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

5   [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)

10   **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

15   corresponds to ~~millions of~~approximately 2.7 million river segments (i.e., flowlines) in the National Hydrography Dataset Plus (~~NHDplus~~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~~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

20   D50 data and the corresponding climate, hydrology, geology and other attributes retrieved from the ~~NHDplus~~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, Abeshu et al., 2021).

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~~2008). 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~~2008). 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 in the Engelund-Hansen equation (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~~3644) include bed-material sediment particle data for rivers over the contiguous U.S., and 0.~~5~~6% (2277) have complete percentiles of bed-material sediment particle data ~~to compute~~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 U.S. Such correlations can then be applied across the U.S. for predicting sediment particle size.

60   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).

65   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.

70   **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~~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 Figure 1a. There are 1367 USGS stations

75   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; 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

80   between Head of Passes, Louisiana and Grafton, Illinois. We exclude the 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~~at different ~~years~~times, although almost half of the locations are sampled only once (see Figure 1b~~). Figure 1c shows~~ for the ~~years that~~ histogram and Figure S1a for the spatial map). For about 94% of these stations, the latest samples

85   were taken. ~~About 94% of these locations have been sampled~~ after ~~1970~~the 1970s (see Figure 1c for the histogram and Figure S1b for the spatial map). We calculated the coefficient of variation (CV) for the 760 stations that have at least 5 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 Figure S2). 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 ~~Krumbein phi~~Ψ scale (~~Krumbein, 1934~~Parker and Andrews, 1985). 2) A linear interpolation is performed between the percentiles smaller and larger than the 50th 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. 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 & 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 in the ~~Krumbein phi~~Ψ scale. About 75% of these D50 values are between 0.0625 mm and 2.0 mm. ~~Garcia (2008)~~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~~.~~ (Garcia, 2008). The D50 values computed from the observed sediment particle size distributions thus ~~dominantly~~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 (Figure S3a, b) and upstream drainage area (Figure S3c, d) as the indicators of stream size and examine the distributions of flowline lengths (Figure S3a, c) and D50 samples (Figure S3b, 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 ~~NHDplus~~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~~NHDPlus, there are about 2.7 million stream segments (average length of 1.93 km, denoted as flowlines from now on). ~~NHDplus~~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~~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~~ Supplementary Table S1 and four illustrative maps in Supplementary Figure ~~S1~~S4.

125 We then establish the spatial correspondence between the observed D50 values and the 76 predictive variables. In ~~NHDplus~~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~~that have observed D50 values~~)~~ with stations in ~~NHDplus~~NHDPlus. Some of the 2277 USGS stations are not included in ~~NHDplus~~NHDPlus, so we obtain 1530 matching

130 stations. The 300 USACE sampling locations are collocated with the flowlines via their geographic coordinates. ~~Several USGS stations or USACE~~There are 12 flowlines with 2 sampling locations ~~may be on the same flowline.~~and 2 flowlines with 3 sampling locations. In ~~that case~~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

135 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~~, as shown in Figure 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

140 algorithm known for its high efficiency and superior performance in recent years (Chen and Guestrin, 2016; Zheng et al., 2019; Fan et al., 2021). 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, ~~2016~~2017) for feature importance analysis and thus feature selection. We

145 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, 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

150 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 $Y_i \in R$. The tree ensemble model employs M additive functions to predict the output of interest as

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

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~~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, $D = \{D_1, D_2, D_3, \ldots, D_k\}$. 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.

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

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 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
185 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~~2017). In this study, we use the mean absolute SHAP values as feature importance measures.

190 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.~~

195 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 the 2.5th and 97.5th percentiles to represent the lower and higher ends of ranges in the available data. We do not directly use the absolute min/max values to avoid the impacts of outliers. We then calculate the percentage of the nationwide data below the 2.5th percentile 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
200 nationwide data above the 97.5th percentile 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. Taking together, for any feature, if more than 80% of the nationwide data are located within the 2.5th and 97.5th percentiles of the available data, we consider that the available data is sufficiently representative of the nationwide data for this specific feature. Otherwise, a feature is considered non-representative thus is removed from model development. Lastly, the remaining features are utilized for tuning the final optimal set of
205 hyperparameter values.

3.4 General Steps

The general steps of the model development procedure are as follows and illustrated in Figure 3.

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~~suppress the
210   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.

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~~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 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~~NHDPlus flowlines. Similar to all observed D50 values in Figure ~~1d~~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 sections 3.3 and 3.4), ~~13~~12 out of 76 predictive variables, or predictors, are eventually selected ~~and shown in Table 1. These~~. 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. 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~~TOT_KFACT~~), followed by average annual wet day (Ann_wet_days~~TOT_WDANN~~) and mean annual snow as a percent of total precipitation (~~TOT_PRSNOW~~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 (~~Pathlength~~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 (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~~Manning's equation, ~~Chezy's~~Chezy's

245   law, etc., and channel bed roughness can be related to bed sediment size (Garcia, 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 ~~parameters~~features such as ~~TOT_WDANN~~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

250   the correlation coefficients between D50 and predictors, as shown in Figure ~~3. TOT_KFACT~~4. Soil_erod_factor and ~~TOT_WDANN~~Ann_wet_days, the two most important features in Table 1, have correlation coefficients of only 0.08 and 0.06, respectively. ~~TOT_PRSNOW~~Ann_snow_perc has the strongest correlation with D50, with a correlation coefficient of 0.29. The _individual_ scatter plots between D50 and ~~all~~each of the selected features do not show any apparent relationship between D50 and any single feature (see Supplementary Figure ~~S2). This indicates~~S5), indicating that ~~the XGBoost model can~~

255   ~~reveal~~there might exist some higher-order interactions among the predictors ~~for better predictions~~which the traditional regression analysis cannot reveal.

We further examine the representativeness of the 1691 flowlines with observed D50 values of the rest 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 as _**D50-available**_, and the whole NHDPlus database as _**nationwide**_. For most of the 13

260   features, the percentages of the nationwide data that are beyond the lower or higher ends of the D50-availableare 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

265   is less than 0.5 for most of the features, again except for "Channel_mean_slope". For a better visual illustration, Figure 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 Figure S6. 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 U.S., hence geographically representative. Therefore, we conclude that the

270   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)

275   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 of 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 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. 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 on the likely causal, yet highly complicated relationships between D50 and the land features, and hopefully inspire future studies to shed light on the underlying mechanisms.

4.2.2 Model Hyperparameters and Performance

Table 23 shows the tuned hyperparameters of the best XGBoost model that is trained using the 1312 selected predictors and 70% of the training data set. For a detailed explanation of the hyperparameters please refer to Chen and Guestrin (2016). Figure 46 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, adand 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 0.79475 for training and 0.513528 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 UncertaintySensitivity Analysis

We carry out further analysis to shed light on how the modellingmodeling 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 utilizeutilizes the modellingmodeling 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, theFor Step 3, we test the sensitivity on the choices of model performance metrics and k value, respectively. For the model performance metrics, we have also tried NSE and R^2 and found that using KGE gave better model performance (Figure

S7) 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 one of the ~~13~~12 variables at a time and repeating the same ~~modelling~~modeling procedure for the remaining 12 variables. Figure ~~5~~7 shows that dropping the variables leads to the model performance dropping below KGE ~~=~~0.~~475 in~~5 during the ~~training~~testing phase ~~for~~in most ~~features except for~~cases. The change in testing phase KGE ranges between 4 – 22 %. The largest changes are observed when dropping R_factor~~TOT_SATOF~~ and ~~TOT_WBM_TAV~~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.

4.3 National Map

Using the developed machine learning model and ~~NHDplus~~NHDPlus channel/basin characteristics data, we are able to produce a national map of bed sediment D50 values (Figure ~~6~~8). ~~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~~8 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~~2 and ~~6~~8 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. To our best knowledge, it is the first-of-its-kind D50 data for the whole contiguous U.S. Such a D50 map is mostly valuable to support the parameterization of large-scale sediment modelling 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 U.S. (i.e., consistent spatial patterns between Figures ~~2~~3 and ~~6~~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, Figure ~~4~~6 shows the model tends to overestimate D50 for smaller D50 values (particularly < 0.25 mm) and underestimate D50 for larger D50 values

340 (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.~~ 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

345 lakes and reservoirs but rather assumed these effects have been indirectly reflected in the ~~NHDplus hydrologic attributes adopted in the predictive model.~~ NHDPlus hydrologic attributes adopted in the predictive model. (4) The bed sediment at the gage 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. Gage stations are established at cross-sections in the stream where flow measurements are convenient and with conditions conducive to high-quality flow

350 measurements – the issue of whether the bed sediment composition represents the reach is generally not taken into account when the gage 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

355 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. For example, we have tested the usage of the new D50 dataset within a large-scale suspended sediment modeling framework (Li, Tan et al., 2021), 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 ~~these maps~~this map sourced from the original D50 observations and

360 ~~NHDplus~~NHDPlus attributes, the XGBoost modeling, and the spatial extrapolation process. This uncertainty should be taken into account while ~~evaluating the uncertainties in the model simulation~~utilizing this map for regional-scale assessment or modeling.

~~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~~

365 ~~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

370 The national D50 map is freely available at http://doi.org/10.5281/zenodo.4921987 (Li, Abeshu et al., 2021). The input data are obtained from the USGS water quality portal (https://nwis.waterdata.usgs.gov/usa/nwis/qwdata), ~~NHDplus~~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

375 channel and watershed characteristics from ~~NHDplus~~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

380 writing.

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

385 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.

390

395

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.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

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.~~https://doi.org/10.1029/2020WR027517.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: A Python library for model selection and hyperparameter optimization. Computational Science and Discovery, 8(1). https://doi.org/10.1088/1749-4699/8/1/014008

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~~'16, 163, 785–794.

415 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*

Dalu, T., Tshivhase, R., Cuthbert, R. N., Murungweni, F. M., & Wasserman, R. J. (2020). Metal Distribution and Sediment
420 Quality Variation across Sediment Depths of a Subtropical Ramsar Declared Wetland. Water, 12(10), 2779. https://doi.org/10.3390/w12102779

Edwards, T. K., & Glysson, G. D. (1999). Field methods for measurement of fluvial sediment: US geological survey techniques of water-resources investigations. In T. K. Edwards & G. D. Glysson (Eds.), Applications of hydraulics (pp. 1– 89). Reston, Virginia: United States Geological Survey. https://doi.org/10.3133/ofr86531

425 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
430 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, M. (Ed.) (2008), *Sedimentation Engineering: Processes, Measurements, Modeling, and Practice*, 1150 pp., Am.
435 Soc. of Civ. Eng., Reston, VaVA.

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.

440 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.2009.08.003https://doi.org/10.1016/j.jhydrol.2009.08.003

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

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.

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-
450 23-4323-2019

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

Li, H.-Y.;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
455 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 over CONUS (Version 1.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4921987.http://doi.org/10.5281/zenodo.4921987.

Li, H.-Y., Tan, Z., Ma, H., Zhu, Z., Abeshu, G., Zhu, S., Cohen, S., Zhou, T., Xu, D., & Leung, R. L. (2021). A new large-
460 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

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems, 32(2).

He, B; Wijesiri, B.; Ayoko, G.A.; Egodawatta, P.; Rintoul, L.; Goonetilleke, A. (2020) Influential factors on microplastics
465 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.

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

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

Parker, G., & Andrews, E. D. (1985). Sorting of bed load sediment by flow in meander bends. *Water Resources Research*, *21*(9), 1361-1373.

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.

475   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), 1613-480   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.

485   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.

490   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

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.

495

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

Predictor		Description	Mean Absolute SHAP Value (w/ Slope)	Mean Absolute SHAP Value (w/o Slope)
~~TOT_KFACT~~**Name used in this study**	**Name in NHDPlus**~~Soil erodibility factor of Universal Soil Loss Equation~~			
Basin_slope	TOT_~~WDANN~~Basin_slope	Average topographic slope within the upstream drainage area ~~Average annual number of wet days~~	0.~~46~~30	0.42
Ann_runoff	TOT_~~PRSNOW~~RUN	~~Mean~~Average annual ~~snow as a percent of total precipitation~~runoff within the upstream drainage area	0.~~37~~23	0.41
~~Slope~~		~~Channel bed slope for each flowline~~	~~0.36~~	
Chan_length	Pathlength	Distance from the downstream end of a flowline to the end of the network (river mouth)	0.34	0.39
~~TOT_RFACT~~		~~R factor of Universal Soil Loss Equation~~	~~0.34~~	
Ann_snow_perc	TOT_~~SATOF~~PRSNOW	~~Annual saturation overland flow~~Mean annual snow as a percent of total ~~runoff~~precipitation	0.~~33~~37	0.37
~~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~~	
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_~~RUN~~ELEV_MEAN	Average ~~annual runoff~~surface elevation within the upstream drainage area	0.~~23~~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~~NHDPlus attribute tables, but moderately revise the description using terminologies that can be understood by a broader audience.

505

~~Table 2.~~**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		Relative difference in percentiles between the D50-available and nationwide data		
	below 2.5th of the D50-available data	above 97.5th 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

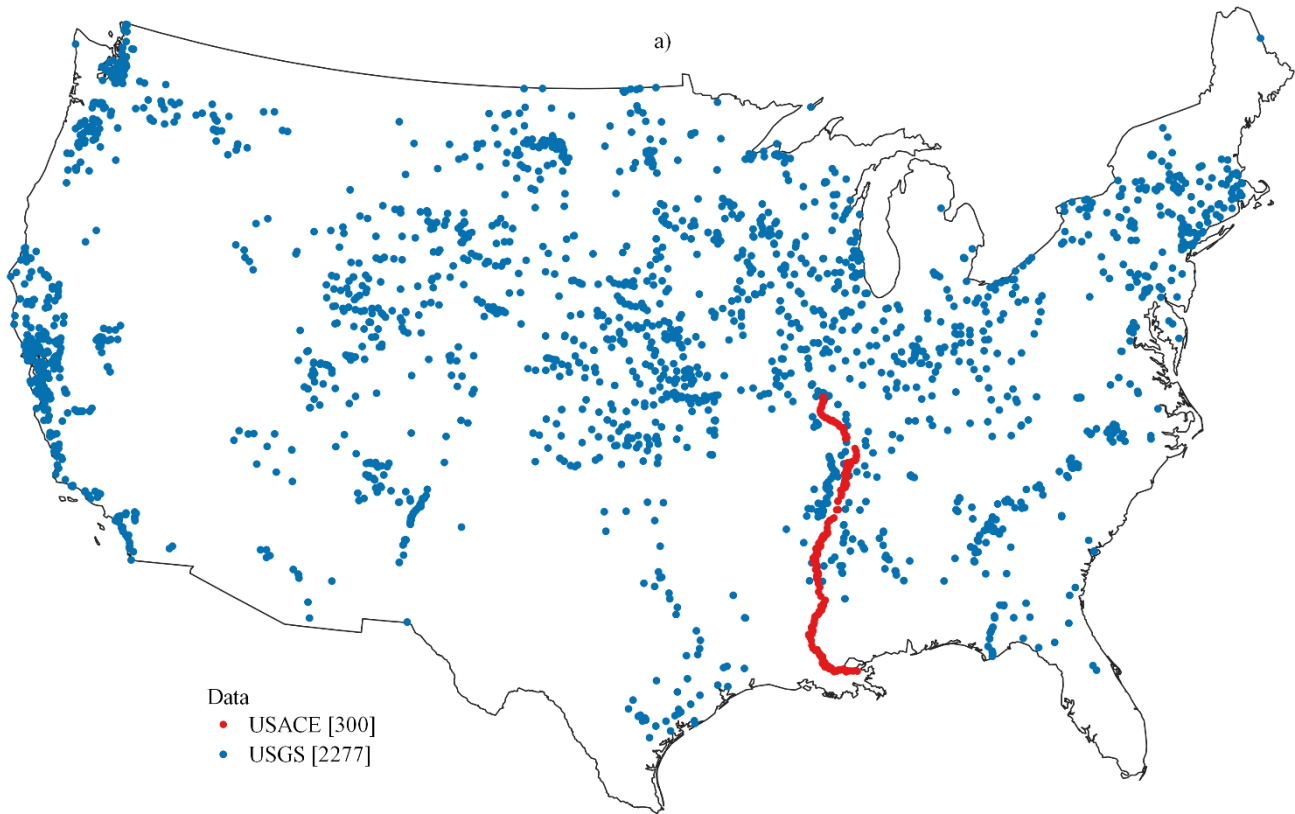
Ann_runoff	4.3	1.6	0.00	0.28	0.23
Qsat_to_Qtotal	0.0	7.1	2.00	0.13	0.38
T_Qsub	6.2	3.1	0.68	0.59	0.37
Basin_slope	9.3	4.8	0.17	0.04	0.10
Mean_elev	9.9	2.5	0.35	0.27	0.22
Aridity_index	2.6	6.5	0.07	0.14	0.03
Mean_temp	4.5	8.6	0.02	0.12	0.18

510

515

520 **Table 3.** **Optimal value of the XGBoost model hyperparameters**

Hyperparameter	Optimal Value	Tuning Range
learning_rate	0.~~825~~442	[0,1]
min_split_loss	~~1~~11	[0,∞]
max_depth	~~6~~7	[0,∞]

min_child_weight	~~58~~21	[0,∞]
max_delta_step	~~22~~41	[0,∞]
subsample	0.~~695~~408	[0,1]
colsample_bytree	0.~~712~~741	[0,1]
reg_lambda	~~26.821~~0.311	[0,∞]
reg_alpha	~~2.561~~4.054	[0,∞]
n_estimators	~~155~~178	[1,∞]

525

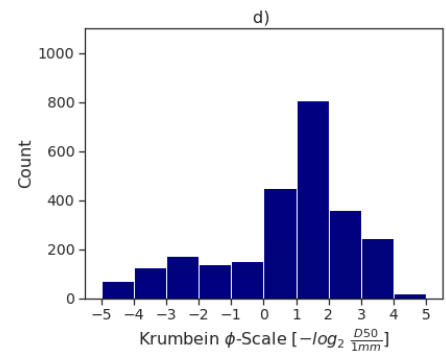
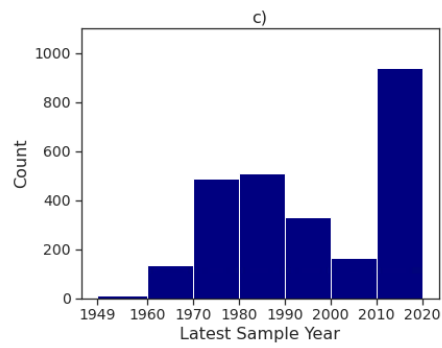
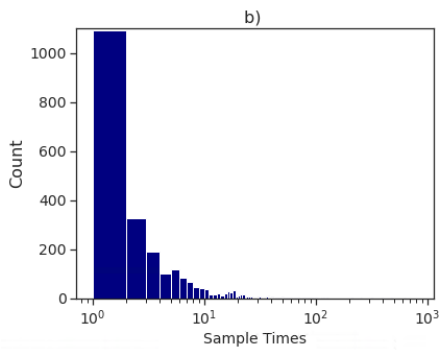
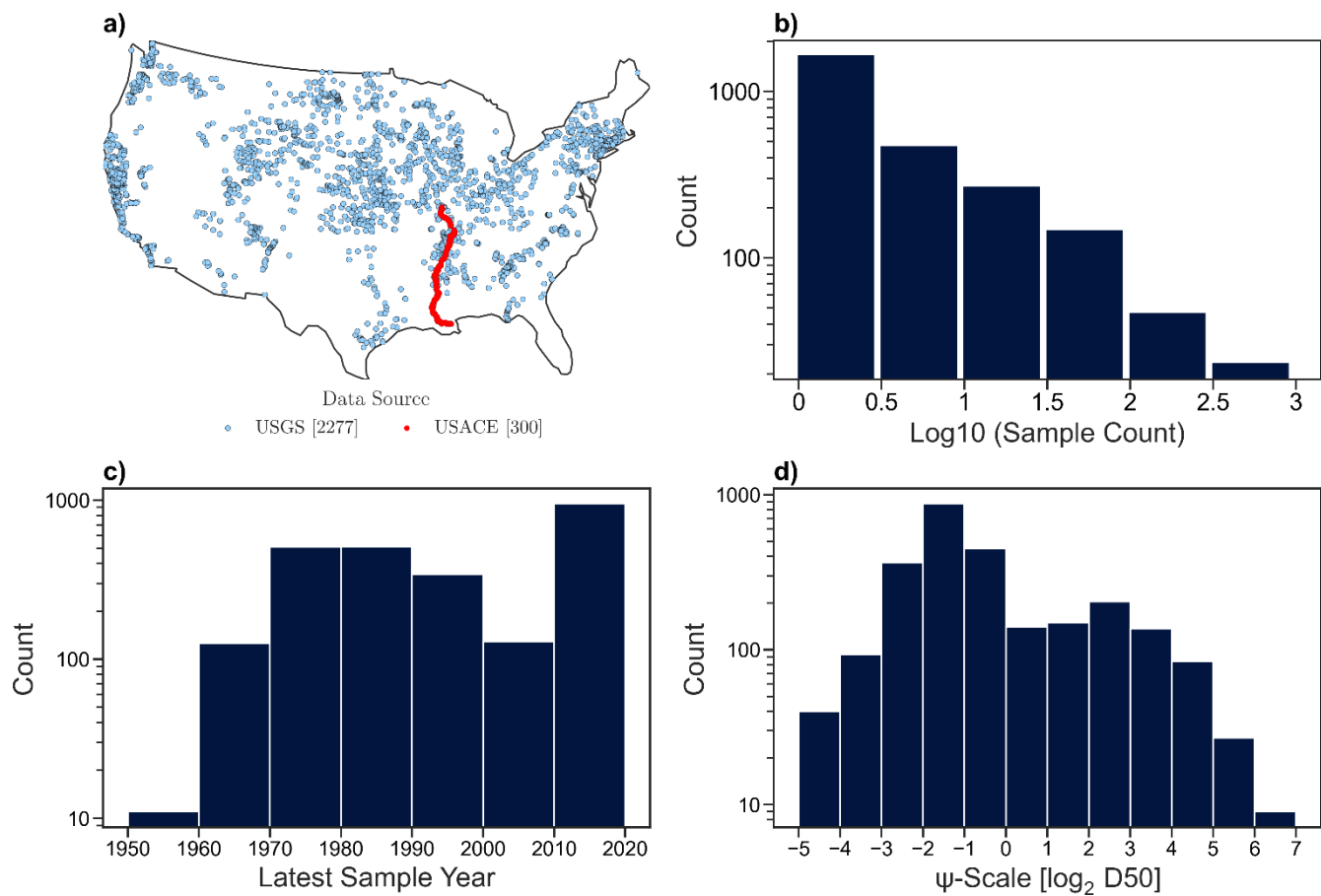
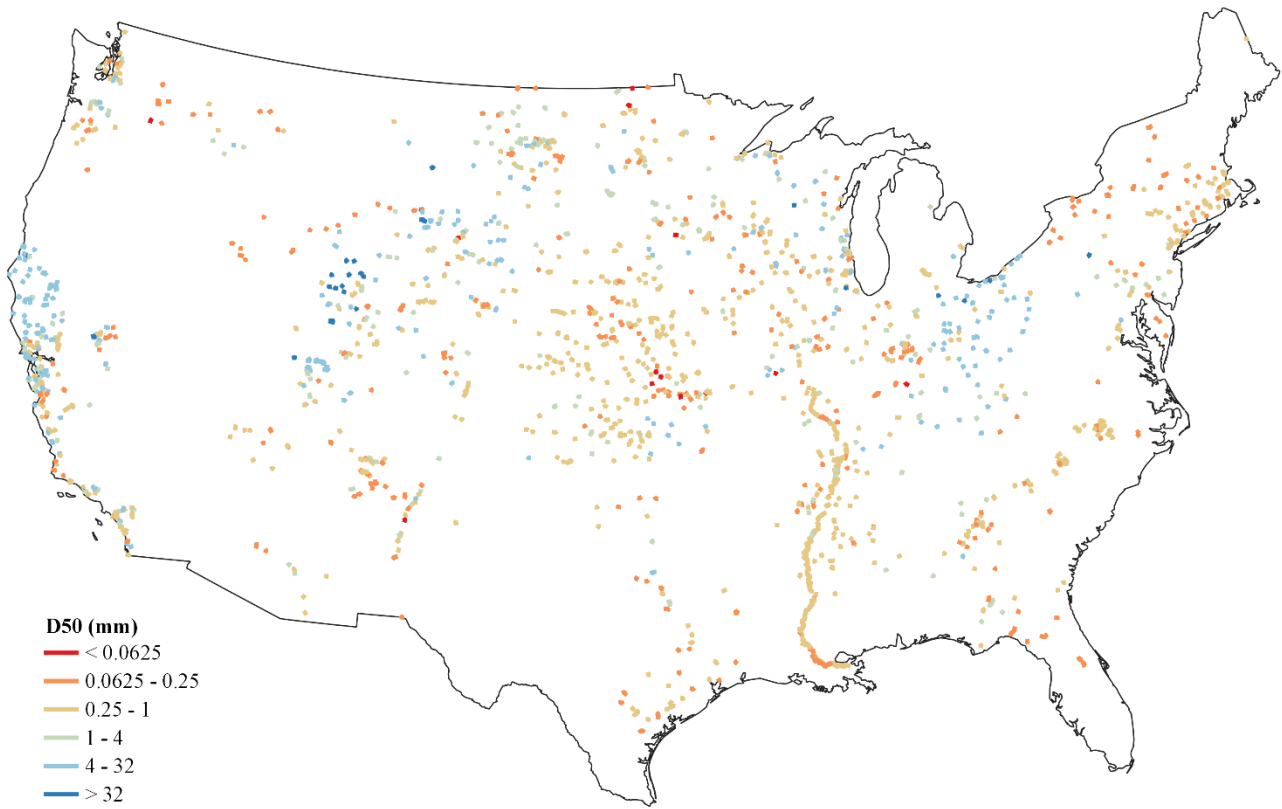



Figure 1: Sediment sample stations. a. ~~Location~~Locations of 2277 USGS stations and 300 USACE sampling locations; **b.** Number of samples at each station/location; **c.** Histogram of latest sample ~~year at each station/location~~years; **d.** Histogram of D50 values~~.~~ in the Ψ scale (log₂D50, D50 in mm).

535

540

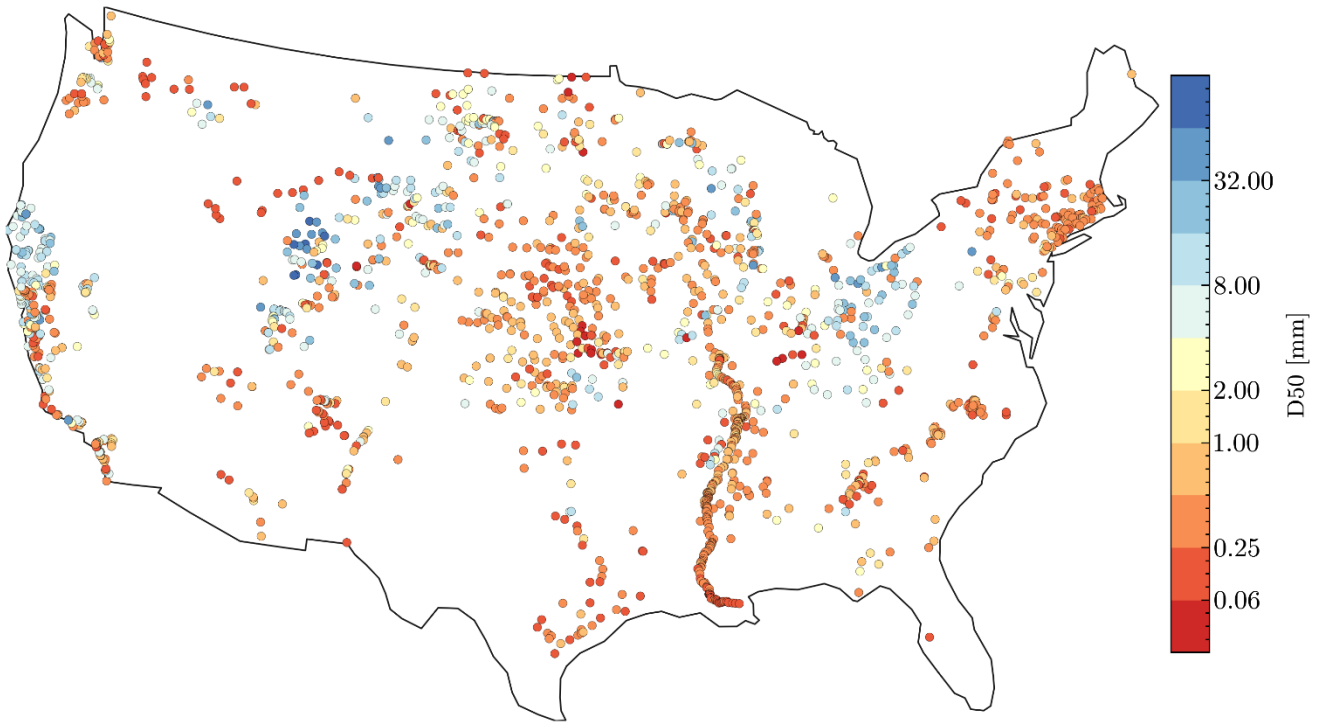
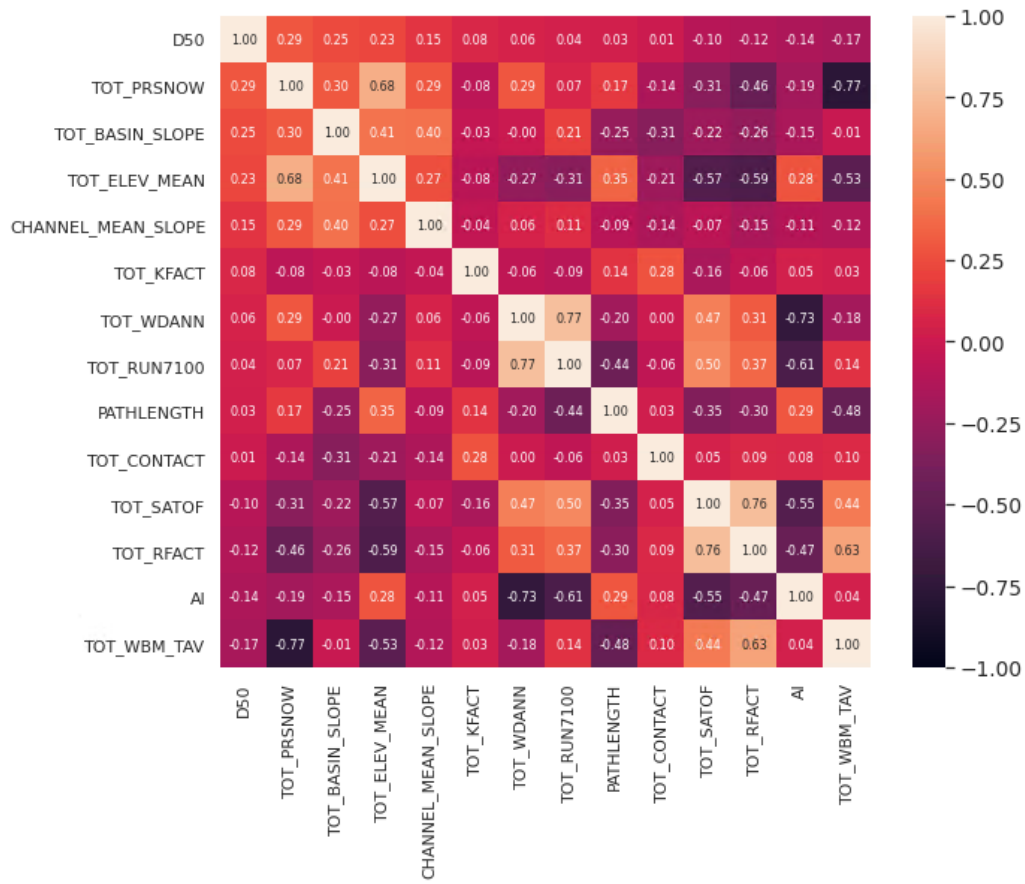
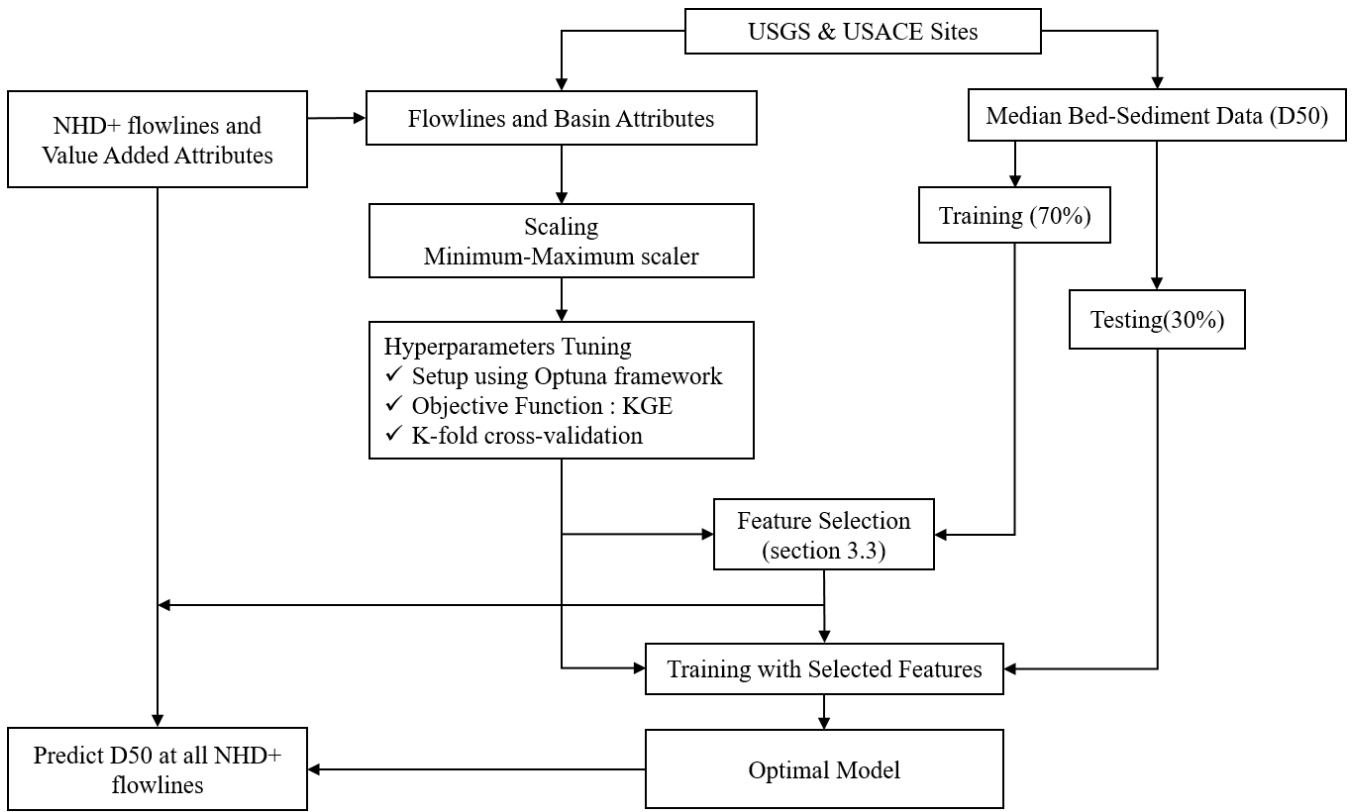
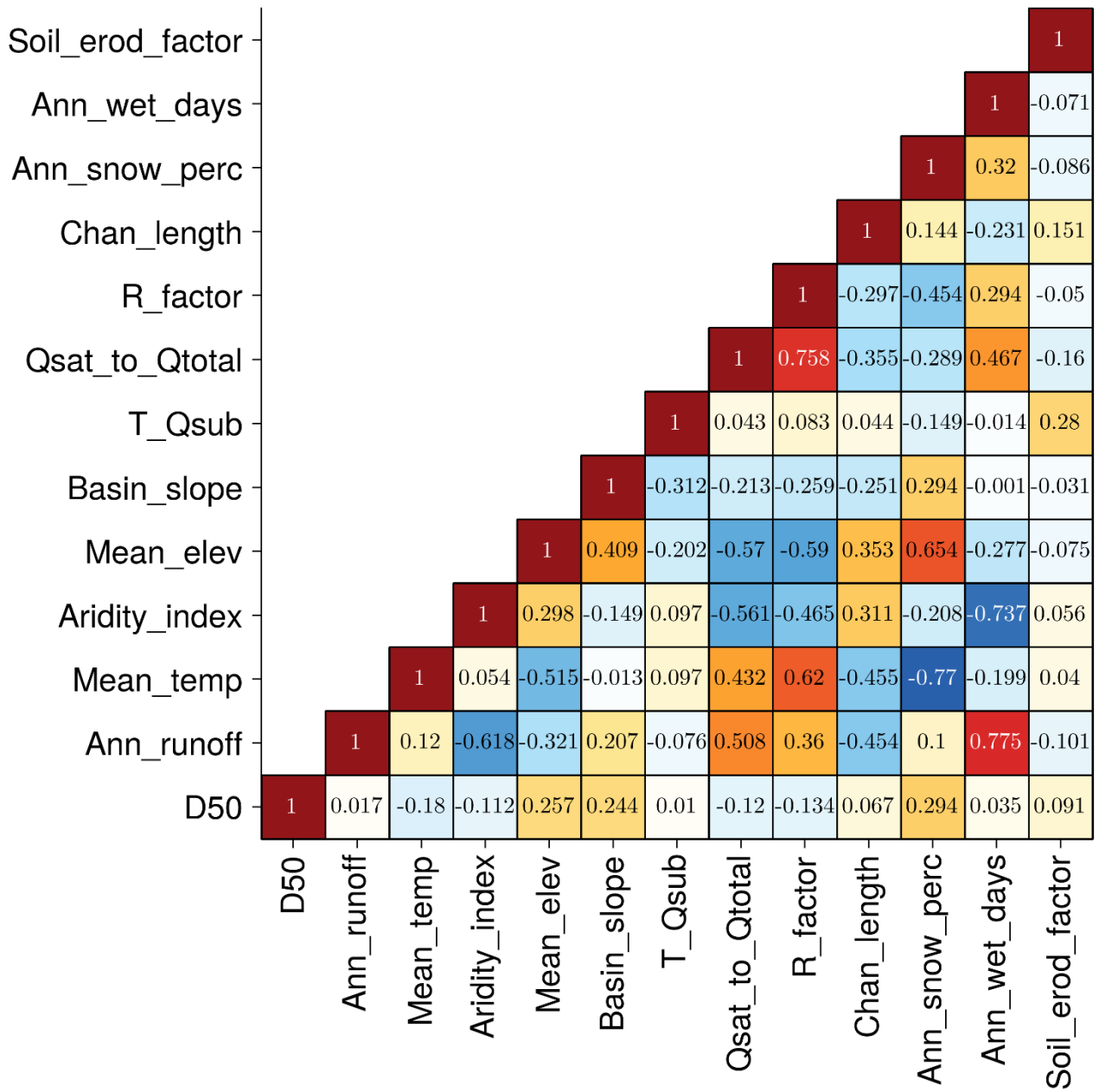



Figure 2. ~~1,691~~1691 flowlines with measured D50.

545

550   **Figure 3: Flowchart for XGBoost training and prediction.**

555

560   **Figure 4**. Correlation coefficients among D50 and the ~~13~~12 selected predictors.

565

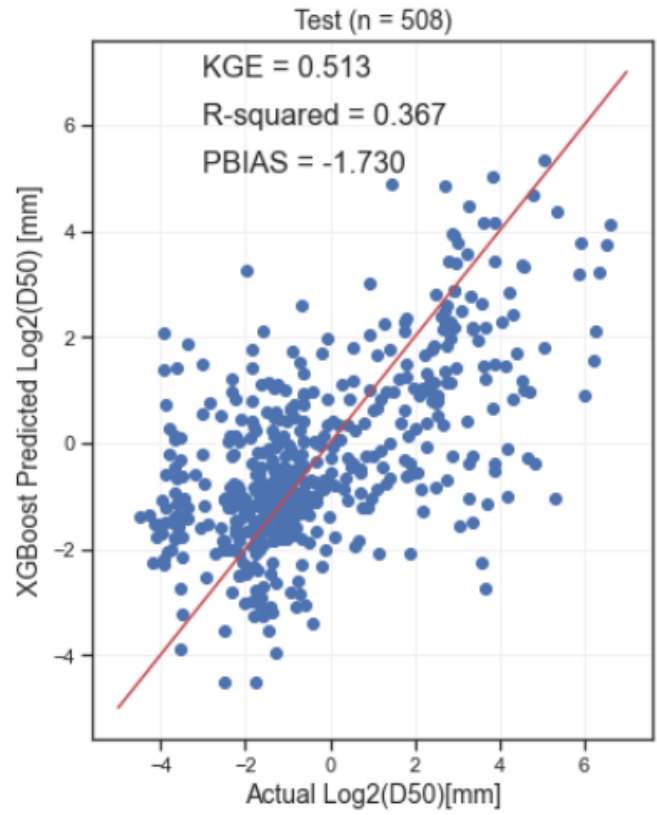
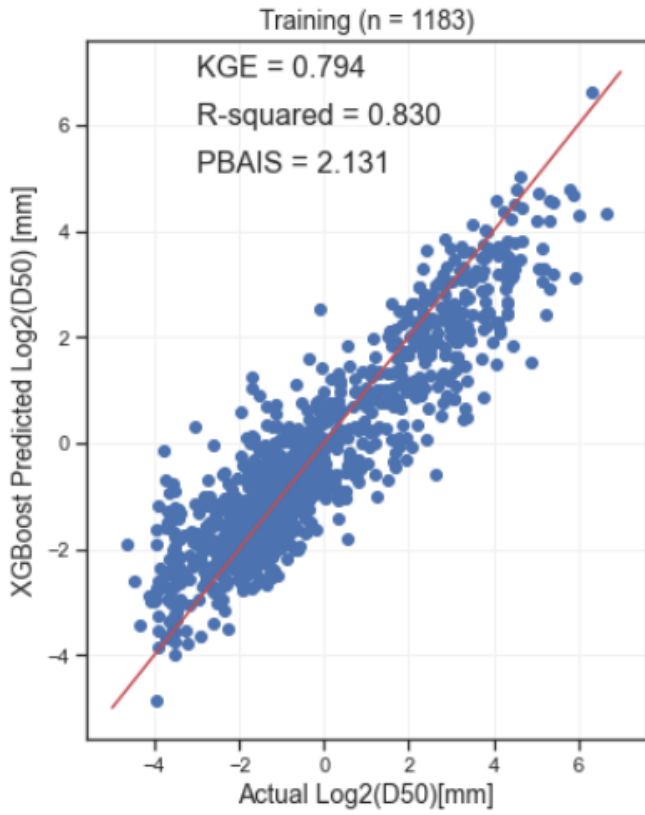
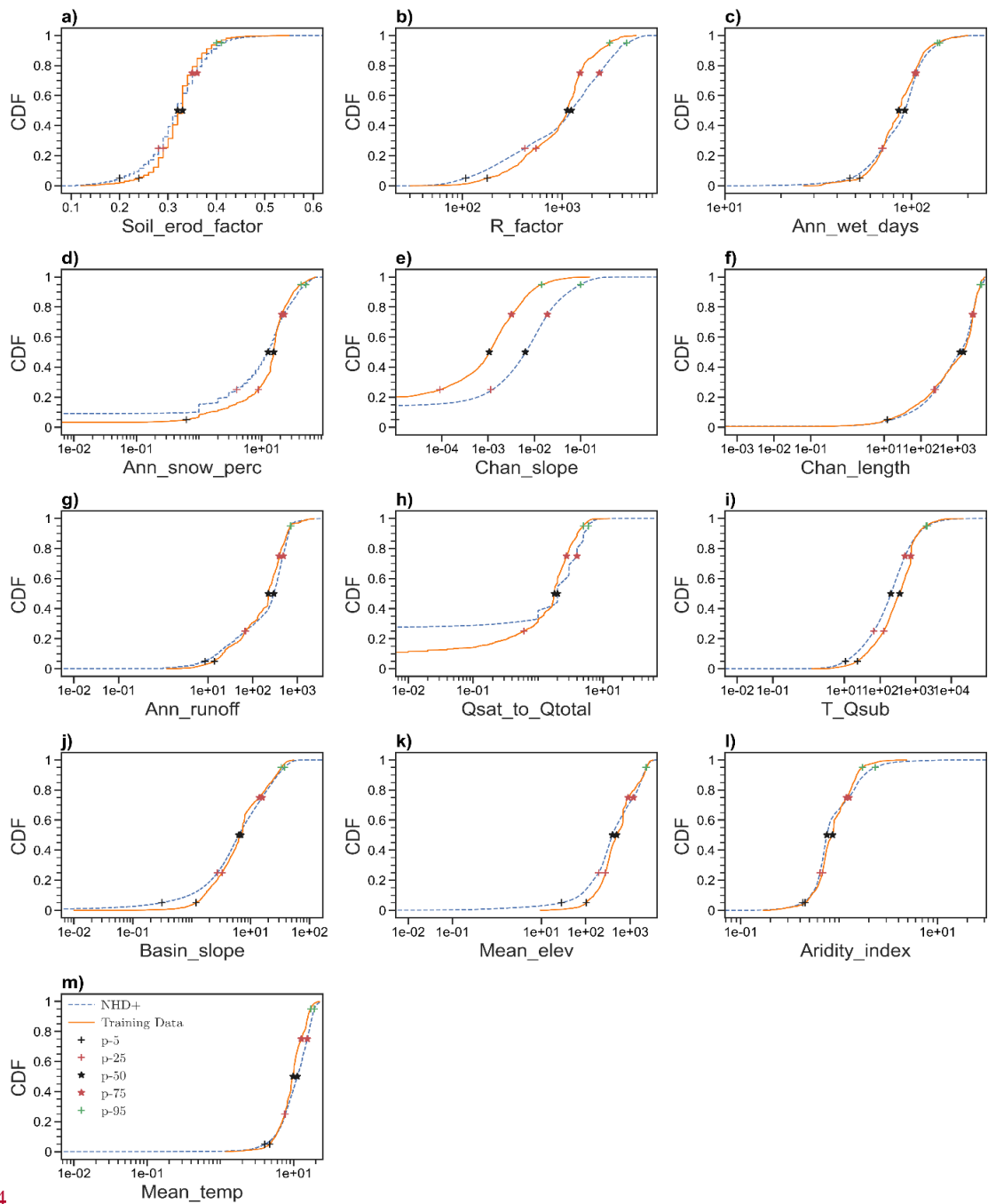
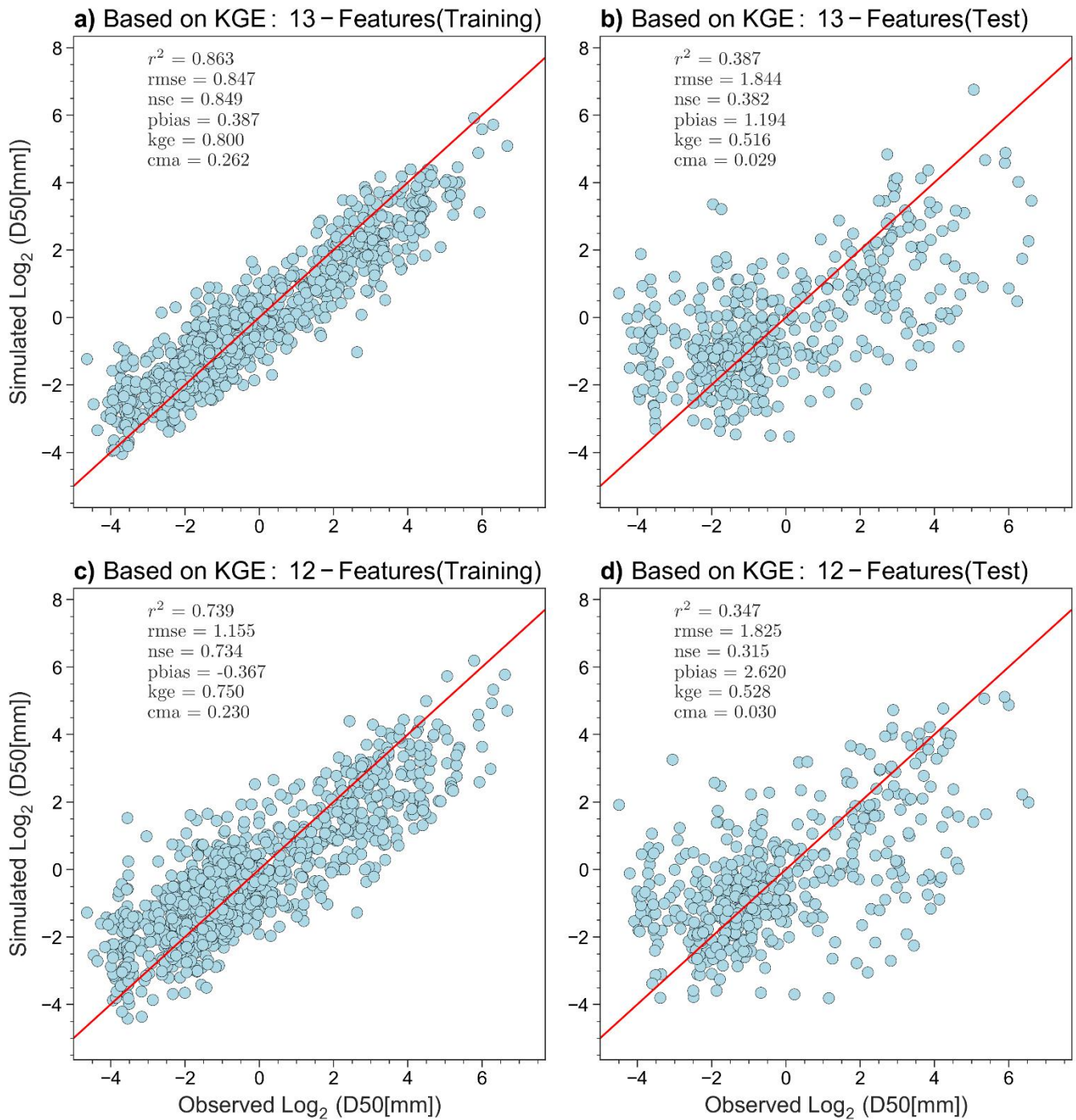
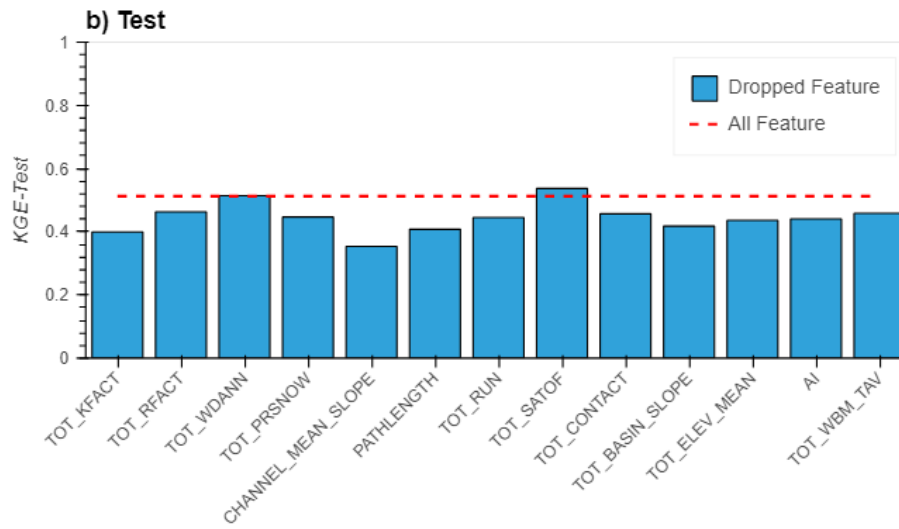
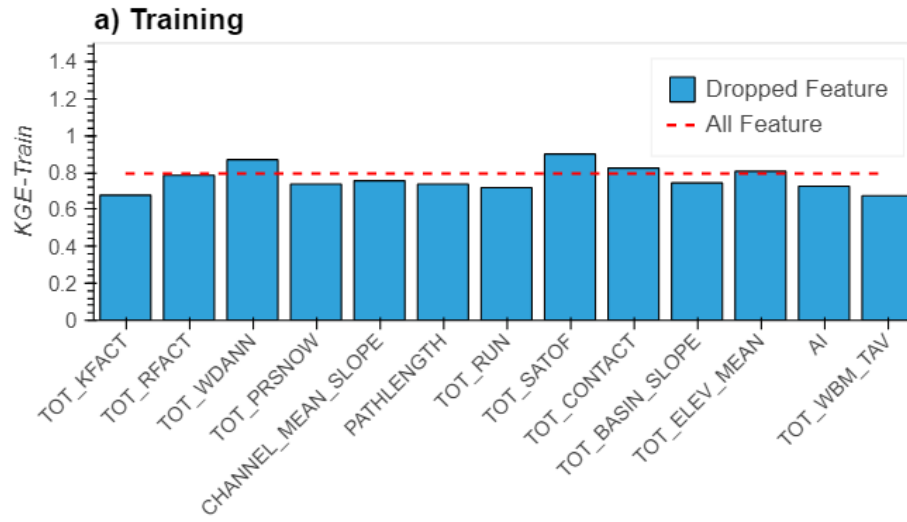
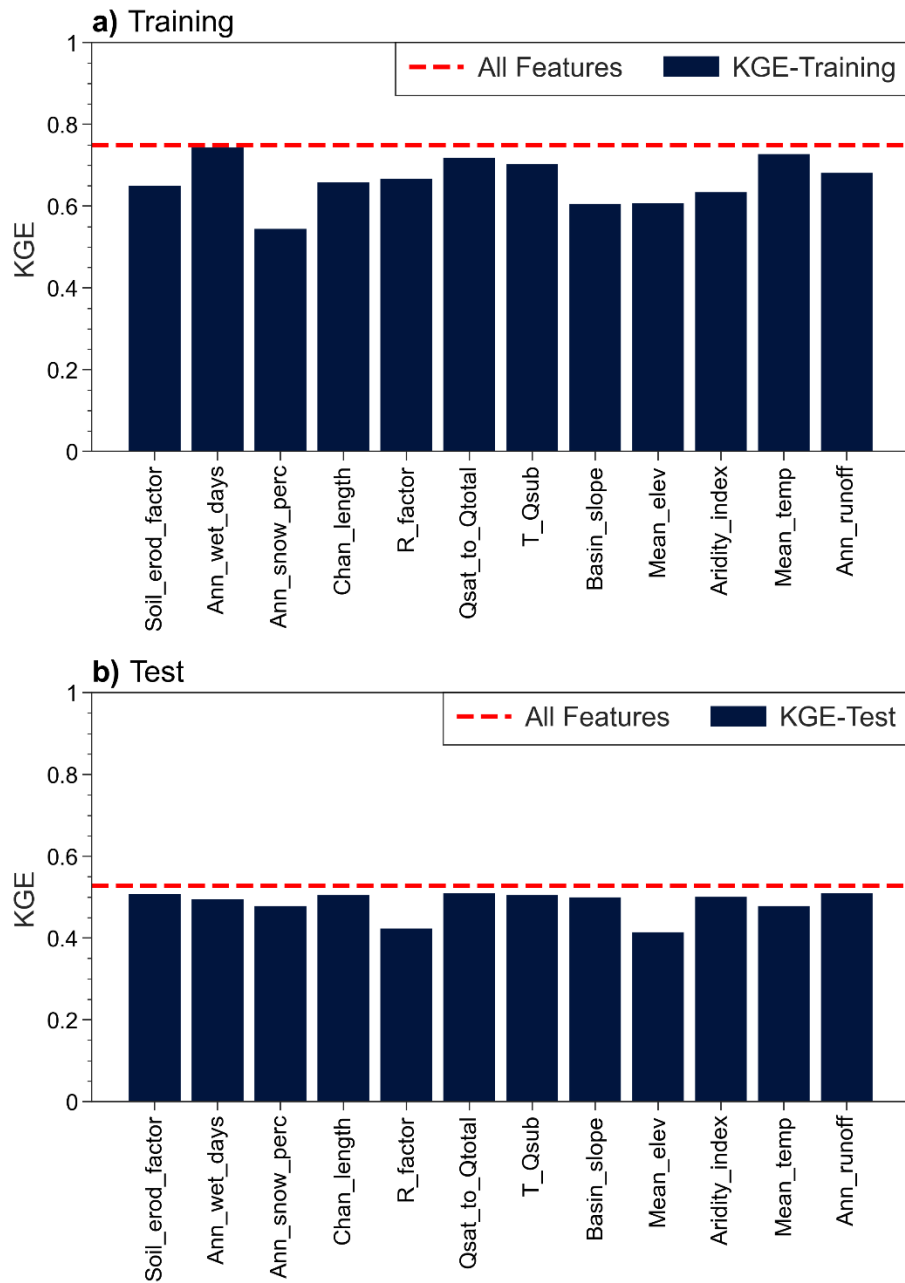



Figure 5: Comparison of the cumulative distribution function (CDF) of 13 features between training data and all flowlines (i.e., NHDPlus).

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

575   **Figure 5̶7. 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.**

D50 [mm]
- 0.01 - 0.0625
- 0.0625 - 0.25
- 0.25 - 1
- 1 - 4
- 4 - 32
- > 32

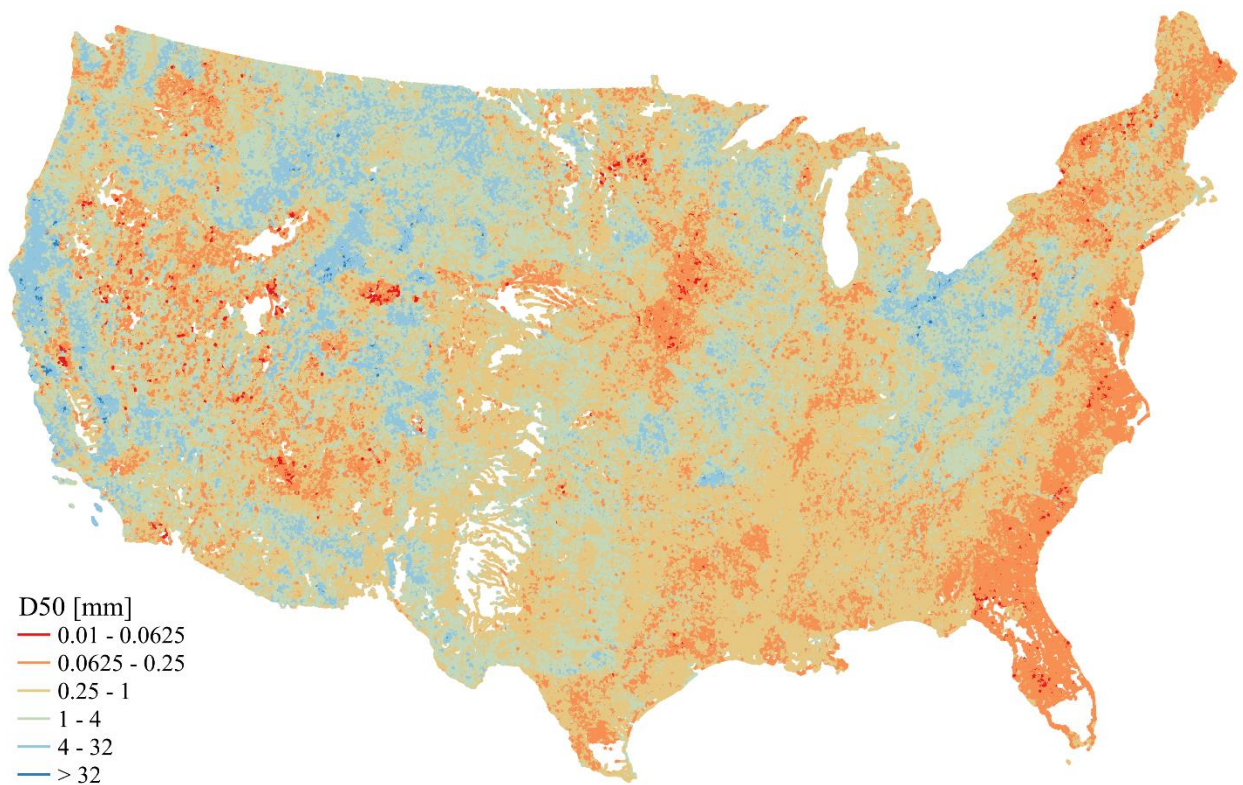
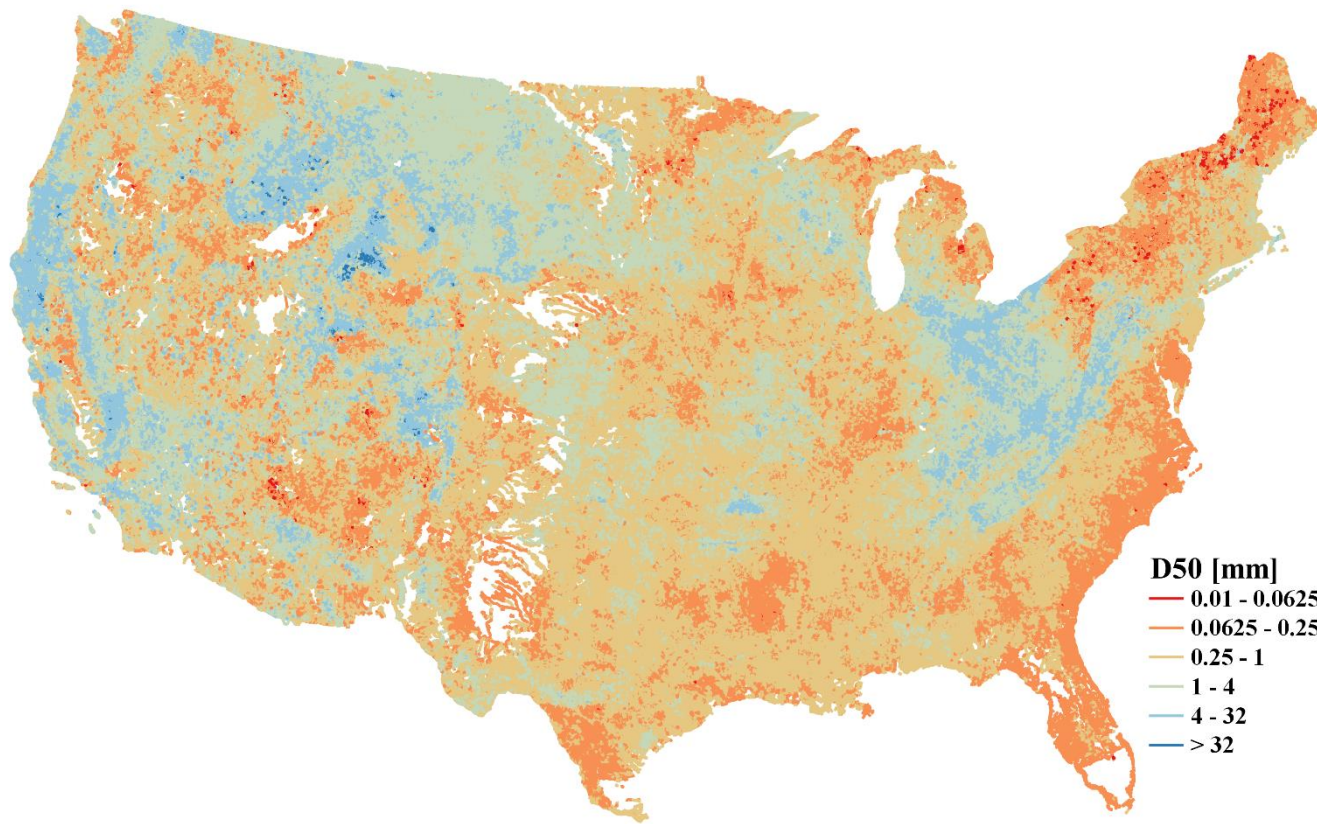



Figure ~~6~~8. Predicted D50 in ~2.7 million flowlines across the contiguous U.S. using the XGBoost model.