Constraining Remote River Discharge Estimation Using Reach-Scale Geomorphology

C.B. Brinkerhoff1\*, C.J. Gleason1, M. Durand2, D. Feng1, P. Lin3

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

2 School of Earth Sciences and Byrd Polar and Climate Research Center, The Ohio State University, Columbus, OH

3 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 is now 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), and preliminary global assessments between 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 geomorphology 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 without restrictive assumptions. 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 remotely sensed 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; Hagemann et al. 2017; **for discharge, xxx those papers that use WMB like pepsi xx**) and/or from external training data of geomorphic and hydraulic variables (e.g. Canova et al. 2016). 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 rougheness 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), in the first test of McFLIs on real-world SWOT-comparable altimetry data, found that McFLI estimation bias is sensitive to the prior on discharge. 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 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 *Q* and other hydraulics to calibrate the rating curves for their stream gauges and all are freely available. Some McFLIs have used a subset of these measurements known as HYDRoSWOT (Canova et al. 2016) to train their priors. 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), here named the ‘Hydraulic Observations with Geomorphology and Geography’ (HOGG) dataset.

In this context, 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 HOGG and 2) constructing a river classification framework to dimensionally reduce hydraulic variation to 10 geomorphically distinct river types. We test these interventions to produce river discharge from Landsat on 23,799 river reaches in the Mackenzie River basin (validated at 95 gauges) and on SWOT-simulated dataset representing 19 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 Methods

Section 2.1 describes BAM. Section 2.2 steps through our river classification and prior estimation methods. Section 2.3 describes validation procedures and test datasets.

**Figure 1.** Flowchart of the methodology undertaken in this study. SWOT refers to the simulated test rivers which mimic the types of observations the SWOT satellite will provide when launched in 2021.

2.1 McFLI Description

This study used the BAM algorithm to estimate *Q* from simulated and observed RS data. BAM (Hagemann et al. 2017) probabilistically estimates discharge via Manning’s equation and/or AMHG, Bayesian inference, and a Hamiltonian Monte Carlo sampling scheme. It returns a full posterior distribution of predicted discharge. BAM allows for easy substitution of priors and thus makes it a convenient test case for our tests. This flexible framework also allows for different flow laws to be ran. Hagemann et al. (2017) enabled BAM to run Manning’s equation and/or AMHG. One can also designate a ‘switch’, where AMHG is turned on and ran for rivers when AMHG is deemed suitably strong. This study used only AMHG in the Mackenzie River basin to estimate discharge from a current RS platform and a manning-AMHG switch for the SWOT-simulated rivers. This switch is detailed in Text S1.

We updated the physical flow law for AMHG to reflect new findings on the physical basis of AMHG (Brinkerhoff et al. 2019). This is detailed in Text S1 and is hereafter termed ‘new AMHG’. Equations 1 and 2 are the Bayesian likelihood functions inverted within BAM to estimate *Q*. Equation 1 is Manning’s and equation 2 is AMHG.

1

2

*i* represents space-varying and *t* represents time-varying quantities. *S* is bed slope, *A0* is the median channel cross-sectional area, *n* is Manning’s roughness coefficient, *b* is the width AHG scaling exponent, *r* is a channel shape parameter, and *Wb* and *Db* are bankfull width and depth, respectively. The AMHG global parameter *Wc* is also in Equation 2, and epsilon is an error term defined as a normal random variable. Epsilon for AMHG was estimated by Hagemann et al. (2017) to have a mean of 0 and a standard deviation of 0.22. Epsilon for Manning’s equation is defined by a mean of and standard deviation of , where is the standard deviation of epsilon for Manning’s. (Hagemann et al. 2017).

2.2 McFLI Interventions

We made two interventions to how BAM estimates prior river knowledge: obtaining priors from a larger and more geomorphically-varied dataset than previous work (Section 2.2.1), and 2) constructing river classification frameworks to dimensionally reduce hydraulic variation to geomorphically distinct river types (Section 2.2.2).

* + 1. New Training Data

We started with Brinkerhoff et al. (2019)’s HOGG dataset, further filtering it and calculating new geomorphic attributes. HOGG features 730,072 unique measurements for cross-sections with at least 20 measurements. We then filtered out impossible measurements (i.e. *Q* < 0), measurements identified by the USGS as ‘poor’, and measurements with AHG exponents > 1 or < 0), ultimately arriving at the dataset used in this study (with cross-section locations mapped in Figure 1) which features 372,109 unique measurements across 190 rivers in the continental United States.

**Figure 2.** Map of geographic locations of cross-sections in training dataset overlaid upon major American rivers.

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. This is the primary reason we chose to run our study on an Arctic/Subarctic watershed (Mackenzie River), thus testing how flexible our interventions are to a region poorly represented by the training data.

Next, we used HOGG to calculate river and landscape geomorphic variables for each hydraulic measurement. Table 1 shows the fifteen variables ultimately calculated for each measurement or each cross-section, depending upon the nature of the variable. Of these fifteen variables, *W, D, V, De, Fb,* , , and *n* were calculated for every measurement, while *S, SO, Dd, HRT, WB, r,* and *DA* were calculated for each cross-section.

**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) |

* + 1. Redefining Priors

With new training data in hand, we now redefine the priors inputted into BAM. In the BAM context, priors are truncated, lognormal, parametric distributions and are defined using a distribution center (), standard deviation (σ), and upper and lower bounds. Priors are needed on *Wb*, *Db*, *r*, *n, A0,* and *b*. *Wb* and *Db* were calculated using a return period of 2 years, acknowledging that *Qb* has been associated with a range of return periods given local geomorphology (Petit & Pauquet, 1997; Williams, 1978). However, a 1.5-2-year return period is a standard statistical definition for bankfull flow and was used here to align with existing literature. *r* was defined empirically following Dingman (2007) as , where *f* is the exponent in the *D*~*Q* AHG relation. We did not change the *Q* prior as we considered this term as separate from the prior knowledge on geomorphology that we are interested in studying. All priors were confirmed to have lognormal distributions in HOGG. , σ, and bounds were derived two different ways: 1) using the unsupervised clustering algorithm ‘density-based spatial clustering of applications with noise’ (DBSCAN), and 2) using a novel supervised classification framework. We also trained priors using Hagemann, et al. (2017)’s method and HOGG: consult Text S2 for details.

DBSCAN (Ester et al. 1996) is a density-based clustering algorithm that groups observations together if they are close to one another in the multi-dimensional feature space. Unlike simpler unsupervised clustering algorithms, DBSCAN does not assume all clusters have a convex shape in the variable 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 not in sufficiently dense areas of the variable space. While determining its own number of clusters, the user must provide 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. We ran DBSCAN on non-dimensional forms of the fifteen geomorphic variables in Table 1; this yielded 9 clusters in approximately 75% of the data and 15% as ‘noise’.

We also developed a bespoke supervised classification framework for assigning river types. This approach uses principal component analysis (PCA) to identify the primary drivers of geomorphic variation across our dataset following similar approaches used by Dallaire et al. (2018) to derive global hydro-climatic river types and Olden et al. (2012) to derive hydrologic flow regimes. We used PCAs to dimensionally reduce our dataset and create multivariate, non-dimensional variables reflective of geomorphic variation. These so-called ‘PC axes’ are each responsible for some amount of the variance experienced in the fifteen-dimension feature space. A PCA was ran on non-dimensional forms of the fifteen variables calculated using HOGG. We then selected the five most influential PC axes, which were responsible for 73% of the variance in the feature space. For each HOGG measurement we aggregated the PC axis values to create a ‘fluvial index’ reflective of these five dimensions of geomorphology and explicitly defined for every HOGG measurement. Finally, we binned the fluvial index into 8 classes using quantiles as group thresholds. Eight classes, or ‘types’, were chosen subjectively as the maximum number of classes wherein the resulting within-class distributions for observed *W* did not have notable overlap across classes.

Next, we identified two river types that are not well represented using this approach: ‘big’ rivers and ‘highly width-variable’ rivers. For some very large rivers, HOGG has very few measurements in rivers of similar size. ‘Big’ rivers were defined as those with a mean log-transformed *W* greater than 6.5 meters. This is approximately equal to the sum of the widest river type’s median and standard deviation of observed *W*. ‘Highly width-variable’ rivers are those with significant variability in observed *W*; these are generally braided rivers and some single-channel rivers. We set the threshold for ‘highly width-variable’ as a standard deviation of at-a-station log-transformed *W* > 0.45.

Because both classification frameworks are fundamentally based on geomorphic measurements that are not remotely sensible, we developed globally scalable methods to assign river types using only RS measurements. The supervised framework was defined in such a way that the observed *W* distributions for each class are significantly different (see Figure Sx). We were able to assign a river type simply by identifying which of the types’ *W* distributions in HOGG were centered the closest to the river’s mean *W*. Using the procedure outlined for the supervised framework, we mapped river types onto every reach in MERIT hydrography (Lin et al. 2019), extracted from the MERIT digital elevation model (Yamazaki et al. 2019), for the Mackenzie River basin (Figure 3). Also included are the major lakes present in the basin. Our river classification successfully identified the two distinct landscapes in the Mackenzie basin, borne out by the two distinct river type regimes in Figure 3: the Rocky Mountains and Continental Divide in the West and the low-lying boreal forests and lakes atop the Canadian Shield to the East.

**Figure 3:** Map of Mackenzie River basin hydrography (Lin et al. 2019) with 23,799 reaches classified according to the supervised framework and assigned using only mean observed W. Major lakes are included as well. Class 100 corresponds to ‘big’ reaches. As MERIT Hydro provides only mean W, ‘highly-width-variable’ rivers are not classified.

This method, however, did not work for the DBSCAN classes as *W* was not a strong predictor for those types. So, we trained a random forest classifier using 80% of HOGG to assign river types by mean at-a-station *W* and the standard deviation of at-a-station *W*. ‘Noise’ was included as a 10th class. This model, when tested on the remaining 20% of HOGG, yielded a 90% accuracy in class assignment.

Finally, McFLI priors were estimated by extracting prior distribution parameters from HOGG, after binning by river type. This was done for both classification systems. For example, the *b* prior for river type 7 was defined as follows: the was the median *b* for all measurements classified as type 7, σ was the standard deviation of *b* for all measurements classified as type 7, and upper and lower bounds were the maximum and minimum *b* values for all measurements classified as type 7. For rivers that were ultimately assigned a ‘noise’ class in DBSCAN, and for ‘big’ rivers in the supervised framework, their priors were estimated using the Hagemann et al. (2017) functions trained on HOGG and outlined in Text S2. For ‘highly-width-variable’ rivers, priors were built from only those HOGG cross-sections with an *r* less than 1. Per Dingman (2007)’s definition for *r,* *W* variability is greater than *D* variability and reflects patterns commonly seen in braided rivers. See Figure Sx for boxplots of the prior distributions by river type in HOGG.

**Figure 4** Truncated, lognormal, parametric distributions of BAM priors as defined using new training data and the supervised classification framework. The ‘global/100’ distributions are the observed distributions for the entire training dataset.

Consult Figure 4 for the truncated, lognormal, parametric distributions of each prior, defined using the 10 river types from the supervised framework, where the ‘global/100’ distributions are for all of the training data. It was this data that the Hagemann et al. (2017) functions were trained on and used for ‘big’ and ‘noise’ rivers. We now have two experimental ‘prior sets’, defined and estimated using DBSCAN and a supervised framework, as well as updated global functions following Hagemann et al. (2017); all priors are now defined in space, systematically vary with geomorphology, and are modelled using only remotely sensible *W* or *S*.

2.3 Evaluating McFLI Interventions

Two test cases were ran using our experimental prior sets: the Mackenzie River basin and the SWOT-simulated rivers. The construction of these tests is outlined first (Section 2.3.1), followed by *Q* estimation (Section 2.3.2).

2.3.1 Mackenzie River Basin: Uncalibrated *Q* Estimation using *W*

The Mackenzie River basin is representative of Arctic hydrology and a good test for applying our interventions in a climatic region not represented in our training data. We first extracted multi-temporal river *W* for every reach 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. 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 (see Table Sx). Then, (2) orthogonal lines were constructed for every cross-section; orthogonal line lengths were defined as 2 \* mean *W* for each cross section. The mean *W* for each cross section was extracted from the MERIT Hydro digital elevation model (Yamazaki et al., 2019).

Then, (3) these orthogonal lines were used as inputs to RivWidthCloud, an automated algorithm for river *W* 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 S3. After filtering for clear-sky images (cloud clover < 25% as defined by Landsat 5 surface reflectance tier 1), the classified water mask was intersected with the orthogonal lines to estimate wetted *W* at each cross-section. This process was performed on all visible reaches (7,522) in the Mackenzie River Basin using Landsat 5 imagery from 1984-1988. 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.

2.3.2 SWOT-Simulated Rivers: Uncalibrated *Q* Estimation using *W* and *S*

We collected 17 simulated rivers to test and validate our interventions on SWOT-like data, using tests cases from Durand et al. (2016). The details of these rivers are outlined in Table Sx. 2/17 rivers are multi-threaded (Tanana, Platte). These river models were developed explicitly to simulate SWOT observables under perfect measurement conditions, i.e. with no measurement error. As outlined in Durand et al. (2016), these are not actual river observations but rather the outputs of hydraulic models, constituting simulated values from forcing observed flows and river bathymetry through said models. They represent ‘best case scenarios’ for what we will receive from SWOT and thus serve as ideal test cases, though they are not observed data. In this study we used reach-averaged observations to estimate discharge, as SWOT will provide reach-scale measurements. The number of reaches per simulation ranged from 3 to 16 and the number of days per simulation ranged from 24 to 595.

2.3.3 Discharge Estimation and Performance Metrics

*Q* estimation was performed using different implementations of the BAM algorithm. Consult Figure 5 for the set of tests that we ultimately performed with different versions of prior estimation methods, depending upon which of the interventions were implemented: ‘new AMHG’, ‘new training data’, ‘time-varying n’ and the prior estimation methods: ‘global-scope’, ‘DBSCAN’, and ‘supervised’. The last row in Table 4 (in dark green) is the control case, and the first row (in dark brown) is the version also ran on the Mackenzie, termed ‘geoBAM’. For SWOT-simulated tests, the *Q* prior was found using the output from a water balance model (Wisser et al. 2010) and Manning-AMHG BAM with a switch as outlined in Text S1 was ran. For the Mackenzie tests, the *Q* prior used the stream gauge present at each validation reach and AMHG-only BAM was ran. Time in-variant *S* from MERIT hydrography was used for these tests.

**Figure 5.** Tests ran on McFLIs with various implementations for estimating prior river knowledge outlined using green checkmarks. Colors for algorithms ran correspond to colors in results figures. AMHG interventions did not apply to MetroMan tests.

*Q* prediction accuracy was quantified using error metrics first outlined in Hagemann et al. (2017) and reprinted in Table 2 (along with the definition for percent change used in this study). 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 was included as the de facto standard streamflow modeling performance metric; it represents the amount of variance in the observed data that the model explains.

**Table 2:** Performance metrics used in this study, as detailed in Hagemann et al. (2017). We chose to use the same performance metrics to fall in line with previous efforts to quantify Q prediction in the McFLI context. Also included is our definition of percent change, which is used frequently in this study.

|  |  |  |
| --- | --- | --- |
| **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
   1. Mackenzie River Basin: Uncalibrated *Q* Estimation using *W*

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)

**Figure 6.** 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: Uncalibrated *Q* Estimation using *W* and *S*

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 7.**  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).

MetroMan results go here.

**Figure 8.** hydrographs Representative hydrographs for BAM and MetroMan ran on the SWOT-simulated rivers, grouped into notable improvement, similar performance, and degraded performance. Observed Q (dashed black) is plotted alongside default McFLI (green) and geoBAM/MetroMan with geoBAM priors.

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*. The rest of the hydrographs are presented in Figure Sx. ‘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).

1. 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.

5 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

**For Supplement**

**Text S1**

Before testing the crux of our interventions on BAM and MetroMan were run, we needed to update the flow law underlying BAM considering recent work providing explicit physical expressions for AMHG (Brinkerhoff et al. 2019). This amounted to three interventions: a new flow law for AMHG, a physically-based ‘AMHG switch’, and defining a spatially-varying channel roughness *n* prior. Consult Eq. S1 for the AMHG flow law currently inverted to estimate *Q* in BAM.

Brinkerhoff et al. (2019) analyzed 155 rivers in the continental United States, finding that AMHG strength is a direct result of a river slope profile’s strength of fit to any slope-roughness model. They were also able to show that empirically-derived *Wc* and *Qcw* coexist with Dingman (2007)’s HG model. Consult that paper for a more thorough treatment on AMHG, but overall their analysis yields a new AMHG expression (equation S2) defined by slope, roughness, and bankfull geometry. We take this and derive a novel flow law (equation S3) by substituting equation S2 into equation S1 and then substituting Manning’s constants for the generalized velocity-depth relation exponent *p* and the generalized roughness coefficient *K* into the resulting expression.

S1

S2

S3

Here, *Wb* is bankfull width, *Db* is bankfull depth, and *r* is a channel shape parameter. *K* is equal to 1/*n* when adhering to Manning’s relation. Consult Dingman (2007) for an explanation of his generalizations of Chezy’s and Manning’s expressions for HG relations. It must be stressed that values for *p* and *K* have been shown to vary widely and not adhere to Manning’s (or Chezy’s) constants due to different river morphologies (e.g. Bjerklie et al. 2005; Dingman & Afshari, 2018; Dingman, 2007; Dingman & Sharma, 1997; Ferguson, 2010; Knighton, 1975); with that said, this new formulation of the AMHG flow law is flexible and in theory allows for future generalized implementations. Note also that with our new definition for AMHG every hydraulic prior is defined as space-varying except for *n*, and so we bring *n* in line with the rest of the priors.

With this flow law in hand, we can now derive a Bayesian likelihood function that will be introduced to BAM. This was done by log-transforming Equation 4 and solving for observed *W*, resulting in Equation S4.

S4

Note that observed *S* is still on the right-hand side of the equation out of algebraic necessity (and that AMHG is now defined by both observed *W* and *S*). Also note that through our redefinition of AMHG we have removed the *Qc* prior while maintaining the *Wc* prior even though algebraically simplifying equation S3 to its fullest extent would remove *Wc* from the expression. This was done deliberately to preserve the AMHG hydraulic pair *Wc*, *Qc* in this physical model. The mathematical definition for *Wc* is robust (Gleason & Wang 2015) and was kept here to force the physical model to reflect *Qc* and no other *Q* possibly experienced in a given river. Equation S4 is thus substituted into BAM in place of the AMHG likelihood function developed in Hagemann et al. (2017) and is hereafter referred to as ‘new AMHG’.

The second component of ‘new AMHG is a physically-based AMHG switch. As previously stated, BAM allows for a switch to be defined, ‘turning on’ AMHG when it is suitably strong. The definition of ‘suitably strong’ used in Hagemann et al. (2017) was a standard deviation of log-transformed *W* > 0.1 and a concurrent ‘percentage of rating curve intersections’ *pint* > 0.15. We redefined ‘suitably strong’ AMHG following Brinkerhoff et al. (2019)’s definition, wherein observed slope strongly fits a river-wide slope model. Here, we use the regime theory model defined by Henderson (1965) and reprinted as Equation S5. Degree-of-fit is defined by the coefficient of determination (r2), and strong fit was defined as an r2 > 0.90.

S5

*De* is the minimum grain size for entrainment, calculated as 11­*DbS* (Henderson, 1965), and *Qb* is bankfull discharge. In order to keep this methodology viable using just remotely sensible observations, *Qb* and *Db* were predicted using simple regression models where observed *W* is the predictor variable, explained in detail in Section 2.1.2.

Finally, a global *n* prior in McFLIs is known to be both physically inaccurate in many scenarios and poorly reflective of the variation in *n* experienced in both space and time (Tuozollo et al. 2019). We thus implement both space-varying (equation S4) and space-and-time-varying *n* variants (equation S4 with *n* defined in time as well) into BAM to test sensitivity to variability in *n*.

**Text S2**

Aside from our experimental classification approaches to prior estimation, we also sought to train the models devolped in Hagemann et al. (2017) on the new and larger training dataset used in this study. It was also used for DBSCAN’s ‘noisy’ rivers and supervised’s ‘big’ rivers. Mean log-transformed *W* and/or the standard deviation of log-transformed *W* (at-a-station), as predictor variables, were trained to predict the prior centers/hats using simple linear regressions. Note that the *n* prior was modeled using *S*, contrary to the other priors, as *W* was found to be a poor predictor of *n* in our training data. For the sake of fitting equation 5, we trained similar models to estimate *Qb* and *De* (recall Section 3.1.1), still using just *W* as the predictor variable. Figure Sx (in the supplementary material) shows the simple linear regressions used to model our six initial prior estimates, with r2 for these models ranging from 0.42-0.98. Sigma for each prior was defined as the residual standard error of each hat model. Finally, in order to test the sensitivity to new AMHG only (i.e. using ‘old training data’), we defined the *Wb, Db,* and *r* priors using HYDRoSWOT. The models used in Hagemann et al. (2017) were used for *n*, *A0*, and *b*. This set of prior definitions are hereafter termed ‘global priors’.

**Figure Sx.** Training of global prior estimation models using just remotely-sensible W and S. In order to increase predictive ability, some models used both mean of log-transformed at-a-station W and the standard deviation of log-transformed at-a-station W; some used only one. For A0, and b, predictors used were identical to those used in Hagemann et al. (2017) as detailed in the main text.

**Figure Sx.** Boxplots of training data priors and widths segregated by river type for supervised.

**Figure Sx.** Boxplots of training data priors and widths segregated by river type for DBSCAN.

**Table Sx:** Varying interval distances between cross-sections, dependent upon mean observed river widths from the MERIT Hydro digital elevation model (Yamazaki et al. 2019), used for extracting mean river widths from Landsat imagery.

|  |  |
| --- | --- |
| **Reach-Defined Mean River Width [m]** | **Cross-section Interval Length [m]** |
| 40-120 | 300 |
| 120-1000 | 500 |
| 1000-2000 | 1000 |
| 2000-3000 | 2000 |
| 3000-4000 | 3000 |
| 4000-5000 | 4000 |
| 5000-10000 | 5000 |

**Figure Sx.** Reprint from Feng et al. (2019) showing (a) cross-sections along a reach, (b)orthogonal lines and buffers, and (c) ultimate width extraction for Yukon River as Eagle, AK. For this study, the equally spaced cross sections are replaced with a varying interval length as detailed in the main text and shown in Table S1.

**Text S3**

To extract observed river widths from cloud-free Landsat imagery, we used equation S6, where *Hw* is the half-length of each orthogonal, is the ratio of overlapped water mask and buffer area, and *W* is wetted width. Consult Figure Sx, a reprint from Feng et al. (2019), showing how these parameters are defined for each cross-section. Panel c) defines *hw* and *W* graphically as based upon the extracted orthogonals to each cross-section.

S6

In order to do this, RivWidthCloud classifies water using the following approach, with a reported accuracy of 97% (Zou et al. 2018): (mNDWI > EVI or mNDWI > NDVI) and (EVI < 0.1). These indexes are defined below in equations S7-S9, where bandgreen, bandred, bandblue, bandswir1, and bandnir are cell values for the green, red, blue, shortwave infrared, and near infrared bands for Landsat images, respectively.

S7

S8

S9

**Table Sx.** Descriptions of 34 simulated rivers representing best-case scenarios of the data SWOT will give us upon launch (and used to estimate *Q* in this study).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| River | **Median Q [m3/s]** | **Number of Days Simulated** | **Number of Reaches Simulated** | **Location** | **Model** | **Reference** |
| Arial Khan | 417 | 334 | 10 | Bangladesh | HEC-RAS |  |
| Cumberland | 933 | 162 | 4 | USA | HEC-RAS | Adams et al. (2010) |
| Garonne Upstream | 480 | 365 | 8 | France | HEC-RAS | Besnard and Goutal (2008); Larnier (2010) |
| Garonne Downstream | 128 | 365 | 16 | France | MASCARET |  |
| Iowa | 69 | 366 | 8 | USA |  |  |
| Jamuna | 13,465 | 250 | 4 | Bangladesh |  |  |
| Kanawha | 510 | 162 | 4 | USA | HEC-RAS | Adams et al. (2010) |
| Kushiyara | 732 | 366 | 5 | Bangladesh | HEC-RAS |  |
| Mississippi Downstream | 14,199 | 162 | 6 | USA | HEC-RAS | Adams et al. (2010) |
| Mississippi Upstream | 4,869 | 162 | 3 | USA | HEC-RAS | Adams et al. (2010) |
| Missouri Downstream | 862 | 595 | 5 | USA |  |  |
| Missouri Midsection | 908 | 595 | 4 | USA |  |  |
| Missouri Upstream | 912 | 595 | 4 | USA |  |  |
| Ohio | 3,444 | 220 | 5 | USA | HEC-RAS | Adams et al. (2010) |
| Ohio Section 1 | 1,335 | 220 | 9 | USA | HEC-RAS |  |
| Ohio Section 2 | 2,065 | 220 | 8 | USA | HEC-RAS |  |
| Ohio Section 3 | 2,488 | 220 | 14 | USA | HEC-RAS |  |
| Ohio Section 4 | 2,847 | 220 | 5 | USA | HEC-RAS |  |
| Ohio Section 5 | 3,273 | 220 | 8 | USA | HEC-RAS |  |
| Ohio Section 7 | 4,705 | 220 | 7 | USA | HEC-RAS |  |
| Ohio Section 8 | 6,327 | 220 | 6 | USA | HEC-RAS |  |
| Padma | 23,605 | 327 | 5 | Bangladesh | HEC-RAS |  |
| Platte | 114 | 22 | 14 | USA | BreZo (2D) | Schubert et al. (2015) |
| Po | 1,009 | 367 | 16 | Italy | HEC-RAS | Di Baldassarre et al. (2009) |
| Sacramento Downstream | 213 | 154 | 9 | USA | HEC-RAS | Rogers (2014) |
| Sacramento Upstream | 181 | 305 | 7 | USA | HEC-RAS | Rogers (2014) |
| Seine | 200 | 365 | 4 | France | ProSe | Vilmin et al. (2015) |
| Seine Downstream | 221 | 365 | 4 | France | ProSe |  |
| Seine Upstream | 137 | 365 | 4 | France | ProSe |  |
| Severn | 62 | 88 | 4 | UK | LISFLOOD-FP | Neal et al. (2015) |
| St. Lawrence Downstream | 9,037 | 139 | 4 | Canada | H2D2 (2D) | Heniche et al. (2000) |
| St. Lawrence Upstream | 9,037 | 139 | 4 | Canada | H2D2 (2D) | Heniche et al. (2000) |
| Tanana | 1,450 | 100 | 9 | USA | LISFLOOD-FP | Humphries et al. (2014) |
| Wabash | 842 | 162 | 4 | USA | HEC-RAS | Adams et al. (2010) |

**Figure Sx.** Hydrographs for the 34 test rivers, with observed discharge plotted as a dashed black line, default BAM priors in green, and new prior information using the supervised classification framework in beige. The y-axis (flow) is normalized by mean observed discharge while the x-axis (time) is normalized by the maximum timestep for the river. The majority of rivers show substantial improvement in mapping the magnitude (and sometimes dynamics) of the observed hydrograph. A few rivers show worsened prediction.

**Table Sx.** Performance Metrics using MERIT hydrography GRADES daily routed *Q* for the *Q* prior on 17 SWOT-simulated rivers. There is good median improvement across all four metrics if the Missouri river is removed, however the Missouri was so poorly estimated that error metrics across all rivers show net performance degradation. Mean scores are also generally worse, indicating significant outliers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Median** | **Mean** | **Median (exc. Missouri rivers)** | **Mean (exc. Missouri rivers)** |
| **NSE % change** | -1.80 | -730 | 17.5 | -36.8 |
| **NRMSE % change** | 40.8 | 60.3 | -26.9 | 61.9 |
| **RRMSE % change** | 78.2 | 91.9 | -15.6 | 92 |
| **rBIAS % change** | 142 | 102 | 14.7 | 64.9 |

**References**

Feng, D., Gleason, C. J., Yang, X., & Pavelsky, T. M. (2019). Comparing discharge estimates made via the BAM algorithm in high‐order Arctic rivers derived solely from optical CubeSat, Landsat, and Sentinel‐2 data. *Water Resources Research*, *55*(9), 7753-7771.

Zou, Z., Xiao, X., Dong, J., Qin, Y., Doughty, R. B., Menarguez, M. A., et al. (2018). Divergent trends of open-surface water body area in the contiguous United States from 1984 to 2016. *Proceedings of the National Academy of Sciences*, *115*(15), 3810–3815. <https://doi.org/10.1073/pnas.1719275115>