Published in final edited form as:
J Am Water Resour Assoc. 2023 October 01; 59(5): 1099–1114. doi:10.1111/1752-1688.13116.

Random Forest models to estimate bankfull and low flow channel widths and depths across the conterminous United States

Jessie M. Doyle,
Ryan A. Hill,
Scott G. Leibowitz,
Joseph L. Ebersole

Oak Ridge Institute for Science and Education c/o Center for Public Health and Environmental Assessment, Pacific Ecological Systems Division (Doyle), U.S. Environmental Protection Agency, Corvallis, Oregon USA; Center for Public Health and Environmental Assessment, Pacific Ecological Systems Division (Hill, Leibowitz, Ebersole), U.S. Environmental Protection Agency, Corvallis, Oregon, USA

Abstract

Channel dimensions (width and depth) at varying flows influence a host of instream ecological processes, as well as habitat and biotic features; they are a major consideration in stream habitat restoration and instream flow assessments. Models of widths and depths are often used to assess climate change vulnerability, develop endangered species recovery plans, and model water quality. However, development and application of such models require specific skillsets and resources. To facilitate acquisition of such estimates, we created a dataset of modeled channel dimensions for perennial stream segments across the conterminous U.S. We used random forest models to predict wetted width, thalweg depth, bankfull width, and bankfull depth from several thousand field measurements of the National Rivers and Streams Assessment. Observed channel widths varied from <5 m to >2000 m and depths varied from <2 m to >125 m. Metrics of watershed area, runoff, slope, land use, and more were used as model predictors. The models had high pseudo R-squared values (0.70 to 0.91) and median absolute errors within ±6% to ±21% of the interquartile range of measured values across ten stream orders. Predicted channel dimensions can be joined to 1.1 million stream segments of the 1:100K resolution National Hydrography Dataset Plus (version 2.1). These predictions, combined with a rapidly growing body of nationally available data, will further enhance our ability to study and protect aquatic resources.

(Correspondence to Hill: hill.ryan@epa.gov).

Research Impact Statement: We developed random forest models that successfully modeled and predicted low flow and bankfull widths and depths for 1.1 million stream segments across the conterminous U.S.

SUPPLEMENTAL MATERIALS
Additional supporting information may be found online under the Supporting Information tab for this article: Supplemental Materials S.1: Correlation matrix of model predictor variables. Supplemental Materials S.2: Data dictionary of response and predictor variables used in modeling. Supplemental Materials S.3: R code used to aggregate National Rivers and Streams Assessment measurements with StreamCat covariates for modeling. Supplemental Materials S.4: Table of observed low flow and bankfull channel widths and depths paired with StreamCat covariates used to develop random forest models. Supplemental Materials S.5: R code of random forest modeling and results.

Keywords

StreamCat; National Rivers and Streams Assessment; Channel Dimensions; Thalweg Depth; Wetted Width; Bankfull Width; Bankfull Depth; Random Forest

INTRODUCTION

Assessing habitat extent and availability (i.e., habitat quantity) is a cornerstone of effective and efficient management of freshwater stream biota. The width and depth of rivers and streams are fundamental measures used to assess aquatic habitat extent and availability. Along with channel slope, these dimensions reflect geomorphic processes that shape physical habitats of streams. The resultant structure and form of rivers and streams influence the abundance and distribution of stream biota, composition of biotic assemblages, as well as an array of other in-stream physical and chemical processes (Angermeier and Winston 1998; Lamouroux et al. 1999; Fausch et al. 2002). Despite the importance of the physical dimensions of streams, they are often unavailable and require substantial field work to collect at scales that are relevant to regional planning and management. To fill this gap, models of width and depth are often used for habitat quantity and quality assessments and as inputs to hydrologic models when field measurements are unavailable. Such spatial models are being increasingly used for applications such as restoration prioritization (Roni et al. 2018), climate change vulnerability assessments for freshwater fishes (Sloat et al. 2017), recovery strategies for threatened and endangered fish species (FitzGerald et al. 2021) and for water quality and streamflow modeling (Mohamoud and Parmar 2006; White et al. 2017; Han et al. 2019). For example, Bond et al. (2019) used estimates of static channel dimensions to assess fish habitat capacity in the Columbia River Basins. In other cases, such as Sloat et al. (2017), estimates of channel dimensions are calculated from regional hydrological models and then used to evaluate increased flood magnitudes on fish spawning habitat. Models such as the Hydrologic and Water Quality System (HAWQS) - which is an application of the Soil and Water Assessment Tool (SWAT) – uses estimates of width and depth at ungaged sites to quantify various water quality endpoints like nutrient and sediment loadings (Ghimire et al. 2021).

Depending on study purpose, models of stream width and depth must be applicable to at least two important flow stages: bankfull and low flow. Bankfull flows are those that reach the transition between the channel and its adjacent floodplain during large storm events or snowmelt (Fig. 1; Leopold et al. 1964). These flows move sediments and form channel features that set the stage for habitats, such as riffles and pools, during non-flood conditions (Leopold 1994; Hey 2006; Parker et al. 2007; Modrick and Georgakakos 2014). In contrast, low flows (also known as baseflows) occur during dry seasons, are sustained by groundwater, and are what organisms experience during much of the year (Humphries and Baldwin 2003; McMahon and Finlayson 2003; Ledford et al. 2020). During low flows, the habitats available for organisms depend on the channel configuration, the sediments moved during bankfull stage, and the water currently available to fill them (Finkenbine et al. 2000; Menció and Boix 2018). For example, channel configuration, structure, and dimensions can

EPA Author Manuscript

EPA Author Manuscript

determine carrying capacity for stream fishes during low or high flows (Fausch et al. 2002; Rosenfeld 2003).

Beyond habitat availability, channel dimensions strongly influence other key features of habitat quality, such as water temperature (Klein 1979; Menció and Mas-Pla 2010; Price 2011; Mayer 2012), nutrient cycling (Alexander et al. 2000; Peterson et al. 2001), bank stability and fine sediments (Menció and Mas-Pla 2010; Price 2011; Ledford et al. 2020). The influence of channel dimensions on habitat and water quality, in turn, affects the distribution of fish and invertebrate species and the composition of stream communities (Rashleigh et al. 2005; Rolls et al. 2012). Furthermore, alterations to channel dimensions due to human activity can have consequences for stream ecosystems, such as altered availability of useable habitat (Beechie et al. 1994) and reduced resiliency in biological communities to disturbances (Detenbeck et al. 1992; Mažeika et al. 2004; Sullivan et al. 2006).

Measurements of width and depth are routinely collected by practitioners in the field to monitor geomorphic and hydrologic conditions, as well as to parameterize models for water quality and fisheries management. However, collecting channel width and depth data across the thousands of sites typically required for robust large-scale modeling can be costly and time consuming. In such cases, modelers may make scaling assumptions between unmeasured and measured channels within the mainstem and tributaries of a stream network (e.g., Neitsch et al. 2005). For example, a common approach to estimate unmeasured channel dimensions is to apply the power functions of Leopold and Maddock (1953) to drainage area (e.g., Stewardson 2005; Johnson and Fecko 2008; Han et al. 2019). However, reliance on drainage areas alone limits the geographic scope over which such relationships can be applied, often constraining their application to specific hydroclimatic regions (Johnson and Fecko 2008; Bieger et al. 2015). Additionally, using the power functions of Leopold and Maddock (1953) requires researchers to assume a channel cross-sectional shape from which they can infer wetted widths and depths (Allen et al. 1994; Ouarda et al. 2008; Ames et al. 2009). Other models, such as those that relate fish distributions to channel dimensions, have included factors in addition to drainage area (e.g., climate, geology, and soils) to improve estimates of channel dimensions and expand the geographic scope over which models can be applied (Stewardson 2005; Faustini et al. 2009; Bond et al. 2019). A recent example of such modeling is from Morel et al. (2020) who used a machine learning technique (random forests) to model widths and depths at mean flows for streams in France and New Zealand. These models performed as well as models from each country alone, indicating the transferability of their approach to other locations. Judes et al. (2021) then applied these predictions to study hydropeaking on fish assemblages within French streams.

Despite the potential transferability of these modeling approaches, practitioners may face several challenges when applying methods to estimate channel dimensions. Management of aquatic resources could benefit substantially from estimates of channel dimensions that are derived from methods and data sources that are open and readily accessible to practitioners (Beck et al. 2020). See et al. (2021) highlights how practitioners often need access to spatially-continuous channel dimension data across large geographic extents to extrapolate estimates of fish carrying capacity to other sites of concern. In this article, we describe the development of conterminous U.S. (CONUS) models of low flow and bankfull channel dimensions (i.e., wetted width, thalweg depth, bankfull width, and bankfull depth). Our goal is to provide a publicly available dataset for managers and researchers that require, but currently lack, channel dimension estimates for large scale modeling applications. With each of these models, we interpolated predictions of channel dimensions to 1.1 million streams across the CONUS. We produced the models and predictions with open and transferrable methods; all code and predicted values produced by this study are provided in Supplemental Materials and online.

METHODS

Channel Dimension Measurements

We compiled channel dimension measurements from the National Rivers and Streams Assessment (NRSA) 2008–09 and 2013–14 surveys (Fig. 2). NRSA is a collaborative survey conducted by the United State Environmental Protection Agency (USEPA) with state and tribal partners that reports on the condition of rivers and streams throughout the CONUS (USEPA 2016; 2020). These surveys generally occur during April to September and characterize the conditions of streams during low flows. Survey sites are selected from a spatially balanced, random sampling design (Stevens and Olsen 2004). USEPA partners collected measurements from approximately 2000 rivers and streams during each of the two-year surveys. The field protocols of the 2008–09 and 2013–14 NRSA provide definitions and methodology on how channel dimension measurements were collected at low flows (USEPA 2016; 2020; Fig. 1). Briefly, wetted width represents the distance of the water's edge from left to right bank. Bankfull width represents the distance from left to right bank at bankfull stage, i.e., the distance from the top of the left and right banks where the potential water height would spill outside of the channel and into the floodplain. Thalweg depth measurements were taken systemically at the deepest point in the channel cross section from the bottom substrate to the water surface. Bankfull depth is thalweg depth plus bankfull height, which is the height from the water surface to the bankfull stage (Fig. 1). The NRSA mean bankfull width-to-depth ratio is defined as the ratio of the mean bankfull width to the mean bankfull depth.

EPA conducts thorough quality assurance of field-collected NRSA data, and we removed measurements that were flagged by this procedure as having quality control issues (e.g., missing observations). In addition to these internal checks, we examined the distributions of predicted values to identify extreme outliers. Through this process, we determined that stream segments identified as "tidal" by the National Hydrography Dataset Plus version 2 (NHDPlusV2; McKay et al. 2012) tended to have large outliers and were remove from subsequent analyses because they are subject to different geomorphic processes than unidirectional channels (Torres 2017). When a site was visited more than once within a year, we used observations from the first site visit for modeling. In addition, 30–40% of sites are revisited among the two surveys and we selected one observation from these repeated visits. The final channel measurements were log-transformed to improve model fit, as non-constant variation of residuals can give greater weight to higher variance data in random forest models (De'ath and Fabricius 2000; Walsh et al. 2017).

To conduct an assessment of our models, we obtained an independent dataset (Kauffman, personal communication, 2020) that included approximately 3000 measurements of bankfull width taken between 1994 and 2000 from the Mid-Atlantic Highland wadeable and boatable surveys (USEPA 2000), Oregon and Washington State Regional Environmental Monitoring and Assessment Program (USEPA 1999), the Environmental Monitoring and Assessment Program – West (EMAP; Lazorchak et al. 2000), and Region 7 Regional Environmental Monitoring and Assessment Program (REMAP; Angradi 2006). We removed sites from these data that were on NHDPlusV2 tidal streams or had missing data (i.e., latitudes, longitudes, or channel measurements).

Independent Watershed Metrics

We used landscape watershed summaries as covariates within the models. These watershed summaries were modified from the USEPA StreamCat dataset, which includes a large suite of anthropogenic and natural watershed features (e.g., soils, land use, and precipitation; Hill et al. 2016). StreamCat data are available for download at https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset. We limited these watershed metrics to factors that have been shown to influence, or that we hypothesized could influence channel dimensions (Leopold et al. 1964; Faustini et al. 2009; Kaufmann et al. 2009; Morel et al. 2020). Although correlations among predictor variables generally do not affect RF predictions (Fox et al. 2017), they can affect the ranking of variables that were important to the model. Therefore, we calculated Pearson correlations among the watershed summaries to better understand the correlations structure of the underlying data (Supplemental Materials S.1). In addition to Pearson correlations, we calculated distance correlations which quantifies non-linear relationships between variables (Székely et al. 2007). Distance correlations range between 0 and 1, in contrast to −1 to 1 for Pearson's r or Spearman's ρ, because they do not assume monotonically increasing or decreasing relationships. In some cases, we combined covariates, such as yearly precipitation into two-year summaries, as well as landcover metrics, to produce aggregate land classes of the National Land Cover Database (NLCD; Homer et al. 2007; Fry et al. 2011). For example, mixed, coniferous, and deciduous forest classes were combined to estimate the percent of each watershed composed of forested area (Supplemental Materials S.2). These combined covariates were then related to the most appropriate NRSA survey years (e.g., NLCD 2006 was related to NRSA 2008–09, whereas NLCD 2011 was related to NRSA 2013–14). Some additional covariates, such as slope and stream order, were obtained from the NHDPlusV2 datasets (Supplemental Materials S.2; McKay et al. 2012). Several metrics known to influence channel formation (e.g., presence of woody debris; Van Sickle and Gregory 1990; Kauffman et al. 1997) could not be included due to the lack of available CONUS-wide coverage. We then joined the NRSA channel dimension measurements (i.e., wetted width, thalweg depth, bankfull width, or bankfull depth) and watershed summaries (StreamCat) for modeling (Supplemental Materials S.3, table available as Supplemental Materials S.4).

Modeling

We developed Random Forest (RF) models for each channel dimension measurement. RF is an ensemble modeling approach that builds many individual regression trees from randomized subset of the data (i.e., bootstrapping) and randomized subsets of covariates at each branch of each a tree (Fig. 3). The randomized subset of data used to calibrate each tree are called “in-bag” observations (typically about two-thirds of the data) while those withheld from modeling are called “out-of-bag” observations. RF has become a popular modeling technique over the last two decades because it requires very little, if any, tuning and can capture non-linear relationships with response variables and interactions among covariates (Breiman 2001; Cutler et al. 2007).The main tuning parameters within RF are the number of trees included in the model, the number of randomly selected variables included as candidate predictors at each node of each tree, and the minimum node size beyond which no further division of the tree is done. However, numerous studies have shown that tuning of RF parameters has negligible effects on RF outcomes (Palmer et al. 2007; Fox et al. 2017; Hurskainen et al. 2019; Tian et al. 2022). An additional advantage is that RF can accept many correlated predictor variables and produce good model performance without features selection (Cutler et al. 2007; Fox et al. 2017) while feature selection can produce negligible improvements in model performance (Hurskainen et al. 2019). In fact, Fox et al. (2017) showed that feature selection can cause instabilities in RF output that may indicate overfitting to data. Expansion of its use consists of numerous examples in hydrology and fluvial geomorphology, including modeling reference condition flows (Carlisle et al. 2009), the potential impacts of climate change on ecologically-important flow metrics (Dhungel et al. 2016) and more recently to estimate hydraulic geometries in New Zealand and France (Morel et al. 2020). Rigorous comparisons of RF with linear modeling techniques, artificial neural nets, and boosted regression trees have shown that similar outcomes can be achieved with far less, if any, tuning or feature selection (Ogutu et al. 2011; Yang et al. 2016; Ouedraogo et al. 2019; Benkendorf and Hawkins 2020; Jun 2021).

To develop each model, we used channel dimension measurements as the response variable (e.g., wetted widths) and StreamCat metrics as explanatory variables with default settings of the randomForest function in the R package of the same name (Liaw and Wiener 2002). Once each RF model was developed, we then made predictions (or model-based interpolations) of the channel dimensions to unsampled rivers and streams throughout the CONUS. To do so, RF produces predictions for each regression tree within the RF model. A final RF prediction is the mean of predictions from each individual tree in the case of regression (Fig. 3).

RF produces model performance metrics from the out-of-bag (withheld) observations, which are a reasonable estimate of model fit (Cutler et al. 2007). Specifically, RF reports the pseudo r-squared which conveys the variation explained in the response variable as $1 - (MSE / s^2)$, where MSE is the mean squared error of RF out-of-bag predictions and s^2 is the variance in the response variable (i.e., width or depth). As an additional measure of model performance, we calculated the median absolute error (MdAE) for the CONUS and for each Strahler stream order. In addition, MdAE values were standardized by the respective interquartile range (IQR) of channel dimension measurements to allow comparison between models (MdAE/IQR). MdAE was chosen as a performance metric because it is more robust to outliers compared to other metrics such as root mean square error (Probst and Boulesteix 2017). Further, IQR is a measure of statistical dispersion that is less sensitive to non-normal distributions and outliers. To estimate the tendency of the models to over- or underestimate values relative to observations, we calculated percent bias (PBIAS) of each model (Moriasi et al. 2015). PBIAS is typically used to assess simulations of hydrologic time series for single sites or watersheds, and no formal guidance exists for models that account for variation in a response variable among sample sites rather than over time. However, Moriasi et al. (2015) suggest that PBIAS <±5, <±10, <±15, and ≥±15 represent very good, good, satisfactory, and not satisfactory models of hydrologic flow, respectively. Finally, RF estimates the importance of each variable to the model by permuting a variable's values while holding other variables constant. By doing so, it can calculate the change in the model mean squared error when a variable's values are randomized and effectively "removed" from the model (Breiman 2001). Lastly, the five most important variables were plotted for each model based on the RF variable importance measure (Supplemental Materials S.5); however, we offer limited interpretation because several predictor variables were highly correlated (Supplemental Materials S.1) and correlations among variables limit the interpretability of importance rankings in RF.

Other studies have demonstrated that predictions made by regional models can outperform CONUS models (e.g., Blackburn-Lynch et al. 2017). However, for ease of application and consistency, a single, CONUS model is desirable, especially if it could match or outperform regional models. In addition, interpolations of regional models can produce distinct shifts in predicted values at regional boundaries which may hinder practioners' ability to use such predictions at large scales (Hill et al. 2017). Thus, models of wetted width were created for each US Geological Survey Physiographic Division (Fig. 2; Fenneman and Johsnon 1946) to assess how a single CONUS wetted width model compared to regionalized models. We used Physiographic Divisions because these regions were recently used to develop bankfull width models that are also available for comparison (see below). All regional models were developed using the randomForest package in R (Liaw and Wiener 2002; R Development Core Team 2019) and methods described above. We compared the regional and national models with pseudo-r^2, MdAE, MdAE/IQR, and PBIAS.

The width-to-depth (W/d) ratio is often used to assess channel stability and instream habitat availability (Rosgen 1998; Dunham et al. 2002). To evaluate the use of our models for such assessments, we also calculated bankfull W/d ratios from the two RF model estimates of bankfull width and bankfull depth – RF(W)/RF(d). Additional channel dimension products could be created from these model predictions such as a the ratio of wetted width to thalweg depth; however, we did not compare or evaluate error estimates for these other possible combinations.

Finally, we compared predicted bankfull width values from recent regression models developed by Bieger et al. (2015) to our own estimates using the independent dataset described previously (Kauffman, personal communication, 2020). The models of Bieger et al. (2015) are often used as inputs to the process-based SWAT model (e.g., White et al. 2017). These comparisons were done by Physiographic Division (Fig. 2) with the r-squared and regression slopes between predicted and observed values (Piñeiro et al. 2008), MdAE, MdAE/IQR, and PBIAS. For accurate models, the slope between predicted and observed values should be near one. Bieger et al. (2015) did not produce model equations for two of the Physiographic Divisions due to small sample sizes; therefore, comparisons could only be made for the Appalachian Highlands, Atlantic Plain, Intermontane Plateau, Interior Plains,

EPA Author Manuscript

EPA Author Manuscript

Pacific Mountain System, and Rocky Mountain System (Fig. 2). We also used scatter plots to visually compare model predictions by region to observed measurements.

RESULTS

Channel Dimension Models

The wetted width model had a pseudo r^2 of 0.91 and an overall MdAE of 3.0 m. The MdAEs represent ±6% of the IQR of measured values across all stream orders. The RF predictions agreed well with known values, however, a few outliers were present (Fig. 4). Generally, the predictions captured the variability of observed values (Fig. 5). However, the RF predictions tend to underestimate observed values across all stream orders (i.e., negative PBIAS; Table 1), especially within stream orders 1–3, and 7. MdAE values increased with Strahler stream order but were lowest relative to IQR in orders 5–9 (Table 1). The most important covariates for this model were mean runoff within the watershed (mm), watershed area (km^2), average 30-year precipitation (mm), stream order, and average 2-year precipitation (mm) (Table 2; see Supplemental Materials S.2 for variable definitions). Several of the most important covariates were correlated with one another, specifically precipitation and runoff (Supplemental Materials S.1). This pattern was the same whether examined with linear Pearson's r or non-linear distance correlations (compare correlation matrices in Supplemental Materials S.1). Performance of the CONUS and Physiographic Division models were similar, indicating that the CONUS models performed as well or better in most regions (Supplemental Materials S.5).

The thalweg depth model had a psuedo r^2 of 0.82 and an overall MdAE of 0.16 m. The MdAEs represent ±15% of the IQR of observed values. Like wetted width, the thalweg depth model tended to underpredict values (PBIAS < 0), but especially in streams of order 1 (Table 1). The RF predictions captured most of the variability in the observations for each stream order but often failed to estimate the largest depths in higher order rivers (Fig. 5). MdAE values started at 0.06 m for stream order one and increased with stream order up to 2.9 m for order ten (Table 1). The top five important covariates for this model were mean runoff (mm), watershed area (km^2), channel slope (%), stream order, and average 30-year precipitation (mm; Table 2). As with wetted width, the single CONUS-wide model of thalweg depth performed as well as regional models (Supplemental Materials S.5)

The bankfull width model explained 90% of the variation in observed values and had an overall MdAE of 4 m (Table 1). MdAE represented ±8% of the IQR of measured values across all stream orders. As with the other models, RF predictions of bankfull width were closely related to observed values (Fig. 4 and 5), but also tended to have negative PBIAS values (i.e., underprediction; Table 1). MdAE values increased with stream order but were lowest relative to IQR in stream orders 6–9 (Table 1). The top five covariates were mean watershed area (km^2), runoff (mm), average 30-year precipitation (mm), stream order, and mean watershed elevation (m) (Table 2). Models constructed with CONUS and Physiographic Division data performed similarly (Supplemental Materials S.5).

The bankfull depth model explained a lower percentage of the variation in measured values (pseudo r^2 value of 0.70) and had an overall MdAE of 0.36 m. The MdAE value was 0.19

m at stream order one and increased to 4 m by stream order 10 (Table 1). The MdAEs were within ±21% of the IQR of observed values. Despite being the lowest performing model, the PBIAS of the bankfull depth model did not differ substantially from the other models and indicates underprediction, especially in streams of order 2 (Table 1). The top five covariates for this model were watershed area (km^2), slope (%), mean runoff (mm), stream order, and average 30-year precipitation (mm) (Table 2). The single CONUS-wide model had a similar performance when compared with models built with data from Physiographic Divisions (Supplemental Materials S.5).

Finally, when compared against observed bankfull W/d ratios, the model-estimated ratios explained less than half of this observed variation (pseudo $r^2 = 0.47$) and had an overall MdAE of 3.20. The MdAEs are ±22.1% of the IQR of observed W/d ratios. We did not calculate PBIAS for W/d ratios.

Model Comparisons

Both our RF bankfull width model and the models of Bieger et al. (2015) performed well within the Physiographic Divisions when applied to the independent dataset (Table 3). The Bieger et al. (2015) models explained from 40% to 73% of the variation in bankfull widths across the Physiographic Divisions and had MdAE values of 2.1 to 8.1 m (MdAE/IQR = 0.23–0.44). PBIAS of these models ranged from −13.5% to 47.4%, with an overall PBIAS of 21.2 for the CONUS. Across the same regions, the RF model explained 61% to 76% of the variation in bankfull width with MdAE values from 2.2 to 3.4 m (MdAE/IQR = 0.19–0.37). The RF r-squared and MdAE values are respectively lower and higher than those reported by the internal RF performance metrics from out-of-bag observations. PBIAS of the RF models was −6.1% for the CONUS and was smaller in most regions (−13.3% to 16.5%), with some exceptions (e.g., the Intermontane Plateau). PBIAS values observed with the independent data were often better than those produced from the RF out-of-bag predictions. The model of Bieger et al. (2015) slightly outperformed the RF model within the Atlantic Plain region. In the remainder of regions, the RF model performed better, especially in the Appalachian Highlands, Interior Plains, and Pacific Mountain System regions (Table 3). With the exceptions of the Appalachian Highlands and Pacific Mountain System, regression slopes between predicted and observed values were closer to one for the RF model (Table 3). The RF models generally did not overpredict observed values; however, in a few regions and stream orders there were some under predictions (Figs. 6–7). The Bieger et al. (2015) models and RF models had similar residual values when compared to measurements (Fig. 7). The residual errors for both models increased with higher order streams (Fig. 7). The Bieger et al. (2015) models underpredict bankfull width values in some streams in stream orders 2–4, and 8 (Figs. 6–7).

DISCUSSION

We developed models to provide channel dimension estimates for 1.1 million perennial stream segments in the US. These models were applied to perennial streams, which cover large swaths of the CONUS and will be added to the existing suite of attributes in the StreamCat dataset (https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset). These predictions will be of value for aquatic resource managers and can contribute to the assessment and modeling of freshwater habitats in lotic ecosystems where estimates of channel dimensions are needed. The approach presented here contrasts with previous models by providing predicted values at stream segments across the CONUS, rather than regression coefficients. Although the predictions are specific to the streams of the CONUS, the methods could be used to construct and predict values in areas outside of the U.S. if similar hydrographic data are available. Below, we consider how the performance of our models compare with previous modeling efforts, the use and limitations of the predicted values, potential applications to aquatic resource management, as well as model improvements that could be made in the future.

Model Performance

By including additional explanatory variables and using a popular machine learning technique that can account for non-linear relationships and interactions, we were able to develop satisfactory single, CONUS-wide models of channel dimensions that performed well when compared with regional models. Although our models performed well overall, the models of bankfull depth and bankfull width-to-depth ratio did not perform as well as the other channel dimension models. This lower performance may be due to challenges in finding bankfull stage in the field, especially during the dry season. In addition, depending on the geometry of side banks, the unit change in depth between non-flood to bankfull depth can be much greater than the unit changes in width (Rhodes 1977; Knighton 1998). Thus, errors in identifying bankfull stage will translate to greater errors in bankfull depth measurements than for bankfull width. However, when standardizing MdAE across stream order by the IQR of observed values, the bankfull depth models appear to perform similarly to the model of thalweg depth models in stream orders ≥3 (Table 1). Overall, MdAE values - regardless of model - were lowest in stream orders one through three (Table 1). The larger residual errors in higher order streams (eighth through tenth order) were due to the higher variability in observed values and did not represent poorer performance when considered proportionally to the IQR of observed values (Table 1).

Throughout the literature and within other modeling approaches, drainage area and slope are often cited as indicator of overall hydraulic energy available is cited as driving channel formation processes (Leopold et al. 1964; Faustini et al. 2009). Other factors, such as regional climatic regime or differences in geology, are often accounted for by constraining the extent of models to within regions. Thus, it was not surprising that indicators of streams size (watershed area and stream order), slope, and climate were important for predicting channel width and depth. The inclusion of precipitation in the models likely helped to account for how flow can differ among sites of a given drainage area, but of differing climatic regimes. Although it is unclear how much the additional covariates (e.g., agriculture, urbanization, and geology) contributed to the overall performance of the models, the aggregate effect of their inclusion may have contributed to models that performed as well as regional RF models, making predictions from a single model possible.

The ratio of bankfull width to bankfull depth has been used extensively as an indicator of geomorphic processes that form channels and habitats, and in multiple geomorphic classification systems (Buffington and Montgomery 2013). Therefore, a possible extension of our models could be the use of these ratios to inform regional or national classification systems. Thus, determining how the ratios of our bankfull models perform relative to NRSA measurements was important. When modeled separately and combined into the bankfull W/D ratios, the predicted values did not perform as well when assessed against measurements. All models contain prediction error, and the errors of our bankfull width and depth models may interact to affect the final accuracy of ratioed values. Thus, care must be taken when using combinations of the predicted values in applications that require ratios, and the level of accuracy needed for these applications must be considered.

Both the Bieger et al. (2015) and the RF bankfull width models performed well when applied to the independent dataset. However, the improved performance of the RF model may be due to the inclusion of additional explanatory variables relative to Bieger et al. (2015), which used only drainage area. The simplicity of the model produced by Bieger et al. (2015) is attractive; however, our models offer at least two advancements. First, beyond improvements in overall performance (Table 3), inclusion of additional covariates better accounted for variations in spatial drivers of channel width and depth. Including these factors allowed us to develop a single nationwide model that performed as well as both individual regional RF and traditional regression models (i.e., Bieger et al. 2015). This outcome contrasts with previous work that found regionalized rating curves based on drainage area outperformed a national model (Blackburn-Lynch et al. 2017). Second, the spatial extent of NRSA data allowed us to develop and apply models for the two Physiographic Divisions that were not modeled by Bieger et al. (2015) because of inadequate data. This advancement provides estimated channel dimensions that are directly relevant to these regions and avoids the misapplication of models that were developed elsewhere.

Use and Limitations of Predicted Values

We will make the model output available as part of a publicly available dataset, which differs from how models of width or depth have traditionally been delivered. Previously, regression or power law parameters were reported, and practitioners calculated the appropriate covariates to apply the models for their needs. However, such an approach could lead to application of results outside of the conditions used to develop the models, which could cause inaccurate estimations of width or depth. Since we will be making predicted channel dimensions available for download, we can constrain estimates to stream segments that are within the sampling frame of the NRSA, thereby ensuring model applications to appropriate stream segments (Wenger and Olden 2012). Due to this constraint, some sections of the CONUS, such as parts of the southwestern United States and the arid foothills of Montana (white areas in Fig. 8), produced fewer predicted values. These regions have more intermittent and ephemeral stream systems and, although these streams are important aquatic resources (Mazor et al. 2014), were not part of this study. Additionally, tidal streams (as defined by the NHDPlusV2) were not included in modeling because they substantially increased model errors. Our difficulty in modeling tidal streams is likely due to the different processes controlling channel dimensions of tidal systems relative to unidirectional inland streams (Torres 2017). Lastly, our models may underestimate the impact of water impoundments (e.g., dam density) as there may not have been enough sites below dams due to the randomized placement of NRSA sample sites.

Our models may not be appropriate for every management application. For many uses, field measurement of channel width and depth may be more appropriate, especially when the application might be sensitive to the model errors reported here. Whether the accuracy and precision of our models meets the needs of a user will depend on the scale and purpose of the application (e.g., watershed-scale SWAT model versus regional fisheries management plan). We encourage the careful consideration of such factors and examination of model performance metrics within stream orders and regions (Supplemental Materials S.5) before using the predicted values provided by the RF models. However, if widths and depths are needed from many locations or from larger streams, which can be difficult and costly to measure, the model errors may become an acceptable trade-off. The low flow models may not be applicable or useful for studies that require temporally dynamic estimates of wetted width and thalweg depth because they were built on single measurements at low flow conditions. However, the open-source code we provide may help researchers and resource managers apply this methodology to their own temporally dynamic datasets (Supplemental Materials S.5). Lastly, although we validated the models with an independent dataset of several thousand streams, some portion of the 1.1 million stream segments will likely have poor predictions with unacceptably high error. We encourage practitioners to ground truth predictions in their study area where possible to determine if model predictions are acceptable for their application.

Model Applications and Future Work

Stream width and other aspects of channel dimensions are frequently used in fisheries habitat or production models (e.g., Shallin Busch et al. 2013; See et al. 2021) that support aquatic species recovery planning (Bond et al. 2019), prioritization of habitat restoration efforts (Roni et al. 2018), and the establishment of environmental flows for habitat needs across broad regions (Dunbar et al. 2012; Spurgeon et al. 2019). Depending on available resources and data, researchers and managers may have access to a variety of methodologies for estimating channel dimensions and habitat capacity, ranging from sub-meter resolution ground-based GPS surveys (e.g., Tonina et al. 2019) to broad regional regression models (e.g., Faustini et al. 2009; Bieger et al. 2015). Thus, our models may be of particular value to regional and national habitat applications where headwater stream data are unavailable through remote sensing (e.g., LIDAR) or when managers are constrained by current computational limitations or costs (Robinson et al. 2022).

Our modeled values can inform regional and national-scale assessments and planning and provide several potential benefits for practioners and scientists. For example, stream size can offer a useful stratum for hierarchical classifications of stream habitats that include both regional and local factors and allow the comprehensive application of models across large regions. These results will also be useful for planning. For instance, the level of effort required to adequately assess stream fish communities is a direct function of channel width (Reynolds et al. 2003). Estimates of channel width and stream size can help crews evaluate time and resource needs when planning national or regional aquatic assessments or fisheries surveys (Hughes et al. 2011) and stratifying sites for regional and national scale aquatic monitoring designs (Olsen and Peck 2008). In such cases, our modeled values could serve as an initial estimate of width or depth or be used to expand the locations to which estimates of carrying capacity can be applied.

Predictions of channel width and depth will also be useful for model-based water quality models, such as SWAT, which is an important tool for conservation and policy managers throughout the US. Currently, SWAT uses a variety of methods to calculate bankfull channel characteristics (width, depth, and others), such as regression equations (Bieger et al. 2015) and the Leopold and Maddock (1953) power functions (White et al. 2017). Recently, Han et al. (2019) evaluated predicted values made from these equations against aerial photography and found that they often overestimate channel bankfull widths. Although we cannot use Han et al. (2019) to assess our models, they provide a useful alternative especially if aerial photography is limited or channels are obscured by vegetation.

Future iterations of these models could include adding new years of NRSA data as they become available. Doing so will increase the number of observations for model development to ensure that the models represent perennial streams and rivers of the US. Further, additional years of data could capture greater variation in climatic conditions within regions and provide the basis for understanding how these variations affect the availability of fish habitat. Such understanding will be increasingly critical in the face of climate change and when considering how to mitigate these impacts on fisheries resources (Paukert et al. 2021). However, to make improvements to the model, sources of potential error will need to be considered. Some portion of this error likely comes from difficulties in identifying the extent of bankfull flows in the field as discussed above. Another source of error could be the single summertime site visit of low flows. The RF models presented here cannot account for temporal variation in flows during this period, contributing to the overall error within the models. For example, Bellmore et al. (2018) found that within-season temporal variability account for up to 17% of the variation in NRSA nitrogen measurements, thereby defining the maximum performance a model could achieve. Finally, identification and inclusion of new geospatial predictor variables that better represent the physical processes that drive channel dimensions could also improve model performances.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We gratefully acknowledge Phil Kaufmann for his assistance and for providing our independent dataset. Thanks to Lou Reynolds, Mike Cyterski, and Patti Meeks for their helpful suggestions on different versions of this manuscript. The information in this document has been funded entirely by the U.S. Environmental Protection Agency, in part through appointments to the Internship/Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. These models and datasets have been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

DATA AVAILABILITY STATEMENT

Modeled channel dimension estimates are available through the USEPA StreamCat database at https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset (Hill et al. 2016).

LITERATURE CITED

Alexander RB, Smith RA and Schwarz GE, 2000. Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. Nature 403:758–761. [PubMed: 10693802]

Allen PM, Arnold JC and Byars BW, 1994. Downstream Channel Geometry For Use In Planning-Level Models 1. JAWRA Journal of the American Water Resources Association 30:663–671.

Ames DP, Rafn EB, Van Kirk R and Crosby B, 2009. Estimation of stream channel geometry in Idaho using GIS-derived watershed characteristics. Environmental Modelling & Software 24:444–448.

Angermeier PL and Winston MR, 1998. Local vs. regional influences on local diversity in stream fish communities of Virginia. Ecology 79:911–927.

Angradi TR. (editor), 2006. Environmental Monitoring and Assessment Program: Great River Ecosystems, Field Operations Manual. EPA/620/R-06/002. U.S. Environmental Protection Agency. Washington, D.C.

Beck MW, O'Hara C, Stewart Lowndes JS, Mazor RD, Theroux S, Gillett DJ, Lane B and Gearheart G, 2020. The importance of open science for biological assessment of aquatic environments. PeerJ 8:e9539. [PubMed: 32742805]

Beechie T, Beamer E and Wasserman L, 1994. Estimating Coho Salmon Rearing Habitat and Smolt Production Losses in a Large River Basin, and Implications for Habitat Restoration. North American Journal of Fisheries Management 14:797–811.

Bellmore RA, Compton JE, Brooks JR, Fox EW, Hill RA, Sobota DJ, Thornbrugh DJ and Weber MH, 2018. Nitrogen inputs drive nitrogen concentrations in U.S. streams and rivers during summer low flow conditions. Science of The Total Environment 639:1349–1359. [PubMed: 29929300]

Benkendorf DJ and Hawkins CP, 2020. Effects of sample size and network depth on a deep learning approach to species distribution modeling. Ecological Informatics 60:101137.

Berriri M, Djema S, Rey G and Dartigues-Pallez C, 2021. Multi-Class Assessment Based on Random Forests. Education Sciences 11:92.

Bieger K, Rathjens H, Allen PM and Arnold JG, 2015. Development and evaluation of bankfull hydraulic geometry relationships for the physiographic regions of the United States. JAWRA Journal of the American Water Resources Association 51:842–858.

Blackburn-Lynch W, Agouridis CT and Barton CD, 2017. Development of Regional Curves for Hydrologic Landscape Regions (HLR) in the Contiguous United States. JAWRA Journal of the American Water Resources Association 53:903–928.

Bond MH, Nodine TG, Beechie TJ and Zabel RW, 2019. Estimating the benefits of widespread floodplain reconnection for Columbia River Chinook salmon. Canadian Journal of Fisheries and Aquatic Sciences 76:1212–1226.

Breiman L, 2001. Random Forests. Machine Learning 45:5–32.

Buffington J and Montgomery D, 2013. Geomorphic classification of rivers. In: Shroder J; Wohl E, ed. Treatise on Geomorphology; Fluvial Geomorphology, Vol. 9. San Diego, CA: Academic Press. p. 730–767.:730–767.

Carlisle D, Falcone J and Meador M, 2009. Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds. Environmental Monitoring and Assessment 151:143–160. [PubMed: 18493861]

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J and Lawler JJ, 2007. Random forests for classification in ecology. Ecology 88:2783–2792. [PubMed: 18051647]

De'ath G and Fabricius KE, 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Detenbeck NE, DeVore PW, Niemi GJ and Lima A, 1992. Recovery of temperate-stream fish communities from disturbance: a review of case studies and synthesis of theory. Environmental management 16:33.

Dhungel S, Tarboton DG, Jin J and Hawkins CP, 2016. Potential Effects of Climate Change on Ecologically Relevant Streamflow Regimes. River Research and Applications 32:1827–1840.

Dunbar MJ, Alfredsen K and Harby A, 2012. Hydraulic-habitat modelling for setting environmental river flow needs for salmonids. Fisheries Management and Ecology 19:500–517.

Dunham JB, Cade BS and Terrell JW, 2002. Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. Transactions of the American Fisheries Society 131:86–98.

Fausch KD, Torgersen CE, Baxter CV and Li HW, 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes: a continuous view of the river is needed to understand how processes interacting among scales set the context for stream fishes and their habitat. BioScience 52:483–498.

Faustini JM, Kaufmann PR and Herlihy AT, 2009. Downstream variation in bankfull width of wadeable streams across the conterminous United States. Geomorphology 108:292–311.

Fenneman NM and Johsnon DW, 1946. Physiographic Divisions of the Conterminous US. U.S. Geological Survey Special Map Series, Scale 1:7,000,000, http://water.usgs.gov/GIS/metadata/usgswrd/XML/physio.xml.

Finkenbine JK, Atwater J and Mavinic D, 2000. Stream health after urbanization 1. JAWRA Journal of the American Water Resources Association 36:1149–1160.

FitzGerald DAM, Boughton DD, Fuller MJ, John MSN, Martin DBT, Harrison DLR and Mantua DNJ, 2021. Physical and biological constraints on the capacity for life-history expression of anadromous salmonids: an Eel River, California, case study. Canadian Journal of Fisheries and Aquatic Sciences 10.1139/cjfas-2021-0229.

Fox EW, Hill RA, Leibowitz SG, Olsen AR, Thornbrugh DJ and Weber MH, 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment 189:316. [PubMed: 28589457]

Fry JA, Xian G, Jin S, Dewitz JA, Homer CG, Yang L, Barnes CA, Herold ND and Wickham JD, 2011. Completion of the 2006 national land cover database for the conterminous United States. PE&RS, Photogrammetric Engineering & Remote Sensing 77:858–864.

Ghimire SR, Corona J, Parmar R, Mahadwar G, Srinivasan R, Mendoza K and Johnston JM, 2021. Sensitivity of Riparian Buffer Designs to Climate Change—Nutrient and Sediment Loading to Streams: A Case Study in the Albemarle-Pamlico River Basins (USA) Using HAWQS. Sustainability 13:12380.

Han J, Lee D, Lee S, Chung S-W, Kim SJ, Park M, Lim KJ and Kim J, 2019. Evaluation of the effect of channel geometry on streamflow and water quality modeling and modification of channel geometry module in SWAT: A case study of the Andong Dam Watershed. Water 11:718.

Hey RD, 2006. Fluvial Geomorphological Methodology For Natural Stable Channel Design 1. JAWRA Journal of the American Water Resources Association 42:357–386.

Hill RA, Fox EW, Leibowitz SG, Olsen AR, Thornbrugh DJ and Weber MH, 2017. Predictive Mapping of the Biotic Condition of Conterminous-USA Rivers and Streams. Ecological Applications 27:2397–2415. [PubMed: 28871655]

Hill RA, Weber MH, Leibowitz SG, Olsen AR and Thornbrugh DJ, 2016. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. Journal of the American Water Resources Association (JAWRA) 52:120–128.

Homer C, Dewitz J, Fry J, Coan M, Hossain N, Larson C, Herold N, McKerrow A, VanDriel JN and Wickham J, 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. Photogrammetric Engineering and Remote Sensing 73:337–341.

Hughes RM, Kaufmann PR and Weber MH, 2011. National and regional comparisons between Strahler order and stream size. Journal of the North American Benthological Society 30:103–121.

Humphries P and Baldwin DS, 2003. Drought and aquatic ecosystems: an introduction. Freshwater Biology 48:1141–1146.

Hurskainen P, Adhikari H, Siljander M, Pellikka PKE and Hemp A, 2019. Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes. Remote Sensing of Environment 233:111354.

Johnson PA and Fecko B, 2008. Regional channel geometry equations: a statistical comparison for physiographic provinces in the eastern US. River Research and Applications 24:823–834.

Judes C, Gouraud V, Capra H, Maire A, Barillier A and Lamouroux N, 2021. Consistent but secondary influence of hydropeaking on stream fish assemblages in space and time. Journal of Ecohydraulics 6:157–171.

Jun M-J, 2021. A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area. International Journal of Geographical Information Science 35:2149–2167.

Kauffman JB, Beschta RL, Otting N and Lytjen D, 1997. An ecological perspective of riparian and stream restoration in the western United States. Fisheries 22:12–24.

Kaufmann PR, Larsen DP and Faustini JM, 2009. Bed stability and sedimentation associated with human disturbances in pacific northwest streams. Journal of the American Water Resources Association 45:434–459.

Klein RD, 1979. Urbanization and stream quality impairment 1. JAWRA Journal of the American Water Resources Association 15:948–963.

Knighton D, 1998. Fluvial forms and processes: a new perspective. London, UK, Routledge,

Lamouroux N, Capra H, Pouilly M and Souchon Y, 1999. Fish habitat preferences in large streams of southern France. Freshwater Biology 42:673–687.

Lazorchak JM, B.H. H, D.K. A, D.V. P and D.J. K, 2000. Environmental Monitoring and Assessment Program - Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Non-Wadeable Rivers and Streams. U.S. Environmental Protection Agency (EPA/620/R-00/007). Washington D.C.

Ledford SH, Zimmer M and Payan D, 2020. Anthropogenic and biophysical controls on low flow hydrology in the southeastern U.S. Water Resources Research 56.

Leopold LB, 1994. A View of the River, Harvard University Press, ISBN 0674937325

Leopold LB and Maddock T, 1953. The hydraulic geometry of stream channels and some physiographic implications, US Government Printing Office,

Leopold LB, Wolman MG and Miller JP, 1964. Fluvial Processes in Geomorphology. New York, Dover Publications,

Liaw A and Wiener M, 2002. Classification and Regression by randomForest. R News 2:18–22.

Mayer TD, 2012. Controls of summer stream temperature in the Pacific Northwest. Journal of Hydrology 475:323–335.

Mažeika S, Sullivan P, Watzin MC and Hession WC, 2004. Understanding stream geomorphic state in relation to ecological integrity: evidence using habitat assessments and macroinvertebrates. Environmental management 34:669–683. [PubMed: 15633026]

Mazor RD, Stein ED, Ode PR and Schiff K, 2014. Integrating intermittent streams into watershed assessments: applicability of an index of biotic integrity. Freshwater Science 33:459–474.

McKay L, Bondelid T, Dewald T, Johnston J, Moore R and Reah A, 2012. NHDPlus Version 2: User Guide. (Available from: http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php),

McMahon T and Finlayson B, 2003. Droughts and anti-droughts: the low flow hydrology of Australian rivers. Freshwater biology 48:1147–1160.

Menció A and Boix D, 2018. Response of macroinvertebrate communities to hydrological and hydrochemical alterations in Mediterranean streams. Journal of Hydrology 566:566–580.

Menció A and Mas-Pla J, 2010. Influence of groundwater exploitation on the ecological status of streams in a Mediterranean system (Selva Basin, NE Spain). Ecological Indicators 10:915–926.

Modrick TM and Georgakakos KP, 2014. Regional bankfull geometry relationships for southern California mountain streams and hydrologic applications. Geomorphology 221:242–260.

Mohamoud YM and Parmar RS, 2006. Estimating Streamflow and Associated Hydralic Geometry, The Mid-Atlantic Region. JAWRA Journal of the American Water Resources Association 42:755–768.

Morel M, Booker DJ, Gob F and Lamouroux N, 2020. Intercontinental predictions of river hydraulic geometry from catchment physical characteristics. Journal of Hydrology 582:124292.

Moriasi DN, Gitau MW, Pai N and Daggupati P, 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Transactions of the ASABE 58:1763–1785.

Neitsch SL, Arnold JG, Kiniry JR and Williams JR, 2005. Soil and Water Assessment Tool Theoretical Documentation (Version 2005).

Ogutu JO, Piepho H-P and Schulz-Streeck T, 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5:S11.

Olsen AR and Peck DV, 2008. Survey design and extent estimates for the Wadeable Streams Assessment. Journal of the North American Benthological Society 27:822–836, 815.

Ouarda TB, Charron C and St-Hilaire A, 2008. Statistical models and the estimation of low flows. Canadian Water Resources Journal 33:195–206.

Ouedraogo I, Defourny P and Vanclooster M, 2019. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeology Journal 27:1081–1098.

Palmer DS, O'Boyle NM, Glen RC and Mitchell JBO, 2007. Random Forest Models To Predict Aqueous Solubility. Journal of Chemical Information and Modeling 47:150–158. [PubMed: 17238260]

Parker G, Wilcock PR, Paola C, Dietrich WE and Pitlick J, 2007. Physical basis for quasi-universal relations describing bankfull hydraulic geometry of single-thread gravel bed rivers. Journal of Geophysical Research: Earth Surface 112.

Paukert C, Olden JD, Lynch AJ, Breshears DD, Christopher Chambers R, Chu C, Daly M, Dibble KL, Falke J and Issak D, 2021. Climate Change Effects on North American Fish and Fisheries to Inform Adaptation Strategies. Fisheries 46:449–464.

Peterson BJ, Wollheim WM, Mulholland PJ, Webster JR, Meyer JL, Tank JL, Martí E, Bowden WB, Valett HM and Hershey AE, 2001. Control of nitrogen export from watersheds by headwater streams. Science 292:86–90. [PubMed: 11292868]

Piñeiro G, Perelman S, Guerschman JP and Paruelo JM, 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecological Modelling 216:316–322.

Price K, 2011. Effects of watershed topography, soils, land use, and climate on baseflow hydrology in humid regions: A review. Progress in physical geography 35:465–492.

Probst P and Boulesteix A-L, 2017. To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research 18:6673–6690.

R Development Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rashleigh B, Parmar R, Johnston JM and Barber MC, 2005. Predictive Habitat Models for the Occurrence of Stream Fishes in the Mid-Atlantic Highlands. North American Journal of Fisheries Management 25:1353–1366.

Reynolds L, Herlihy AT, Kaufmann PR, Gregory SV and Hughes RM, 2003. Electrofishing Effort Requirements for Assessing Species Richness and Biotic Integrity in Western Oregon Streams. North American Journal of Fisheries Management 23:450–461.

Rhodes DD, 1977. The bfm diagram; graphical representation and interpretation of at-a-station hydraulic geometry. American journal of science 277:73–96.

Robinson HE, Henderson MJ, Perry RW, Goodman DH and Som NA, 2022. From site to system: Approaches for producing system-wide estimates of fish habitat in large rivers. River Research and Applications 38:119–130.

Rolls RJ, Leigh C and Sheldon F, 2012. Mechanistic effects of low-flow hydrology on riverine ecosystems: ecological principles and consequences of alteration. Freshwater Science 31:1163–1186.

Roni P, Anders PJ, Beechie TJ and Kaplowe DJ, 2018. Review of Tools for Identifying, Planning, and Implementing Habitat Restoration for Pacific Salmon and Steelhead. North American Journal of Fisheries Management 38:355–376.

Rosenfeld J, 2003. Assessing the Habitat Requirements of Stream Fishes: An Overview and Evaluation of Different Approaches. Transactions of the American Fisheries Society 132:953–968.

Rosgen D 1998. The reference reach: a blueprint for natural channel design. Pages 1009–1016 Engineering Approaches to Ecosystem Restoration.

See KE, Ackerman MW, Carmichael RA, Hoffmann SL and Beasley C, 2021. Estimating carrying capacity for juvenile salmon using quantile random forest models. Ecosphere 12:e03404.

Shallin Busch D, Sheer M, Burnett K, McElhany P and Cooney T, 2013. Landscape-level model to predict spawning habitat for lower Columbia River fall Chinook salmon (Oncorhynchus tshawytscha). River Research and Applications 29:297–312.

Sloat MR, Reeves GH and Christiansen KR, 2017. Stream network geomorphology mediates predicted vulnerability of anadromous fish habitat to hydrologic change in southeast Alaska. Global Change Biology 23:604–620. [PubMed: 27611839]

Spurgeon J, Pegg M, Parasiewicz P and Rogers J, 2019. River-Wide Habitat Availability for Fish Habitat Guilds: Implications for In-Stream Flow Protection. Water 11:1132.

Stevens DL and Olsen AR, 2004. Spatially Balanced Sampling of Natural Resources. Journal of the American Statistical Association 99:262–278.

Stewardson M, 2005. Hydraulic geometry of stream reaches. Journal of hydrology 306:97–111.

Sullivan SM, Watzin P,MC and Hession WC, 2006. Influence of stream geomorphic condition on fish communities in Vermont, USA. Freshwater Biology 51:1811–1826.

Székely GJ, Rizzo ML and Bakirov NK, 2007. Measuring and Testing Dependence by Correlation of Distances. The Annals of Statistics 35:2769–2794.

Tian Z, Liu F, Liang Y and Zhu X, 2022. Mapping soil erodibility in southeast China at 250 m resolution: Using environmental variables and random forest regression with limited samples. International Soil and Water Conservation Research 10:62–74.

Tonina D, McKean JA, Benjankar RM, Wright CW, Goode JR, Chen Q, Reeder WJ, Carmichael RA and Edmondson MR, 2019. Mapping river bathymetries: Evaluating topobathymetric LiDAR survey. Earth Surface Processes and Landforms 44:507–520.

Torres R, 2017. Channel geomorphology along the fluvial-tidal transition, Santee River, USA. GSA Bulletin 129:1681–1691.

USEPA (US Environmental Protection Agency), 1999. Regional Environmental Monitoring and Assessment Program - Region 10 1994–1995 Washington/Oregon Coastal Streams and Yakima Basin Streams Fish and Amphibian Count Data.

USEPA (US Environmental Protection Agency) Office of Water and Office of Research and Development, 2000. Mid-Atlantic Highlands Streams Assessment (EPA/903/R-00/015). Region 3. Philadelphia, PA.

USEPA (US Environmental Protection Agency) Office of Water and Office of Research and Development, 2016. National Rivers and Streams Assessment 2008–2009: A Collaborative Survey (EPA/841/R-16/007). Washington, DC.

USEPA (US Environmental Protection Agency) Office of Water and Office of Research and Development, 2020. National Rivers and Streams Assessment 2013–2014: A Collaborative Survey (EPA 841-R-19–001). Washington, DC.

Van Sickle J and Gregory SV, 1990. Modeling inputs of large woody debris to streams from falling trees. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere 20:1593–1601.

Walsh ES, Kreakie BJ, Cantwell MG and Nacci D, 2017. A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. PLOS ONE 12:e0179473. [PubMed: 28738089]

Wenger SJ and Olden JD, 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260–267.

White MJ, Beiger K, Gambone M, Haney E, Arnold J and Gao J, 2017. Development of a Hydrologic Connectivity Dataset for SWAT Assessments in the US. Water 9:892.

Yang R-M, Zhang G-L, Liu F, Lu Y-Y, Yang F, Yang F, Yang M, Zhao Y-G and Li D-C, 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators 60:870–878.

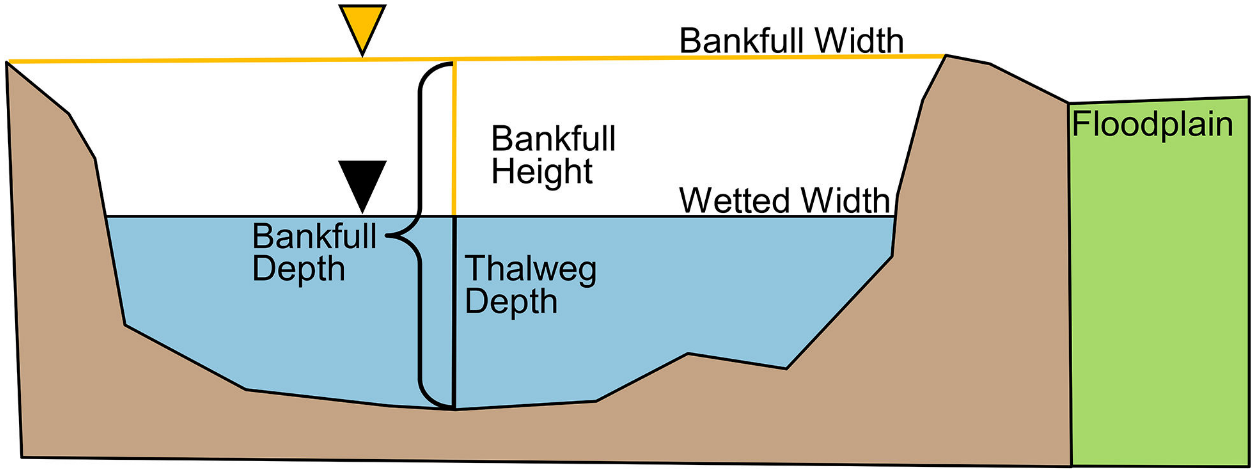

FIGURE 1.

Definitions of channel dimensions: Wetted width represents the distance from left to right bank of the stream channel filled with water. Bankfull width represents the distance from left to right bank at the bankfull stage (i.e., water level at which a stream will overtop its banks and enter the floodplain). Thalweg depth represents the depth at the deepest point in the channel cross section from the bottom substrate to the water surface. Bankfull depth represents thalweg depth added to the bankfull height from the water surface to the bankfull stage. Above the bankfull stage, water will no longer be confined solely within the channel but will begin to spill into the floodplain.

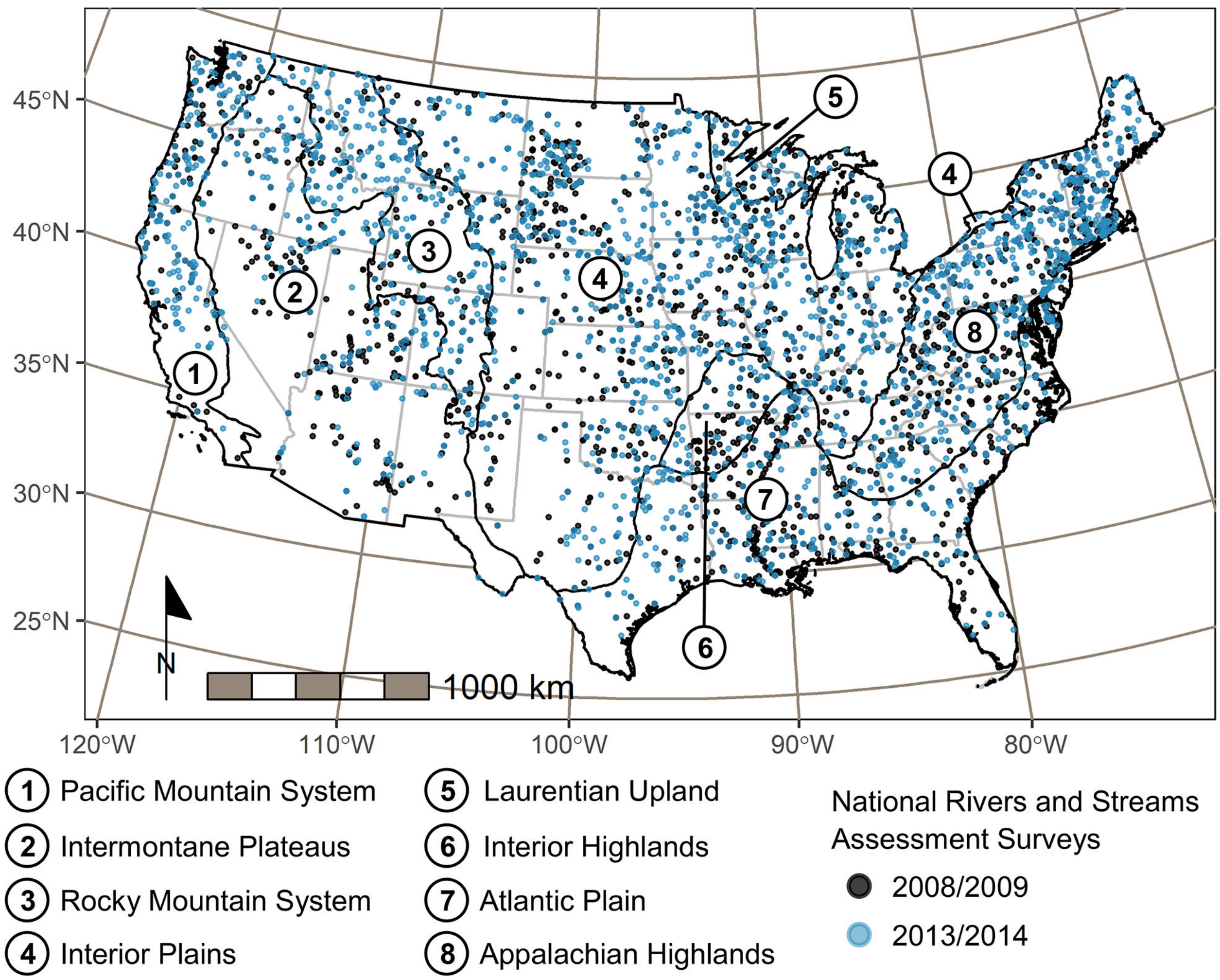

FIGURE 2.

Sampling locations of the 2008/2009 and 2013/2014 National Rivers and Streams Assessments. Physiographic Divisions of the United States (Fenneman and Johsnon 1946).

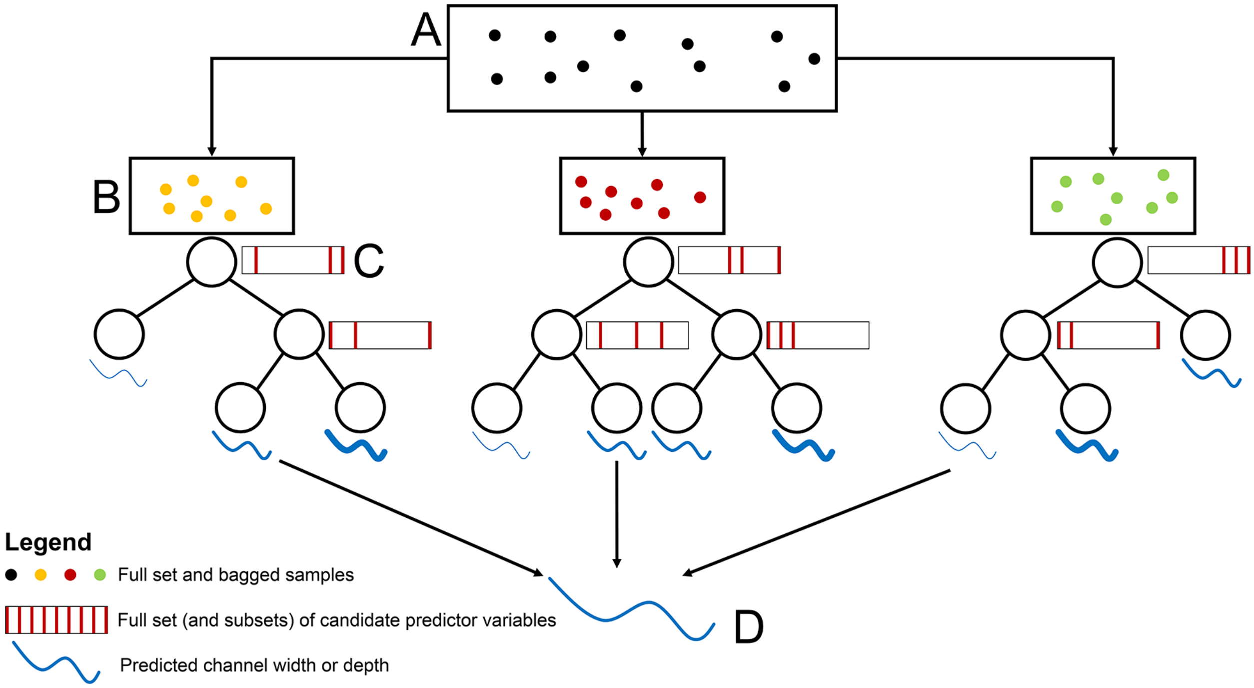

FIGURE 3.

Diagram depicting random forest modeling (modified from Berriri et al. 2021). (A) The full set of observations are bagged (i.e., subsetted) to create (B) random subsets of the data (also called "in-bag" samples) to construct each regression tree of the random forest. In-bag samples are typically about two-thirds of the full set of observations. During construction of a tree, (C) a randomized subset of candidate predictors is tested at each node (split) of the tree. The default number of predictors tested at each node for many random forest regression applications is p/3, where p is the number of available predictors. The (D) final prediction for a new observation is made by averaging predictions across individual regression trees. While the illustration shows a random forest constructed from just three trees, typical applications of random forest construct several hundred to thousands of individual trees.

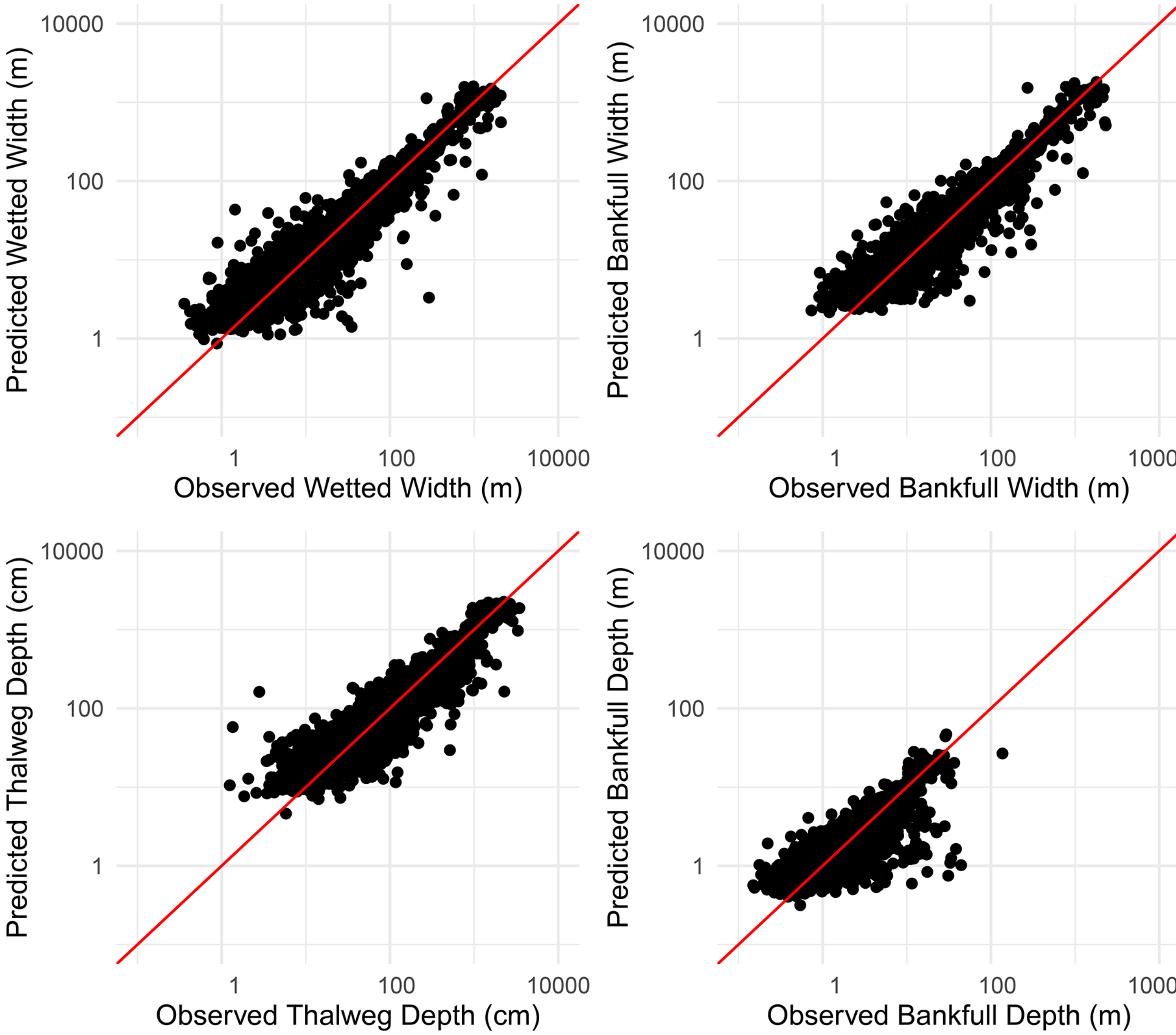

FIGURE 4.

Predicted versus observed values of wetted width, thalweg depth, bankfull width, and bankfull depth.

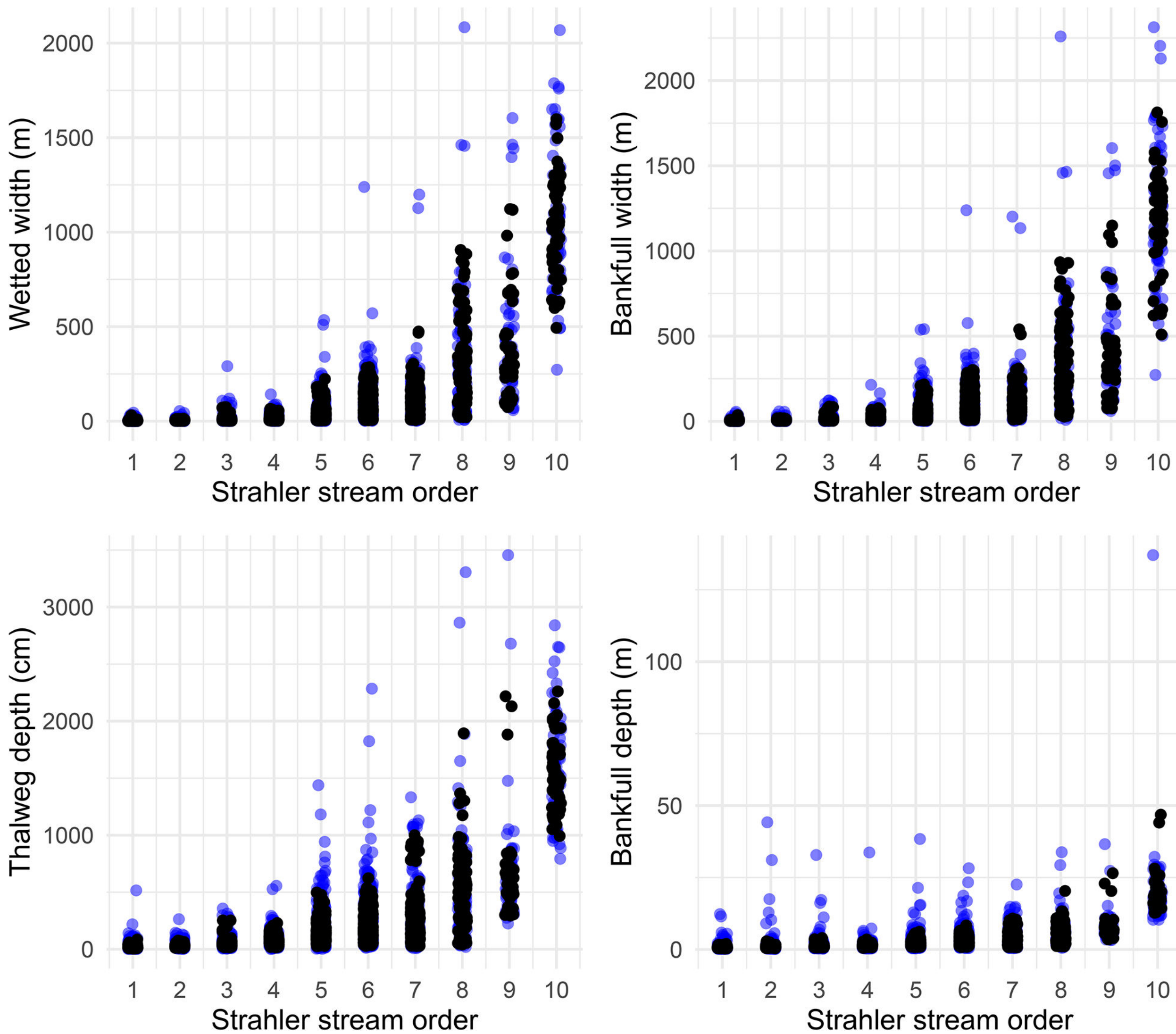

FIGURE 5.

Distribution of predicted (black) and observed (blue) channel dimensions across Strahler stream orders. Pseudo r-squared values for each model are as follow: wetted width = 0.91, thalweg depth = 0.82, bankfull width = 0.90, and bankfull depth = 0.70.

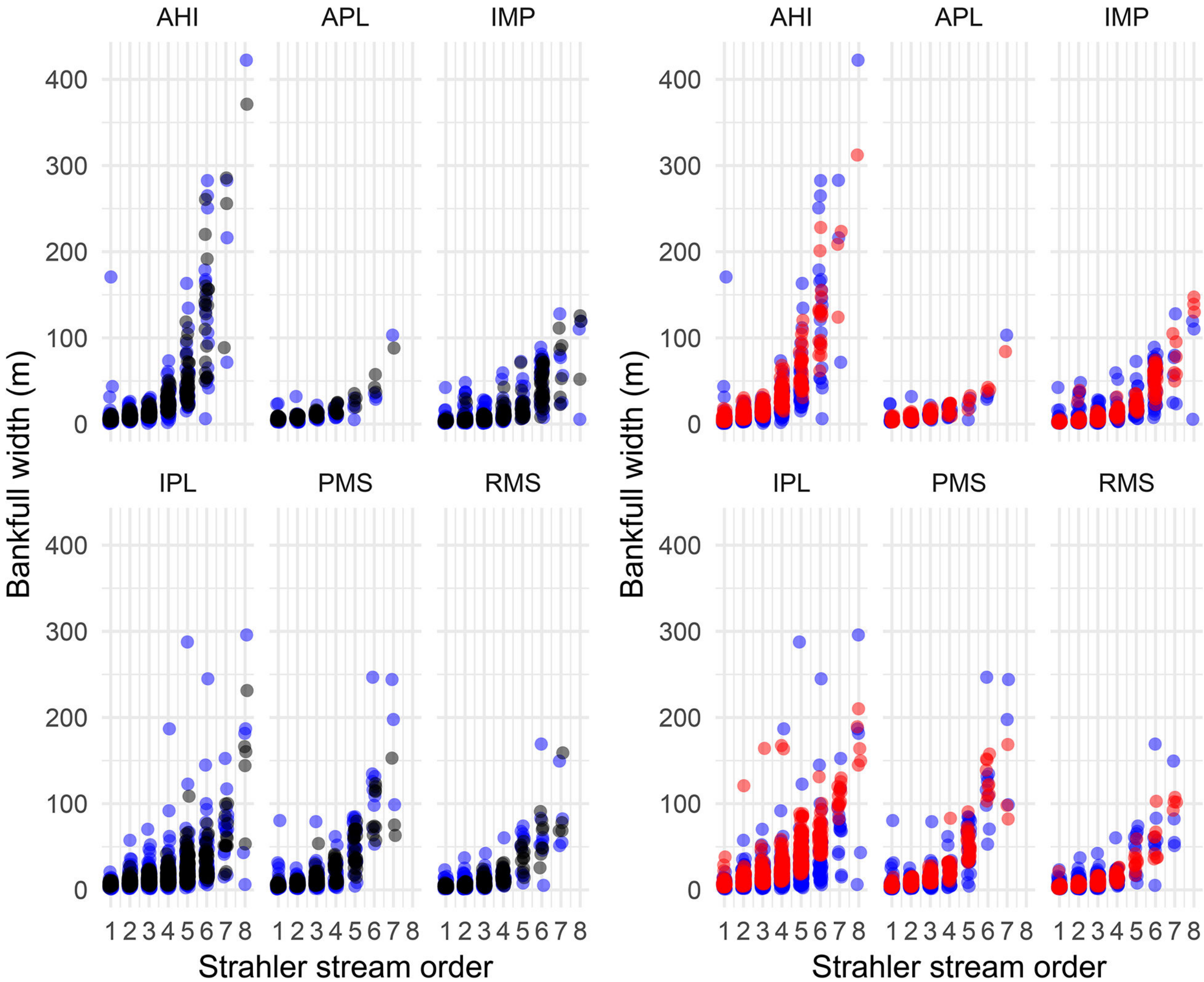

FIGURE 6.

Comparison of observed bankfull width values from 1994 to 2004 (blue) with predicted values using random forest (black; left) and watershed area-based regression coefficients (Bieger et al. 2015) (red; right) by Physiographic Division and stream order. AHI: Appalachian Highlands, APL: Atlantic Plains, IMP: Intermontane Plateaus, IPL: Interior Plains, PMS: Pacific Mountain System, and RMS: Rocky Mountain System.

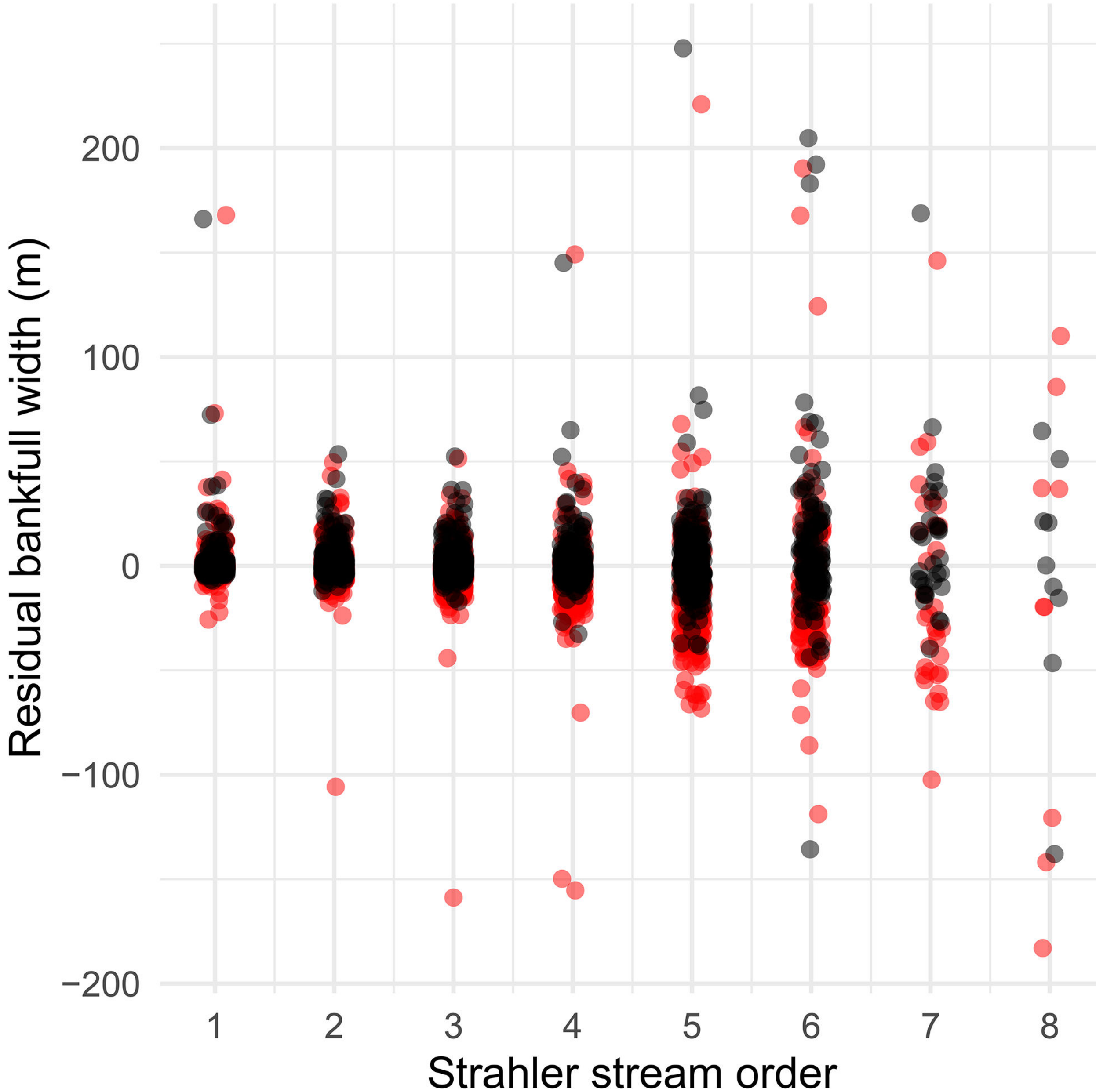

FIGURE 7.
Distribution of residual errors of modeled bankfull width by Strahler stream order. Black = random forest model. Red = watershed area-based regression (Bieger et al. 2015).

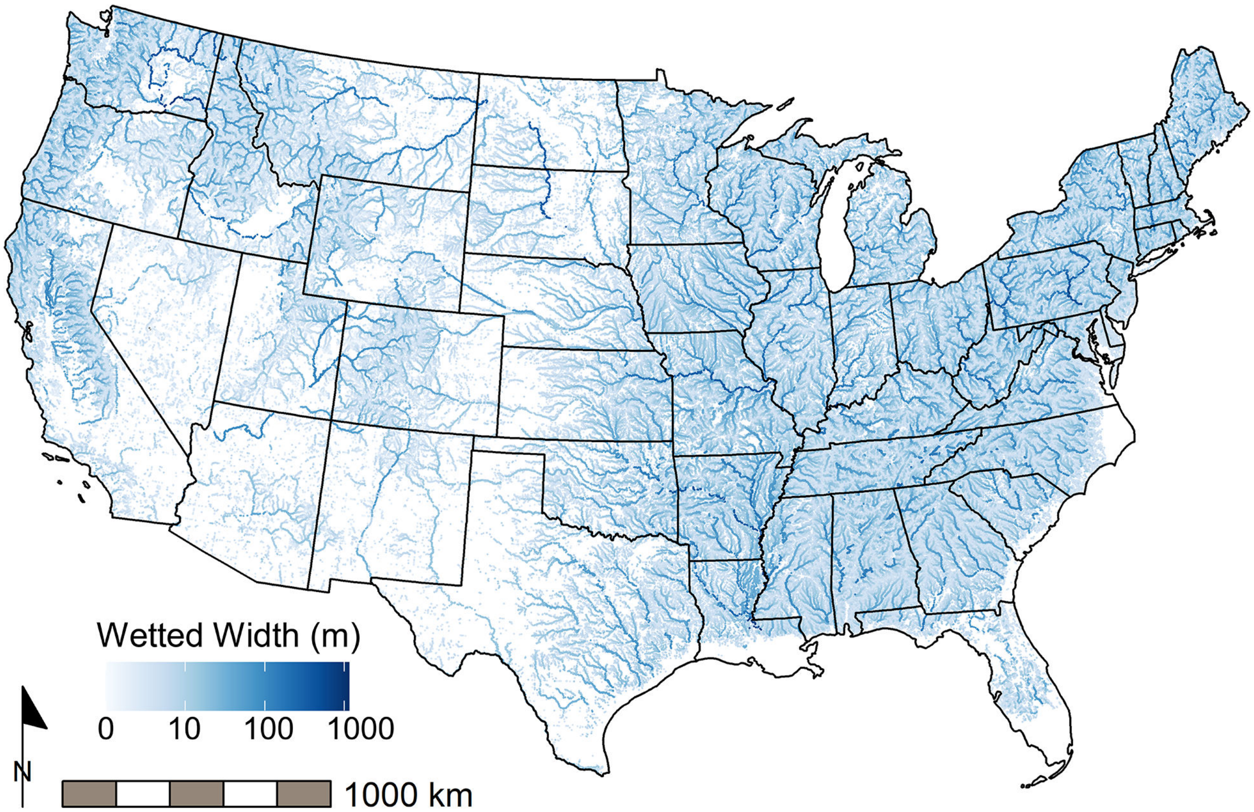

FIGURE 8.
Map of predicted wetted width values across the conterminous U.S.

TABLE 1.

Median Absolute Error (MdAE) for each model by NHDPlusV2 stream order and the conterminous US (CONUS). The MdAE (m) divided by the interquartile range of observed values for each stream order are provided within brackets (unitless) to standardize MdAE values across models. Percent bias (PBIAS) quantifies the tendency of a model to overestimate (positive) or underestimate (negative) observed values.

Stream Order	Wetted Width		Thalweg Depth		Bankfull Width		Bankfull Depth	
	MdAE	PBIAS	MdAE	PBIAS	MdAE	PBIAS	MdAE	PBIAS
1	0.69 [0.31]	−19.4	0.06 [0.33]	−21.2	1.23 [0.36]	−15.6	0.19 [0.41]	−15.4
2	1.10 [0.30]	−12.1	0.09 [0.35]	−13.0	1.80 [0.35]	−9.5	0.24 [0.42]	−24.6
3	1.62 [0.24]	−18.8	0.11 [0.35]	−9.4	2.57 [0.28]	−14.8	0.26 [0.37]	−13.8
4	2.84 [0.18]	−10.7	0.15 [0.35]	−10.2	4.12 [0.24]	−10.6	0.33 [0.37]	−9.4
5	4.92 [0.14]	−10.3	0.25 [0.28]	−12.1	6.50 [0.17]	−11.3	0.44 [0.32]	−10.0
6	8.79 [0.12]	−10.3	0.30 [0.19]	−14.5	10.84 [0.13]	−9.3	0.56 [0.27]	−11.2
7	12.67 [0.12]	−14.2	0.38 [0.17]	−15.0	16.77 [0.15]	−13.2	0.57 [0.16]	−13.5
8	29.22 [0.09]	−8.5	0.68 [0.16]	−10.6	37.23 [0.11]	−9.3	0.98 [0.21]	−7.8
9	27.27 [0.07]	−14.8	0.87 [0.27]	−8.3	28.67 [0.07]	−13.9	0.80 [0.23]	−5.8
10+	173.38 [0.35]	−4.8	2.92 [0.45]	−3.3	204.05 [0.35]	−4.2	4.04 [0.41]	−7.6
CONUS	2.96 [0.06]	−9.7	0.16 [0.15]	−11.1	4.27 [0.08]	−9.2	0.36 [0.20]	−11.3

EPA Author Manuscript

EPA Author Manuscript

TABLE 2.

Ranking of top 5 most important covariates from each model (1: most important, 5: least important).

Important Covariates	Wetted Width	Thalweg Depth	Bankfull Width	Bankfull Depth
Mean Runoff (mm)	1	1	2	3
Watershed Area (km^2)	2	2	1	1
Average 30-year Precipitation (mm)	3	5	3	5
Average 2-year Precipitation (mm)	5			
Slope (%)		3		2
Stream Order	4	4	4	4
Mean Watershed Elevation (m)			5	

EPA Author Manuscript

EPA Author Manuscript

TABLE 3.

Comparison of random forest and Bieger et al. (2015) models of bankfull width when applied to an independent dataset. Regression r-squared (R^2) and slope between predicted and observed values indicate model precision and accuracy, respectively. Also provided are Median Absolute Error (m) (MdAE) and MdAE standardized by the interquartile range of observed values in brackets (unitless). Percent bias (PBIAS) estimates the propensity of the model to overestimate (positive) or underestimate (negative) observed values (Moriasi et al. 2015).

		Random Forest				Bieger et al. (2015)			
Physiographic Region	**n**	**R^2**	**Slope**	**MdAE**	**PBIAS**	**R^2**	**Slope**	**MdAE**	**PBIAS**
Appalachian Highlands	589	0.76	1.03	2.20 [0.23]	1.0	0.73	0.98	4.15 [0.42]	19.2
Atlantic Plain	114	0.61	1.00	2.30 [0.37]	16.5	0.57	0.79	2.10 [0.34]	8.0
Interior Plains	716	0.63	1.07	3.42 [0.25]	−10.4	0.40	0.68	8.05 [0.59]	47.4
Intermontane Plateaus	350	0.67	1.00	2.05 [0.19]	−13.3	0.61	0.82	2.49 [0.23]	−2.5
Pacific Mountain System	350	0.74	1.12	2.48 [0.27]	−6.9	0.59	0.96	4.04 [0.44]	10.4
Rocky Mountain System	287	0.68	1.02	2.21 [0.26]	−7.4	0.52	0.84	3.06 [0.37]	−13.5
CONUS	2406	0.70	1.03	2.55 [0.24]	−6.1	0.56	0.78	4.20 [0.39]	21.2