

Bankfull and Mean-flow Channel Geometry Estimation through a Hybrid Multi-Regression and Machine Learning Algorithms across the CONTiguous United States (CONUS)

Reihaneh Zarrabi¹, Riley McDermott¹, Seyed Mohammad Hassan Erfani^{2,3}, and Sagy Cohen¹

¹Department of Geography and Environmental Studies, University of Alabama, Shelby Hall 2021, Tuscaloosa, AL, 35487.

²Center for Climate Systems Research, Columbia Climate School, Columbia University, New York, NY, 10025.

³NASA Goddard Institute for Space Studies, New York, NY, 10025.

Corresponding author: Sagy Cohen (sagy.cohen@ua.edu)

Key Points:

- Channel geometry estimation is essential for hydrological, geomorphological, and ecological modeling and analysis.
- A suite of data-driven models is developed to estimate channel width and depth under bankfull and mean-flow conditions.
- The best models are applied for reach-scale estimation of channel geometry for the contiguous United States.

Abstract

Widely adopted models for estimating channel geometry attributes rely on simplistic power-law (hydraulic geometry) equations. This study presents a new generation of channel geometry models based on a hybrid approach combining traditional statistical methods (Multi-Linear Regression (MLR)) and advanced tree-based Machine Learning (ML) algorithms (Random Forest Regression (RFR) and eXtreme Gradient Boosting Regression (XGBR)), utilizing novel datasets. To achieve this, a new preprocessing method was applied to refine an extensive observational dataset, namely the HYDRoacoustic dataset supporting Surface Water Oceanographic Topography (HYDRoSWOT). This process improved data quality and identified observations representing bankfull and mean-flow conditions. A compiled dataset, combining the preprocessed dataset with datasets containing additional catchment attributes like the National Hydrography Dataset Plus (NHDplusv2.1), was then used to train a suite of models to predict channel width and depth under bankfull and mean-flow conditions. The analysis shows that tree-based ML algorithms outperform traditional statistical methods in accuracy and handling the data but face limitations in prediction capabilities for streams with characteristics outside the training range. Consequently, a hybrid method was selected, combining XGBR for streams within the dataset range and MLR for those outside it. Two tiers of models were developed for each attribute using discharges derived from distinct sources (HYDRoSWOT and NHDplusV2.1, respectively), where the second tier of models offers applicability across approximately 2.6 million streams within NHDplusv2.1. Comprehensive independent evaluations are conducted to assess the capability of the developed models in providing stream/reach-averaged (rather than at-a-station) predictions for locations outside the training and testing datasets.

1. Introduction

Rivers, dynamic features of the earth's natural system, play a significant role in the lives of humans, flora, and fauna (Gleason, 2015; Wilby & Gibert, 1996). Estimating river hydraulic characteristics, such as width and depth, is crucial in analyzing river channel geomorphology (Harrelson et al., 1994; Monegaglia & Tubino, 2019; Naito & Parker, 2020; Zhou et al., 2022), the stream's ecology and water quality state (Walling & Webb, 1975; Rice et al., 2001; Thoms, 2003; Sobotka & Phelps, 2017), river management (Rosgen, 1994; Andrews & Nankervis, 1995;

Clerici et al., 2015), and flood forecasting and management (Orlandini & Rosso, 1998; Neal et al., 2015; Dey et al., 2022; Heldmyer et al., 2022).

Hydraulic geometry is critical in refining hydrological models, particularly within operational forecasting frameworks such as the National Oceanic and Atmospheric Administration (NOAA) National Water Model (NWM). These models often oversimplify key attributes, which limits their ability to accurately capture the intricate dynamics and routing of natural water systems. Consequently, this simplification undermines the accuracy of streamflow predictions. The NOAA Office of Water Prediction (OWP) relies on NWM-forecasted streamflow to produce Flood Inundation Mapping (FIM) via the Height Above Nearest Drainage (HAND) method. Additionally, models of channel geometry can be utilized to develop a refined Digital Elevation Model (DEM) that accurately represents both topography and the unique characteristics of river channels. This refined DEM can further enhance the accuracy of HAND-FIM predictions.

Leopold and Maddock Jr (1953) proposed a set of power-law equations to predict the mean hydraulic geometry attributes based on mean-flow discharge. This set of equations can be employed to predict bankfull hydraulic geometry attributes by replacing bankfull flow discharge with the previously considered mean-flow discharge (Leopold et al., 1964). Bankfull channel geometry is frequently used in hydrological modeling and analysis (Wolman & Leopold, 1957; Leopold et al., 1964; Williams, 1978; Radecki-Pawlik, 2002; Navratil et al., 2006; Charlton, 2007; Naito & Parker, 2019; Keast & Ellison, 2022). A similar methodology, known as Regional Hydraulic Geometry Curves (RHGC), was proposed by Dunne and Leopold (1978) to estimate the bankfull hydraulic attributes based on drainage area. This approach effectively resolved the challenge of restricting the utilization of hydraulic geometry solely to rivers and streams with recorded flow discharge by substituting flow discharge with drainage area (Ames et al., 2009). However, these equations were not widely utilized due to the lack of available measured channel dimensions necessary for their development over extensive geographic areas (Bieger et al., 2015). To address this limitation, various studies proposed to localize the regional curves for different regions across the United States, such as New York state (Mulvihill & Baldigo, 2012), Pennsylvania, and selected areas of Maryland (Chaplin, 2005), North Carolina's coastal plain (Sweet & Geratz, 2003), and the Pacific Northwest of the USA (Castro & Jackson, 2001).

In (2015), Bieger et al. established bankfull hydraulic geometry relationships that covered eight physiographic divisions, including 22 physiographic provinces as subdivisions across the USA, by utilizing an extensive dataset compiled from over 50 publications. The accuracy of channel bankfull prediction was further improved by Blackburn-Lynch et al. (2017) those developed hydraulic geometry equations for 20 Hydrologic Landscape Regions (*HLR*) across the USA. *HLR* classification was proposed by Wolock et al. (2004) for the CONtiguous United States (*CONUS*) based on geology, hydrology, climate, and soil characteristics. These calibrated equations are now employed to estimate reach-averaged bankfull channel geometry in the NOAA operational hydrological forecasting framework, the NWM (Gochis et al., 2020).

Despite the ongoing improvements in estimating channel geometry, accuracy remains limited by factors such as poor dataset quantity and quality, variations in spatial and temporal characteristics, and a lack of incorporation of catchment and reach attributes. Availability of large datasets, such as the HYDRoacoustic dataset in support of the Surface Water Oceanographic Topography (*HYDRoSWOT*) (Canova et al., 2016; Bjerklie et al., 2020) and the National Hydrography Dataset Plus (*NHDplusv2.1*) (McKay et al., 2012), containing extensive and wide-ranging data on catchment and reach properties, as well as the proliferation of machine learning algorithms offer new pathways for considerably enhancing the accuracy of channel geometry estimation.

A recent instance of such modeling is presented in the work of Doyle et al. (2023), where they explored the potential of employing the random forest algorithm and incorporating channel and watershed parameters to predict bankfull and low-flow hydraulic attributes of channels within the CONUS. While their models demonstrated acceptable accuracy, it is important to note that their application is confined to 1.1 million river segments from NHDPlusV2.1 within the sampling frame of the National Rivers and Streams Assessment (NRSA) datasets utilized in developing the models. This limitation results in the exclusion of significant regions, such as parts of the southwestern US and the arid foothills of Montana. Furthermore, the models may underestimate the impact of water impoundments (e.g., dam density) since the randomized placement of NRSA sample sites might not include sufficient sites below dams.

In this paper, we develop and test new CONUS-wide bankfull and mean-flow channel width and depth datasets. We compare a suite of machine learning algorithms and multi-regression models. Key methodological novelties introduced in this study include extensive data quality control and the identification of bankfull and mean-flow observations from cross-sectional surveyed data through the Acoustic Doppler Current Profiler (*ADCP*). A validation process is conducted to assess the efficacy of the developed models. Furthermore, an independent evaluation procedure is used to evaluate the accuracy of reach-averaged width and depth using an independent dataset derived from bathymetry surveys. Finally, geospatial datasets of bankfull and mean-flow width and depth for over 2.6 million reaches across CONUS are presented and analyzed.

2. Materials and Methods

2.1. Datasets and Pre-processing

The HYDRoSWOT dataset consists of 223,022 observations of channel and flow attributes obtained using an ADCP at more than 10,081 unique United States Geological Survey (*USGS*) stream gages sites, resulting in an average of 22 observations per sit, from the 1940s to 2014 (Canova et al., 2016). Key attributes included in this dataset include discharge, mean depth, maximum depth, width, cross-sectional area, mean velocity, and maximum velocity. Even though the data have received approval from the USGS, many records within the dataset contain blank fields, and a comprehensive examination for outliers or potentially erroneous data entries has not been carried out (Bjerklie et al., 2020).

For this study, a comprehensive procedure is implemented to enhance the quality of the HYDRoSWOT dataset. The process begins by filtering out observations containing zero, null, or negative values in any fields related to drainage area, discharge, mean depth, stream width, mean velocity, and maximum velocity. Canova et al. (2016) categorized gauge sites into 13 distinct categories, including atmosphere (AT), estuary (ES), diversion (FA-DV), outfall (FA-OF), QC lab (FA-QC), lake (LK), coastal (OC-CO), GW drain (SB-GWD), spring (SP), stream (ST), canal (ST-CA), ditch (ST-DCH), and tidal SW (ST-TS). Following this classification, gauge sites not identified as "stream (ST)" are excluded from further consideration. Then, observations

wherein the mean depth surpasses the maximum depth, or the mean velocity exceeds the maximum velocity are removed.

The discharge measurements obtained through the ADCP technique may contain errors, which could arise from inaccuracies in measuring flow velocity, errors in extrapolating discharge through unmeasured subsections, and variations in velocity along the river (Marsden & Ingram, 2004). To ensure the quality of the discharge obtained using this method, another filtration is considered to identify and eliminate observations that exhibit a discrepancy exceeding 5% within each pair of discharge values. These discharge values include the discharge value (q_va), the measured discharge value ($meas_q_va$), and the calculated discharge derived from the cross-sectional area multiplied by the mean velocity ($q2_xsec_area_X_mean_vel_va$). After implementing the filtration steps, the total number of observations decreased to 38,191 from 4,607 unique sites with an average of 8 observations per site. The dataset following this cleaning procedure is designated as *HYDRoSWOT_init* for future reference.

Analyzing the plot of the observed width/depth ratio against discharge for at-a-station channel geometry observations aids in identifying observations that can be classified as bankfull conditions. Initially, as discharge increases, both channel width and depth increase. However, within the channel, depth tends to increase more rapidly than width, resulting in a decrease in the width/depth ratio with increasing flow discharge. This trend shifts when the flow reaches channel banks, where even a small increase in channel depth results in a significant increase in the channel width as water spills over the channel banks onto the floodplain. This sharp increase in channel width results in an increase in the width/depth ratio with increasing flow discharge. The breakpoint in the trend, where the relationship changes, can be regarded as a quantitative indicator of the bankfull condition, as illustrated by Keast and Ellison (2022).

The trend of decreasing width/depth ratio with increasing discharge before the breakpoint is consistent. However, there is a significant deviation from this general pattern after the breakpoint. Hence, the data following the breakpoint can be regarded as outliers. The method proposed in this project for automating the identification of breakpoints in the width/depth ratio versus discharge plots relies on detecting outliers through the interquartile range (*IQR*) method. By applying this method to the *HYDRoSWOT_init* dataset, an upper limit for channel width is

established. This limit is defined as $Q_3 + 1.5 \times IQR$, where Q_3 represents the third quartile and IQR is the difference between the first quartile (Q_1) and Q_3 . Observations with width values exceeding this limit were considered as overbank and excluded. From the remaining data, the observation with the maximum discharge value was selected as the closest representation of the bankfull condition for each site.

To extract the observation associated with the mean-flow condition for each site, the observation in the *HYDRoSWOT_init* dataset with flow discharge that is closest to the NHDPlusV2.1 Mean Annual Flow from gage adjustment (QE_MA) attribute is selected. A new dataset is then created from the selected data for each site. Figure 1 illustrates the observations for USGS site number 06818000 after the filtration and identification process for bankfull and mean-flow conditions. This aims to enhance the comprehension of how the parameters for bankfull and mean-flow are selected in data preprocessing. More detail and additional examples are provided in Text S1 and Figure S1.

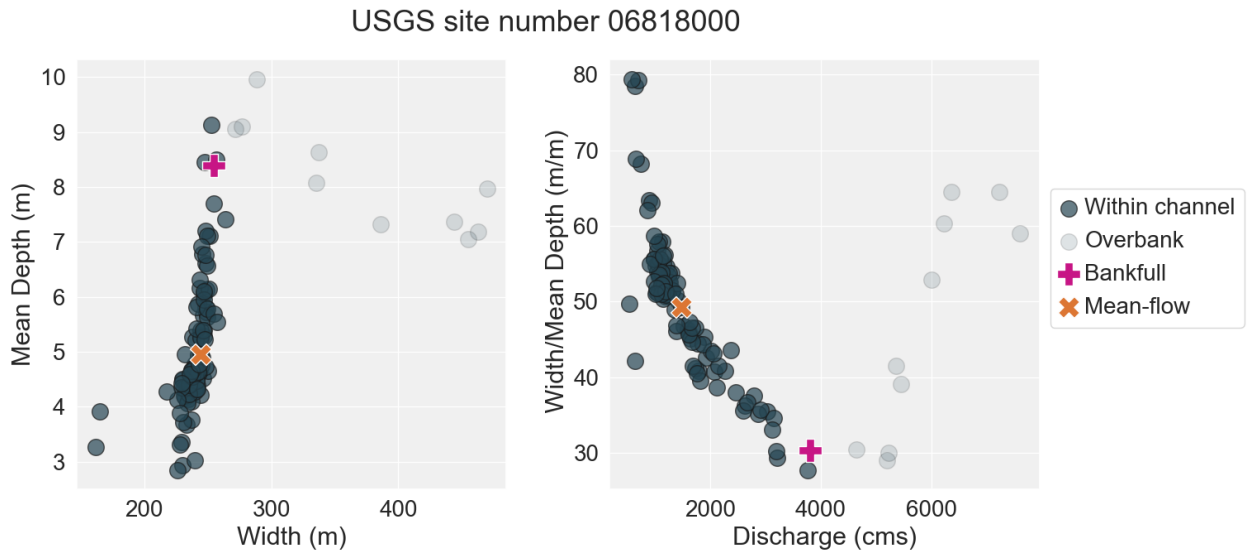


Figure 1. Visualization of within channel, overbank, bankfull, and mean-flow observations in the United States Geological Survey (USGS) site number 06818000.

The NHDPlusV2.1 dataset contains catchment and stream properties for more than 2.6 million reaches across the United States. This dataset is published by the USGS National Water-Quality Assessment Project (*NAWQA*), which is part of the USGS National Water Quality Program (*NWQP*) (McKay et al., 2012). The reaches are categorized into six groups within

NHDPlusV2.1, including StreamRiver, CanalDitch, ArtificialPath, Pipeline, Coastline, and Connector (Figure 2). For this study, those are categorized as StreamRiver, CanalDitch, and ArtificialPath are only considered for further analysis and application. In addition to the original NHDPlusV2.1, there is a metadata record that contains 13 various themes of datasets of natural and anthropogenic landscape features linked to the NHDPlusV2.1 (Wieczorek et al., 2018). Some river and catchment characteristics related to population infrastructure, soil, land cover, and hydrologic modification themes are selected from this metadata.

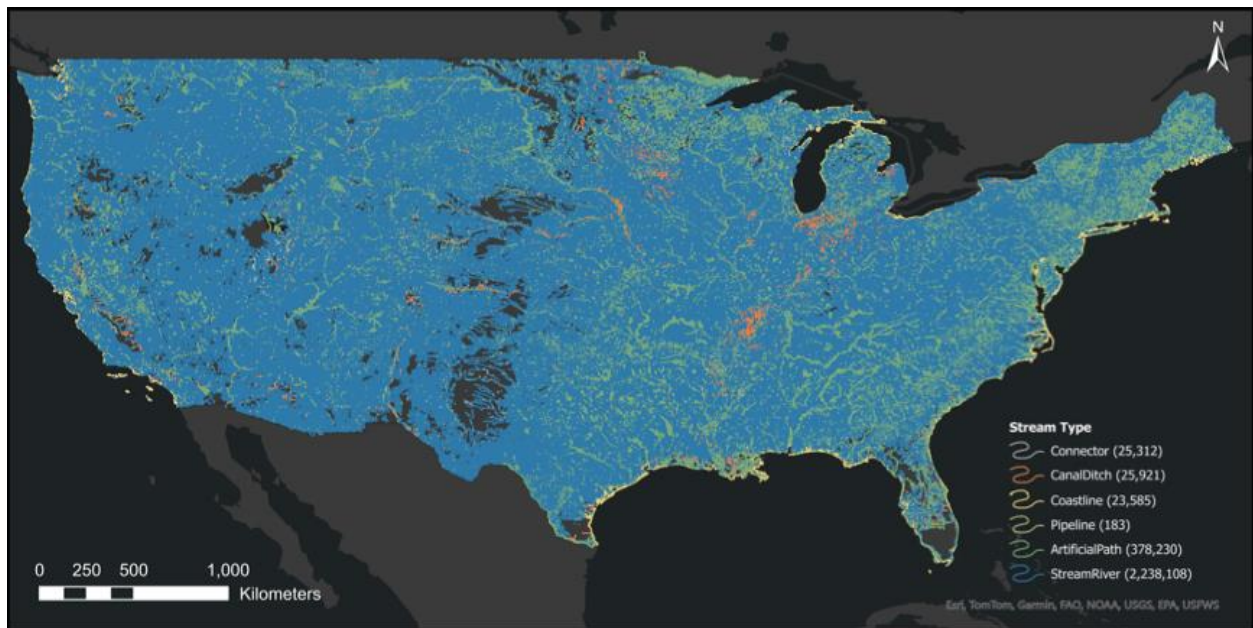


Figure 2. Map of stream/reach types in the National Hydrography Dataset Plus Version 2.1 (NHDplusv2.1).

The median bed-material sediment particle size (D_{50}) dataset (Abeshu et al., 2022) is presented in a vector format aligned with approximately 2.7 million river flowlines from the NHDPlusV2.1 dataset. The Global Aridity Index (Global-Aridity) dataset is a high-resolution global raster climate data at 30 arc seconds (~ 1 km at the equator) related to evapotranspiration processes and rainfall deficit for potential vegetative growth (Trabucco & Zomer, 2019). The Mean Aridity Index value was derived for each NHDPlusV2.1 flow stream using Geographic Information System (*GIS*). All the mentioned datasets are merged to compile the input dataset for model development. Table 1 presents all attributes along with their related descriptions, data sources, units of measurement, and some descriptive statistics.

Table 1. Dataset and attributes used for model development.

Source	Attribute name	Attribute description	Unit	Min	Max	Mean	Std
(Canova et al., 2016)	<i>Site no</i>	USGS site number	—	—	—	—	—
	<i>Lat</i>	Decimal latitude	<i>Degrees</i>	—	—	—	—
	<i>Long</i>	Decimal longitude	<i>Degrees</i>	—	—	—	—
	Q_{bnk}	Bankfull flow discharge	m^3/s	0.41	33,195.4	329.52	1,670.38
	d_{bnk}	Bankfull depth	m	0.30	27.9	2.54	2.02
	w_{bnk}	Bankfull width	m	3.78	1,816.6	63.62	93.84
	Q_{mf}	Mean-flow flow discharge	m^3/s	0.02	18,228.6	113.65	778.65
	d_{mf}	Mean-flow depth	m	0.19	27.9	1.57	1.62
	w_{mf}	Mean-flow width	m	3.35	2,124.5	55.08	90.54
(McKay et al., 2012)	<i>SO</i>	Modified Strahler stream order	—	1	10	4.79	1.35
	<i>DA</i>	Total upstream catchment area from the downstream end of the flowline	km^2	4.34	2,881,390	19,110.9	138,658
	<i>Z</i>	Smoothed minimum elevation	cm	3	269,497	24,842.3	32,354.4
	<i>S</i>	Slope of flowline based on smoothed elevations	m/m	0.00001	0.08803	0.0018	0.00426

(Wieczorek et al., 2018)	Q_E	Mean annual flow from gage adjustment/Best EROM estimate of actual mean-flow	m^3/s	0.00017	19,022.01	109.32	739.43
	ND	Accumulated number of dams built on or before 2010 based on total upstream accumulation	<i>Count</i>	1	41,971	248.56	1,955.87
	PD	Catchment population density from U.S. block-level population density rasters for 2010	<i>Persons</i> $/km^2$	0.01	4,478.56	215.47	442.14
	EVI_{fa}	Catchment mean Enhanced Vegetation Index value for the fall season 2011 (OND)	—	0.01	0.43	0.24	0.06
	EVI_{wi}	Catchment mean Enhanced Vegetation Index value for the winter season 2012 (JFM)	—	0.01	0.44	0.19	0.06
	EVI_{sp}	Catchment mean Enhanced Vegetation Index value for the spring season 2012 (AMJ)	—	0.02	0.62	0.39	0.09
	EVI_{su}	Catchment mean Enhanced Vegetation Index value for the summer season 2012 (JAS)	—	0.02	0.61	0.41	0.09
	Cl	Catchment average percent of clay	%	2.13	68.36	23.31	11.24
	Si	Catchment average percent of silt	%	4.13	77.24	43.44	13.63
	Sa	Catchment average percent of sand	%	3.04	92.80	33.25	19.86
	Dv	Estimated percent of catchment that contains the land-use and land-cover type developed	%	0.03	99.92	22.94	25.51

	<i>Fr</i>	Estimated percent of catchment that contains the land-use and land-cover type forest	%	0.01	96.85	29.78	25.15
	<i>Ag</i>	Estimated percent of catchment that contains the land-use and land-cover type agriculture	%	0.01	95.31	25.68	24.62
(Abeshu et al., 2022)	<i>D₅₀</i>	Median sediment particle size	<i>mm</i>	0.029	89.275	1.349	3.152
(Trabucco & Zomer, 2019)	<i>AI</i>	Mean Aridity Index	—	0.07	2.51	0.79	0.23

Once compiled, this dataset is subsequently shuffled and distributed randomly into training and testing sets with a split ratio of 75:25%. The final dataset contains 2626 observations collected from 2626 USGS gauge sites across the CONUS. The training and testing datasets size is 1969 and 657, respectively. Figure 3 shows the spatial distribution of the training and testing datasets over the CONUS.

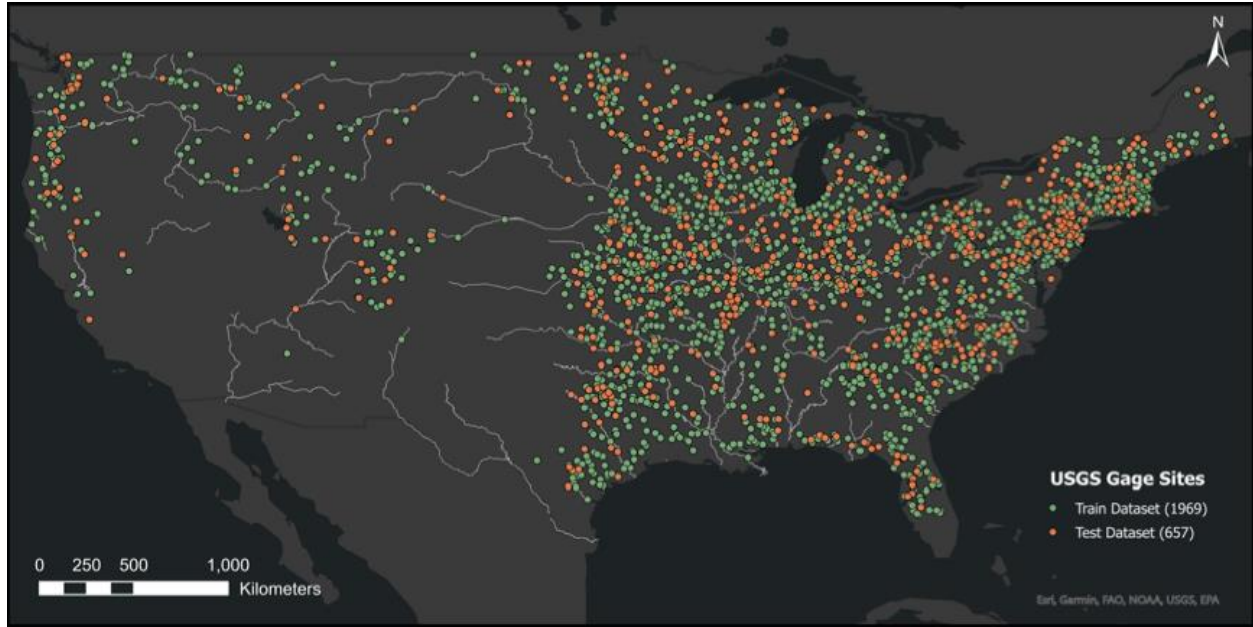


Figure 3. Spatial distribution map of the training and testing datasets utilized for model development.

2.2. Model Development

Two types of models are developed for each dependent variable (bankfull width, bankfull depth, mean-flow width, and mean-flow depth) using a suite algorithm (detailed below). The first tier of models uses HYDRoSWOT-derived discharge, denoted as Q_{bnk} for bankfull condition and Q_{mf} for mean-flow condition. Although these models exhibit notable performance, their applicability is limited to gauged rivers where bankfull and mean-flow discharges exist. Therefore, a second tier of models is developed, in which NHDPlusV2.1-derived mean annual flow, denoted as Q_E , is used.

2.2.1. Multi-Linear Regression (MLR)

Multi-Linear Regression (MLR) relates the target (dependent) variable to a set of independent variables. All variables undergo logarithmic transformation to fulfill the

assumptions inherent to regression modeling due to the skewness observed (Holder, 1986). This methodology has been used widely in developing prediction and forecasting models in water-related sciences (J et al., 2020; Bastola & Diplas, 2023). In this research, optimized models are developed for each specific target variable by implementing forward stepwise regression, which aids in identifying significant variables for modeling efficacy.

2.2.2. *Random Forest Regression (RFR)*

The Random Forest Regression (*RFR*) technique is a decision tree-based supervised model (Breiman, 2001). Due to its ability to handle a wide range of variables, large datasets, non-linearity among variables, complex higher-order interactions, and missing data (Boulesteix et al., 2012; Ziegler & König, 2014; Boulesteix et al., 2015; Biau & Scornet, 2016), this algorithm can be employed to model water-related attributes (Shortridge et al., 2016; Worland et al., 2018; Doyle et al., 2023).

2.2.3. *eXtreme Gradient Boosting Regression (XGBR)*

Introduced by Chen and Guestrin (2016), eXtreme Gradient Boosting Regression (*XGBR*) is another supervised algorithm that utilizes decision trees within the gradient boosting framework. This model demonstrates superior robustness, improving accuracy and computation time, achieved through parallel tree construction and learning from past errors to create a more powerful learner (Zakaria et al., 2023). Some limited studies have been developed in the water science area using *XGBR* algorithms (Ni et al., 2020; Nguyen et al., 2021).

2.3. Performance Metrics

Five metrics are utilized as performance indicators to assess the models' performance and uncertainties, comparing observed and predicted river geometry parameters. These metrics include the coefficient of determination (R^2), the Root Mean Square Error (*RMSE*), the Absolute Percent Bias (*APB%*), the Nash Sutcliffe Efficiency (*NSE*), and Kling-Gupta Efficiency (*KGE*). For more information about the definition of these metrics, refer to Krause et al. (2005) and Booker & Woods (2014).

2.4. Independent Evaluation

250 The models are initially developed and validated using the dataset extracted from the
251 HYDRoSWOT dataset at selected USGS gauge sites. However, the filtering and the bankfull and
252 mean-flow width and depth identification procedure may have introduced systematic biases
253 resulting in reduced accuracy of the model's predictions. These developed models are utilized to
254 predict reach-average channel geometry parameters for constructing the CONUS-scale database
255 (utilizing NHDplusv2.1 in this study). To ensure its applicability, we conduct an evaluation for
256 locations that were not included in either the training or testing datasets, referred to here as
257 independent evaluation.

To independently assess the mean-flow width and depth, we generate a new dataset by averaging reach-averaged width and depth using the US Army Corps of Engineers *eHydro* survey database, accessible at <https://www.sam.usace.army.mil/Missions/Spatial-Data-Branch/eHydro/> and <https://www.arcgis.com/apps/dashboards/4b8f2ba307684cf597617bf1b6d2f85d>. The bathymetric survey data in this repository is collected via single-beam or multi-beam sonar, from small or large ships, and occasionally from planes. Representative mean-flow depth and width values are extracted from the survey bathymetric raster and assigned to individual NHDplusv2.1 reach IDs (*COMID*). The calculation of representative mean-flow depth involves performing zonal statistics to obtain the mean of each depth value pixel within the NHDplusv2.1 catchment boundary, which is then assigned to the corresponding reach. Determining the mean reach width follows a three-step process. Initially, zonal statistics are applied to sum all the depth pixel values, calculating the total volume of the bathymetric survey within each NHDplusv2.1 catchment. This volume is then divided by the length of the NHDplusv2.1 reach, yielding the mean cross-sectional area for that reach. Finally, this cross-sectional area is divided by the mean depth calculated in the first step, providing a representative value for the stream width of the corresponding reach. In total, 60 surveys are used to extract data for 394 NHDplusv2.1 reaches for 25 rivers (refer to Table 2). We calculate the average value of adjacent reaches along a river path to mitigate spatial autocorrelation within rivers. Consequently, of the 394 reaches, 76 locations are used (check Figure S2 for a spatial distribution map of locations). The model-predicted parameters are subsequently averaged to the same averaged/joint reaches for the evaluation analysis.

Table 2. Summary of independent evaluation dataset (*eHydro* Surveys) for mean-flow condition including descriptive statistics.

River names	Number of Reaches	w_{mean}				d_{mean}			
		Min	Max	Mean	Std	Min	Max	Mean	Std
Ohio	82	76.37	781.72	372.73	118.10	3.64	12.24	6.99	2.24
Arkansas	77	34.79	645.99	370.64	93.24	2.02	10.86	5.61	1.99
Monongahela	36	59.99	305.11	172.60	40.32	3.45	8.13	5.51	1.23
Illinois	22	28.75	268.90	148.18	55.38	2.24	4.36	3.26	0.57

Missouri	20	98.79	222.80	178.36	34.33	3.36	5.04	4.35	0.41
Colorado	20	3.80	213.77	80.03	36.42	3.41	6.13	4.49	0.74
Kaskaskia	20	60.75	106.06	83.86	12.00	3.28	5.24	4.06	0.53
Tombigbee	16	1.20	149.15	78.64	42.45	2.82	7.27	4.63	1.17
Vermilion	14	34.26	60.61	44.96	8.27	1.19	2.77	2.03	0.46
Allegheny	12	137.45	464.52	270.05	91.76	3.30	7.57	5.15	1.31
Choptank	12	17.75	109.75	56.23	32.95	1.49	2.14	1.79	0.20
Mississippi	11	66.37	655.00	366.59	166.44	3.57	6.30	4.58	0.91
Kanawha	8	152.68	200.41	180.56	17.09	3.29	4.41	3.83	0.39
San Bernard	6	44.47	54.70	47.76	4.14	3.41	4.38	4.01	0.36
Broad Creek	6	18.30	46.01	35.25	10.83	1.69	2.24	1.98	0.24
James	6	105.39	120.90	115.77	5.76	5.04	6.34	5.61	0.58
Buffalo Bayou	4	89.05	194.44	142.79	43.72	11.96	13.36	12.51	0.67
Columbia	4	392.65	957.46	650.88	232.31	4.90	13.56	9.50	4.34
Arroyo Colorado	3	52.65	57.50	55.19	2.43	3.90	4.38	4.07	0.27
Green	3	69.04	87.01	80.39	9.88	6.90	7.19	7.00	0.16
Delaware	3	124.50	204.10	164.74	39.80	3.30	6.98	5.62	2.03
Cocheco	3	36.52	47.22	42.81	5.59	1.89	2.06	1.98	0.09
Willamette	2	329.32	329.86	329.59	0.38	14.48	14.64	14.56	0.12
Alabama	2	141.58	169.77	155.67	19.93	4.50	5.37	4.94	0.62
Mackeys Creek	2	100.71	140.16	120.44	27.90	3.04	3.18	3.11	0.09

For the evaluation of bankfull width and depth, an observational dataset is compiled from 11 published sources (Table 3). These sources include cross-sectional surveys and various hydraulic attribute measurements conducted at USGS gage sites, including bankfull width, depth, and discharge. From this dataset, data related to USGS gages that are not included in the models' training or testing datasets is used. While this dataset is somewhat similar to HYDRoSWOT (cross-sectional observation at USGS gages), the bankfull geometry measurements are independent of our extraction procedure. The resulting evaluation dataset only includes small rivers and streams, with a maximum width and depth of 85.1 and 4.39 meters, respectively (Table 3).

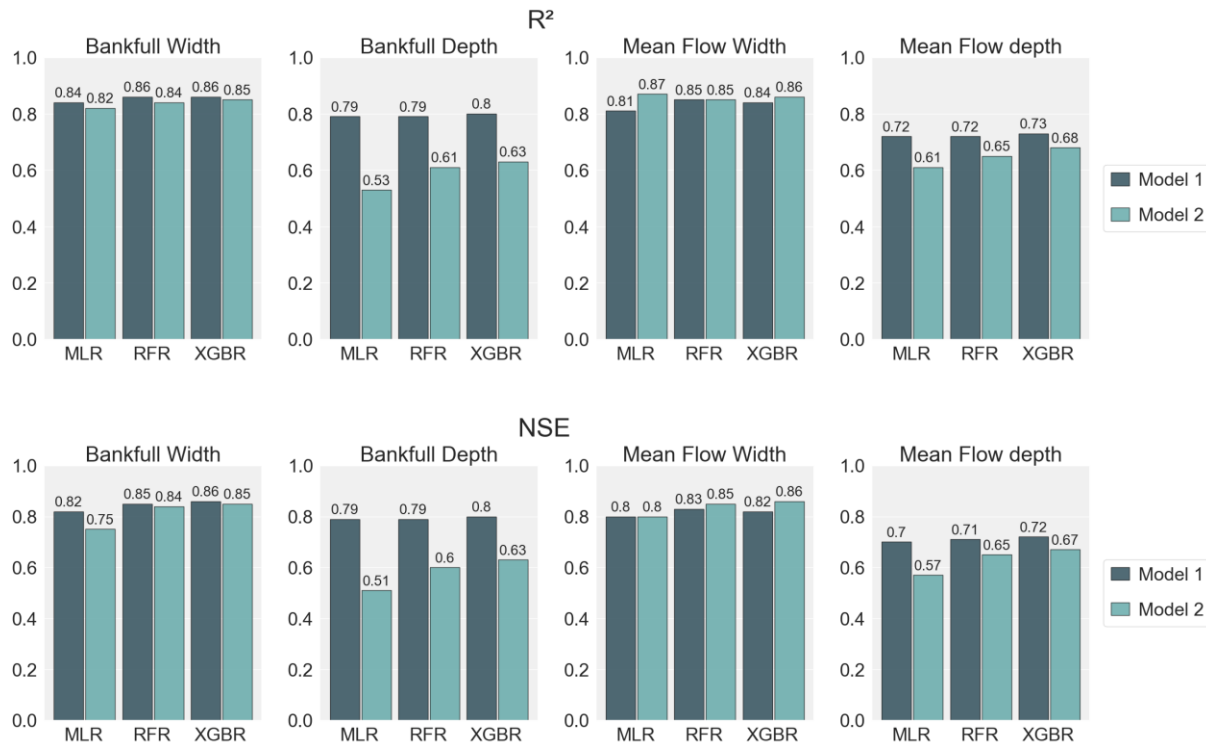
291 **Table 3.** Summary of independent evaluation dataset (gathered from 11 different sources) for bankfull
 292 condition including descriptive statistics.

Source	Number of Reaches	w_{bnk}				d_{bnk}			
		Min	Max	Mean	Std	Min	Max	Mean	Std
(Mulvihill et al., 2009)	31	14.68	68.99	30.8	15.36	0.73	2.79	1.25	0.5
(Keaton et al., 2005)	21	13.35	40.84	27.92	8.22	0.76	1.62	1.23	0.27
(Dutnell, 2000)	20	12.85	85.1	37.06	19.07	0.71	4.39	1.71	0.85
(Moody et al., 2003)	19	13.66	52.12	29.16	9.63	0.73	1.43	1.04	0.21
(McCandless & Everett, 2002)	11	12.31	26.27	19.34	4.22	0.79	1.83	1.32	0.31
(Brockman, 2010)	10	13.65	35.86	21.35	6.02	0.71	1.88	1.01	0.32
(Lotspeich, 2009)	6	13.81	41.15	24.51	9.6	0.76	2.04	1.35	0.52
(Parola et al., 2007)	6	17.83	37.49	25.42	6.23	1.17	3.89	2.32	0.92
(Metcalf, 2004)	5	14.17	40.63	19.79	10.42	1.37	2.44	1.79	0.39
(Chase, 2004)	4	34.14	44.81	39.32	4.52	0.91	1.22	1.1	0.12
(Mccandless, 2003)	3	19.42	38.34	26.28	8.55	0.85	0.98	0.91	0.05

3. Results and Discussion

3.1. Channel Geometry Modeling

The channel width models show strong prediction capabilities for the testing subset, with R^2 values ranging between 0.81 to 0.87, averaging at 0.85 (Figure 4). In contrast, the channel depth models result in lower predictive capability and a wider range of R^2 values, from 0.53 to 0.80, averaging at 0.69 (Figure 4). Additionally, the NSE and KGE values are higher for the width models, underscoring their proficiency compared to depth models. This discrepancy in performance is attributed to the superior quality of the width dataset employed in model development relative to the depth observations. The depth measurement presents inherent challenges, including interference from local obstructions such as debris or vegetation, water turbulence, and complexities in channel bathymetry. On the other hand, measuring width is comparatively more straightforward as it can be visually observed.



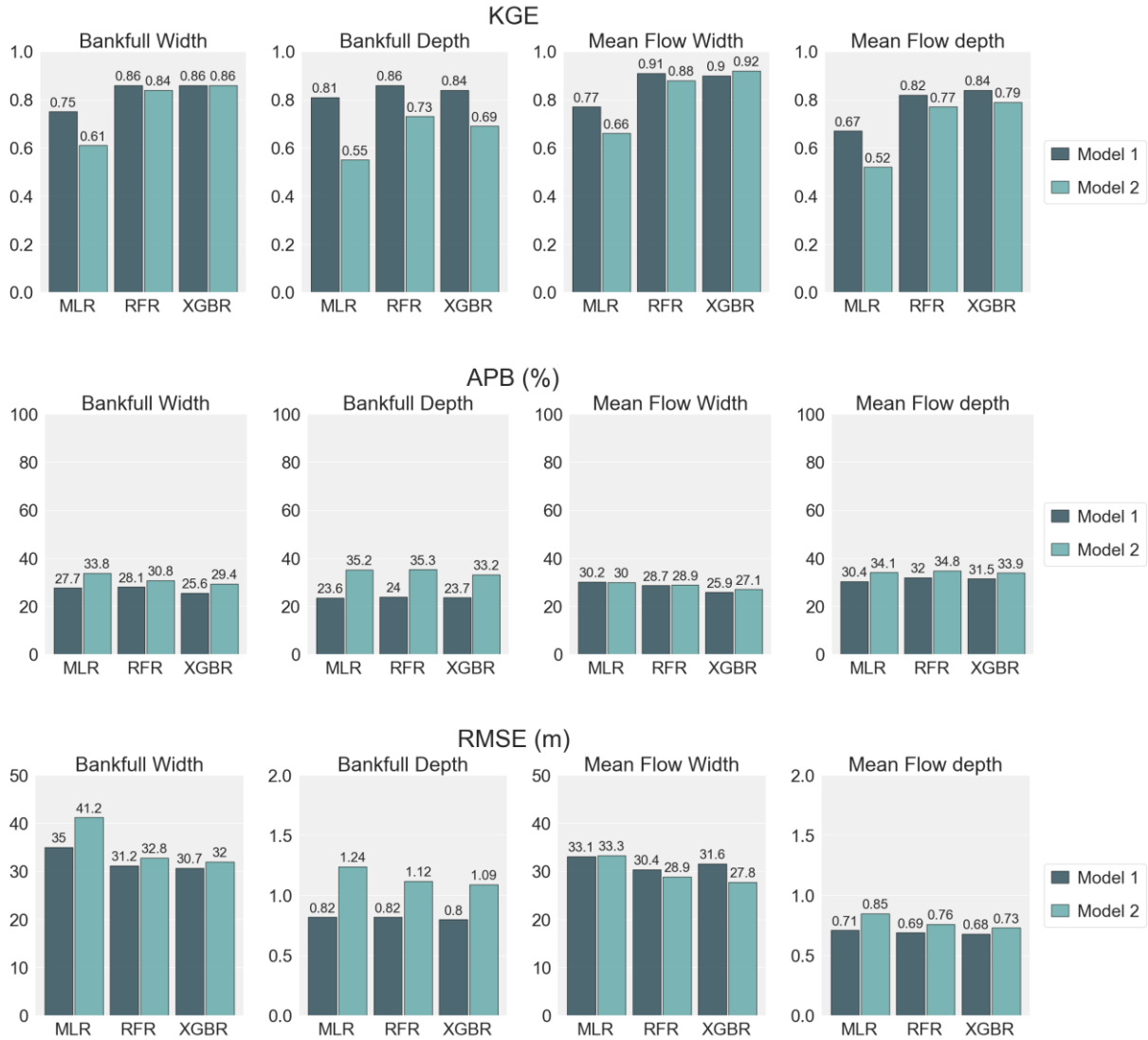


Figure 4. Performance metrics for the first and second tiers of models (Model 1 (HYDRoSWOT discharge) and Model 2 (NHDplusV2.1 discharge), respectively) using Multi-Linear Regression (MLR), Random Forest Regression (RFR), and eXtreme Gradient Boosting Regression (XGBR) algorithms on the test dataset to estimate of bankfull width, bankfull depth, mean-flow width, and mean-flow depth.

In the context of predicting width, both tiers of models yield nearly identical results in terms of accuracy in both bankfull and mean-flow conditions. To illustrate, when employing the XGBR algorithm, the R^2 values for the first tier of models are 0.86 and 0.84 for bankfull and mean-flow conditions, respectively (Figure 4). Similarly, for the second tier of models, the corresponding R^2 values are 0.85 and 0.86 for bankfull and mean-flow conditions (Figure 4), showcasing high consistency between the two models. In contrast, for depth predictions, the first

tier of models produces more robust results in the bankfull state than the mean-flow state, with R^2 values obtained by XGBR being 0.80 and 0.73, respectively (Figure 4). Conversely, the second tier of models delivers better results for the mean-flow condition than the bankfull condition, with R^2 values obtained by XGBR of 0.68 and 0.63, respectively (Figure 4).

Comparing various metrics values reported in Figure 4, it becomes clear that MLR models yield less accurate results across almost all attributes, with R^2 ranging from 0.53 to 0.87 and an average of 0.75. This lower accuracy is attributed to the MLR models' limited ability to capture non-linear and intricate relationships. In contrast, both RFR and XGBR models, being tree-based, exhibit more accuracy by adeptly handling non-linearity and complexity. Notably, models generated by the XGBR algorithm demonstrate the most robust outcomes, with R^2 ranging from 0.63 to 0.86 and an average of 0.78, due to their inherently robust algorithms that can learn from preceding steps.

It is important to note that the data used for creating MLR models was log-transformed to satisfy the primary assumptions necessary for MLR models. In contrast, the data was not log-transformed for the RFR and XGBR, as these models do not require preprocessing. This presents another advantage of utilizing tree-based models like RFR and XGBR over MLR in addition to their superior accuracy. However, when extending the application of models to all streams in the CONUS, a limitation emerges with RFR and XGBR. These models need help predicting values for streams where one or more river and catchment attributes (independent variables) fall outside the range covered by the training dataset. This often leads to the generation of negative values. To address this issue, a new approach is adopted: the XGBR model is selected for application to streams with independent variable values within the range of those in the training datasets. In contrast, MLR models are applied for streams with independent variable values outside this range. The MLR power-law equations for the first and second tiers of models (Model 1 and Model 2, respectively) for each attribute are reported as follows (see Table 1 for annotation):

$$w_{bnk, model1} = 5.36 Q_{mf}^{0.29} DA^{0.18} AI^{0.31} D50^{0.03} Agr^{-0.02} SI^{-0.08} \quad (1a)$$

$$w_{bnk, model2} = 11.58 Q_{mf}^{0.35} EVI_{fa}^{-0.27} ND^{0.03} AI^{0.17} Fr^{-0.02} \quad (1b)$$

$$d_{bnk, model1} = 3.53 Q_{bnk}^{0.31} PD^{-0.02} Z^{-0.09} S^{-0.03} Fr^{-0.03} Ag^{0.02} Si^{-0.17} Sa^{-0.15} \quad (1c)$$

$$d_{bnk, model2} = 1.76 EVI_{wi}^{-0.13} DA^{0.19} AI^{0.28} Z^{-0.08} S^{-0.04} Fr^{-0.02} Si^{-0.09} Sa^{-0.18} \quad (1d)$$

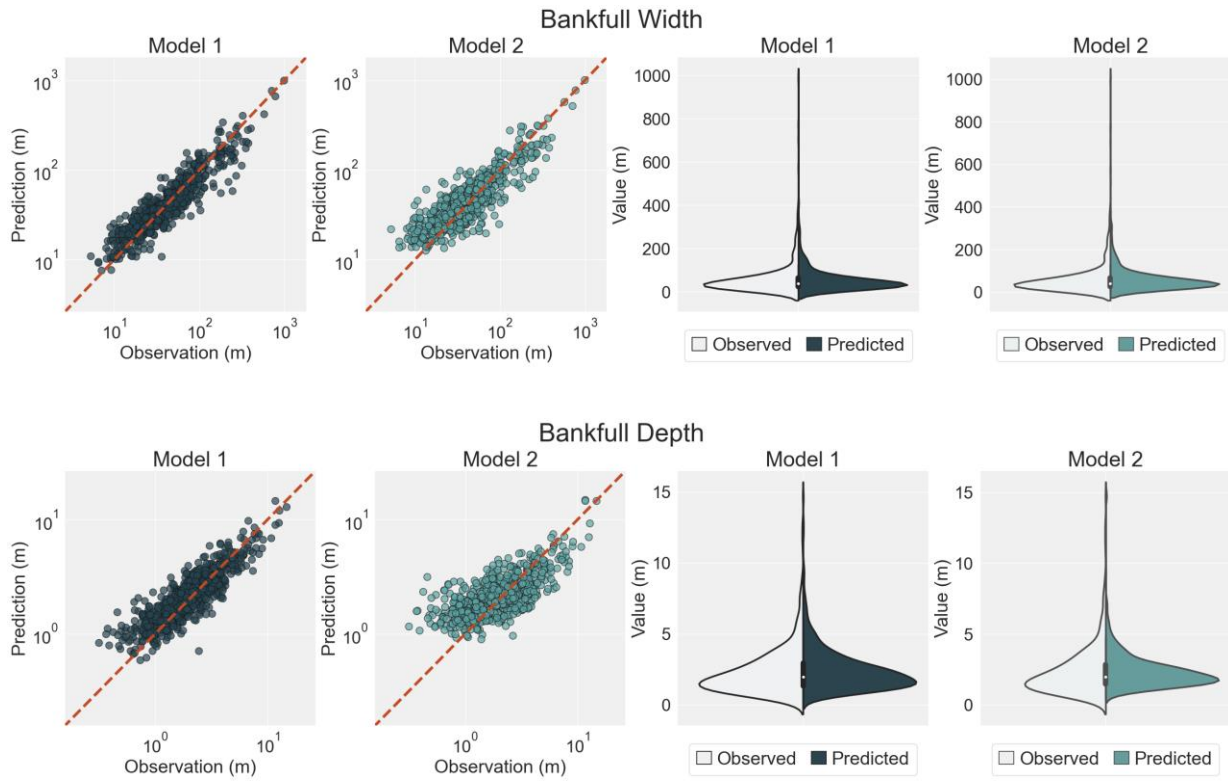
$$w_{mean, model1} = 5.81 Q_{mf}^{0.37} EVI_{wi}^{-0.16} ND^{0.02} SO^{0.49} AI^{0.25} D_{50}^{0.04} Dv^{0.022} Si^{-0.08} \quad (1e)$$

$$w_{mean, model2} = 7.76 Q_E^{0.41} EVI_{wi}^{-0.22} AI^{0.15} Dv^{0.03} \quad (1f)$$

$$d_{mean, model1} = 5.48 Q_{mf}^{0.28} PD^{-0.02} Z^{-0.11} S^{-0.02} Fr^{-0.03} Si^{-0.21} Sa^{-0.15} \quad (1g)$$

$$d_{mean, model2} = 7.41 Q_E^{0.24} ND^{0.04} PD^{-0.02} SO^{-0.28} Z^{-0.10} S^{-0.02} Fr^{-0.04} Si^{-0.19} Sa^{-0.14} \quad (1h)$$

Both tiers of models adeptly captured the central tendencies of the data under both bankfull and mean-flow conditions, as illustrated in Figure 5 and outlined in detail in Table 4. However, it is noteworthy that the second tier of models demonstrates slightly better performance than the first. Similarly, both models exhibit an enhanced capability to predict maximum values accurately. However, the trend differs from that observed for central tendencies. In bankfull conditions, Model 1 outperforms Model 2. However, under mean-flow conditions, Model 2 performs better than Model 1.



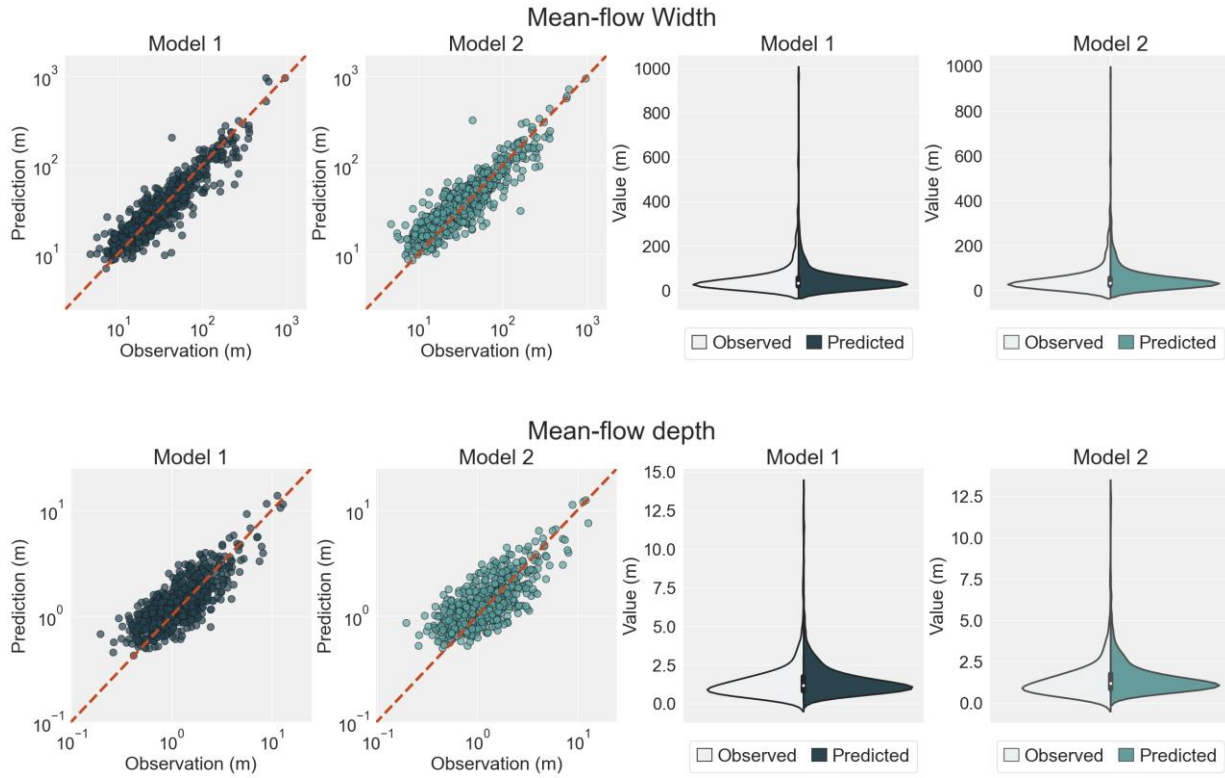


Figure 5. Scatter and violin plots of observations against predictions obtained by the first and second tiers of models (Model 1 (HYDRoSOT discharge) and Model 2 (NHDplusV2.1 discharge), respectively)) using the eXtreme Gradient Boosting Regression (XGBR) algorithm on the test dataset to estimate bankfull width, bankfull depth, mean-flow width, and mean-flow depth.

Table 4. Summary of observation and predicted values ranges, and bias in percent (in parenthesis) obtained by the first and second tiers of models (Model 1 (HYDRoSOT discharge) and Model 2 (NHDplusV2.1 discharge), respectively)) using the eXtreme Gradient Boosting Regression (XGBR) algorithm on the test dataset to estimate bankfull width, bankfull depth, mean-flow width, and mean-flow depth.

Attribute	Observation			Model 1			Model 2		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Bankfull Width (m)	5.2	987.6	61.6	7.7 (46.3%)	992.1 (0.5%)	58.2 (-5.6%)	12.7 (142.0%)	1009.9 (2.3%)	60.8 (-1.3%)
Bankfull Depth (m)	0.3	14.7	2.4	0.6 (88.2%)	14.6 (-0.9%)	2.5 (2.6%)	0.9 (186.9%)	15.0 (1.8%)	2.4 (0.4%)

Mean-flow				6.9	965.1	50.8	8.6	952.3	53.7
Width (m)	4.7	957.1	53.4	(47.0%)	(0.8%)	(-4.9%)	(81.3%)	(-0.5%)	(0.7%)
Mean-flow				0.4	13.8	1.5	0.5	12.5	1.5
Depth (m)	0.2	12.8	1.5	(118.0%)	(8.0%)	(5.6%)	(155.7%)	(-2.3%)	(5.5%)

Despite effectively estimating central tendencies and maximum values, both models demonstrate limitations in accurately predicting minimum values. The first tier of models displays biases of 46.3% for bankfull width, 88.2% for bankfull depth, 47% for mean-flow width, and 118% for mean-flow depth. The second tier of models yielded even more significant biases, with values increasing to 142%, 186.9%, 81.3%, and 155.7%, respectively. These findings underscore that while XGBR effectively handles non-linear relationships, it might encounter challenges when dealing with small values that deviate significantly from the general trends in most of the data. This highlights the importance of understanding the specific characteristics of the data and considering potential model limitations when relying on the XGBR to make predictions.

Feature importance analysis (Figure 6) shows that discharge (Q_{bnk} , Q_{mf} , and Q_E) plays the most significant role in predicting the channel geometry parameters. If discharge is removed from the feature sets used for developing models, the loss function, a mean squared error for the XGBR algorithm, will increase significantly. Furthermore, the importance of discharge features is higher for the first tier of models. For instance, in predicting bankfull width using the first tier of models, the significance of bankfull discharge (Q_{bnk}) is calculated at 55.33%. In contrast, in the second tier of models, the importance of the flow discharge feature (Q_E) decreases to 46.75%. This discrepancy in the importance of discharge attributes between the two models stems from the specific attributes used to develop each tier. For the first tier, Q_{bnk} and Q_{mf} are used, derived from the ADCP (HYDRoSOT) measurements. In contrast, the NHDplusV2.1-derived discharge (Q_E) attribute is used in the second tier of models.

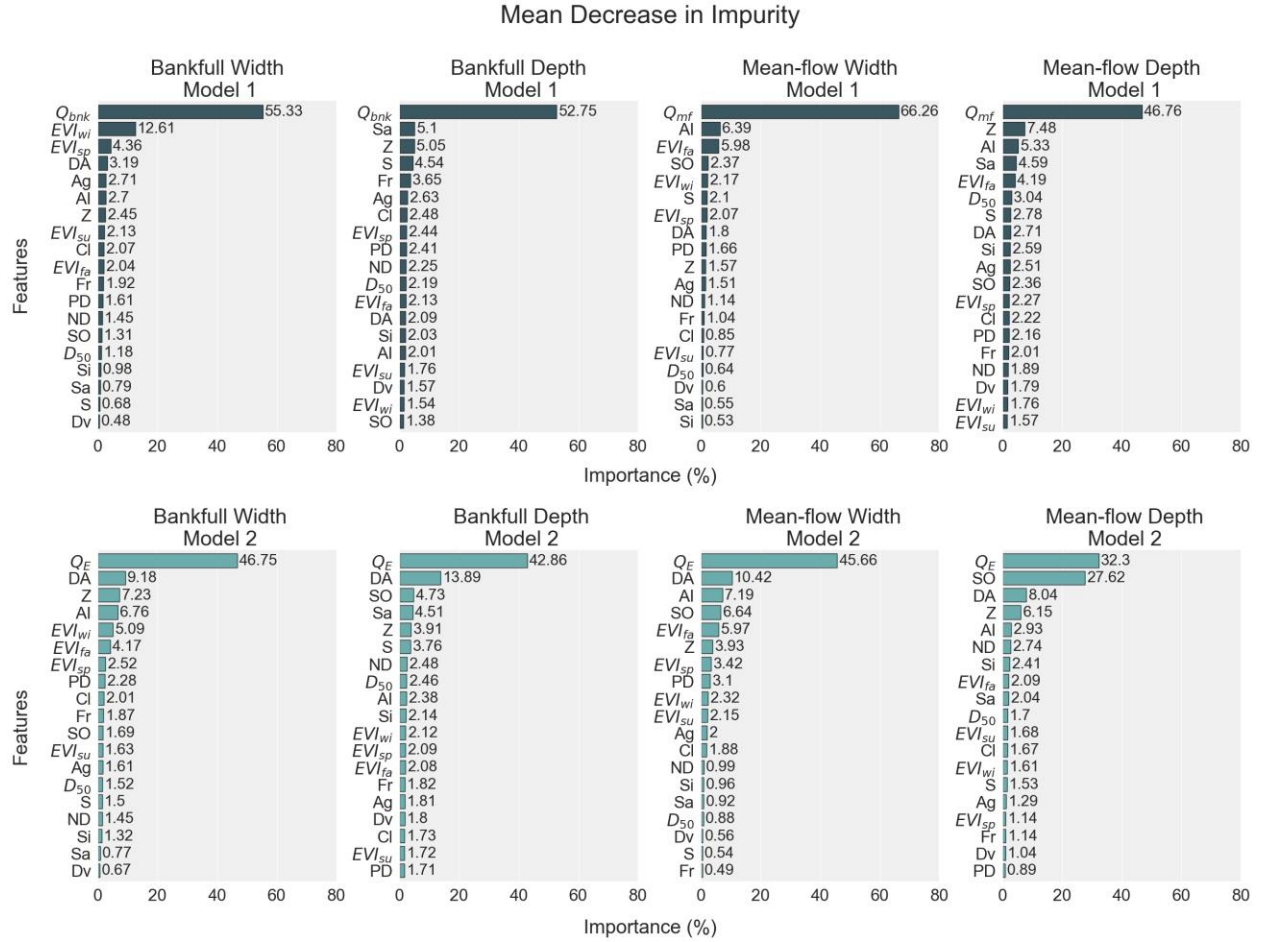


Figure 6. Results of the Mean Decrease in Impurity (MDI) analyses were obtained by applying developed first and second tiers of models (Model 1 (HYDRoSWOT discharge) and Model 2 (NHDplusV2.1 discharge), respectively)) using the eXtreme Gradient Boosting Regression (XGBR) algorithm on the test dataset to estimate bankfull width, bankfull depth, mean-flow width, and mean-flow depth.

The second and third most significant features vary across different models and attributes. Notably, drainage area (DA), aridity index (AI), stream order (SO), minimum elevation (Z), catchment average percentage of sand (Sa), and enhanced vegetation index (EVI) emerge as the second and third most influential features for different models. The contribution of these parameters can be explained by considering the fundamental principles of river hydrology and geomorphology and the spatial dynamics of channel characteristics from headwaters to river mouths.

Higher elevations are often associated with steeper slopes, fostering more energetic flows contributing to channel erosion and sediment transport. The composition of bed materials, like

the catchment-averaged percentage of sand, directly influences erosion and sediment transport. This factor contributes to the dynamics of channel morphology and sedimentation patterns. Furthermore, upstream river areas typically have more natural and intact vegetation cover, as they are generally less affected by human activities like agriculture or urbanization. This vegetation cover can influence sediment transport rates, acting as a stabilizing factor. The combination of channel elevation, climate features, bed-material composition, and vegetation cover highlights the complex interplay between natural forces and human activities that shape river systems' hydrological and morphological aspects along their course, resulting in substantial modifications to river channel geometry.

Although the first tier of models exhibits better accuracy, their applicability is restricted to USGS gage sites due to the requirement for Q_{bnk} and Q_{mf} , which is only available for some streams in the CONUS. Consequently, the second tier of models, developed using Q_E derived from NHDPlusV2.1 and through a combined approach that incorporates both MLR and XGBR, are chosen as the final model to predict bankfull width, bankfull depth, mean-flow width, and mean-flow depth (Figure 7). Maps (Figure 7) are provided for reaches/streams with drainage areas greater than 100 km² to enhance visualization. However, the final predicted dataset resulting from this research encompasses values of predicted width and depth under both mean-flow and bankfull conditions for 2,642,259 reaches in NHDPlusV2.1.



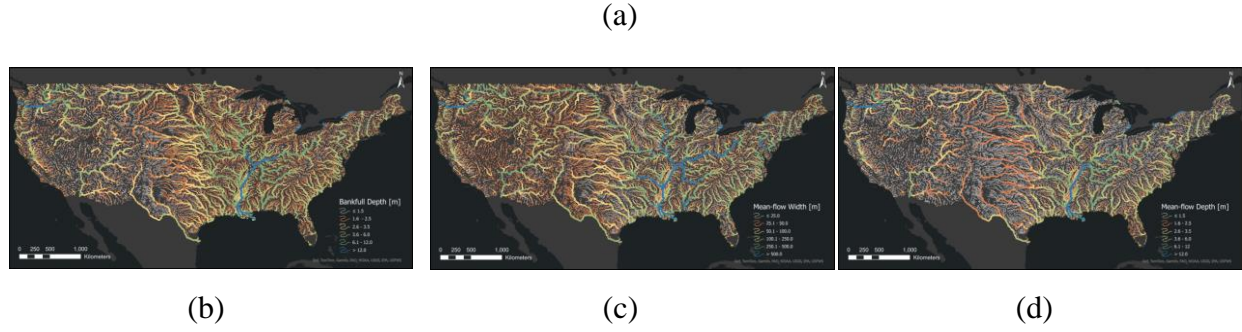


Figure 7. Maps of predicted values of (a) bankfull width, (b) bankfull depth, (c) mean-flow width, and (d) mean-flow depth over CONTiguous United States (CONUS) for reaches/streams in the National Hydrography Dataset Plus Version 2.1 (NHDplusv2.1) with drainage area greater than 100 km².

3.2. Independent Evaluations

3.2.1. Mean-flow Condition

The NHDplusV2.1 reach-scale channel geometry estimation (using the XGBR-MLR coupling) compared against data derived from bathymetry surveys (eHydro database) shows $R^2 = 0.32$ for depth and $R^2 = 0.84$ for width (Figure 8).

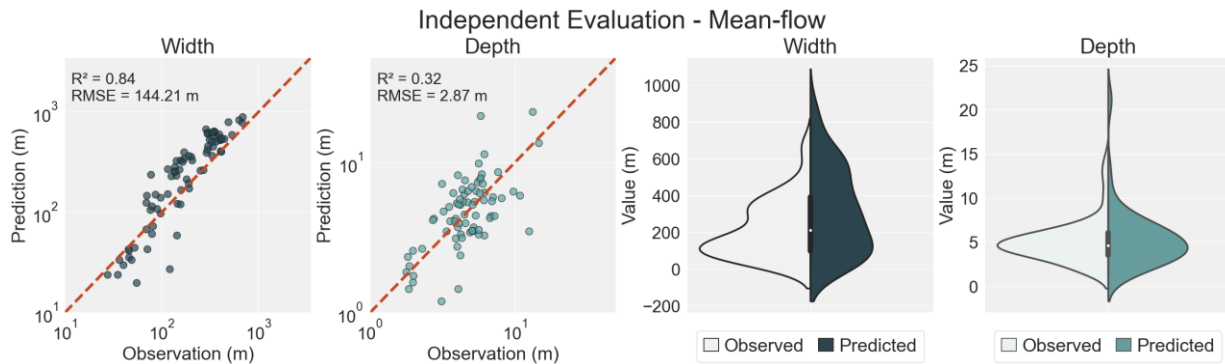


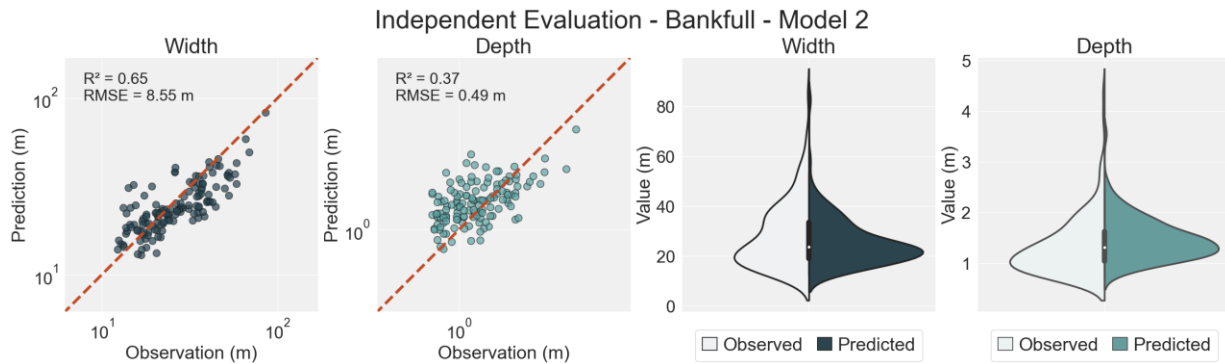
Figure 8. Scatter and violin plots of observations obtained by independent datasets (eHydro surveys) against predictions derived from the second tier of models (Model 2 (NHDplusV2.1 discharge)), utilizing a hybrid approach combining eXtreme Gradient Boosting Regression (XGBR) and Multi-Linear Regression (MLR) algorithms, for the mean-flow condition.

The less accurate results in the mean-flow condition can be attributed to several factors. First, eHydro surveys typically focus on the middle of the stream, which is accessible to navigable boats. This leads to a lack of coverage towards the banks and limits the surveys to

large channels, which creates a bias toward larger stream orders (between 4 and 9). Second, the spatial distribution of survey locations is concentrated predominantly east of the Mississippi River (Figure S2). Third, the surveyed lengths may not align with the corresponding NHDPlusV2.1 reach. Consequently, the extracted values from eHydro for a reach may only represent a portion of an NHDPlusV2.1 reach. Fourth, most surveys were conducted from 2017 to 2023, whereas the predictive models are based on data recorded until 2014. This up to nine-year difference may introduce a bias in the results, as the nature of rivers and their surrounding environments, which can influence river geometry, undergo substantial changes over time. Fifth, the surveys have not consistently been conducted during mean-flow conditions, potentially resulting in extracted values that do not accurately represent the channel geometry attributes at mean-flow conditions.

3.2.2. Bankfull Condition

The evaluation of NHDplusV2.1 reach-scale bankfull channel geometry estimation (implemented with the XGBR-MLR coupling) against data gathered from at-a-station bankfull observations at 11 diverse sources (Check Table 3) yielded a R^2 of 0.37 for depth and a R^2 of 0.65 for width (Figure 9).



(a)

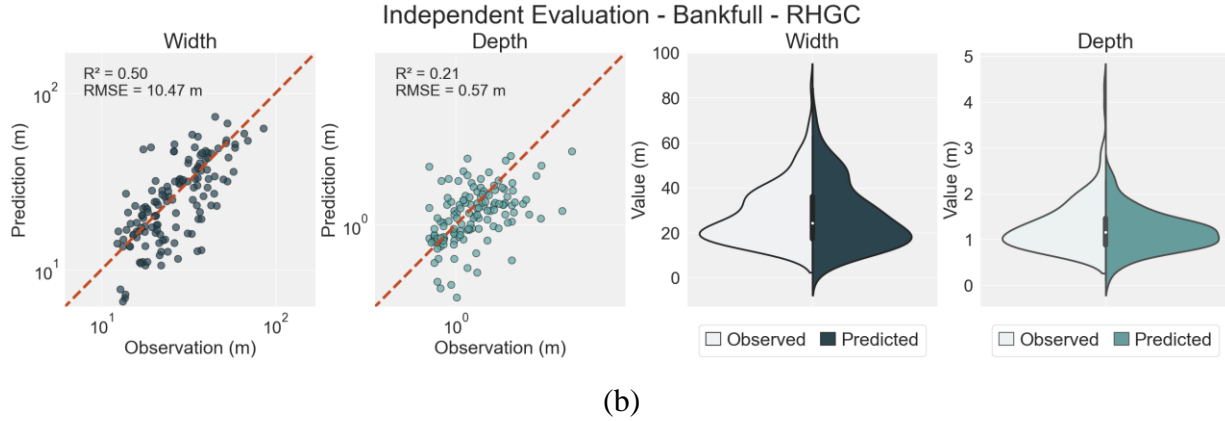


Figure 9. Scatter and violin plots of observations obtained by independent datasets (from 11 different sources) against (a) predictions obtained from applying the second tier of models (Model 2 (NHDplusV2.1 discharge)), utilizing a hybrid approach combining eXtreme Gradient Boosting Regression (XGBR) and Multi-Linear Regression (MLR) algorithms, for bankfull condition (b) predictions obtained by applying Regional Hydraulic Geometry Curves (RHGC) model.

The results for bankfull independent evaluation fall short of achieving very high accuracy. One contributing factor is that the independent evaluation dataset spans from 2000 to 2010. In contrast, the models were developed using measurements up to 2014 and reach and catchment attributes from 2011 to 2012. Additionally, discrepancies in the definition of bankfull condition may exist compared to our considerations. Also, the predicted values of bankfull attributes are reach-averaged while those considered observations come from at-a-station measurements, which are singular points rather than reach-averaged. Moreover, the observational dataset exclusively consists of smaller rivers, with a maximum width and depth of 68.99 m and 4.39 m, respectively. This falls within a range where we understand that the developed model may not offer precise predictions.

A comparison between Figure 9 (a) and Figure 9 (b) reveals a decrease in R^2 values for both width and depth, with width decreasing from 0.65 to 0.50 and depth decreasing from 0.37 to 0.21. This illustrates that the developed models demonstrate greater accuracy than the widely used RHGC method for predicting bankfull width and depth.

4. Conclusions

This research focuses on developing more accurate models for predicting channel width and depth under bankfull and mean-flow conditions. Flow discharge features (Q_{bnk} , Q_{mf} , and Q_E) emerge as the most significant parameters in the models developed, aligning with foundational river hydraulics principles that link flow discharge to the channel cross-sectional area. The primary models, incorporating ADCP-measured flow discharge features (Q_{bnk} and Q_{mf}), extracted from the HYDRoSWOT observational dataset through rigorous pre-processing, outperform secondary models that rely on derived mean annual flow from gage adjustment (Q_E) extracted from the NHDPlusV2.1 dataset. Additional hydraulic and catchment attributes beyond discharge and drainage area, such as elevation (Z), stream order (SO), and Aridity Index (AI), were shown to contribute significantly to the model's performance. The significant influence of these attributes underscores the complexity of river geometry spatial dynamics, affected by factors such as land cover and climate characteristics.

The XGBR algorithm stands out for its power in predicting attributes, showcasing superior accuracy, adeptness in handling non-linearity, and independence from data preprocessing. However, limitations arise when applying XGBR to the NHDPlusV2.1 reaches, with negative values returned for reaches beyond the training range. Consequently, a novel approach is proposed—a combination of MLR and XGBR as the final model—to address this limitation and enhance overall predictive capabilities.

An independent evaluation analysis was conducted to quantify the final model's predictive accuracy against datasets not associated with the assessed training and testing (HYDRoSWOT). By comparing the mean-flow geometry estimation with the reach-averaged channel geometry from eHydro surveys, we can assess how realistic our model (NHDplusV2.1) is when applied to reach-averaged data. The evaluation was challenging due to the limited quality of the datasets, which led to less accurate results. However, under bankfull conditions, the developed models performed better than the RHGC method, indicating improved prediction accuracy. Furthermore, width prediction consistently proves more accurate across all evaluations than depth. This discrepancy is attributed to the higher quality of the dataset used for width model development, as measuring river width is less controversial and complex than measuring river depth.

The outcomes of the applied developed models on NHDPlusV2.1 reaches are presented as a dataset and four maps. These data and maps are valuable resources for water-related experts, enabling further investigations to gain a deeper understanding of river channel evolution. These insights can significantly impact the development of water-related and river studies, including flood inundation mapping and modeling, river channel geomorphology, ecological investigations, and biological studies.

Acknowledgments

Funding for this project was provided by the National Oceanic & Atmospheric Administration (NOAA), awarded to the Cooperative Institute for Research to Operations in Hydrology (CIROH) through the NOAA Cooperative Agreement with The University of Alabama (NA22NWS4320003). We would like to thank Chance Jones for assisting with data collection in this project.

Open Research

The datasets utilized for model development are available in Canova et al., (2016), McKay et al., (2012), Wiczorek et al., (2018), Abeshu et al., (2022), and Trabucco & Zomer, (2019). The US Army Corps of Engineers eHydro survey database is accessible at <https://www.sam.usace.army.mil/Missions/Spatial-Data-Branch/eHydro/> and <https://www.arcgis.com/apps/dashboards/4b8f2ba307684cf597617bf1b6d2f85d>. The bankfull independent evaluation dataset is gathered from Mulvihill et al., (2009), Keaton et al., (2005), Dutnell, (2000), Moody et al., (2003), McCandless & Everett, (2002), Brockman, (2010), Lotspeich, (2009), Parola et al., (2007), Metcalf, (2004), Chase, (2004), and Mccandless, (2003). Multi-Linear Regression analysis has been done through JMP Pro®, Version 16.0.0. SAS Institute Inc., Cary, NC, 1989–2023. The generated datasets, maps, codes, and scripts are available on GitHub (link will be provided) and CIROH DocHub (link will be provided).

References

Abeshu, G. W., Li, H.-Y., Zhu, Z., Tan, Z., & Leung, L. R. (2022). Median bed-material sediment particle size across rivers in the contiguous US [Dataset]. *Earth System Science Data*, 14(2), 929–942. <https://doi.org/10.5194/essd-14-929-2022>

- Ames, D. P., Rafn, E. B., Van Kirk, R., & Crosby, B. (2009). Estimation of stream channel geometry in Idaho using GIS-derived watershed characteristics. *Environmental Modelling & Software*, 24(3), 444–448. <https://doi.org/10.1016/j.envsoft.2008.08.008>
- Andrews, E. D., & Nankervis, J. M. (1995). Effective Discharge and the Design of Channel Maintenance Flows for Gravel-Bed Rivers. In *Natural and Anthropogenic Influences in Fluvial Geomorphology* (pp. 151–164). American Geophysical Union (AGU). <https://doi.org/10.1029/GM089p0151>
- Bastola, H., & Diplas, P. (2023). Modeling Bankfull Channel Geometry Based on Watershed and Precipitation Characteristics Using Dimensionless Parameters. *Water Resources Research*, 59(6), e2022WR032688. <https://doi.org/10.1029/2022WR032688>
- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
- Bieger, K., Rathjens, H., Allen, P. M., & Arnold, J. G. (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(3), 842–858. <https://doi.org/10.1111/jawr.12282>
- Bjerklie, D. M., Fulton, J. W., Dingman, S. L., Canova, M. G., Minear, J. T., & Moramarco, T. (2020). Fundamental Hydraulics of Cross Sections in Natural Rivers: Preliminary Analysis of a Large Data Set of Acoustic Doppler Flow Measurements. *Water Resources Research*, 56(3), e2019WR025986. <https://doi.org/10.1029/2019WR025986>
- Blackburn-Lynch, W., Agouridis, C. T., & Barton, C. D. (2017). Development of Regional Curves for Hydrologic Landscape Regions (HLR) in the Contiguous United States. *JAWRA Journal of the American Water Resources Association*, 53(4), 903–928. <https://doi.org/10.1111/1752-1688.12540>
- Booker, D., & Woods, R. (2014). Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, 508, 227–239. <https://doi.org/10.1016/j.jhydrol.2013.11.007>

- 544 Boulesteix, A.-L., Janitza, S., Hapfelmeier, A., Van Steen, K., & Strobl, C. (2015). Letter to the
545 Editor: On the term ‘interaction’ and related phrases in the literature on Random Forests.
546 *Briefings in Bioinformatics*, 16(2), 338–345. <https://doi.org/10.1093/bib/bbu012>
- 547 Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest
548 methodology and practical guidance with emphasis on computational biology and
549 bioinformatics. *WIREs Data Mining and Knowledge Discovery*, 2(6), 493–507.
550 <https://doi.org/10.1002/widm.1072>
- 551 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
552 <https://doi.org/10.1023/A:1010933404324>
- 553 Brockman, R. R. (2010). *Hydraulic Geometry Relationships and Regional Curves for the Inner*
554 *and outer Bluegrass regions of Kentucky* [Dataset, PhD Thesis, University of Kentucky
555 Libraries].
556 [https://uknowledge.uky.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1055&context=](https://uknowledge.uky.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1055&context=gradschool_theses)
557 [gradschool_theses](https://uknowledge.uky.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1055&context=gradschool_theses)
- 558 Canova, M. G., Fulton, J. W., & Bjerklie, D. M. (2016). *USGS HYDRoacoustic dataset in*
559 *support of the Surface Water Oceanographic Topography satellite mission (HYDRoSWOT)*
560 [dataset]. [object Object]. <https://doi.org/10.5066/F7D798H6>
- 561 Castro, J. M., & Jackson, P. L. (2001). Bankfull Discharge Recurrence Intervals and Regional
562 Hydraulic Geometry Relationships: Patterns in the Pacific Northwest, Usa1. *JAWRA Journal of*
563 *the American Water Resources Association*, 37(5), 1249–1262. [https://doi.org/10.1111/j.1752-](https://doi.org/10.1111/j.1752-1688.2001.tb03636.x)
564 [1688.2001.tb03636.x](https://doi.org/10.1111/j.1752-1688.2001.tb03636.x)
- 565 Chaplin, J. J. (2005). *Development of Regional Curves Relating Bankfull-Channel Geometry and*
566 *Discharge to Drainage Area for Streams in Pennsylvania and Selected Areas of Maryland*.
567 <https://pubs.usgs.gov/sir/2005/5147/>
- 568 Charlton, R. (2007). *Fundamentals of Fluvial Geomorphology*. Routledge.
569 <https://doi.org/10.4324/9780203371084>

- Chase, K. J. (2004). *Channel-morphology data for the Tongue River and selected tributaries, southeastern Montana, 2001-02* [dataset]. <https://pubs.er.usgs.gov/publication/ofr20041260>
- 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*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Clerici, A., Perego, S., Chelli, A., & Tellini, C. (2015). Morphological changes of the floodplain reach of the Taro River (Northern Italy) in the last two centuries. *Journal of Hydrology*, 527, 1106–1122. <https://doi.org/10.1016/j.jhydrol.2015.05.063>
- Dey, S., Saksena, S., Winter, D., Merwade, V., & McMillan, S. (2022). Incorporating Network Scale River Bathymetry to Improve Characterization of Fluvial Processes in Flood Modeling. *Water Resources Research*, 58(11), e2020WR029521. <https://doi.org/10.1029/2020WR029521>
- Doyle, J. M., Hill, R. A., Leibowitz, S. G., & Ebersole, J. L. (2023). Random forest models to estimate bankfull and low flow channel widths and depths across the conterminous United States. *JAWRA Journal of the American Water Resources Association*, 59(5), 1099–1114. <https://doi.org/10.1111/1752-1688.13116>
- Dunne, T., & Leopold, L. B. (1978). *Water in Environmental Planning*. Macmillan.
- Dutnell, R. C. (2000). *Development of bankfull discharge and channel geometry relationships for natural channel design in Oklahoma using a fluvial geomorphic approach* [Dataset, PhD Thesis, University of Oklahoma]. <http://www.riverman-engineering.com/thesis-text.pdf>
- Gleason, C. J. (2015). Hydraulic geometry of natural rivers: A review and future directions. *Progress in Physical Geography: Earth and Environment*, 39(3), 337–360. <https://doi.org/10.1177/0309133314567584>
- Gochis, D., Barlage, M., Cabell, R., Casali, M., Dugger, A., FitzGerald, K., McAllister, M., McCreight, J., RafieeiNasab, A., Read, L., Sampson, K., Yates, D., & Zhang, Y. (2020). *The WRF-Hydro® modeling system technical description, (Version 5.1.1)*. <https://wrf-hydro.readthedocs.io/en/latest/>

- 596 Harrelson, C. C., Rawlins, C. L., & Potyondy, J. P. (1994). Stream channel reference sites: An
 597 illustrated guide to field technique. *Gen. Tech. Rep. RM-245. Fort Collins, CO: U.S. Department*
 598 *of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 61 p.,*
 599 *245.* <https://doi.org/10.2737/RM-GTR-245>
- 600 Heldmyer, A., Livneh, B., McCreight, J., Read, L., Kasprzyk, J., & Minear, T. (2022).
 601 Evaluation of a new observationally based channel parameterization for the National Water
 602 Model. *Hydrology and Earth System Sciences*, 26(23), 6121–6136. [https://doi.org/10.5194/hess-](https://doi.org/10.5194/hess-26-6121-2022)
 603 [26-6121-2022](https://doi.org/10.5194/hess-26-6121-2022)
- 604 Holder, R. L. (1986). Multiple Regression in Hydrology. *The Statistician*, 35(5), 566.
 605 <https://doi.org/10.2307/2987976>
- 606 J, O. F., A, T. M., & C.c, A. (2020). Multiple Linear Regression (MLR) Model: A Tool for
 607 Water Quality Interpretation. *Momona Ethiopian Journal of Science*, 12(1), Article 1.
 608 <https://doi.org/10.4314/mejs.v12i1.8>
- 609 Keast, D., & Ellison, J. C. (2022). Evaluation of bankfull stage from plotted channel geometries.
 610 *Journal of Hydrology: Regional Studies*, 41, 101052. <https://doi.org/10.1016/j.ejrh.2022.101052>
- 611 Keaton, J. N., Messinger, T., & Doheny, E. J. (2005). *Development and analysis of regional*
 612 *curves for streams in the non-urban valley and ridge physiographic province, Maryland,*
 613 *Virginia, and West Virginia (2005–5076) [dataset].* <https://doi.org/10.3133/sir20055076>
- 614 Krause, P., Boyle, D. P., & Båse, F. (2005). Comparison of different efficiency criteria for
 615 hydrological model assessment. *Advances in Geosciences*, 5, 89–97.
 616 <https://doi.org/10.5194/adgeo-5-89-2005>
- 617 Leopold, L. B., & Maddock Jr, T. (1953). The hydraulic geometry of stream channels and some
 618 physiographic implications. In *Professional Paper (252)*. U.S. Government Printing Office.
 619 <https://doi.org/10.3133/pp252>
- 620 Leopold, L. B., Wolman, M. G., & Miller, J. P. (1964). *Fluvial processes in geomorphology.*
 621 <https://pubs.usgs.gov/publication/70185663>

- 622 Lotspeich, R. R. (2009). *Regional curves of bankfull channel geometry for non-urban streams in*
623 *the Piedmont physiographic province, Virginia* [dataset]. US Geological Survey.
624 <https://pubs.usgs.gov/publication/sir20095206>
- 625 Marsden, R. F., & Ingram, R. G. (2004). Correcting for Beam Spread in Acoustic Doppler
626 Current Profiler Measurements. *Journal of Atmospheric and Oceanic Technology*, 21(9), 1491–
627 1498. [https://doi.org/10.1175/1520-0426\(2004\)021<1491:CFBSIA>2.0.CO;2](https://doi.org/10.1175/1520-0426(2004)021<1491:CFBSIA>2.0.CO;2)
- 628 McCandless, T. (2003). *Maryland Stream Survey: Bankfull Discharge and Channel*
629 *Characteristics in the Allegheny Plateau and the Valley and Ridge Hydrologic Region* [dataset].
630 https://www.researchgate.net/publication/237472244_Maryland_Stream_Survey_Bankfull_Disc
631 [harge_and_Channel_Characteristics_in_the_Allegheny_Plateau_and_the_Valley_and_Ridge_Hy](https://www.researchgate.net/publication/237472244_Maryland_Stream_Survey_Bankfull_Disc)
632 [drologic_Region/stats](https://www.researchgate.net/publication/237472244_Maryland_Stream_Survey_Bankfull_Disc)
- 633 McCandless, T. L., & Everett, R. A. (2002). *Bankfull Discharge and Channel Characteristics of*
634 *Streams in the Piedmont Hydrologic Region* [dataset].
635 <https://corpora.tika.apache.org/base/docs/govdocs1/308/308564.pdf>
- 636 McKay, L. D., Timothy R. Bondelid, Thomas G. Dewald, Craig M. Johnston, Richard B. Moore,
637 & Alan H. Rea. (2012). *NHDPlus Version 2: User Guide* [dataset].
638 https://www.epa.gov/system/files/documents/2023-04/NHDPlusV2_User_Guide.pdf
- 639 Metcalf, C. (2004). *Regional channel characteristics for maintaining natural fluvial*
640 *geomorphology in Florida streams* [dataset]. <https://trid.trb.org/View/748355>
- 641 Monegaglia, F., & Tubino, M. (2019). The Hydraulic Geometry of Evolving Meandering Rivers.
642 *Journal of Geophysical Research: Earth Surface*, 124(11), 2723–2748.
643 <https://doi.org/10.1029/2019JF005309>
- 644 Moody, T., Wirtanen, M., & Yard, S. N. (2003). *Regional relationships for bankfull stage in*
645 *natural channels of the arid southwest* [dataset].
646 https://efotg.sc.egov.usda.gov/references/Public/AZ/Arid_SW_Report_Regional_Curves.pdf

- 647 Mulvihill, C. I., & Baldigo, B. P. (2012). Optimizing Bankfull Discharge and Hydraulic
 648 Geometry Relations for Streams in New York State1. *JAWRA Journal of the American Water*
 649 *Resources Association*, 48(3), 449–463. <https://doi.org/10.1111/j.1752-1688.2011.00623.x>
- 650 Mulvihill, C. I., Baldigo, B. P., Miller, S. J., DeKoskie, D., & DuBois, J. (2009). Bankfull
 651 discharge and channel characteristics of streams in New York State [dataset]. In *Scientific*
 652 *Investigations Report* (2009–5144). U.S. Geological Survey. <https://doi.org/10.3133/sir20095144>
- 653 Naito, K., & Parker, G. (2019). Can Bankfull Discharge and Bankfull Channel Characteristics of
 654 an Alluvial Meandering River be Cospecified From a Flow Duration Curve? *Journal of*
 655 *Geophysical Research: Earth Surface*, 124(10), 2381–2401.
 656 <https://doi.org/10.1029/2018JF004971>
- 657 Naito, K., & Parker, G. (2020). Adjustment of self-formed bankfull channel geometry of
 658 meandering rivers: Modelling study. *Earth Surface Processes and Landforms*, 45(13), 3313–
 659 3322. <https://doi.org/10.1002/esp.4966>
- 660 Navratil, O., Albert, M.-B., Hérouin, E., & Gresillon, J.-M. (2006). Determination of bankfull
 661 discharge magnitude and frequency: Comparison of methods on 16 gravel-bed river reaches.
 662 *Earth Surface Processes and Landforms*, 31(11), 1345–1363. <https://doi.org/10.1002/esp.1337>
- 663 Neal, J. C., Odoni, N. A., Trigg, M. A., Freer, J. E., Garcia-Pintado, J., Mason, D. C., Wood, M.,
 664 & Bates, P. D. (2015). Efficient incorporation of channel cross-section geometry uncertainty into
 665 regional and global scale flood inundation models. *Journal of Hydrology*, 529, 169–183.
 666 <https://doi.org/10.1016/j.jhydrol.2015.07.026>
- 667 Nguyen, D. H., Le, X. H., Heo, J.-Y., & Bae, D.-H. (2021). Development of an Extreme
 668 Gradient Boosting Model Integrated With Evolutionary Algorithms for Hourly Water Level
 669 Prediction. *IEEE Access*, PP, 1–1. <https://doi.org/10.1109/ACCESS.2021.3111287>
- 670 Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., & Liu, J. (2020). Streamflow forecasting
 671 using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of*
 672 *Hydrology*, 586, 124901. <https://doi.org/10.1016/j.jhydrol.2020.124901>

- Orlandini, S., & Rosso, R. (1998). Parameterization of stream channel geometry in the distributed modeling of catchment dynamics. *Water Resources Research*, 34(8), 1971–1985. <https://doi.org/10.1029/98WR00257>
- Parola, A. C., Vesely, W. S., Croasdaile, M. A., Hansen, C., & Jones, M. S. (2007). *Geomorphic characteristics of streams in the Bluegrass physiographic region of Kentucky* [dataset]. <https://eec.ky.gov/Environmental-Protection/Water/Reports/Reports/NPS0010-Bluegrass.pdf>
- Radecki-Pawlik, A. (2002). Bankfull discharge in mountain streams: Theory and practice. *Earth Surface Processes and Landforms*, 27(2), 115–123. <https://doi.org/10.1002/esp.259>
- Rice, S. P., Greenwood, M. T., & Joyce, C. B. (2001). Tributaries, sediment sources, and the longitudinal organisation of macroinvertebrate fauna along river systems. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(4), 824–840. <https://doi.org/10.1139/f01-022>
- Rosgen, D. L. (1994). A classification of natural rivers. *CATENA*, 22(3), 169–199. [https://doi.org/10.1016/0341-8162\(94\)90001-9](https://doi.org/10.1016/0341-8162(94)90001-9)
- Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7), 2611–2628. <https://doi.org/10.5194/hess-20-2611-2016>
- Sobotka, M. J., & Phelps, Q. E. (2017). A Comparison of Main and Side Channel Physical and Water Quality Metrics and Habitat Complexity in the Middle Mississippi River. *River Research and Applications*, 33(6), 879–888. <https://doi.org/10.1002/rra.3061>
- Sweet, W. V., & Geratz, J. W. (2003). Bankfull Hydraulic Geometry Relationships and Recurrence Intervals for North Carolina’s Coastal Plain1. *JAWRA Journal of the American Water Resources Association*, 39(4), 861–871. <https://doi.org/10.1111/j.1752-1688.2003.tb04411.x>
- Thoms, M. (2003). Floodplain-river ecosystems: Lateral connections and the implications of human interference. *Geomorphology*, 56, 335–349. [https://doi.org/10.1016/S0169-555X\(03\)00160-0](https://doi.org/10.1016/S0169-555X(03)00160-0)

- Trabucco, A., & Zomer, R. J. (2019). *Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2* [dataset]. figshare. <https://doi.org/10.6084/m9.figshare.7504448.v3>
- Walling, D. E., & Webb, B. W. (1975). Spatial Variation of River Water Quality: A Survey of the River Exe. *Transactions of the Institute of British Geographers*, 65, 155–171. <https://doi.org/10.2307/621615>
- Wieczorek, M. E., Jackson, S. E., & 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. 4.0, August 2023)* [dataset]. [object Object]. <https://doi.org/10.5066/F7765D7V>
- Wilby, R., & Gibert, J. (1996). Hydrological and hydrochemical dynamics. In G. E. Petts & C. Amoros (Eds.), *The Fluvial Hydrosystems* (pp. 37–67). Springer Netherlands. https://doi.org/10.1007/978-94-009-1491-9_3
- Williams, G. P. (1978). Bank-full discharge of rivers. *Water Resources Research*, 14(6), 1141–1154. <https://doi.org/10.1029/WR014i006p01141>
- Wolman, M. G., & Leopold, L. B. (1957). River flood plains: Some observations on their formation. In *Professional Paper* (282–C; pp. 87–109). U.S. Government Printing Office. <https://doi.org/10.3133/pp282C>
- Wolock, D. M., Winter, T. C., & McMahon, G. (2004). Delineation and Evaluation of Hydrologic-Landscape Regions in the United States Using Geographic Information System Tools and Multivariate Statistical Analyses. *Environmental Management*, 34(1), S71–S88. <https://doi.org/10.1007/s00267-003-5077-9>
- Worland, S. C., Farmer, W. H., & Kiang, J. E. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, 101, 169–182. <https://doi.org/10.1016/j.envsoft.2017.12.021>
- Zakaria, M. N. A., Ahmed, A. N., Abdul Malek, M., Birima, A. H., Hayet Khan, M. M., Sherif, M., & Elshafie, A. (2023). Exploring machine learning algorithms for accurate water level

forecasting in Muda river, Malaysia. *Heliyon*, 9(7), e17689.

<https://doi.org/10.1016/j.heliyon.2023.e17689>

Zhou, Z., Tuo, Y., Li, J., Chen, M., Wang, H., Zhu, L., & Deng, Y. (2022). Effects of hydrology and river characteristics on riverine wetland morphology variation in the middle reaches of the Yarlung Zangbo–Brahmaputra river based on remote sensing. *Journal of Hydrology*, 607, 127497. <https://doi.org/10.1016/j.jhydrol.2022.127497>

Ziegler, A., & König, I. R. (2014). Mining data with random forests: Current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1), 55–63. <https://doi.org/10.1002/widm.1114>