Constraining Remote River Discharge Estimation Using Reach-Scale Geomorphology

C.B. Brinkerhoff1\*, C.J. Gleason1, D. Feng1, P. Lin2

1 Department of Civil & Environmental Engineering, University of Massachusetts, Amherst MA

2 Department of Civil & Environmental Engineering, Princeton University, Princeton NJ

\*Corresponding author: Craig Brinkerhoff (cbrinkerhoff@umass.edu)

Key Points:

* Remote sensing of river discharge is highly sensitive to the quality of prior information the algorithms receive
* Discharge prediction accuracy is substantially increased by improving the quality of priors, tested on both SWOT-simulated data and Landsat imagery for two algorithms
* Statistical clustering techniques can be used to better make sense of river geomorphology and further improve algorithm performance

**Keywords:** remote sensing, river discharge, fluvial geomorphology, open-channel flow, McFLI

Abstract

Remote sensing of river discharge is presently possible in any river on Earth, but the algorithms used to do rely upon prior understandings of river geomorphology and hydraulics. Our current methods for estimating prior knowledge on rivers is not necessarily reflective of the physics experienced in a river. To address this problem, we studied the sensitivity of algorithm performance due to the quality of the prior information received by two algorithms capable of estimating discharge in uncalibrated settings: Bayesian-AMHG-Manning’s (BAM- Hagemann et al. 2017). We trained new models for estimating prior river knowledge on a dataset of over 370,000 field hydraulics measurements and statistically clustered geomorphology in order to prescribe prior river knowledge curated specifically for the river at hand, using only remotely-sensed observations. These interventions were tested by running BAM on the entire Mackenzie River basin, where median relative improvement in Nash-Sutcliffe Efficiency (NSE) was 78%; we also ran BAM and MetroMan on NASA Surface Water and Ocean Topography (SWOT)-simulated rivers and median NSE across rivers (for BAM) improved by 9%. For BAM, almost all improvement came in high flow events, as opposed to baseflow. These findings are implementable in any related algorithm and will play a vital role in both current attempts at global remotely-sensed river discharge as well as future discharge estimates from SWOT, where prior knowledge is essential to estimating flow in ungauged rivers where nothing is known *a priori*.

1 Introduction

In recent decades, remote sensing (RS) of rivers has flourished as a sub-field within fluvial geomorphology and hydrology. At the global scale, RS of rivers is changing current perceptions of rivers and their role in the earth system: there exists globally modeled hydrography at fine-spatial scales (Yamazaki et al. 2019; Lehner et al. 2008), daily runoff routed through almost 3 million river reaches over 30 years (Lin et al. 2019), assessments of rivers and climate (Yang et al. 2020), water quality (Ross et al. 2019), surface area (Allen & Pavelsky, 2018), and hydrological connectivity (Grill et al. 2019). These examples, along with similar recent work quantifying global fluvial geomorphic patterns (Chen et al. 2019; Frasson et al. 2019), suggest that RS is coming of age in its ability to provide global scale data that honors local differences in rivers. These ideas will be further explored with the launch of the Surface Water and Ocean Topography (SWOT) satellite in 2021, which is expected to provide measurements of water surface elevation and extent at unprecedented spatial and temporal resolutions (Biancamaria et al. 2016).

A particular subset of this literature is showing that global RS of river discharge (RSQ) is presently possible with some gauging information in hand and should be globally possible in ungauged basins in the near future (Gleason and Durand, *in review*). In basins with stream gauges or extensive field-measurements, RSQ approaches calibrate RS to local channel hydraulics (e.g. Brackenridge et al. 2007; LeFavour & Alsdorf 2005; Pavelsky, 2014; Pavelsky & Smith, 2009; Tarpanelli et al. 2013) or introduce RS data into hydrologic or hydraulic models (e.g. Bjerklie et al. 2005; Chandanpurker et al., 2017; King et al. 2018; Lin et al. 2019; Neal et al. 2009; Silvestro et al., 2015; Siquera et al., 2018; Zhang et al., 2016). These approaches (i.e., merging *in situ* and RS data) yield good predictive accuracy and can extend existing gauge records in space and time. In ungauged settings however, there are no gauge records to extend. Ground-based knowledge would improve RSQ accuracy in these scenarios, but in lieu of such information these methods must produce reasonably accurate results without relying on *in situ* knowledge (Gleason & Durand, *in review*). This makes ungauged RSQ particularly attractive for global applications. In ungauged settings, standard practice is again to introduce RS data into hydrologic models (e.g. Emery et al 2018; Sun et al. 2015) or hydraulic models (e.g. Andreadis et al. 2007; Biancamaria et al. 2011; Durand et al. 2008; Yoon et al. 2012). The most recent and sophisticated methods for assimilating RS into hydraulic models (Larnier et al. 2019; Oubanas et al. 2018a; b) are highly accurate in ungauged settings but computationally burdensome for global application (Gleason & Durand, *in review*).

A recent branch of RSQ has emerged with global application, SWOT, and ungauged basins in mind. This approach is termed Mass Conserved Flow-Law Inversion or McFLI (Gleason et al., 2017). McFLIs assume a river reach is mass conserved and then inversely solve for the unknown parameters in a flow law given some set of RS observations. This means that no hydrologic or hydraulic model is necessary, and that *Q* is exclusively estimated from RS by inverting basic geomorphic theories. McFLIs are therefore defined by their flow laws. To date, all published McFLIs have used either Manning’s equation (Bjerklie et al. 2018; Durand et al. 2014; Garambois & Monnier 2015; Hagemann et al. 2017; Sichangi et al. 2018) or at-many-stations hydraulic geometry (AMHG; Gleason et al. 2014; Hagemann et al. 2017; Feng et al., 2019) as a flow law, where AMHG reflects relationships between at-a-station hydraulic geometry (AHG) parameters along a river’s course (Gleason & Smith, 2014).

At the core of McFLI inversion is a reliance on initial guesses for parameters not observable from RS, termed ‘priors’ in Bayesian parlance. For example, to invert Manning’s equation priors are generally needed for discharge, channel roughness, and channel cross-sectional area. Priors have previously been estimated from global hydrologic model output (Durand et al. 2016; Bonnema et al., 2016; Feng et al., 2019) and/or from external training data of geomorphic and hydraulic variables (e.g. Canova et al. 2016; Hagemann et al., 2017). Priors take the form of a probability distribution of these RS-unobservable parameters. If *in situ* data are available, then priors have extremely low variance: we know, for example, channel roughness a priori. The less certain we are about a parameter a priori, the wider the distribution. Durand et al. (2016) found that McFLIs are sensitive to their priors in a test of five McFLIs on simulated SWOT observations (as SWOT has not launched, McFLIs are tested on ‘SWOT-like’ simulated data). In a similar comparison of algorithms using simulated rivers, Bonnema et al. (2016) found that AMHG inversion is particularly sensitive to its priors, and Tuozzolo et al. (2019) found that McFLI estimation bias is sensitive to the prior on discharge in the first test of McFLIs on real-world SWOT-comparable altimetry data. This is logical- the more we know about a river, the better we can invert discharge.

These findings indicate that priors play a pivotal role in McFLI discharge accuracy, yet despite the geomorphic foundations of the McFLI paradigm, present McFLIs use the same priors for every river on earth, regardless of differences in planform geometry, hydraulics, and river size. This means that McFLIs use the same expectations for, e.g., Manning’s *n*, width/depth ratios, and AHG exponents in a braided river and a canal. No study to date has explicitly explored the sensitivity of McFLIs to the *quality* of their priors, where ‘quality’ refers to the hydraulic and geomorphic representativeness of a prior for a given river. High-quality priors would be accurate, river-specific knowledge that closely approximates field measurements and contextualizes discharge inversion to the specific hydraulics of the river.

Global RS of hydrology has advanced to the point where this information is available at the global scale, but there are also troves of existing prior data that can be mapped onto these RS products. The United States Geological Survey (USGS) makes periodic field measurements of discharge and other hydraulics to calibrate the rating curves for their stream gauges, and all of these are freely available. These measurements are easily joined to existing hydrographic datasets, thereby providing reach-scale geomorphic attributes for every hydraulic measurement and dramatically expanding the scope of the data. For example, Brinkerhoff et al. (2019) joined over 730,000 of these measurements to USGS hydrography (USGS, 2019), building on earlier work (HYDRoSWOT- Bjerklie et al. 2020; Canova, et al. 2016). This recently published dataset comprises a subset of the measurements ultimately used by Brinkerhoff, et al. (2019), designed specifically for providing channel hydraulics for SWOT-related research. In the McFLI context, datasets of *in situ* measurements like HYDRoSWOT have been mapped to RS observations by Hagemann, et al. (2017). They trained simple regression models on HYDRoSWOT such that hydraulic priors could be predicted using just river width. A potential alternative to these models that addresses geomorphic differences in rivers is statistical classification. Distinct river classes, or types, should exhibit fundamentally different hydraulics that can be assigned to rivers using RS observations as predictors. By assigning representative values from the *in situ* datasets to each river type, global hydraulic priors are reduced to look-up tables of ballpark hydraulic estimates for each river type. However, scaling these river types to every river in a network poses a significant challenge (e.g. Guillon et al. 2020).

Merging global RS of rivers, existing high quality *in situ* prior data, and McFLI thus seems a fruitful way to improve RSQ. We hypothesize that McFLI performance can be improved by acknowledging geomorphic differences between rivers and assigning different priors to different rivers, building on recent global RS of rivers and decades of detailed *in situ* work mapped to the global scale. Further, we hypothesize that this intervention alone should be sufficient to improve accuracies, and no new RS-observations or updates to McFLI algorithms are needed to make better predictions of discharge. We use the Bayesian-AMHG-Manning’s (BAM) algorithm (Hagemann et al. 2017) as a case study for McFLIs. We provide BAM with improved prior river knowledge by 1) obtaining priors from the largest known repository of *in situ* data joined to hydrography (Brinkerhoff et al., 2019) and 2) constructing a river classification framework to dimensionally reduce hydraulic variation to geomorphically distinct river types. We test these interventions to produce river discharge from Landsat on 7,522 river reaches in the Mackenzie River basin (validated at 95 gauges) and on SWOT-simulated data representing 17 rivers from Durand et al (2016). Ultimately, we provide a method for improving discharge estimation that is globally scalable using only RS observations and could theoretically be applied to any river on Earth.

2 Data

We sought to improve the quality of priors introduced to McFLIs and assess the resulting sensitivity of our interventions on RSQ. So, this study required constructing three distinct datasets: 1) training data for generating new prior river knowledge (Section 2.1), 2) RS observations for RSQ in the Mackenzie River basin (Section 2.2), and 3) RS observations for RSQ using SWOT-simulated rivers (Section 2.3).

**2.1 Hydraulic dataset used for prior river knowledge**

As one of two interventions made to how we can estimate prior river knowledge, we gathered the most varied set of ‘training data’ we could find. We started with Brinkerhoff et al. (2019)’s dataset of field-measured hydraulics spatially joined to the USGS’s National Hydrography Dataset (USGS, 2019), further filtering it and calculating new geomorphic attributes. The dataset contains 730,072 unique measurements for cross-sections with at least 20 measurements. We filtered out impossible measurements (i.e. *Q* < 0) and measurements identified by the USGS as ‘poor’, ultimately arriving at the dataset used in this study (with cross-section locations mapped in Figure 1). This features 372,109 unique measurements across 190 rivers in the continental United States. To train our geomorphology classification frameworks, we calculated river and landscape geomorphic variables for each observation in the training data (Table 1).



**Figure 1.** Map of geographic locations of cross-sections in training dataset overlaid upon major American rivers.

**Table 1.** The 15 geomorphic variables used to define river types in this study.

|  |  |  |  |
| --- | --- | --- | --- |
| **Channel Width**  **(*W*)** | Observed channel width | **Minimum Grain Size Entrained**  **(*De*)** | From Henderson (1966) |
| **Channel Velocity**  **(*V*)** | Observed mean channel velocity | **USGS Waterbody Type**  **(*WB*)** | Artificial channel, intermittent river, perennial river, or lake/reservoir/wetland |
| **Channel Depth**  **(*D*)** | Observed mean channel depth | **Unit Power**  **()** |  |
| **Slope**  **(*S*)** | Observed slope | **Channel Shape**  **(*r*)** |  |
| **Manning’s n**  **(*n*)** |  | **Drainage Area**  **(*DA*)** | Observed drainage area |
| **Stream Order**  **(*SO*)** | Strahler stream order | **Froude Number**  **(*Fb*)** |  |
| **Distance Downstream**  **(*Dd*)** | Distance from headwater to current reach | **Shear Stress**  **()** |  |
| **Hydraulic Residence Time**  **(*HRT*)** |  |  | (At mean annual flow) |

We acknowledge that the continental United States is not reflective of all physical landscapes experienced globally. With that said, this is to our knowledge the largest freely available fluvial geomorphology dataset and covers a wide range of geographies, from temperate and semi-arid climates to deserts and sub-tropical regions and is a best-case scenario for our analysis. Specifically, our training data are missing observations reflective of equatorial and Arctic/subarctic regions. With the aim of global RSQ at the center of this study, we chose to use one of these poorly represented regions as a case study: the Mackenzie River basin in the Canadian Arctic.

**2.2 Mackenzie River basin**

The Mackenzie River basin is representative of Arctic hydrology and a good test for applying our new prior river knowledge at a large network scale. To run a McFLI here, we need RS observations of river widths. Using Landsat imagery, multi-temporal widths were extracted for every cross-section in the basin following the process outlined in Feng et al. (2019), with two main deviations from their approach in how centerlines and cross-section intervals were defined (Text S1). In short, we (1) used MERIT hydrography (Lin et al. 2019) to define river centerlines and generated cross sections at varying intervals along the centerlines (Table Sx). Then, (2) orthogonal lines were constructed for every cross-section, and (3) these orthogonal lines were used as inputs to RivWidthCloud, an automated algorithm for river width extraction using the Google Earth Engine (Yang et al. 2019). RivWidthCloud classifies a pixel as water using a novel algorithm detailed in Yang et al. (2019), Feng et al. (2019), and Text S1. After filtering for clear-sky images (Text S1), the classified water mask was intersected with the orthogonal lines to estimate wetted width at each cross-section. This process was performed on 228,659 cross-sections in 7,522 Landsat-visible reaches in the Mackenzie River Basin using Landsat imagery from 1984-2013. For this time period, there are 95 reaches with a stream gauge we can use for validation, and so we collected observed streamflow for these time periods from gauge records.

**2.2 SWOT-simulated rivers**

Most McFLIs invert Manning’s equation, which relies on observations of bed slope. SWOT will measure water surface slope and under steady flow this is assumed equal to bed slope. This means that once SWOT launches, Manning’s-based McFLIs can be ran globally. However, SWOT-like data is currently difficult to process and use for inverting discharge (Tuozollo et al. 2019a), making it presently difficult to test these Manning’s-based McFLIs on real-world RS data. To circumvent this difficulty, we used simulated SWOT data to test our new prior river knowledge on Manning’s-based McFLIs. Simulated SWOT data is simply reach-averaged hydraulic model outputs with water surface slopes and river widths outputted as ‘RS observations’. SWOT observations will exist at the reach scale due to known measurement errors (Frasson et al. 2017), and so these simulated rivers are reach-averaged to mimic this. By removing the effects of clouds and limited observations, these simulated rivers mimic perfect measurement conditions and represent the best-case scenarios of what SWOT will provide to hydrologists. For this test, we used 17 test rivers from Durand et al. (2016), outlined in Table Sx and covering the United States, Canada, Great Britain, France, and Italy. Median discharge ranged from 62-14,199 m3/s and ‘observation’ windows ranged from 22-365 days over 3-16 reaches.

**3 Methods**

**A picture containing text, map

Description automatically generated**

**Figure 1.** Flowchart of the methods for this study.

**3.1 River Classification**

Recall our hypothesis that contextualizing priors to their specific rivers will improve McFLI performance. To test this: we 1) obtained prior river knowledge from a larger and more geomorphically-varied dataset (Section 2.1) than previous work, and 2) constrain prior river knowledge using supervised and unsupervised classifications to extract geomorphically distinct river types.

Statistical learning is generally binned into unsupervised and supervised approaches (Hastie et al. 2009). Both use a suite of variables extracted from a training dataset to define a feature space and then identify patterns in that space. Unsupervised learning identifies these patterns and clusters observations from the training data, while supervised learning uses predictors to model ‘target variables’ that are known a priori. Commonly, supervised learning refers to regression if the target variable is continuous, or classification if it is discrete. In this study, we have no a priori river types in the training data, and so supervised learning refers to using our expert, domain-specific knowledge on river geomorphology to guide statistical methods of classification.

As a representative unsupervised clustering approach, we used the ‘density-based spatial clustering of applications with noise’ (DBSCAN: Ester et al. 1996) algorithm. DBSCAN is a density-based clustering algorithm that groups observations together if they are close to one another in the multi-dimensional feature space. Distance between points is determined using Euclidean distance. Unlike simpler unsupervised clustering algorithms, DBSCAN does not assume all clusters have a convex shape in the feature space and instead uses density to group observations. This means clusters can be arbitrarily shaped or completely surround other clusters. This also permits DBSCAN to identify ‘noise’ points which are outside of the dense areas of the feature space, differing in practice from other simple unsupervised learning methods (e.g. K-Means clustering will assign a cluster to every observation). While determining its own number of clusters, the user must provide a minimum number of points for a cluster and a maximum cluster radius. After iterating through cluster radii from 0.1-0.3, we settled on 0.2 as the best balance between number of clusters, within-cluster variance, and computational efficiency. The same iterative process suggested a minimum cluster size of 1,000 hydraulic measurements. We ran DBSCAN on non-dimensional forms of the fifteen geomorphic variables in Table 1, yielding 9 clusters in approximately 75% of the data and 15% classified as ‘noise’.

We also developed a bespoke supervised classification framework for extracting river types. By using principal component analysis (PCA) as a guiding tool, the approach follows similar methods used to extract global hydro-climatic river types (Dallaire et al. 2018) and hydrologic flow regimes (Olden et al. 2012). A PCA was used to dimensionally reduce our dataset and create multivariate, non-dimensional principal components (PCs) responsible for some amount of geomorphic variation across the feature space (and ultimately the United States). We ran a PCA on non-dimensional forms of the fifteen geomorphic variables in Table 1 and selected the five most influential PCs (cumulatively responsible for 73% of the variance in the feature space). In order of most variance explained, these PCs qualitatively represented 1) stream competency, 2) longitudinal location, 3) velocity/Froude number, 4) waterbody type, and 5) channel shape. For each observation in the training data, we aggregated the PC values to create a ‘fluvial index’ reflective of these five dimensions of geomorphology. Finally, we binned the fluvial index into 8 classes using quantiles as class thresholds. Eight classes, or ‘types’, were chosen to minimize overlap of width distributions between river types (Figure Sx). This becomes important when we map river types to rivers using RS observations of width (Section 3.2).

While this supervised framework is sufficient to segment classic single-channel rivers, its design is unable to parse out unique river geomorphology, and so we must introduce expert-defined river types to account for these. We identified two river types that are not well represented by this framework: ‘big’ rivers and ‘highly width-variable’ rivers. For some very large rivers (e.g. the Mississippi or St. Lawrence rivers), the training data had few measurements in rivers of similar size. We defined ‘big’ rivers as those with a mean log-transformed river width greater than 6.5 meters. ‘Highly width-variable’ rivers are those with significant variability in river width. These are generally multi-threaded or braided systems and some unique single-channel rivers (Text S3). We defined ‘highly width-variable’ as a standard deviation of at-a-station log-transformed width > 0.45 meters.

Both classification frameworks manifest such that channel hydraulics are fundamentally different across river types. Figure 3 plots truncated, parametric distributions of six hydraulic terms in the training data, by supervised river type. *A0* is median cross-sectional area, *Wb* is bankfull width,*Db* is bankfull depth*, n* is Manning’s roughness term, *r* is a channel shape term, and *b* is an AHG exponent (Text S2 details calculating these terms). *A0*, *Wb*, *Db,* and *b* fundamentally vary by river type, while *n* and *r* less so. Per the definition of *r* and *b*, their ‘highly width-variable’ distributions are significantly different from the other types’ distributions. Some types also show greater variance in experienced hydraulics, such as river type 8 for *A0*, *Wb,* and *Db.*

A picture containing text, photo, map, computer

Description automatically generated

**Figure 3** Truncated, lognormal distributions of hydraulic terms as defined using new training data and the supervised classification framework.

**3.2 River Type Mapping**

The goal of this study is to improve ungauged discharge prediction at the global scale. Thus, river types must be assignable to rivers using only RS observations, despite our river types being defined using *in situ* measurements that are not remotely sensible. This presents a unique classification challenge using width and/or slope as the sole predictors of the river types.

For the DBSCAN framework, neither width nor slope were strong predictors of the river types. So, we turned to basic machine learning to assign river types, which has seen some success at the regional scale (Guillon et al., 2020). After testing simpler regression-based approaches with no predictive success, we settled on a random forest classifier. Random forests are ensemble methods for classification (or regression) that use a series of decision trees to assign classes. Individual decision trees are prone to overfitting and so an ensemble approach is used to curtail that tendency (Hastie et al. 2009). Using a classic validation-set approach to model training, we trained this classifier on 80% of the training data using mean and standard deviation of cross-sectional width as predictors. When tested on the remaining 20% of the data, the classification accuracy was 90%.

For the supervised framework, the distributions in the training data of width by river type were sufficiently different from one another (Figure Sx- panel 7). The medians of these distributions were different enough that they were used as ‘characteristic widths’ for each river type. So, river types were determined to be whichever type’s ‘characteristic width’ was closest to that river’s observed mean width. Using this procedure, we mapped river types onto every reach in MERIT hydrography (Lin et al. 2019) for the Mackenzie River basin. The mean width for each reach was extracted from the MERIT Hydro digital elevation model (Yamazaki et al., 2019). Because we only have mean width in MERIT Hydro, the ‘highly width-variable’ class was not included in this mapping exercise.

**3.3 Discharge estimation**

McFLI RSQ was performed using the BAM algorithm (Hagemann et al. 2017). BAM probabilistically estimates discharge via Bayesian inference and a Hamiltonian Monte Carlo sampling scheme. The user chooses to invert Manning’s equation, AMHG, or some combination of both as its flow law. The user also chooses whether to run BAM at every cross-section or to use reach-averaged observations in line with how SWOT will observe rivers. For this study, we ran AMHG on every cross-section in the Mackenzie River basin and ran a ‘switch’ flow law on the reach-averaged SWOT-simulated data. This ‘switch’ always inverts Manning’s equation and inverts AMHG when it is deemed suitably strong (Text Sx). We also updated AMHG’s flow law to reflect new findings on the physical basis of AMHG (Brinkerhoff et al. 2019) and, following recent work on temporally defined roughness terms in McFLIs (Tuozollo et al. 2019b), tested space-varying and space-and-time-varying roughness terms in BAM. We ultimately implemented a space-varying term as a compromise between computational efficiency and increased predictive accuracy (Text Sx).

To run BAM, the user provides width and/or slope RS observations, as well as a prior estimate on discharge. Default BAM uses those width observations to predict the priors on channel hydraulics via functions trained using HYDRoSWOT (Hagemann et al. 2017). It then infers discharge using Bayes theorem. These priors, which are the six terms in Figure 3, flow law errors, and AMHG’s ‘congruent width’ term, are formalized within BAM as truncated, lognormal distributions where ) for , using mean (), standard deviation (σ), and upper () and lower bounds () as parameters. For our tests, we used geomorphic river types to redefine the prior river knowledge on the six hydraulic terms in Figure 3. A river type was mapped to the RS observations following Section 3.2. Prior parameters were then extracted from the distribution of all measurements in the training data pertaining to that river type and inputted into BAM. This workflow differed slightly for DBSCAN’s ‘noisy’ rivers and the supervised framework’s ‘big’ and ‘highly-width variable’ rivers (Text Sx) but they are still assignable using only RS observations. The prior estimate on discharge was defined depending upon the test: for the Mackenzie River basin, we used a stream gauge when available and otherwise used the mean daily discharge estimate from MERIT hydrography (Lin et al 2019). For the SWOT-simulated rivers, output from a water-balance model (Wisser et al. 2010) was used in line with Durand et al. (2016).

**3.4 Experimental Design**

BAM was run on both test datasets using default BAM as a control case. BAM was then run using the new prior river knowledge, termed ‘geoBAM’ (Table 2). geoBAM’s new prior river knowledge refers to the two interventions previously outlined: 1) using a larger and more geomorphically-varied dataset to obtain the knowledge (Section 2.1), and 2) constraining it using either DBSCAN or supervised river types (Section 3.1-2). Ultimately, BAM, geoBAM-DBSCAN, and geoBAM-Supervised were run on the SWOT-simulated rivers, while BAM and geoBAM-Supervised were run on the Mackenzie River basin.

**Table 2:** Experiments ran for both test cases (Mackenzie River basin and SWOT-simulated rivers).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **New Training Data** | **Unsupervised Classification** | **Supervised Classification** |
| geoBAM- Supervised | ✓ |  | ✓ |
| geoBAM-DBSCAN | ✓ | ✓ |  |
| BAM-Control |  |  |  |

Error metrics (Table 3) to quantify RSQaccuracy followed Hagemann et al. (2017). rBIAS and RRMSE define the range and central tendency of prediction errors, respectively. NRMSE is a normalized variant of RRMSE to account for RRMSE’s high sensitivity to errors in low flow estimation, and NSE represents the amount of variance in the observed data that the model explains. An NSE greater than zero indicates that our model estimates better than guessing mean flow every time.

**Table 3:** Error metrics used in this study. Also included is our definition of percent change.

|  |  |  |
| --- | --- | --- |
| **Description** | **Abbreviation** | **Definition** |
| Relative root-mean-square error | RRMSE |  |
| Normalized root-mean-square error | NRMSE |  |
| Relative bias | rBIAS |  |
| Nash-Sutcliffe efficiency | NSE |  |
| Percent Change | % Change |  |

1. Results

*OUT OF DATE NOW*

* 1. Mackenzie River Basin

Figure 6 plots Nash-Sutcliffe Efficiency (NSE) for the 95 validation reaches as empirical cumulative density functions (panel a), boxplots (panel b), and a probability density function of percent change between BAM and geoBAM, defined as (panel c). There is significant improvement in NSE, corroborating our choice of interventions made to BAM. With no new RS and no changes to BAM, we have yielded a median percent increase in NSE (panel c), across all reaches, of 78%; panel a) highlights that most improvement is occurring in reaches that were poorly estimated by default BAM. BAM performed better than a simple mean flow estimate would have (NSE = 0) in less than half of the reaches (42%), despite using the stream gauge present at these reaches to calculate the prior on *Q*; geoBAM, however, outperforms a mean flow estimate across most reaches (62%). In panel b, median NSE across all reaches improved from -0.11 to 0.15 and the spread of prediction accuracy is far more consistent using geoBAM than using default BAM, as the inter-quartile range (IQR) has more than halved (from 1.12 to 0.53)

A close up of a map

Description automatically generated

**Figure 2.** NSE improvement for 95 stream reaches in the Mackenzie River basin. A) Empirical cumulative density functions (CDFs) of the NSE scores for every reach; B) boxplots of the same results; c) probability density function (PDF) of the % change in NSE. Axes are truncated for visualization’s sake, with arrows and number of reaches not plotted provided when necessary.

The axes on these plots are truncated at the lower ends for visibility. For both BAM and geoBAM, 12 and 6 reaches (respectively) exhibited significant blow up of *Q* inversion and NSE scores below -2.3. However, panel c) shows that there is a sharp drop-off in percent change in NSE once 0% change is reached; put another way, the reaches in which geoBAM yielded worsened performance are few (even though some had, for example, -250% degradation in NSE scores).

* 1. SWOT-Simulated Data

Due to computational limitations, we were unable to iterate through all possible approaches to prior estimation on the Mackenzie system. Thus, we did so on the SWOT-simulated data, presented in Figure 7 as boxplots of the test rivers’ error metrics and colored according to ‘intervention’ (i.e. green is default priors, purple and blue are AMHG and training data interventions and brown adds river type classifications to the previous interventions- consult Figure 5). The far-right boxplots in the darkest brown is geoBAM as ran on the Mackenzie system, and far left is the control case.

**Figure 3.**  Boxplots of performance metrics for the 34 test rivers across all of our tests, colored by interventions (see Figure 5) to highlight sensitivity to different implementations of new ways to estiate priors. Axes are truncated for visibility, with the number of rivers not plotted noted for each boxplot.

In Figure 7, we continue to see substantial improvement in NSE for McFLIs using remotely-sensed *S*. For BAM, there is a median NSE increase for all rivers of 0.46 across all interventions (from green to brown boxplots), as well as more consistent performance for all experiments when compared to default/control BAM (IQR shifts from 1.49 to 1.21). Across interventions, the number of outliers changed little across interventions and was, generally around three. The introduction of the supervised classification framework yielded an increase in median NSE of 0.15 (and 0.05 when implementing DBSCAN). There is a small improvement in overall median rBIAS (0.07) and median RRMSE (6%), though no real change in median NRMSE (3%). NRMSE results did get notably more consistent, as the IQR shrunk from 0.50-0.40). There are two notable ‘jumps’ in NSE improvement: the introduction of new AMHG and/or new training data, and then the addition of a river type classification (brown), where more notable improvement occurs. There is little difference in median NSE when permuting options with new/old AMHG and new/old training data, though both rBIAS and NRMSE get slightly better. The most consistently good NSE scores are for old AMHG and new training data (no classification). Finally, the prior definition that includes a time-varying *n* yielded the only positive median rBIAS across all prior sets tested, and yielded the worse consistency in performance across all experiments tested. At its least consistent (NSE), its IQR was 1.58. This justifies our use of a space-only-varying *n* in the river type implementations (brown boxplots).

A close up of a map

Description automatically generated

**Figure 8.** Hydrographs for the SWOT-simulated rivers. Observed Q (dashed black) is plotted alongside default BAM (green) and geoBAM (brown).

Finally, in Figure 8 we show representative hydrographs for the SWOT-simulated rivers using both BAM (top row) and MetroMan (bottom row) and showing observed *Q* (black), default BAM/MetroMan (green), and BAM/MetroMan with the geoBAM set of priors (beige), all normalized by maximum timestep and mean observed *Q*. ‘Similar’ performance was defined as within 10% change in NSE from default BAM/MetroMan, with noteworthy improvement and degradation > and < 10% change, respectively. Overall for BAM, we found 16 rivers to get notably better, 11 rivers to yield similar results, and 7 rivers to get notably worse.

It is generally observed across hydrographs that default BAM performed quite well in low/baseflow periods but much worse in high flow events. Both implementations of BAM yielded adequate recovery of flow dynamics. However, geoBAM appears to have ‘filled in’ those peak events by more accurately reflecting the increased magnitude of flow, at least partially. Some rivers with little to no change in NSE scores still appear to have partially ‘filled in’ the discrepancies in predicted peak flows (e.g. Seine, Missouri, parts of the Ohio).

A close up of a map

Description automatically generated

**Figure 9:** Selected hydrographs for validation cross-sections from the Mackenzie River basin.

**6 Discussion**

The results presented show that prior quality and quantity are significant influencers on *Q* prediction, and that significant performance improvements can be made for McFLIs by simply improving the quality of prior river knowledge introduced to the algorithms. While algorithm improvements will also continue to evolve our understandings of McFLIs and *Q* inversion, we have found that our relatively simple interventions to prior estimation are easy to integrate into existing McFLIs and still yield noteworthy skill improvement, regardless of the algorithm tested.

The success of our interventions in the Mackenzie River basin suggest that this approach is globally implementable. BAM showed poor performance in the basin, but this was significantly improved when implementing geoBAM. This is noteworthy as all reaches used in this study are in the Arctic, where our training data is supposedly unrepresentative (none of the field measurements were made in the Arctic/Subarctic). Further, Feng et al. (2019)’s method for running BAM yielded acceptable *Q* estimation in the narrowest rivers to date with no *in situ* calibration; here we show that a similar approach aimed at global *Q* estimation can be run using geoBAM and yield a strong result, corroborating the argument for using both these prior estimation and remote *W* extraction methods in global applications. This result functionally opens the door for geomorphically-informed *Q* estimation across global-scale river networks.

It is interesting that substantial performance improvement, for the SWOT-simulated rivers, is largely limited to NSE, and not NRMSE, RRMSE, or rBIAS. RRMSE and rBIAS generally track together and are easily inflated due to errors in baseflow prediction. In the SWOT-simulated hydrographs, we observe that our interventions alter baseflow predictions very little, and so it is likely that these two metrics do not change much as baseflow prediction has changed very little. Because our interventions have primarily addressed magnitude issues with priors (i.e. big rivers should not have *A0* bounds of 10m2 as they might in BAM but will not in geoBAM), we see significant improvement in NSE as error due to *Q* magnitude differences is reduced. This is corroborated by studying the hydrographs in Figures 8 and Sx.

We found that classification will yield improvements if classes are assigned correctly, but significantly degrade performance if classified poorly. Because we reduce priors to look-up tables, if the ‘wrong’ prior is assigned then inversion will be considerably off (unlike when using a global function where it can be easier to be close to ‘correct’).For example, section 1 of the Ohio River had an NSE of 0.13 however after assigning river type and constraining the priors, performance degraded to -5.66. Andreadis et al. (*in review*) found a similar result when their planform classification ‘misclassified’ the same Wabash and St. Lawrence Downstream rivers used in this study; however in this study, both of these rivers were well-modeled by our classification and exhibited little performance degradation. Likewise, if performance is already great (i.e. Jamuna river control-case NSE of 0.983) then it does not get much better from our interventions (geoBAM Jamuna river NSE is 0.986). This follows logically, as the default BAM priors were of high quality for these rivers and so a classification approach simply produces similar priors to those estimated using the global functions. It is also worth acknowledging that though th supervised framework outperformed DBSCAN, DBSCAN largely replicated the findings from our more complicated supervised approach. This suggests that unsupervised algorithms can adequately replicate the different types of rivers experienced globally, and that future implementations might equal or outperform a more manual approach.

A similar finding manifests when we test the sensitivity of a McFLI to its initial prior on *Q*,as up until now we have not amended the prior estimation method for *Q*. BAM by default uses *Q*priors centered on the output of a global water balance model (Wisser et al. 2010). We substituted this term ( with new estimates from the Global Reach‐Level A Priori Discharge Estimates (GRADES- Lin et al. 2019) for 17 of our test rivers using mean daily *Q* from GRADES and then reran geoBAM. We observed significant performance degradation in one river (Missouri- median percent NSE decrease across the three Missouri River sections was 1,526%) but improvement in all others (median percent NSE increase, excluding Missouri, was 17.5%). This is presumably due to a poorly informed prior on *Q* provided from GRADES, when compared to the WBM-derived *Q* prior previously used. Consult Table Sx for the full results.

Sometimes, performance is degraded from the control run, but the classification is still yielding improvement; net performance degradation is due to our other interventions. For example, the Saint Lawrence Downstream river had an NSE of -2.2 when running the control case of BAM. After introducing new AMHG and new training data, NSE dropped to -4.98. Then, we introduced the supervised classification atop these interventions and NSE improved to -4.12. Although the overall performance is still quite poor, this suggests that a BAM implementation using older AMHG and/or training data with a classification framework might yield the best results. While these scores are very poor overall, the relative changes successfully highlight the fact that even in rivers with net performance degradation, statistical river classifications are yielding improvement and that there are occasions when our other interventions represent a river poorly. The topic of why some rivers improve from some interventions and others worsen is ripe for future research.

The SWOT-simulated rivers are quite homogenous (18/34 were assigned the same class in geoBAM) and while SWOT will be limited to rivers wider than 100m, they aren’t reflective of the global variation in hydraulics and geomorphology SWOT will encounter. While this has been noted in the past (Durand et al. 2016), until now there has not been a geomorphically explicit way to quantify the homogeneity between SWOT-simulated test cases. What is thus needed is a wider range of river types for validating and testing McFLIs before SWOT launches. These results also highlight the significant variation in McFLI performance within river types (i.e. for river type 8, NSE ranges from -21.7 for the Tanana to 0.94 in the Cumberland). Future work should attempt to parse out specific river types that different McFLIs model quite well, and those that McFLIs model poorly.

Other future work should address the reliance on Manning’s constants and Manning’s equation. The new AMHG flow law is easily generalizable and so is the Manning’s implementation within BAM. Manning’s equation has been shown to not introduce too much prediction error in larger rivers, however if the ultimate goal is network-scale *Q* prediction (as we have done in the Mackenzie basin), generalized flow~roughness relations would ideally be used and hypothetically improve prediction accuracy in smaller, low-order streams.

7 Conclusions

This study presents a first attempt at quantifying the sensitivity of priors on global-scale RSQ in two distinct settings: thousands of Arctic river reaches using just Landsat imagery and simulated rivers representing data the NASA SWOT satellite will provide upon launch in 2021. We found very significant improvement in the accuracy of our *Q* predictions for both test cases, with the median percent improvement of NSE in the Arctic rivers to be 78%. When testing the explicit sensitivity of various prior estimation methods for the SWOT-simulated rivers, we found marginal differences across methods tested, but all yielded significant improvement over the current approaches to prior estimation. Statistically classifying rivers and truncating priors yielded further improvement. These findings are both significant and highlight the importance of prior knowledge in a Bayesian mathematical setting, where we have shown that starting from a more informed understanding of the river at hand yields more accurate results. These priors are implementable in any McFLI and will play a pivotal role in both current efforts to remotely-sense *Q* across global-scale river networks as well as future *Q* estimates from SWOT, where prior knowledge is essential to estimating flow in ungauged rivers where nothing is known *a priori.*

Acknowledgments

C. B. Brinkerhoff was supported by a grant to C. J. Gleason through the NSF RoL FELS RAISE program (PI Peter Raymond, Grant 1840243). Code to run our analysis is available at <https://github.com/craigbrinkerhoff/2019_GeoClass>. geoBAM is available in the R programming language at <https://github.com/craigbrinkerhoff/geoBAM>. The authors would like to thank Mark Hagemann for constructive comments during this project, and for building the original BAM algorithm.

References