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 Erfani², and Sagy Cohen³

¹University of Alabama
²University of South Carolina
³University of Alabama, Tuscaloosa

May 30, 2024

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.

Hosted file


Hosted file

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 Erfani²,³, 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.

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 burned 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 \( q_{\text{va}} \), the measured discharge value \( \text{meas}_q_{\text{va}} \), and the calculated discharge derived from the cross-sectional area multiplied by the mean velocity \( q_2_{\text{xsec}_\text{area}_X\text{mean}_\text{vel}_\text{va}} \). 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 \textit{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 \( \text{IQR} \) method. By applying this method to the \textit{HYDRoSWOT\_init} dataset, an upper limit for channel width is
established. This limit is defined as $Q_3 + 1.5 \times \text{IQR}$, 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).](image-url)
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 (~ 1km 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.

<table>
<thead>
<tr>
<th>Source</th>
<th>Attribute name</th>
<th>Attribute description</th>
<th>Unit</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site no</td>
<td>USGS site number</td>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lat</td>
<td>Decimal latitude</td>
<td>Degrees</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Long</td>
<td>Decimal longitude</td>
<td>Degrees</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q_{bnk}</td>
<td>Bankfull flow discharge</td>
<td>m³/s</td>
<td>0.41</td>
<td>33,195.4</td>
<td>329.52</td>
<td>1,670.38</td>
<td></td>
</tr>
<tr>
<td>d_{bnk}</td>
<td>Bankfull depth</td>
<td>m</td>
<td>0.30</td>
<td>27.9</td>
<td>2.54</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>w_{bnk}</td>
<td>Bankfull width</td>
<td>m</td>
<td>3.78</td>
<td>1,816.6</td>
<td>63.62</td>
<td>93.2</td>
<td></td>
</tr>
<tr>
<td>Q_{mf}</td>
<td>Mean-flow flow discharge</td>
<td>m³/s</td>
<td>0.02</td>
<td>18,228.6</td>
<td>113.65</td>
<td>778.65</td>
<td></td>
</tr>
<tr>
<td>d_{mf}</td>
<td>Mean-flow depth</td>
<td>m</td>
<td>0.19</td>
<td>27.9</td>
<td>1.57</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td>w_{mf}</td>
<td>Mean-flow width</td>
<td>m</td>
<td>3.35</td>
<td>2,124.5</td>
<td>55.08</td>
<td>90.54</td>
<td></td>
</tr>
<tr>
<td>SO</td>
<td>Modified Strahler stream order</td>
<td></td>
<td>–</td>
<td>1</td>
<td>10</td>
<td>4.79</td>
<td>1.35</td>
</tr>
<tr>
<td>DA</td>
<td>Total upstream catchment area from the downstream end of the flowline</td>
<td>km²</td>
<td>4.34</td>
<td>2,881,390</td>
<td>19,110.9</td>
<td>138,658</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Smoothed minimum elevation</td>
<td>cm</td>
<td>3</td>
<td>269,497</td>
<td>24,842.3</td>
<td>32,354.4</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Slope of flowline based on smoothed elevations</td>
<td>m/m</td>
<td>0.00001</td>
<td>0.08803</td>
<td>0.0018</td>
<td>0.00426</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>Units</td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
<td>Standard Deviation</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>--------------------</td>
<td></td>
</tr>
<tr>
<td>$Q_E$</td>
<td>Mean annual flow from gage adjustment/Best EROM estimate of actual mean-flow</td>
<td>$m^3/s$</td>
<td>0.00017</td>
<td>19.022.01</td>
<td>109.32</td>
<td>739.43</td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>Accumulated number of dams built on or before 2010 based on total upstream accumulation</td>
<td>Count</td>
<td>1</td>
<td>41,971</td>
<td>248.56</td>
<td>1,955.87</td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>Catchment population density from U.S. block-level population density rasters for 2010</td>
<td>Persons $/km^2$</td>
<td>0.01</td>
<td>4,478.56</td>
<td>215.47</td>
<td>442.14</td>
<td></td>
</tr>
<tr>
<td>$EVI_{fa}$</td>
<td>Vegetation Index value for the fall season 2011 (OND)</td>
<td>—</td>
<td>0.01</td>
<td>0.43</td>
<td>0.24</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>$EVI_{wi}$</td>
<td>Vegetation Index value for the winter season 2012 (JFM)</td>
<td>—</td>
<td>0.01</td>
<td>0.44</td>
<td>0.19</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>$EVI_{sp}$</td>
<td>Vegetation Index value for the spring season 2012 (AMJ)</td>
<td>—</td>
<td>0.02</td>
<td>0.62</td>
<td>0.39</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>$EVI_{su}$</td>
<td>Vegetation Index value for the summer season 2012 (JAS)</td>
<td>—</td>
<td>0.02</td>
<td>0.61</td>
<td>0.41</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>$Cl$</td>
<td>Catchment average percent of clay</td>
<td>%</td>
<td>2.13</td>
<td>68.36</td>
<td>23.31</td>
<td>11.24</td>
<td></td>
</tr>
<tr>
<td>$Sl$</td>
<td>Catchment average percent of silt</td>
<td>%</td>
<td>4.13</td>
<td>77.24</td>
<td>43.44</td>
<td>13.63</td>
<td></td>
</tr>
<tr>
<td>$Sa$</td>
<td>Catchment average percent of sand</td>
<td>%</td>
<td>3.04</td>
<td>92.80</td>
<td>33.25</td>
<td>19.86</td>
<td></td>
</tr>
<tr>
<td>$Dv$</td>
<td>Estimated percent of catchment that contains the land-use and land-cover type developed</td>
<td>%</td>
<td>0.03</td>
<td>99.92</td>
<td>22.94</td>
<td>25.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Estimated percent of catchment that contains the land-use and land-cover type</td>
<td>%</td>
<td>0.01</td>
<td>96.85</td>
<td>29.78</td>
<td>25.15</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>$Fr$</td>
<td>forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Ag$</td>
<td>Estimated percent of catchment that contains the land-use and land-cover type agriculture</td>
<td>%</td>
<td>0.01</td>
<td>95.31</td>
<td>25.68</td>
<td>24.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Abeshu et al., 2022)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$D_{50}$ Median sediment particle size</td>
<td>mm</td>
<td>0.029</td>
<td>89.275</td>
<td>1.349</td>
<td>3.152</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Trabucco &amp; Zomer, 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$AI$ Mean Aridity Index</td>
<td>–</td>
<td>0.07</td>
<td>2.51</td>
<td>0.79</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>
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.

![Spatial distribution map of the training and testing datasets utilized for model development.](image)

**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
The models are initially developed and validated using the dataset extracted from the HYDROSWOT dataset at selected USGS gauge sites. However, the filtering and the bankfull and mean-flow width and depth identification procedure may have introduced systematic biases resulting in reduced accuracy of the model’s predictions. These developed models are utilized to predict reach-average channel geometry parameters for constructing the CONUS-scale database (utilizing NHDplusv2.1 in this study). To ensure its applicability, we conduct an evaluation for locations that were not included in either the training or testing datasets, referred to here as 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/](https://www.sam.usace.army.mil/Missions/Spatial-Data-Branch/eHydro/) and [https://www.arcgis.com/apps/dashboards/4b8f2ba307684cf597617bf1b6d2f85d](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.

<table>
<thead>
<tr>
<th>River names</th>
<th>Number of Reaches</th>
<th>( w_{\text{mean}} )</th>
<th>( d_{\text{mean}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio</td>
<td>82</td>
<td>Min 76.37</td>
<td>Max 781.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 372.73</td>
<td>Std 118.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Min 3.64</td>
<td>Max 12.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 6.99</td>
<td>Std 2.24</td>
</tr>
<tr>
<td>Arkansas</td>
<td>77</td>
<td>Min 34.79</td>
<td>Max 645.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 370.64</td>
<td>Std 93.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Min 2.02</td>
<td>Max 10.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 5.61</td>
<td>Std 1.99</td>
</tr>
<tr>
<td>Monongahela</td>
<td>36</td>
<td>Min 59.99</td>
<td>Max 305.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 172.60</td>
<td>Std 40.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Min 3.45</td>
<td>Max 8.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 5.51</td>
<td>Std 1.23</td>
</tr>
<tr>
<td>Illinois</td>
<td>22</td>
<td>Min 28.75</td>
<td>Max 268.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 148.18</td>
<td>Std 55.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Min 2.24</td>
<td>Max 4.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean 3.26</td>
<td>Std 0.57</td>
</tr>
</tbody>
</table>
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).
Table 3. Summary of independent evaluation dataset (gathered from 11 different sources) for bankfull condition including descriptive statistics.

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Reaches</th>
<th>$w_{bnk}$ Min</th>
<th>$w_{bnk}$ Max</th>
<th>$w_{bnk}$ Mean</th>
<th>$w_{bnk}$ Std</th>
<th>$d_{bnk}$ Min</th>
<th>$d_{bnk}$ Max</th>
<th>$d_{bnk}$ Mean</th>
<th>$d_{bnk}$ Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Mulvihill et al., 2009)</td>
<td>31</td>
<td>14.68</td>
<td>68.99</td>
<td>30.8</td>
<td>15.36</td>
<td>0.73</td>
<td>2.79</td>
<td>1.25</td>
<td>0.5</td>
</tr>
<tr>
<td>(Keaton et al., 2005)</td>
<td>21</td>
<td>13.35</td>
<td>40.84</td>
<td>27.92</td>
<td>8.22</td>
<td>0.76</td>
<td>1.62</td>
<td>1.23</td>
<td>0.27</td>
</tr>
<tr>
<td>(Dutnell, 2000)</td>
<td>20</td>
<td>12.85</td>
<td>85.1</td>
<td>37.06</td>
<td>19.07</td>
<td>0.71</td>
<td>4.39</td>
<td>1.71</td>
<td>0.85</td>
</tr>
<tr>
<td>(Moody et al., 2003)</td>
<td>19</td>
<td>13.66</td>
<td>52.12</td>
<td>29.16</td>
<td>9.63</td>
<td>0.73</td>
<td>1.43</td>
<td>1.04</td>
<td>0.21</td>
</tr>
<tr>
<td>(McCandless &amp; Everett, 2002)</td>
<td>11</td>
<td>12.31</td>
<td>26.27</td>
<td>19.34</td>
<td>4.22</td>
<td>0.79</td>
<td>1.83</td>
<td>1.32</td>
<td>0.31</td>
</tr>
<tr>
<td>(Brockman, 2010)</td>
<td>10</td>
<td>13.65</td>
<td>35.86</td>
<td>21.35</td>
<td>6.02</td>
<td>0.71</td>
<td>1.88</td>
<td>1.01</td>
<td>0.32</td>
</tr>
<tr>
<td>(Lotspeich, 2009)</td>
<td>6</td>
<td>13.81</td>
<td>41.15</td>
<td>24.51</td>
<td>9.6</td>
<td>0.76</td>
<td>2.04</td>
<td>1.35</td>
<td>0.52</td>
</tr>
<tr>
<td>(Parola et al., 2007)</td>
<td>6</td>
<td>17.83</td>
<td>37.49</td>
<td>25.42</td>
<td>6.23</td>
<td>1.17</td>
<td>3.89</td>
<td>2.32</td>
<td>0.92</td>
</tr>
<tr>
<td>(Metcalf, 2004)</td>
<td>5</td>
<td>14.17</td>
<td>40.63</td>
<td>19.79</td>
<td>10.42</td>
<td>1.37</td>
<td>2.44</td>
<td>1.79</td>
<td>0.39</td>
</tr>
<tr>
<td>(Chase, 2004)</td>
<td>4</td>
<td>34.14</td>
<td>44.81</td>
<td>39.32</td>
<td>4.52</td>
<td>0.91</td>
<td>1.22</td>
<td>1.1</td>
<td>0.12</td>
</tr>
<tr>
<td>(McCandless, 2003)</td>
<td>3</td>
<td>19.42</td>
<td>38.34</td>
<td>26.28</td>
<td>8.55</td>
<td>0.85</td>
<td>0.98</td>
<td>0.91</td>
<td>0.05</td>
</tr>
</tbody>
</table>
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 R² 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 R² 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 R² 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 R² 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):

\[
\begin{align*}
w_{bnk, \text{model1}} &= 5.36 Q_{mf}^{0.29} DA^{0.18} AI^{0.31} D50^{0.03} Agr^{-0.02} SI^{-0.08} \\
w_{bnk, \text{model2}} &= 11.58 Q_{mf}^{0.35} EVI_{fa}^{-0.27} ND^{0.03} AI^{0.17} Fr^{-0.02} \\
d_{bnk, \text{model1}} &= 3.53 Q_{bnk}^{0.31} PD^{-0.02} Z^{-0.09} S^{-0.03} Fr^{-0.03} Ag^{0.02} Si^{-0.17} Sa^{-0.15} \\
d_{bnk, \text{model2}} &= 1.76 EVI_{wi}^{-0.13} DA^{0.19} AI^{0.28} Z^{-0.08} S^{-0.04} Fr^{-0.02} Si^{-0.09} Sa^{-0.18}
\end{align*}
\]
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 (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.

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 (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.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Observation</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
</tr>
<tr>
<td>Bankfull Width (m)</td>
<td>5.2</td>
<td>987.6</td>
<td>61.6</td>
</tr>
<tr>
<td></td>
<td>(46.3%)</td>
<td>(0.5%)</td>
<td>(-5.6%)</td>
</tr>
<tr>
<td>Bankfull Depth (m)</td>
<td>0.3</td>
<td>14.7</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>(88.2%)</td>
<td>(-0.9%)</td>
<td>(2.6%)</td>
</tr>
<tr>
<td>Mean-flow Width (m)</td>
<td>4.7</td>
<td>957.1</td>
<td>53.4</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(47.0%)</td>
<td>(0.8%)</td>
<td>(-0.5%)</td>
</tr>
<tr>
<td>Mean-flow Depth (m)</td>
<td>0.2</td>
<td>12.8</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>(118.0%)</td>
<td>(8.0%)</td>
<td>(5.6%)</td>
</tr>
</tbody>
</table>
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 (HYDRoSWOT) 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 \( Q_{\text{bnk}} \) and \( Q_{\text{mf}} \), 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 km$^2$.

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 combing 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).
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 combing 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**


**References**


https://doi.org/10.1029/GM089p0151


https://doi.org/10.1007/s11749-016-0481-7


https://doi.org/10.1111/jawr.12282


https://doi.org/10.1111/1752-1688.12540


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


forecasting in Muda river, Malaysia. *Heliyon, 9*(7), e17689.
https://doi.org/10.1016/j.heliyon.2023.e17689
