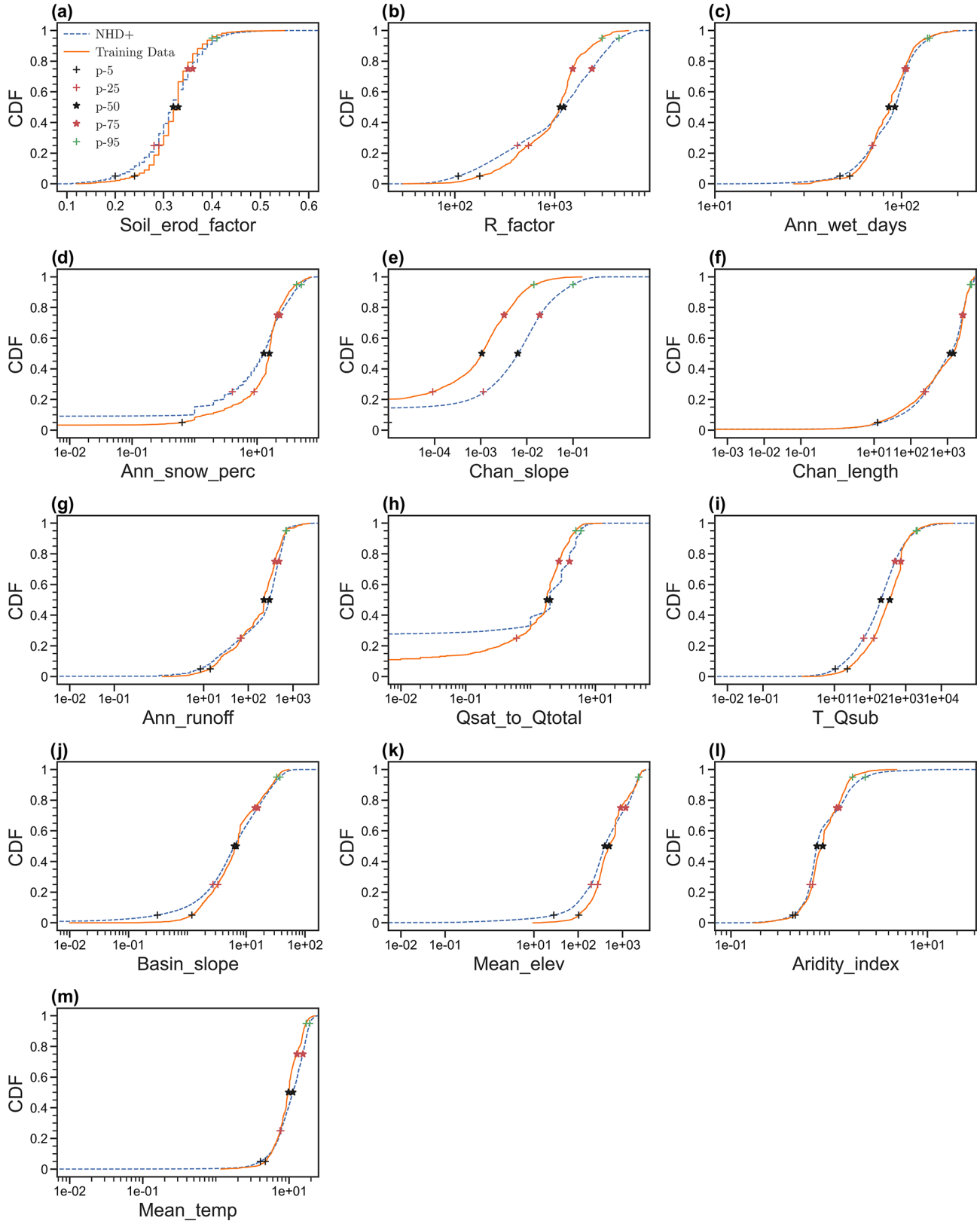

Figure 5. Comparison of the cumulative distribution function (CDF) of 13 features between training data and all flowlines (i.e., NHDPlus).

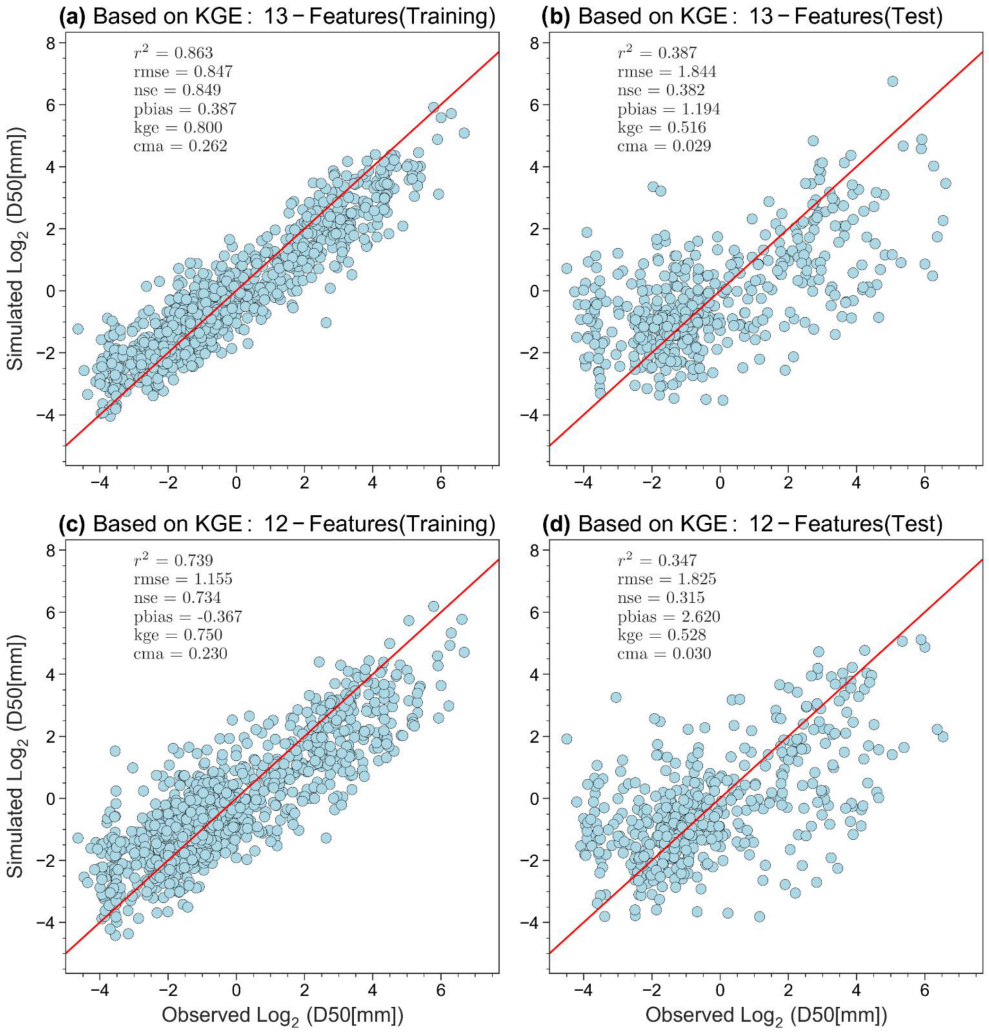

Figure 6. XGBoost model performance with the training **(a, c)** and testing **(b, d)** datasets. Comparison of model performances using 13 features **(a, b)** and 12 features (**c, d**; after removing "Channel_Mean_Slope").

tically 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. In the 12 selected predictors, only one is directly related to the channel processes. The remaining 11 are all land features, and their mechanistic connections with D50 are rather mysterious at this stage, which could be considered as empirical evidence of the likely causal, yet highly complicated, relationships between D50 and the land features and which could hopefully inspire future studies to shed light on the underlying mechanisms.

Table 3. Optimal value of the XGBoost model hyperparameters.

Hyperparameter	Optimal value	Tuning range
learning_rate	0.442	[0,1]
min_split_loss	11	[0,∞]
max_depth	7	[0,∞]
min_child_weight	21	[0,∞]
max_delta_step	41	[0,∞]
subsample	0.408	[0,1]
colsample_bytree	0.741	[0,1]
reg_lambda	0.311	[0,∞]
reg_alpha	4.054	[0,∞]
n_estimators	178	[1,∞]

4.2.2 Model hyperparameters and performance

Table 3 shows the tuned hyperparameters of the best XGBoost model that is trained using the 12 selected predictors and 70 % of the training dataset. For a detailed explanation of the hyperparameters please refer to Chen and Guestrin (2016). Figure 6 shows the performance of the optimal XGBoost model on the respective training and testing datasets. Here we consider an optimal model based on two criteria: (1) the model performance is satisfactory in both the training and testing phases, 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 0.75 for training and 0.528 for testing. The testing value is above 0.5, suggesting satisfactory model performance (Gupta et al., 2009; Knoben et al., 2019). The performance of the testing data is noticeably worse than that of 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 sensitivity analysis

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

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, we test the sensitivity of the choices of model performance metrics and k value. For the model performance metrics, we have also tried Nash–Sutcliffe efficiency (NSE) and R^2 and found that using KGE gave better model performance (Fig. S7 in the Supplement) due to two reasons: (1) a much smaller percentage of bias (PBIAS) and (2) visually better alignment between the simulated and observed D50 values along the 1 : 1 line. 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 1 of the 12 variables at a time and repeating the same modeling procedure for the remaining 12 variables. Figure 7 shows that dropping the variables leads to the model performance dropping below KGE 0.5 during the testing phase in most cases. The change in testing-phase KGE ranges between 4 % and 22 %. The largest changes are observed when dropping R_factor and Mean_elev. 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.

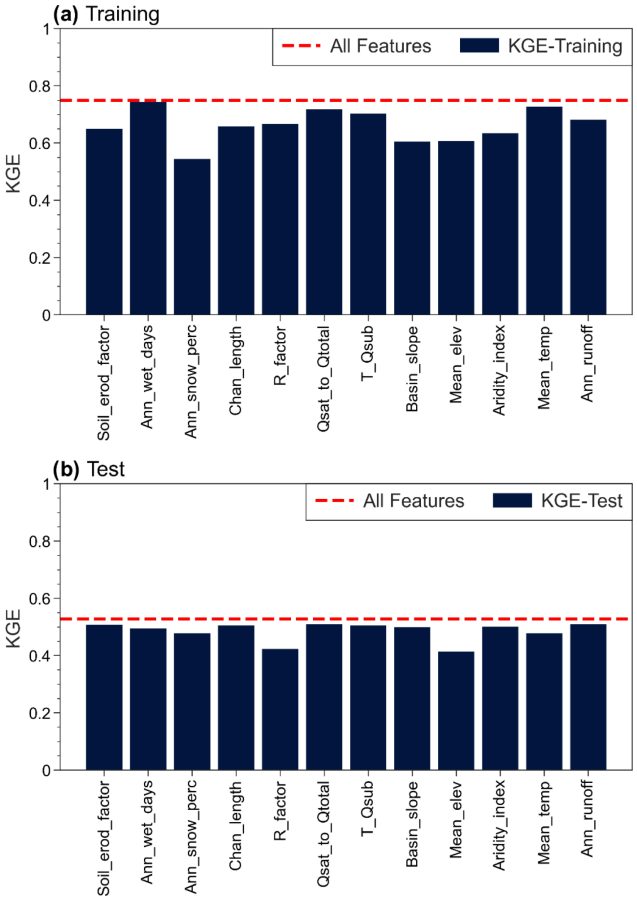

Figure 7. Sensitivity of the XGBoost model to the selected features. The result shown in blue bars are obtained by dropping the corresponding labeled feature from the 13 selected features. The dashed red line represents the model performance with all variables.

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 (Fig. 8). The spatial pattern of D50 in Fig. 8 is generally consistent with the observed D50 in Fig. 2. High D50 values are mostly distributed on the west coast, upper Missouri, and Ohio regions, and low D50 values are concentrated on the east coast. The consistency between Figs. 2 and 8 suggests that the observed D50 data are reasonably representative of the whole contigu-

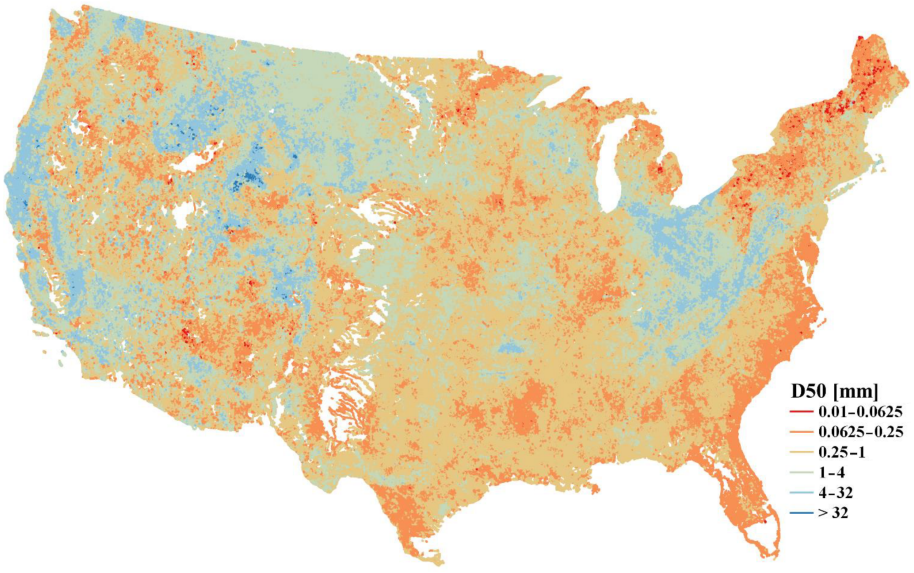

Figure 8. Predicted D50 in ∼ 2.7 million flowlines across the contiguous US using the XGBoost model.

ous US, despite the sparse distribution. Given that the testing dataset 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. To our best knowledge, it is the first-of-its-kind D50 data for the whole contiguous US. Such a D50 map is mostly valuable to support the parameterization of large-scale sediment modeling at the regional or national scale, which has been very challenging, if not impossible, before this map.

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 US (i.e., consistent spatial patterns between Figs. 3 and 8), 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, Fig. 6 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 because it would require (a) a highly integrated, process-based model that considers at least sediment erosion, deposition, and transport processes in both hillslopes and channels, and (b) well-designed numerical experiments to isolate the dominant processes and controlling factors. (3) We have not explicitly incorporated the effects of lakes and reservoirs but rather assumed these effects have been indirectly reflected in the NHDPlus hydrologic attributes adopted in the predictive model. (4) The bed sediment at the gauge station may not always be representative of the reach. Edwards and Glysson (1999) characterized how most of the bed sediment samples were collected and composited at a cross-section by the USGS over the years. Gauge stations are established at cross-sections in the stream where flow measurements are convenient and with conditions conducive to high-quality flow measurements – the issue of whether the bed sediment composition represents the reach is generally not taken into account when the gauge station location is established. 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 US or a major river basin such as the Mississippi River basin. For example, we have tested the usage of the new D50 dataset within a large-scale suspended sediment modeling framework (Li et al., 2021b), and our successful model validation against the USGS-observed suspended sediment load over multiple stations suggests the good value of such a national-scale D50 dataset. There is inevitably some uncertainty embedded in

this map sourced from the original D50 observations and NHDPlus attributes, the XGBoost modeling, and the spatial extrapolation process. This uncertainty should be taken into account when utilizing this map for regional-scale assessment or modeling.

7 Data availability

The national D50 map is freely available at https://doi.org/10.5281/zenodo.4921987 (Li et al., 2021a). The input data are obtained from the USGS water quality portal (https://nwis.waterdata.usgs.gov/usa/nwis/qwdata, U.S. Geological Survey, 2016), NHDPlus (https://www.epa.gov/waterdata/nhdplus-national-data, U.S. Geological Survey and U.S. Environmental Protection Agency, 2020), and ScienceBase (https://doi.org/10.5066/F7765D7V, Wieczorek et al., 2018).

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 had not been feasible previously.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/essd-14-929-2022-supplement.

Author contributions. GWA conducted the analysis with input from HYL, ZZ, ZT, and LRL. HYL and ZZ designed the study. HYL and LRL conceived the idea. All the authors contributed to the writing.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. This research was supported as part of the Energy Exascale Earth System Model (E3SM) project, formerly known as Accelerated Climate Modeling for Energy (ACME), funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. This research was performed in part with resources from the Research Computing Data Core at the University of Houston. The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RLO1830.

Financial support. This research has been supported by the U.S. Department of Energy via Lawrence Livermore National Laboratory (contract no. B633822).

Review statement. This paper was edited by Jens Klump and reviewed by David Waterman and two anonymous referees.

References

Ackers, P. and White, W. R.: Sediment Transport: New Approach and Analysis, J. Hydraul. Div., 99, 2041–2060, https://doi.org/10.1061/JYCEAJ.0003791, 1973.

Afan, H. A., El-shafie, A., Mohtar, W. H. M. W., and Yaseen, Z. M.: Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction, J. Hydrol., 541, 902–913, https://doi.org/10.1016/j.jhydrol.2016.07.048, 2016.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., July 2019, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

An, C., Gong, Z., Naito, K., Parker, G., Hassan, M. A., Ma, H., and Fu, X.: Grain Size-Specific Engelund-Hansen Type Relation for Bed Material Load in Sand-Bed Rivers, With Application to the Mississippi River, Water Resour. Res., 57, e2020WR02751, https://doi.org/10.1029/2020WR027517, 2021.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D.: Hyperopt: A Python library for model selection and hyperparameter optimization, Comput. Sci. Discov., 8, 014008, https://doi.org/10.1088/1749-4699/8/1/014008, 2015.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, August 2016, vol. 42, pp. 785–794, ACM, New York, NY, USA, 2016.

Corcoran, P. L., Belontz, S. L., Ryan, K., and Walzak, M. J.: Factors Controlling the Distribution of Microplastic Particles in Benthic Sediment of the Thames River, Canada, Environ. Sci. Technol., 54, 818–825, https://doi.org/10.1021/acs.est.9b04896, 2020.

Dalu, T., Tshivhase, R., Cuthbert, R. N., Murungweni, F. M., and Wasserman, R. J.: Metal Distribution and Sediment Quality Variation across Sediment Depths of a Subtropical Ramsar Declared Wetland, Water, 12, 2779, https://doi.org/10.3390/w12102779, 2020.

Edwards, T. K. and Glysson, G. D.: Field Methods for Measurement of Fluvial Sediment: U.S. Geological Survey Techniques of Water-Resources Investigations 3–C2, Tech. Water-Resources Investig. U.S. Geol. Surv., Reston, Virginia, 1999.

Einstein, A. H.: The Bed-Load Function for Sediment Transportation in Open Channel Flows United States Department Of Agriculture Soil Conservation Service, Tech. Bull., 1026, United States Department Of Agriculture, Soil Conservation Service, Washington, D. C, available at: https://naldc.nal.usda.gov/download/CAT86201017/PDF (last access: 6 February 2022), 1950.

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

Fan, C., Song, C., Liu, K., Ke, L., Xue, B., Chen, T., Fu, C., and Cheng, J.: Century-Scale Reconstruction of Water Storage Changes of the Largest Lake in the Inner Mongolia Plateau Using a Machine Learning Approach, Water Resour. Res., 57, e2020WR028831, https://doi.org/10.1029/2020WR028831, 2021.

Gaines, R. A. and Priestas, A. M.: Particle Size Distribution of Bed Sediments along the Mississippi River, Grafton, Illinois, to Head of Passes, Louisiana, November 2013, U.S. Army Engineer Research and Development Center, Coastal & Hydraulics Laboratory, 2016.

Garcia, M. and Parker, G.: Entrainment of Bed Sediment into Suspension, J. Hydraul. Eng., 117, 414–435, 1991.

García, M. H.(Ed): Sedimentation Engineering, in: American Society of Civil Engineers, p. 1155, InTech, Reston, VA, 2008.

Glaser, C., Zarfl, C., Rügner, H., Lewis, A., and Schwientek, M.: Analyzing particle-associated pollutant transport to identify in-stream sediment processes during a high flow event, Water (Switzerland), 12, 1794, https://doi.org/10.3390/w12061794, 2020.

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

He, B., Wijesiri, B., Ayoko, G. A., Egodawatta, P., Rintoul, L., and Goonetilleke, A.: Influential factors on microplastics occurrence in river sediments, Sci. Total Environ., 738, 139901, https://doi.org/10.1016/j.scitotenv.2020.139901, 2020.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

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

Li, H.-Y., Tan, Z., Ma, H., Zhu, Z., Abeshu, G., Zhu, S., Cohen, S., Zhou, T., Xu, D., and Leung, L.-Y. R.: A new large-scale suspended sediment model and its application over the United States, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-491, in review, 2021b.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, 31st Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 4768–4777, available at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (last access: 6 February 2022), 2017.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., and Rea, A.: NHD Plus Version 2: User Guide, U.S. Environmental Protection Agency, 2012.

Meyer-Peter, E. and Müller, R.: Formulas for Bed-Load transport, in: Proced. 3rd Meet. IAHR, Stockholm, Sweden, 39–64, 1948.

Niño, Y.: Simple Model for Downstream Variation of Median Sediment Size in Chilean Rivers, J. Hydraul. Eng., 128, 934–941, https://doi.org/10.1061/(asce)0733-9429(2002)128:10(934), 2002.

Parker, G.: Surface-based bedload transport relation for gravel rivers, J. Hydraul. Res., 28, 417–436, https://doi.org/10.1080/00221689009499058, 1990.

Parker, G. and Andrews, E. D.: Sorting of Bed Load Sediment by Flow in Meander Bends, Water Resour. Res., 21, 1361–1373, https://doi.org/10.1029/WR021i009p01361, 1985.

Riecki, L. O., Riecki, L. O., and Sullivan, S. M. P.: Coupled fish-hydrogeomorphic responses to urbanization in streams of Columbus, Ohio, USA, PLoS One, 15, 1–29, https://doi.org/10.1371/journal.pone.0234303, 2020.

Unda-Calvo, J., Ruiz-Romera, E., Fdez-Ortiz de Vallejuelo, S., Martínez-Santos, M., and Gredilla, A.: Evaluating the role of particle size on urban environmental geochemistry of metals in surface sediments, Sci. Total Environ., 646, 121–133, https://doi.org/10.1016/j.scitotenv.2018.07.172, 2019.

U.S. Geological Survey: National Water Information System, Water Quality Data for the Nation [data set], available at: https://nwis.waterdata.usgs.gov/usa/nwis/qwdata (last access: December 2020), 2016.

U.S. Geological Survey and U.S. Environmental Protection Agency: National Hydrography Dataset Plus (NHDPlus) Medium Resolution [data set], available at: https://www.epa.gov/waterdata/nhdplus-national-data, last access: December 2020.

Van Rijn, L. C.: Sediment Transport, Part II: Suspended Load Transport, J. Hydraul. Eng., 110, 1613–1641, 1985.

Wieczorek, M. E., Jackson, S. E., and Schwarz, G. E.: Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States, U. S. Geol. Surv. Data Release [data set], https://doi.org/10.5066/F7765D7V, 2018.

Wu, W., Wang, S. S. Y., and Jia, Y.: Nonuniform sediment transport in alluvial rivers, J. Hydraul. Res., 38, 427–434, https://doi.org/10.1080/00221680009498296, 2000.

Xia, X., Jia, Z., Liu, T., Zhang, S., and Zhang, L.: Coupled Nitrification-Denitrification Caused by Suspended Sediment (SPS) in Rivers: Importance of SPS Size and Composition, Environ. Sci. Technol., 51, 212–221, https://doi.org/10.1021/acs.est.6b03886, 2017.

Zhang, L., Zhang, H., Tang, H., and Zhao, C.: Particle size distribution of bed materials in the sandy river bed of alluvial rivers, Int. J. Sediment Res., 32, 331–339, https://doi.org/10.1016/j.ijsrc.2017.07.005, 2017.

Zhang, W., Wang, H., Li, Y., Lin, L., Hui, C., Gao, Y., Niu, L., Zhang, H., Wang, L., Wang, P., and Wang, C.: Bend-induced sediment redistribution regulates deterministic processes and stimulates microbial nitrogen removal in coarse sediment regions of river, Water Res., 170, https://doi.org/10.1016/j.watres.2019.115315, 2020.

Zheng, Z., Ma, Q., Jin, S., Su, Y., Guo, Q., and Bales, R. C.: Canopy and Terrain Interactions Affecting Snowpack Spatial Patterns in the Sierra Nevada of California, Water Resour. Res., 55, 8721–8739, https://doi.org/10.1029/2018WR023758, 2019.