Predicting hydrologic function with aquatic gene fragments

S. P. Good[1*], D. URycki[1], and B. C. Crump[2]

[1]Department of Biological & Ecological Engineering, Oregon State University, Corvallis, OR, United States.

[2]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, United States

*Corresponding author: Stephen Good (stephen.good@oregonstate.edu)

Key Points:

- The 'genohydrology' approach uses aquatic gene fragments to predict hydrologic function.
- Seasonal variation and recurrence intervals of monthly flows are predicted with 16S rRNA gene sequences.
- Genohydrology predictions outperform estimates based on area-scaled mean specific discharge values in similar rivers.

Abstract

Recent advances in microbiology techniques, such as genetic sequencing, allow for rapid and cost-effective collection of large quantities of genetic information carried within water samples. Here, we posit that the unique composition of aquatic DNA material within a water sample contains relevant information about hydrologic function at multiple temporal scales. In this study, machine learning was used to develop discharge prediction models trained on the relative abundance of bacterial taxa classified into operational taxonomic units (OTUs) based on 16S rRNA gene sequences from six large arctic rivers. We term this approach 'genohydrology,' and show that OTU relative abundances can be used to predict river discharge at monthly and longer timescales. Based on a single DNA sample from each river, the average Nash-Sutcliffe efficiency (NSE) for predicted mean monthly discharge values throughout the year was 0.84, while the NSE for predicted discharge values across different return intervals was 0.67. These are considerable improvements over predictions based only on the area-scaled mean specific discharge of five similar rivers, which had average NSE values of 0.64 and -0.32 for seasonal and recurrence interval discharge values, respectively. The genohydrology approach demonstrates that genetic diversity within the aquatic microbiome is a large and underutilized data resource with benefits for prediction of hydrologic function.

Plain Language Summary

An important task in water resources is prediction of the discharge in rivers and streams at locations where there are no direct measurements. In this study, we show that the flow in a river can be predicted based only on the bacteria that are present in the river. Because different flow conditions create environments in which different groups of bacteria grow, measurements of the diversity of the bacteria community can be used for hydrologic purposes. We call this approach 'genohydrology' and explore different discharge predictions based on streamwater bacteria composition.

1 Introduction

A core objective of contemporary hydrology is the prediction of discharge in un-gauged streams and rivers (Seibert & McDonnell, 2013; Sivapalan et al., 2003). While much success had be achieve through development of many hydrologic models for this purpose, the accurate calibration of these models often requires some minimum quantity of discharge measurements, and equifinality can cause to ambiguity in predictions even if measurements are present (Beven, 2006). When faced with inadequate direct measurements of discharge from a study catchment, the collection of some other type of information-dense dataset during short field campaigns (i.e., 'soft data') can be remarkably useful in understanding hydrologic function (Seibert & McDonnell, 2013). In this study, we explore a new type of hydrologic information: the DNA of aquatic microbes carried in a stream or river, which we evaluate as an emergent property of a catchment as a whole that is useful for quantitative predictions of discharge. This use of DNA-derived information differs from earlier applications of DNA as a hydrologic tracer, e.g., (Dahlke et al., 2015), in which synthetic DNA was released and recaptured downstream. In our application, we examine naturally occurring aquatic bacteria DNA fragments, and relate their variation between rivers to variations in flow regimes.

Here, we focus on bacterial diversity as reflected in 16S rRNA gene fragments from river samples (Crump et al., 2009), although other types of DNA derived data may hold similar potential. The 16s rRNA gene has been used in microbiology since the 1980's to classify bacteria into relative positions in the evolutionary order, i.e. phylum, class, order, family, etc. (Kolbert & Persing, 1999). Much of the bacterial diversity in rivers and streams originates from upslope soil environments and headwater streams (Crump, Adams, Hobbie, & Kling, 2007; Crump, Amaral-Zettler, & Kling, 2012), as well as from groundwater (Sorensen et al., 2013). Although aquatic

64 microorganisms are generally considered passive dispersers, in that dispersal is controlled by the
65 flow of water, evidence indicates that environmental variables have a strong influence in shaping
66 aquatic microbial communities (Crump et al., 2012; Whittaker & Rynearson, 2017). Lower costs
67 and recent advances in molecular biology methods have resulted higher quality freshwater
68 microbial DNA extraction and analysis (Li et al., 2015), making this type of information more
69 accessible to a wider research community for an increasing variety of applications.

70 Recent studies have linked bacterial community composition with hydrologic function,
71 with these studies primarily directed at understanding the microbial ecology of rivers and
72 streams. In the River Thames basin, Read et al. (2015) found a significant relationship between
73 bacterial community composition and cumulative stream length upstream of the community, and
74 concluded that physical and chemical characteristics of the river were less important than hydro-
75 geomorphic parameters in shaping microbial communities. Savio et al. (2015) measured 280
76 individual water quality parameters and found that the bacterioplankton community along the
77 Danube River continuum was primarily correlated with catchment characteristics, including river
78 kilometer, dendritic stream length, mean dendritic length, catchment size, and accumulated
79 dendritic distance. Other studies have linked microbial communities to river flow rate (Crump &
80 Hobbie, 2005; Doherty et al., 2017), and flow conditions have been used to model the abundance
81 of crucial bacterial populations, such as *Vibrio cholera* (Bertuzzo et al., 2008). Furthermore,
82 freshwater microbial communities have demonstrated seasonal shifts, with returns to
83 characteristic “core” seasonal communities (Crump & Hobbie, 2005; Doherty et al., 2017; Savio
84 et al., 2015). These studies suggest that the composition of microbial communities of rivers and
85 streams is influenced by the hydrology of the watersheds in which they are found.

Given that geographically and hydrologically diverse rivers have been shown to host characteristic, seasonally shifting, and predictable microbial communities, and that those communities are shaped by hydrological properties of a watershed, including discharge, we hypothesized that microbial community composition could be used to predict the hydrological characteristics of a basin. We term this approach 'genohydrology.' In this study, we use previously measured estimates of the bacterial community composition of six arctic rivers to make predictions of river flow regimes.

2 Materials and Methods

2.1 Arctic River Bacteria Community Composition

This study evaluates the bacterial and hydrologic characteristics of six arctic rivers: the Yukon, Kolyma, Yenisey, Mackenzie, Lena, and Ob (Figure 1). These rivers range in discharge from ~100 cubic kilometers per year to ~600 cubic kilometers per year, with basin sizes from 0.8 million square kilometers to 2.4 million square kilometers (Table 1). In total, these northern latitude rivers constitute 67% of the Arctic Ocean's drainage area (Holmes et al., 2012), with all six ranked in the world's top 50 largest rivers by discharge (Dai & Trenberth, 2002), and share broad similarities in their discharge patterns, geochemical composition, and bacterial community structure (Crump et al., 2009; Holmes et al., 2012). Discharge observations (Bodo, 2001a, 2001b) were compiled by the Global Runoff Data Center of the Federal Institute of Hydrology, Germany, and the International Hydrological Programme of the United Nations Educational, Scientific, and Cultural Organization. Only years with discharge data for all 12 months of the year were used, and across the six rivers there was an average of 33 years of data per river.

Table 1. Characterization of the six arctic river basins used in this study.

River	Total Area (km^2)	Annual Discharge (km^3/yr)	Gauge Latitude (degrees N)	Gauge Longitude (degrees E)	Discharge (Years)
Yukon	831386	203.2	61.93	-162.88	21
Kolyma	526000	99.3	68.73	158.27	16
Yenisey	2440000	577.4	67.43	86.48	59
Mackenzie	1660000	288.3	67.46	-133.75	21
Lena	2460000	486.1	72.37	-126.80	18
Ob	2430000	397.3	66.63	-66.60	64

The prediction of two key hydrologic flow quantities [m^3/s] was evaluated: (1) the average monthly discharge as it varies throughout the year, and (2) the average discharge expected at different recurrence intervals. Observed monthly discharge volumes were estimated by averaging across all years with 12 months of discharge data (see Table 1 for number of years in each gauge record). Discharge values at different recurrence intervals used data from all months of the year for estimates at 20 logarithmically spaced intervals spanning 0.1 to 10 years, with the recurrence period (T_x) calculated as $T_x=1/P_x$, where P_x is the probability that a discharge of x will be exceeded in the observed record (L. K. Read & Vogel, 2015). Observed flows were compared to predicted flows (see below) using both root means squared errors (RMSE) and Nash-Sutcliffe Efficiently (NSE) metrics. RMSE, which can range from 0 for a perfect model fit to positive infinity, captures both model bias and precision. NSE values can range from positive 1, for a perfect model fit, to negative infinity, with an NSE of zero occurring when predictions are as accurate as the mean of the observed data. Values of the RMSE and NSE for predictions were calculated for each river separately based on either the 12 monthly means or the 20 return intervals, and then averaged across all six rivers.

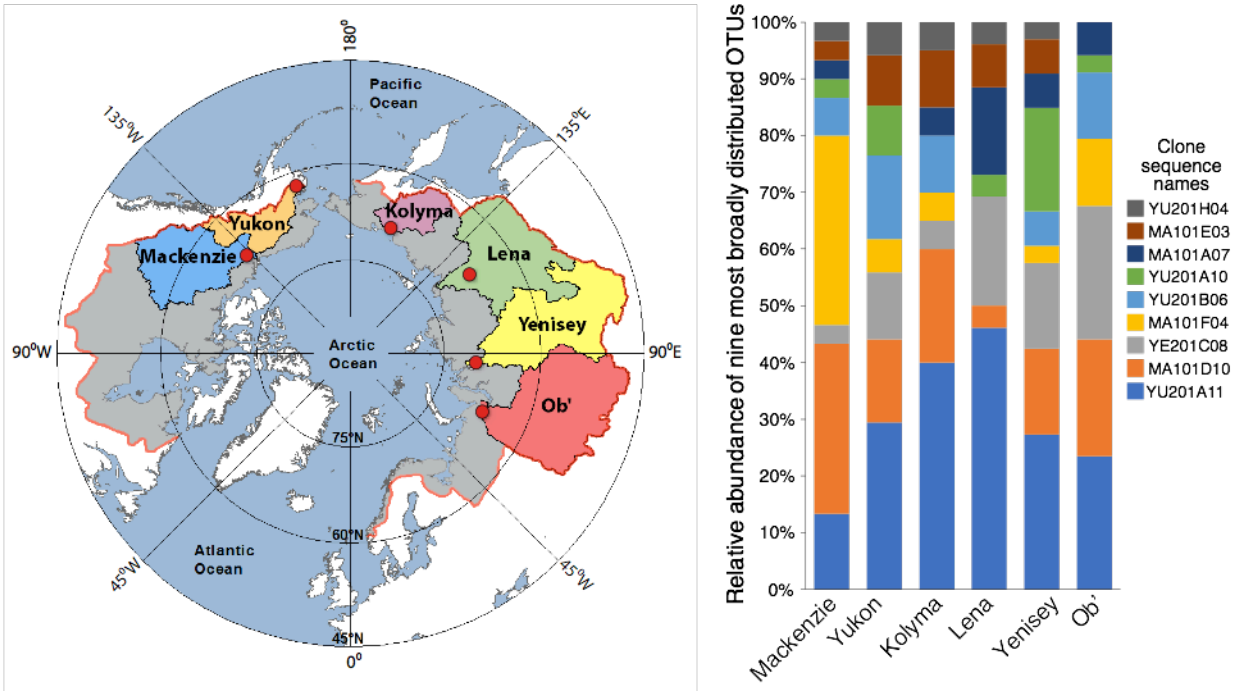

Figure 1 (Left) The watersheds of the six arctic rivers used in this study. Rivers were sampled (red circles) near their outlets into the Arctic Ocean in June of 2004 to obtain the (Right) relative abundances of nine broadly distributed OTUs, based on the number of 16S rRNA gene sequences in clone libraries prepared from DNA samples.

This study is based on DNA samples collected though the Pan-Arctic River Transport of Nutrients, Organic Matter and Suspended Sediments (PARTNERS) program (Holmes et al., 2012), which focused on collection water samples throughout the arctic with consistent sampling and analytical methods. Though water samples were obtained throughout the year through the PARTNERS program, only samples from June of 2004 were analyzed in detail for bacterial community composition (Crump et al., 2009). The composition of bacterial communities in these six rivers was measured in samples collected by the United States Geological Survey (USGS) National Research Program and Alaska Water Science Center, Canada's Department of Indian Affairs and Northern Development, the South Russian Centre for Preparation and Implementation of International Projects, and the Northeast Science Station in Russia (Crump et al., 2009). Water samples were taken from the river mouths following USGS sampling protocols

(Striegl, Aiken, Dornblaser, Raymond, & Wickland, 2005) as cross-sectionally integrated, flow-weighted water samples during June 2004.

Community composition was assessed using DNA sequencing of PCR-amplified and cloned bacterial 16S rRNA genes (i.e., clone library sequencing) (Crump et al., 2009; Crump & Hobbie, 2005). Phylogenetic distances between sequences were calculated with DNADIST using the Jukes-Cantor model, and the DOTUR application was used to group sequences into operation taxonomic units (OTUs) based on 97% DNA sequence similarity (Crump et al., 2009). Taxonomic assignments were made using the Ribosomal Database Project naïve Bayesian rRNA classifier tool using a confidence threshold of 80% (http://rdp.cme.msu.edu). Each OTU in this dataset represents a group of closely-related bacterial species found in each river, and the number of clones belonging to each OTU in each river were published as supplementary Table S2 of Crump et al., (2009). A total of 148 different OTUs were identified, and nine of these OTUs were present in at least five of the six rivers. These nine OTUs represented 20% to 34% of clone library sequences from each river, and comparison with the GenBank global database showed that they are common in freshwater systems world-wide (Crump et al., 2009). The relative abundances of these nine OTUs in each river (Figure 1) were used as input data for model calculations. These nine OTUs were selected because they appear in most (at least 5 of 6) of the studied arctic rivers, and the variation in their relative abundances was explored for predictive purposes.

2.2 Genohydrology Prediction Method

In this study a common machine learning technique, support vector regression (SVR) is used to map measured OTU abundances to hydrologic properties. SVR is a machine learning technique with few parameters (regularization C and kernel width ε) that is able to achieve

results that match or surpass neural-network approaches with minimal tuning (Smola, Sch, & Schölkopf, 2004). In particular, linear SVR is useful when feature counts (here the nine OTUs) exceeds the number of samples (here the six rivers). Linear SVR from the python SciKit-Learn machine learning library (Pedregosa et al., 2012) was used to construct predictor functions, $f_{j,t}(\ )$, that take as input $\mathbf{x}_j$, the array of normalized OTU counts in target river j, and returns as output the estimated log fractional discharge anomalies, $\hat{y}_{j,t}$, during an individual month or recurrence interval t. Discharge values were estimated based on models trained with both the absolute discharge and the specific discharge for each river, with specific discharge then scaled by basin area to estimate a final absolute discharge value.

Predictor functions were trained using an m by n input matrix of the normalized OTU counts for the other rivers (with m the five rivers used in training and n the nine most common OTUs), and m output values of the log anomalies in the discharge in the other five rivers during period t. Anomalies were calculated relative to the log mean discharge of the five other rivers, and thus the predictor functions, $f_{j,t}(\ )$, do not include any information of the discharge in river j during period t or any other time period. Observed log discharge anomalies, $y_{k,t}$, in each river k ($k \neq j$, with j the target river) for interval t that were used for training are calculated as

$$y_{k,t\,(\mathrm{AD})} = \ln\left(d_{k,t}\right) - \ln\left(\frac{1}{m}\sum_{i=1}^{m} d_{i,t}\right) \tag{1a}$$

$$y_{k,t\,(\mathrm{SD})} = \ln\left(\frac{d_{k,t}}{A_k}\right) - \ln\left(\frac{1}{m}\sum_{i=1}^{m} \frac{d_{i,t}}{A_i}\right) \tag{1b}$$

where A_i is the area of catchment i. Equation 1a gives the derived log anomaly of absolute discharge (AD), and Equation 1b gives the SD derived log anomaly of the specific discharge (SD). Note that the m rivers in the summation does not include the target river j.

The five sets of normalized OTU data from the non-target rivers and the five values of $y_{k,t}$ for interval t were used to train the SVR predictor function $f_{j,t}(\ )$. Note that the j subscript in $f_{j,t}(\ )$ specifies that this is the predictor function trained with flow and OTU data from all the rivers that are not j, and is therefore unique to river j because it is the only predictor function that will not contain data from river j. The t subscript in $f_{j,t}(\ )$ signifies that each time interval is predicted independently, in that the same summer OTU is repeatedly used to predicate each separate month and return interval discharge. In the training of the predictor function, OTU data is passed to possible predictor functions to produce estimates of the expected discharge anomaly in that river, i.e. $f_{j,t}(\mathbf{x}_k) = \hat{y}_{k,t}$. The SVR approach seeks the hyperplane, or series of hyperplanes, that have the largest separation between sets of training data (Pedregosa et al., 2012), with a larger margin typically associated with smaller training errors: $\hat{y}_{k,t}$-$y_{k,t}$. Note that separate predictor functions and resulting predictions, were created for both the absolute ($\hat{y}_{j,t\ (\mathrm{AD})}$) and specific ($\hat{y}_{j,t\ (\mathrm{AD})}$) discharge based approaches.

Once the form of a $f_{j,t}(\ )$ has been determined, the OTU data from the target river is passed through the predictor function to produce the SVR estimate of discharge anomalies in the target river j relative to average of the other non-target rivers, i.e. $f_{j,t}(\mathbf{x}_j) = \hat{y}_{j,t}$. The final predicted discharge values are then calculated as

(2a) $$\hat{d}_{j,t} = e^{\hat{y}_{j,t\ (\mathrm{AD})}} \left(\frac{1}{m}\sum_{i=1}^{m} d_{i,t}\right)$$

(2b) $$\hat{d}_{j,t} = e^{\hat{y}_{j,t\ (\mathrm{SD})}} \left(\frac{1}{m}\sum_{i=1}^{m} \frac{d_{i,t}}{A_i}\right) A_j,$$

where Equation 2a gives the predicted discharge based on the absolute discharge and Equation 2b gives the predicted discharge based on the specific discharge. The SciKit-Learn linearSVR routine was used to train each $f_{j,t}(\ )$ function with the regularization penalty C was set as 60 for

the absolute discharge models and 7.1 for the specific discharge models, where these were values selected through trail and error to reduce overall error. For all models the ε value set to 0, as suggested in the SciKit documentation. Training the SVR function to predict log anomalies bounds the exponentially transformed discharge predictions in equation 2 to positive values, and defaults predictions in river *j* to the mean of the other five rivers when $f_{j,t}(\ )$ carries no information and predictions $\hat{y}_{j,t}$ approach zero.

In this study, discharge data from five of six rivers was used to develop the predictive models, which were then used with remaining river OTU distribution for a leave-one-out validation approach. All genohydrology predictions were compared to both observations from river gauges and to predictions obtained using the mean of the five non-target rivers to estimate the discharge in river *j* for interval *t*, as well as to predictions obtained using the mean specific discharge of the five non-target rivers multiplied by the area of the target basin. Our comparison with the mean of the other, non-target, rivers was not intended to suggest that this is good hydrologic practice, but only to assess the added information that the genohydrology approach brings. In cases where OTU data hold no relation to true discharge values, or where our SVR approach cannot discern this relation, the genohydrology predictions will result in error statistics similar to those obtained when comparing the mean discharge of the non-target rivers to that of the target river.

3 Results

After training both the absolute and specific discharge based models, we compare the predicted monthly discharge values from both models against the mean of the five other non-

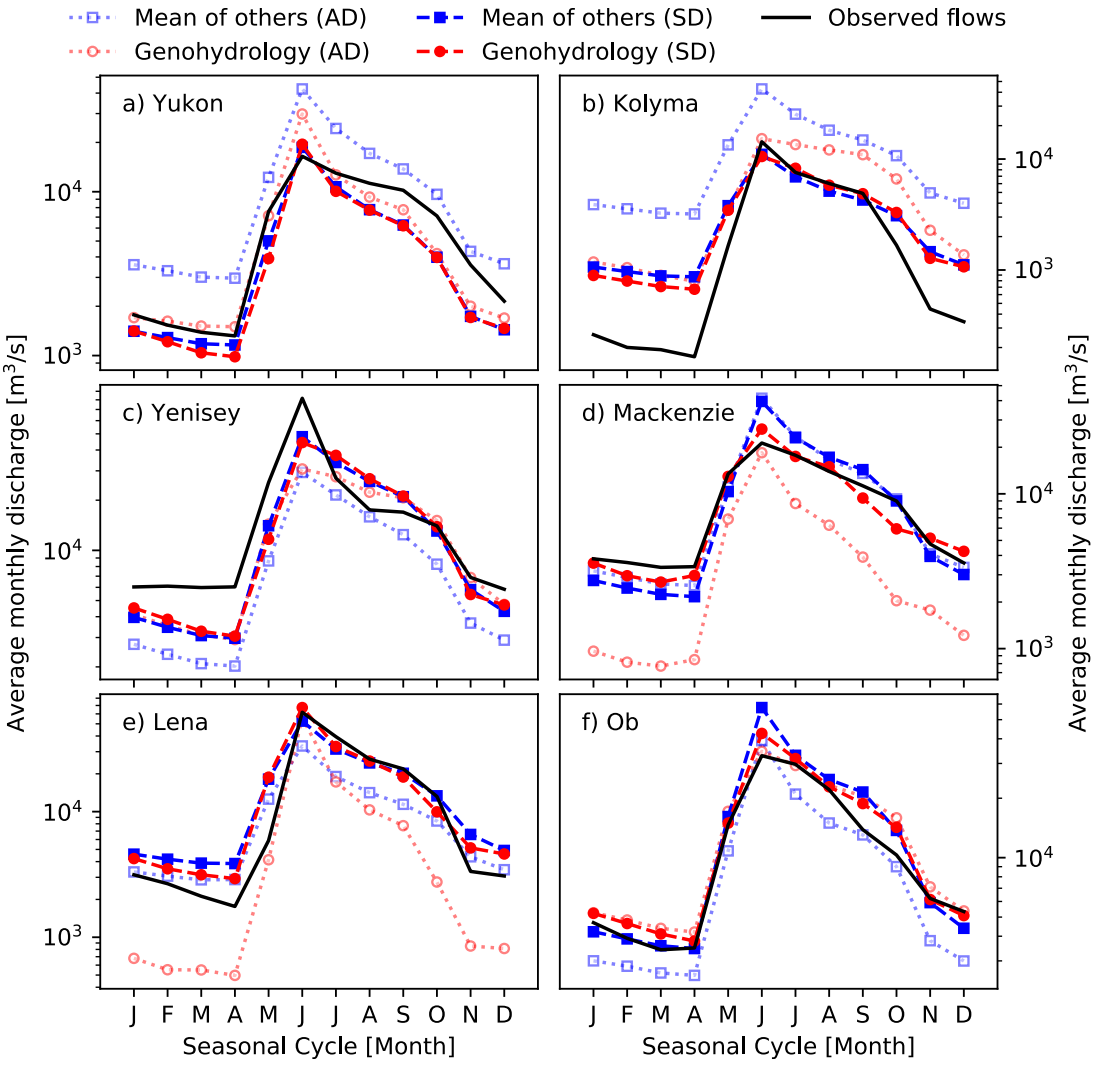

Figure 2. Average monthly discharge in six arctic rivers (a-f). Genohydrology estimated discharge values (circles), and the mean discharge of the other five rivers (squares) are shown based on both absolute discharge (AD) and specific discharge (SD).

target rivers, the area-scaled means specific discharge of the five other non-target rivers, and the observed discharge values (Figure 2). Error metrics for each of the four predictions methods across all months of the year are listed for each river (Table 2). Similarly, we compare the two genohydrology approaches with the two approaches based on the means of the non-target rivers and with the observed data for recurrence intervals from 0.1 to 10 years (Figure 3). Error metrics for each of the four predictions methods across all of the different recurrence intervals are listed for each river (Table 3). Additionally, we also show the cross-plot of each prediction method

232 with observed values for both the monthly flows and the different recurrence intervals (Figure 4).

233

234 **Table 2.** Error statistics for predicted average monthly flows using the genohydrology approach
235 and the mean other rivers based on both absolute discharge (AD) and specific discharge (SD).
236 The best preforming approach is shown in **bold.**

River	Root Mean Squared Error (m^3/s)				Nast-Sutcliffe Efficiency			
	Genohydrology		Mean Of Others		Genohydrology		Mean of Others	
	(AD)	(SD)	(AD)	(SD)	(AD)	(SD)	(AD)	(SD)
Yukon	4084	2485	8622	**2202**	0.35	0.76	-1.97	**0.81**
Kolyma	3463	1385	11818	**1379**	0.31	0.89	-7.00	**0.89**
Yenisey	15245	12338	16313	**10920**	0.45	0.64	0.37	**0.72**
Mackenzie	5313	**1822**	6179	5754	0.24	**0.91**	-0.03	0.11
Lena	9661	**4744**	11386	5255	0.72	**0.93**	0.61	0.92
Ob	2724	**3431**	4083	7655	0.93	**0.88**	0.83	0.42
Average:	6748	**4367**	9734	5527	0.50	**0.84**	-1.19	0.64

237

238 On average, the seasonal discharge predictions in each of the six rivers using the specific
239 discharge trained genohydrology approaches showed a clear improvement in both the root mean
240 squared error (RMSE) and Nash-Sutcliffe efficiency (NSE) over the mean flow of the other, non-
241 target, rivers (Table 2) using both the absolute and specific discharge, and over the absolute
242 discharge trained genohydrology approach. The addition of bacterial information resulted in an
243 average RMSE of 4367 m^3/s, representing a decrease of 21% in the RMSE from predictions
244 based on the area-scaled mean specific discharge of the non-target rivers. While on average the
245 genohydrology improved RMSE values, there were specific months in specific rivers where
246 genohydrology predictions were worse than those predicted from observations of the average
247 specific discharge in the other five rivers (Figure 2). In rivers where the RMSE was best based
248 on the mean of other specific discharge values (the Yukon, Kolyma, and Yenisey rivers), the
249 genohydrology was only slightly worse, with an average difference in RMSE for these rivers of
250 ~600 m^3/s. However, when the genohydrology approach was best (the Mackenzie, Lena, and Ob

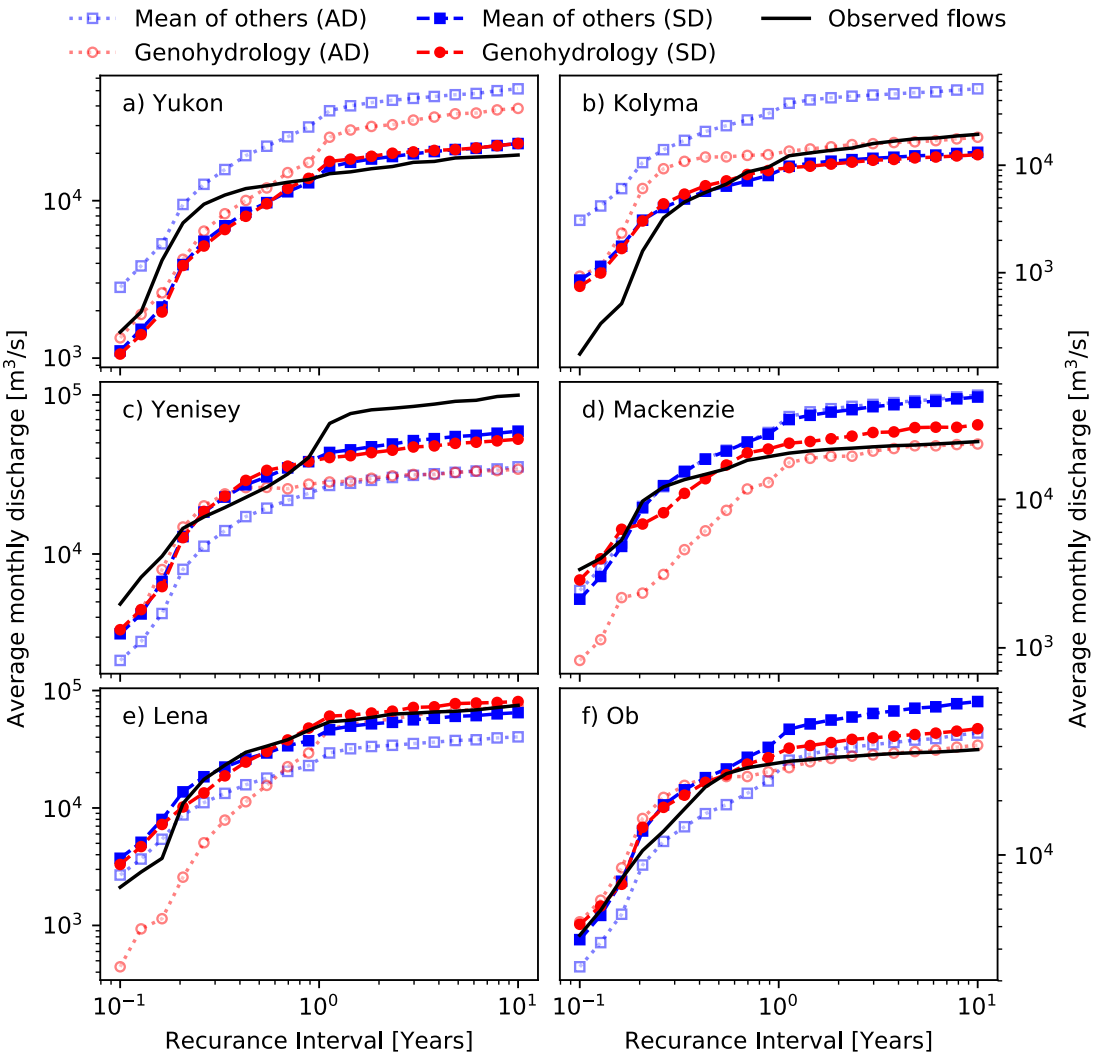

Figure 3. Average discharge for different return intervals in six arctic rivers (a-f). Genohydrology estimated discharge values (circles), and the mean discharge of the other five rivers (squares) are shown based on both absolute discharge (AD) and specific discharge (SD).

rivers) its improvement over the area scaled mean of the specific discharge of the over rivers was much larger (~2900 m^3/s).

The average NSE value of monthly discharges estimated without the bacteria data and only based on the mean of non-target rivers specific discharge was -1.19. This negative value signifies that using a single, average value of the observed flow in the target river across all months, which by definition gives an NSE of zero, would be more accurate than using the mean

257 monthly discharge values of the other, non-target, rivers. If the average specific discharge of the
258 non-target rivers is scaled by the basin area of the target river, the average NSE rises to 0.64.
259 When the bacterial information is also included with the specific discharge, the average NSE
260 value rises to 0.84 and ranged from 0.93 to 0.64 for individual rivers, with predictions in all
261 rivers greater than zero. Predictions based on the state-of-the-art distributed hydrologic model
262 (VIC) of discharge in the Lena (NSE of 0.96), Yenisey (NSE of 0.96) and Ob (NSE of 0.92)
263 (Troy, Sheffield, & Wood, 2011), are higher than our monthly genohydrology predictions.

264

265 **Table 3.** Error statistics for predicted monthly flows of different return intervals using the
266 genohydrology approach and the mean of the five other rivers based on both absolute discharge
267 (AD) and specific discharge (SD).

River	**Root Mean Squared Error (m^3/s)**				**Nast-Sutcliffe Efficiency**			
	Genohydrology		Mean Of Others		Genohydrology		Mean of Others	
	(AD)	(SD)	(AD)	(SD)	(AD)	(SD)	(AD)	(SD)
Yukon	11148	2958	20336	**2636**	-3.11	0.71	-12.68	**0.77**
Kolyma	**3179**	3639	22651	3283	**0.77**	0.70	-10.78	0.75
Yenisey	39077	28008	39358	**24534**	-0.26	0.35	-0.28	**0.50**
Mackenzie	4941	**4259**	15460	14357	0.47	**0.61**	-4.15	-3.44
Lena	9451	**5544**	22105	6206	0.85	**0.95**	0.19	0.94
Ob	**2924**	6650	5534	18498	**0.94**	0.69	0.78	-1.41
Average:	11789	**8510**	20907	11586	-0.06	**0.67**	-4.48	-0.32

268

269 Predictions of the discharge across return intervals ranging from 0.1 to 10 years using the
270 specific discharge genohydrology approach were also better on average than similar predictions
271 based on the mean of the other rivers (Table 3). When the bacterial information was included the
272 RMSE decreased by 26%, with the average dropping from 11586m^3/s to 8510m^3/s. Similar to the
273 seasonal predictions, even though there was a large improvement overall, predictions of
274 individual return intervals in individual rivers were at times worse than predictions from the
275 mean of the non-target rivers (Figure 4). However, as above, decreases in RMSE for the specific
276 discharge based genohydrology approach over the area scaled mean specific discharge were

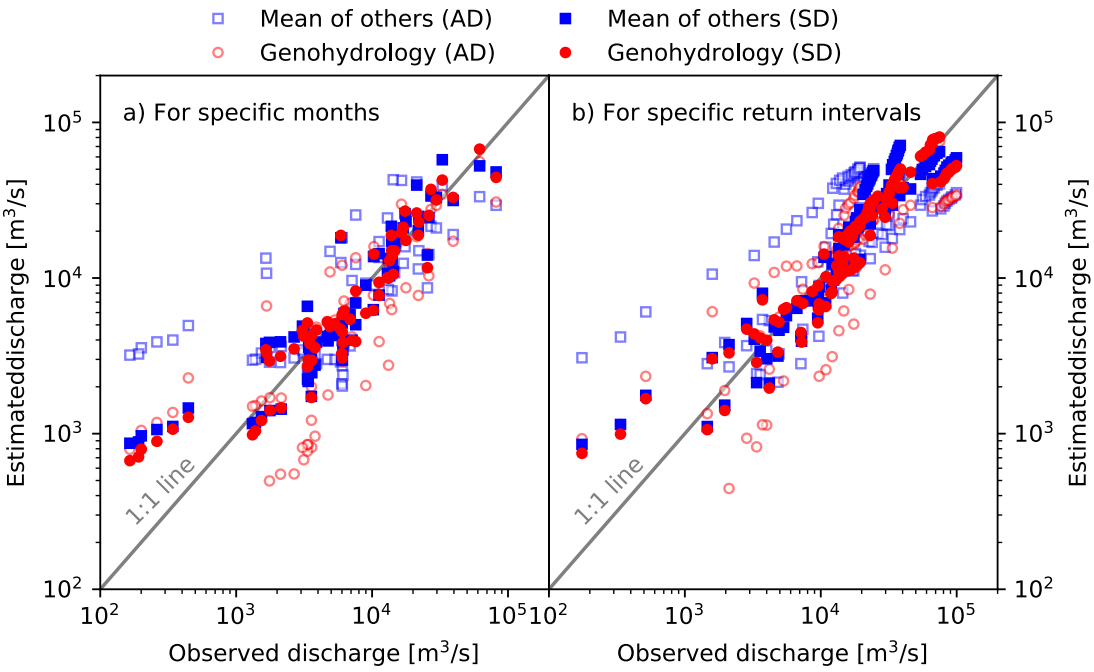

Figure 4. Cross-plot of observed and estimated river discharge for (a) specific months, and (b) for specific return intervals using the different approaches.

much larger than increases in RMSE. Interestingly, there were two specific cases (the Kolyma and Ob rivers) where the absolute discharge based genohydrology approach preformed best.

On average, the predicted NSE values for return intervals improved when adding the bacterial data (Table 3), though predictions for this hydrologic quantity were less accurate then for predictions of monthly means. With the exception of the specific discharge based genohydrology approach, all average NSE values were negative. As with the monthly mean predictions, the differences between the specific discharge genohydrology approach and the area scaled specific discharge mean of others were strongly skewed. Adding the bacteria community information either resulted in in large improvements or small weakening in predictions.

All the DNA samples used in the study were collected during the month of June, but predictions were made for all months in order to evaluate the utility of summer bacterial community for predictions during other periods. When viewed at the monthly timescale,

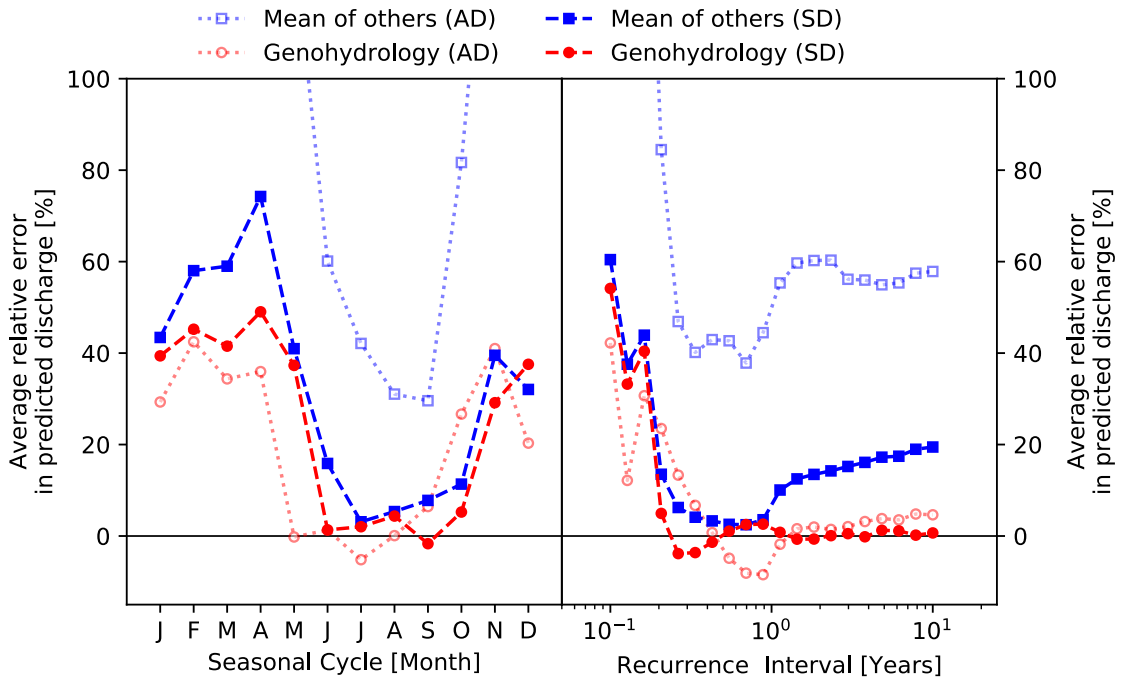

Figure 5. Average relative error in predictions of (left panel) monthly discharge values and (right panel) the discharge for different recurrence intervals.

predictions from June to September were very accurate (Figure 5) with relative errors near zero. Predictions in the low-flow, colder months showed larger errors and were biased high. On average, monthly predictions using the non-target means were biased high during all months for both specific and absolute discharge means. The overall relative error in these approaches also decreased in the summer. For the return intervals, all approaches had larger errors at shorter time intervals than at longer ones. Below 0.5 years, all approaches were biased high. At time scales larger than a year, the two genohydrology approaches demonstrate very low relative errors in predicted discharges.

4 Discussion

The objective of this study was to explore the hydrologic information contained within aquatic bacterial DNA fragments. While multiple previous studies (D. S. Read et al., 2015; Savio

et al., 2015) have suggested that bacterial composition is influenced by hydrologic flows, here we attempted quantitative macro-scale flow predictions based on this genetic information. When compared to observed flows, our accuracy varied considerably between the six rivers examined and between the hydrologic quantities that were predicted. However, we demonstrated overall improvement over predictions based only on flow information from other similar rivers.

While the accuracy of the genohydrology approach for these arctic rivers is below that obtained from advanced hydrologic models, this study demonstrates that non-trivial hydrologic information can be obtained from river DNA. In comparison of the absolute and specific discharge approaches, it was expected that the inclusion of basin area would be highly informative. It is expected that other basic hydrologic properties such as basin-averaged precipitation would improve our results further. However, the objective of this study was to test if bacteria alone, without any other ancillary data about the hydrologic system, carry hydrologic information. There are many possible genohydrology approaches for incorporation of DNA derived data into predictive macro-scale models, and this study is only an initial investigation. We expect the accuracy of genohydrology approaches to improve with more extensive sampling of aquatic bacterial DNA across a larger range of river flow regimes.

For specific rivers, in the cases when the genohydrology approach was not an improvement over the mean of the other rivers, the decrease in model fit was small. Conversely, when the genohydrology approach did improve over the mean of the others, the improvement was much larger. This skewness is likely caused by the fact that genohydrology approach is constructed to predict log relative anomalies from the means of the others (either specific or absolute discharge). Thus, if the DNA carries little information, $\hat{y}_{j,t}$ approaches zero, and predictions do not deviate strongly from the means of the other training rivers. However, when

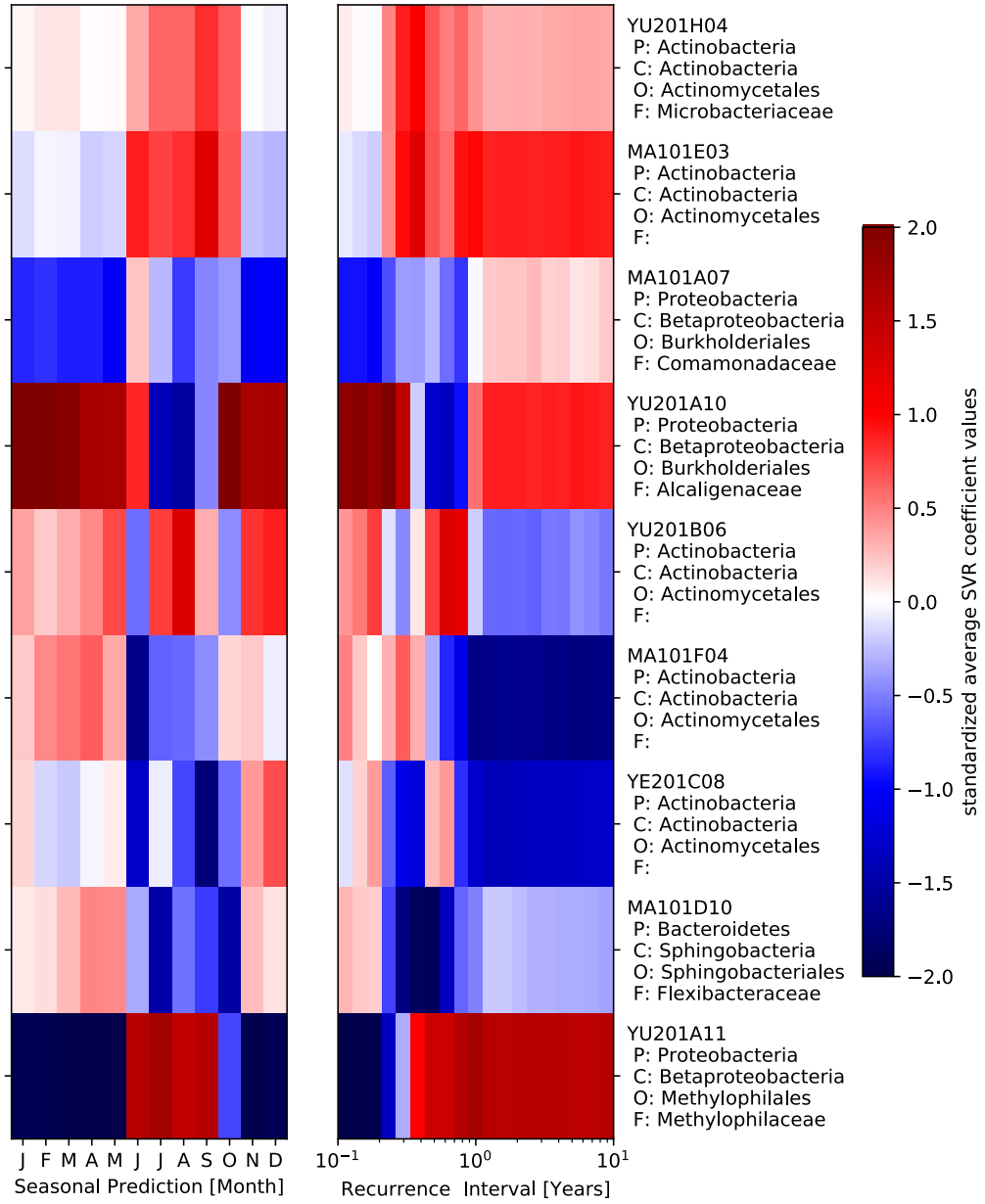

Figure 6. Average of the standardized SVR regression coefficients used in prediction of discharge for different months of the year (left panel), and for different recurrence intervals (right panel). The phylum (P), class (C), order (O), and family (F) of OTUs are listed when known.

the DNA does contain information about the hydrologic system, these improvements can be quite large.

The genohydrology approach was more successful in predicting average monthly flows than predicting flows associated with different return periods. This suggests that the average

328 seasonal variations in river conditions are more influential on bacterial community structure than
329 discharge associated with events of different frequencies. However, when looking at the
330 discharges associated with return intervals greater than one year, the accuracy of our approach
331 improved. These larger discharge values, which occur less frequently, are most likely to occur
332 during the summer months when they do occur. Higher accuracy during summer months and at
333 longer return intervals is likely due to the fact that the DNA was collected in summer. It is
334 possible that winter sampling of DNA would yield improved predictions of discharges associated
335 with winter months and smaller return intervals.

336 The OTU-based genohydrology models used in this study were created using Support
337 Vector Regression, though other machine learning techniques may be applicable. Machine
338 learning techniques can be prone to both over- and under- fitting (Pedregosa et al., 2012), and the
339 removal of superfluous information via data reduction approaches aids in fitting. Given the small
340 number of rivers examined here, we focused on OTUs that appeared in five of six surveyed
341 rivers. We also examined prediction accuracy using the OTUs that appeared in all six rivers
342 (only three OTUs total), and on the OTUs that appeared in less than five of the rivers. Both cases
343 resulted in much worse predictions (results not shown), suggesting that when either too few
344 features or too few samples are used, prediction accuracy decreases. Given the limited number of
345 sampled rivers, we employed a leave-one-out cross validation approach of training prediction
346 models with OTU and flow data from five rivers and testing this on the sixth. This resulted in
347 16% of the observations being used for validation. Further studies, based on DNA information
348 from more rivers may wish to use a higher percentage of observations for cross validation.

349 The six rivers in this study are all found at northern latitudes, and they share broadly
350 similar climate (Arctic), vegetation (tundra and taiga), and natural bacteria communities (Crump

et al., 2009). At present, it is unclear how widely applicable the OTU based prediction models derived here are, because collection methods and DNA analysis varies significantly between surveys of aquatic microbial communities. Comparison of standardized regression coefficients (Figure 6) allows us to diagnose the stability of our prediction models for different prediction intervals. Each vertical column of Figure 6 represents an average of six prediction models. There is some stability in the average prediction model across different prediction months or recurrence intervals. In the case of the monthly coefficients, the summer coefficients often have a different sign than the winter coefficients, suggesting that a different set of OTUs are most informative of flows in different seasons. For the discharge predictions at different recurrence intervals, an inflection point occurs at one year, with distinct sets of coefficients for models at greater than and less than one year. Furthermore, the sub-year return interval coefficients more closely match those of monthly prediction values during wither months only. This is consistent because summer months and longer recurrence intervals both represent periods associated with larger discharge values.

Comparison of standardized average regression coefficient values at different prediction intervals also provides some insight into which bacterial taxa are likely associated with which type of flow. In the models explored here, a positive (or negative) SVR regression coefficient corresponds to larger (or smaller) discharge predictions when those bacteria are more abundant. For both seasonal flow and recurrence interval predictions greater than one year, the SVR regression coefficients had strong consistency in sign, and, to a lesser degree, in magnitude. However, given the limited number of rivers (six) examined here, and the fact that samples were only collected once, it remains difficult to associated specific OTUs with specific hydrologic patterns at this time.

5 Conclusions

In this study, we examined the suitability of using bacterial DNA fragments to predict seasonal discharge dynamics and the discharge expected at various return intervals. Our approach was successful in demonstrating that DNA-derived information, as captured in the relative abundance of different OTUs, contains information about discharge levels. Predictions of discharge volume improved once the OTU data was incorporated. While the number of rivers involved in this study (six), their sampling period (June only), and the sequencing approach (16S rRNA clone libraries), are somewhat limiting, further studies with more sampling points in space and time, as well as improved sequencing techniques will likely expand the applications and improve the precision of the genohydrology approach.

Acknowledgments, Samples, and Data

This work was partially supported by the U.S. National Science Foundation grants DEB-1457794 and DEB-1347042. DU acknowledges the support of XXX. All data used in this study is publicly available in previous publications. Discharge data is available at the National Center for Atmospheric Research (NCAR) at https://rda.ucar.edu/ (Bodo, 2001a, 2001b). DNA derived bacteria community composition was published in (Crump et al., 2009).

References

Bertuzzo, E., Azaele, S., Maritan, A., Gatto, M., Rodriguez-Iturbe, I., & Rinaldo, A. (2008). On the space-time evolution of a cholera epidemic. *Water Resources Research*, *44*(1). http://doi.org/10.1029/2007WR006211

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. http://doi.org/10.1016/j.jhydrol.2005.07.007

Bodo, B. (2001a). Monthly Flow Rates of World Rivers (except former Soviet Union), Version 1.3. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D61G0JGZ. Accessed 21 03 2017.* http://doi.org/10.5065/D6571974

Bodo, B. (2001b). Russian River Flow Data by Bodo, Enhanced, Version 1.1. *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds553.2/. Accessed 21 03 2017.*

Crump, B. C., Adams, H. E., Hobbie, J. E., & Kling, G. W. (2007). Biogeography Of Bacterioplankton In Lakes And Streams Of An Arctic Tundra Catchment. *Ecology*, *88*(6), 1365–1378. http://doi.org/10.1890/06-0387

Crump, B. C., Amaral-Zettler, L. A., & Kling, G. W. (2012). Microbial diversity in arctic freshwaters is structured by inoculation of microbes from soils. *The ISME Journal*, *6*, 1629–1639. http://doi.org/10.1038/ismej.2012.9

Crump, B. C., & Hobbie, J. E. (2005). Synchrony and seasonality in bacterioplankton

communities of two temperate rivers. *Limnology and Oceanography*, *50*(6), 1718–1729. http://doi.org/10.4319/lo.2005.50.6.1718

Crump, B. C., Peterson, B. J., Raymond, P. A., Amon, R. M. W., Rinehart, A., McClelland, J. W., & Holmes, R. M. (2009). Circumpolar synchrony in big river bacterioplankton. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21208–12. http://doi.org/10.1073/pnas.0906149106

Dahlke, H. E., Williamson, A. G., Georgakakos, C., Leung, S., Sharma, A. N., Lyon, S. W., & Walter, M. T. (2015). Using concurrent DNA tracer injections to infer glacial flow pathways. *Hydrological Processes*, *29*(25), 5257–5274. http://doi.org/10.1002/hyp.10679

Dai, A., & Trenberth, K. E. (2002). Estimates of Freshwater Discharge from Continents: Latitudinal and Seasonal Variations. *Journal of Hydrometeorology*, *3*(6), 660–687. http://doi.org/10.1175/1525-7541(2002)003<0660:EOFDFC>2.0.CO;2

Doherty, M., Yager, P. L., Moran, M. A., Coles, V. J., Fortunato, C. S., Krusche, A. V, … Crump, B. C. (2017). Bacterial Biogeography across the Amazon River-Ocean Continuum. *Frontiers in Microbiology*, *8*, 882. http://doi.org/10.3389/fmicb.2017.00882

Holmes, R. M., McClelland, J. W., Peterson, B. J., Tank, S. E., Bulygina, E., Eglinton, T. I., … Zimov, S. A. (2012). Seasonal and Annual Fluxes of Nutrients and Organic Matter from Large Rivers to the Arctic Ocean and Surrounding Seas. *Estuaries and Coasts*, *35*(2), 369–382. http://doi.org/10.1007/s12237-011-9386-6

Kolbert, C. P., & Persing, D. H. (1999). Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Current Opinion in Microbiology*, *2*(3), 299–305.

http://doi.org/10.1016/S1369-5274(99)80052-6

Li, P., Yang, S. F., Lv, B. B., Zhao, K., Lin, M. F., Zhou, S., … Tang, X. M. (2015). Comparison of extraction methods of total microbial DNA from freshwater. *Genetics and Molecular Research*, *14*(1), 730–738. http://doi.org/10.4238/2015.January.30.16

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://doi.org/10.1007/s13398-014-0173-7.2

Read, D. S., Gweon, H. S., Bowes, M. J., Newbold, L. K., Field, D., Bailey, M. J., & Griffiths, R. I. (2015). Catchment-scale biogeography of riverine bacterioplankton. *The ISME Journal*, *9*(2), 516–26. http://doi.org/10.1038/ismej.2014.166

Read, L. K., & Vogel, R. M. (2015). Reliability, return periods, and risk under nonstationarity. *Water Resources Research*, *51*(8), 6381–6398. http://doi.org/10.1002/2015WR017089

Savio, D., Sinclair, L., Ijaz, U. Z., Parajka, J., Reischer, G. H., Stadler, P., … Eiler, A. (2015). Bacterial diversity along a 2600?km river continuum. *Environmental Microbiology*, *17*(12), 4994–5007. http://doi.org/10.1111/1462-2920.12886

Seibert, J., & McDonnell, J. J. (2013). Gauging the Ungauged Basin: The Relative Value of Soft and Hard Data. *Journal of Hydrologic Engineering*, *20*(June), 130607193631000. http://doi.org/10.1061/(ASCE)HE.1943-5584.0000861

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., … Zehe, E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping

454 an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–
455 880. http://doi.org/10.1623/hysj.48.6.857.51421

456 Smola, A. J., Sch, B., & Schölkopf, B. (2004). A Tutorial on Support Vector Regression.
457 *Statistics and Computing*, *14*(3), 199–222.
458 http://doi.org/10.1023/B:STCO.0000035301.49549.88

459 Sorensen, J. P. R., Maurice, L., Edwards, F. K., Lapworth, D. J., Read, D. S., Allen, D., …
460 Williams, P. J. (2013). Using Boreholes as Windows into Groundwater Ecosystems. *PLoS*
461 *ONE*, *8*(7), e70264. http://doi.org/10.1371/journal.pone.0070264

462 Striegl, R. G., Aiken, G. R., Dornblaser, M. M., Raymond, P. A., & Wickland, K. P. (2005). A
463 decrease in discharge-normalized DOC export by the Yukon River during summer through
464 autumn. *Geophysical Research Letters*, *32*(21), L21413.
465 http://doi.org/10.1029/2005GL024413

466 Troy, T. J., Sheffield, J., & Wood, E. F. (2011). Estimation of the terrestrial water budget over
467 Northern Eurasia through the use of multiple data sources. *Journal of Climate*, *24*(13),
468 3272–3293. http://doi.org/10.1175/2011JCLI3936.1

469 Whittaker, K. A., & Rynearson, T. A. (2017). Evidence for environmental and ecological
470 selection in a microbe with no geographic limits to gene flow. *Proceedings of the National*
471 *Academy of Sciences of the United States of America*, *114*(10), 2651–2656.
472 http://doi.org/10.1073/pnas.1612346114

473