Science of the Total Environment 849 (2022) 157911

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

The streamwater microbiome encodes hydrologic data across scales

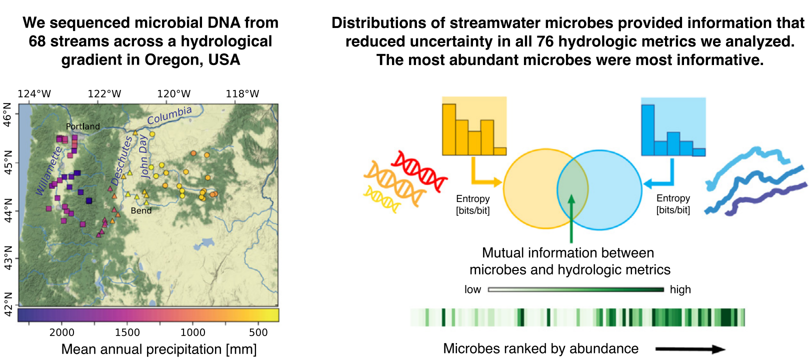
Dawn R. URycki [a,b,c,*], Maoya Bassiouni [d,e], Stephen P. Good [a,b], Byron C. Crump [f], Bonan Li [a,b,g]

[a] Water Resources Graduate Program, Oregon State University, USA
[b] Department of Biological and Ecological Engineering, Oregon State University, USA
[c] Department of Civil Engineering, University of Colorado Denver, USA
[d] Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Sweden
[e] Department of Environmental Science, Policy, and Management, University of California Berkeley, USA
[f] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, USA
[g] Department of Biological and Agricultural Engineering, University of Arkansas, USA

HIGHLIGHTS

- We sequenced microbial DNA from 68 streams across Oregon, USA.
- We used information metrics to quantify relationships between microbes and hydrology.
- Streamwater microbes encode information about all 76 hydrologic metrics we analyzed.
- An average of 9.6 % of common microbes were related to each hydrologic metric.
- Hydrologic information was concentrated among the most abundant microbes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ewa Korzeniewska

Keywords:
Watershed hydrology
16S metabarcoding
Information theory
Mutual information
Pacific Northwest

ABSTRACT

Many fundamental questions in hydrology remain unanswered due to the limited information that can be extracted from existing data sources. Microbial communities constitute a novel type of environmental data, as they are comprised of many thousands of taxonomically and functionally diverse groups known to respond to both biotic and abiotic environmental factors. As such, these microscale communities reflect a range of macroscale conditions and characteristics, some of which also drive hydrologic regimes. Here, we assess the extent to which streamwater microbial communities (as characterized by 16S gene amplicon sequence abundance) encode information about catchment hydrology across scales. We analyzed 64 summer streamwater DNA samples collected from subcatchments within the Willamette, Deschutes, and John Day river basins in Oregon, USA, which range 0.03–29,000 km^2 in area and 343–2334 mm/year of precipitation. We applied information theory to quantify the breadth and depth of information about common hydrologic metrics encoded within microbial taxa. Of the 256 microbial taxa that spanned all three watersheds, we found 9.6 % (24.5/256) of taxa, on average, shared information with a given hydrologic metric, with a median 15.6 % (range = 12.4–49.2 %) reduction in uncertainty of that metric based on knowledge of the microbial biogeography. All of the hydrologic metrics we assessed, including daily discharge at different time lags, mean monthly discharge, and seasonal high and low flow durations were encoded within the microbial community. Summer microbial taxa shared the most information with winter mean flows. Our study demonstrates quantifiable relationships between streamwater microbial taxa and hydrologic metrics at different scales, likely resulting from the integration of multiple overlapping drivers of each. Streamwater microbial communities are rich sources of information that may contribute fresh insight to unresolved hydrologic questions.

* Corresponding author at: North Classroom Building, Denver, CO 80217, USA.
E-mail address: dawn.urycki@ucdenver.edu (D.R. URycki).

http://dx.doi.org/10.1016/j.scitotenv.2022.157911
Received 7 March 2022; Received in revised form 15 July 2022; Accepted 4 August 2022
Available online 6 August 2022

1. Introduction

Hydrology spans earth and life sciences. The living environment is shaped by water and in turn shapes the hydrological cycle (e.g.,Dingman, 2015). A deeper knowledge of these interactions requires data that integrates processes at multiple spatial and temporal scales, beyond direct human observations. Meteorological, eddy covariance, stable isotopes, and remotely sensed earth observation data have made high-resolution hydrologic observations more widely available than ever before, through observation networks such as FLUXNET (Baldocchi et al., 2001), the National Ecological Observatory Network (NEON; Schimel et al., 2007), and the Critical Zone Collaborative Network (CZNet; https://criticalzone.org/). Despite this growing amount of data, the specific mechanisms and interactions of many drivers of hydrologic function are still not well understood. A fundamental process-based understanding of streamflow generation, for example, is necessary to predict water availability and adapt to changing climate and landcover conditions, yet the dynamics of streamflow sources and transit time distribution remain areas of active research (Blöschl et al., 2019). The ability to develop new hydrologic insight may not be limited by the amount of data, but instead by the type of data available for hydrologic study. Patterns that are not always apparent from data traditionally employed in hydrology may emerge when analyzed with biotic data that integrates information about the spatiotemporal dynamics of the hydrological cycle and its drivers (Seibert and McDonnell, 2002).

Microbial communities native to an environment (i.e., microbiomes) constitute a novel type of environmental data that comprise many thousands of taxonomically and functionally diverse groups. These communities are most often assessed taxonomically using DNA sequences of the phylogenetically informative 16S rRNA gene. This approach to assessing microbiomes is common in a range of research fields from oceanography to human health. The taxonomic composition of microbial communities is diagnostic of environment types (e.g., freshwater, soil, seawater, etc.) and is sensitive to perturbations, shifting species composition in response to changes in environment (Thompson et al., 2017). Moreover, these communities are often highly diverse and include hundreds of abundant taxa and many thousands of rare taxa, thus providing a rich dataset of information about biological responses to environmental conditions.

Streamwater microbiomes originate primarily from upslope soil, groundwater, and sediment (Crump et al., 2007, 2012; Sorensen et al., 2013; Hermans et al., 2019; Miller et al., 2021) and subsequently develop, through species-sorting and dispersal, in response to biotic factors such as predation and reproduction, but also to abiotic subsurface and environmental factors, including soil saturation and streamflow rate (Newby et al., 2009), water residence time and network connectivity (Hrachowitz et al., 2016), and interactions with sediment (Droppo et al., 2009). These microscale communities thus reflect a range of macroscale characteristics and processes that also influence and interact with catchment scale hydrology.

Spatial characteristics related to water residence time have been identified as the strongest correlates with streamwater microbial community composition, even among physicochemical factors, along the River Thames (Read et al., 2015) and the Danube River (Savio et al., 2015). Microbial community composition has also been linked to a broader range of landscape scale climatological and geomorphological characteristics (URycki et al., 2020), which are important drivers of hydrologic function. Recent hydrologic studies have employed microbial communities as tracers to elucidate groundwater recharge and flow paths (Sugiyama et al., 2018; Miller et al., 2021), and it has been suggested that microbial information may be useful for hydrologic prediction (Good et al., 2018). However, the breadth of hydrologic data encoded within microbiomes, and the timescales over which microbial communities may be informative, has yet to be quantified and explored.

One major challenge of employing microbial communities as hydrologic observations is identifying informative constituents among thousands or even millions of taxa subject to complex ecosystem interactions that remain poorly understood. The tools of information theory, based on Shannon's entropy (Shannon, 1948), capture linear and nonlinear relationships; analysis of information flows are thus a powerful framework to understand complex patterns in Earth systems science (Goodwell et al., 2020). Here, we sought to leverage information theory to explore the extent to which microbial communities inform the hydrology of the ecosystems in which they occur. We analyze microbial community samples from 64 gauged streams across three major watersheds in the state of Oregon, USA. Our objective is to quantify the information shared between microbial taxa and a set of hydrologic metrics that represent watershed function and characteristic water balance dynamics. Results of this analysis can contribute to a broader understanding of the relationships between hydrology and streamwater microbial communities and offer insights about the value of microbial communities as a novel source of information to answer open hydrological questions.

2. Data and methods

2.1. Study area

The Willamette, Deschutes, and John Day basins, three similarly sized adjacent watersheds, are exceptional in that together they span a wide range of ecoclimatic conditions yet all lie within close proximity in the state of Oregon, USA; although all three basins experience wet winters and dry summers characteristic of the Pacific Northwest, the effect diminishes eastward. Westernmost of the three, the Willamette Basin, encompassing 29,000 km^2 with a mean elevation of 560 m, is bounded by the Oregon Coast Range to the west and the Cascade Range to the east (United States Geological Survey, 2017). The Willamette Basin is the wettest and most temperate of the three watersheds, receiving 1640 mm precipitation annually (246 mm in January and 20 mm in July, on average) and with mean annual minimum and maximum temperatures of 4 °C and 15 °C, respectively (Fig. 1). The Deschutes Basin, encompassing 27,700 km^2 to the east of the Cascades, lies at a mean elevation of 1230 m and is considerably drier and cooler than the Willamette Basin, receiving just 530 mm precipitation annually (75 mm [Jan] and 14 mm [Jul]) and with mean annual minimum and maximum temperatures of 0 °C and 14 °C, respectively. To the east of the Deschutes Basin, the John Day Basin encompasses 20,500 km^2 and rises to a mean elevation of 1170 m. Climatically more similar to the Deschutes Basin, the John Day Basin receives 460 mm precipitation annually (53 mm [Jan] and 15 mm [Jul]) with mean annual minimum and maximum temperatures of 1 °C and 14 °C, respectively.

All three basins are generally oriented with water flowing south to north and ultimately drain to the Columbia River at the northern border of Oregon. Median winter (Jan, Feb, Mar [*JFM*]) discharge (50 % flow duration) at the outlet is 1260 m^3/s on the Willamette River, 178 m^3/s on the Deschutes River, and 52 m^3/s on the John Day River; whereas median summer (Jul, Aug, Sep [*JAS*]) discharge is 289 m^3/s on the Willamette River, 132 m^3/s on the Deschutes River, and 3 m^3/s on the John Day River (Risley et al., 2008). The Willamette Basin is the most developed of the three study basins (6 % developed area), containing the city of Portland at the mouth of the Willamette River, as well as some additional urban development along the Willamette Valley; the Deschutes Basin (1 % developed area), including the city of Bend, and the John Day Basin (<1 % developed area) are far less developed (United States Geological Survey, 2017).

2.2. DNA collection and sequencing

We collected DNA samples near active stream gages from 40 sites in the Willamette Basin and 21 sites in the Deschutes Basin between 21 July and 8 August 2017 and from 20 sites in the John Day Basin from 6 to 8 August 2018 (Fig. 1). This mid-summer sampling strategy provided reasonable assurance that samples represent baseflow conditions across all sites, given typically very dry summers in both the Mediterranean climate of western Oregon and semi-arid climate of eastern Oregon. We collected most samples from the approximate center of the waterway by lowering from a bridge a two-gallon bucket that had been sanitized at the beginning

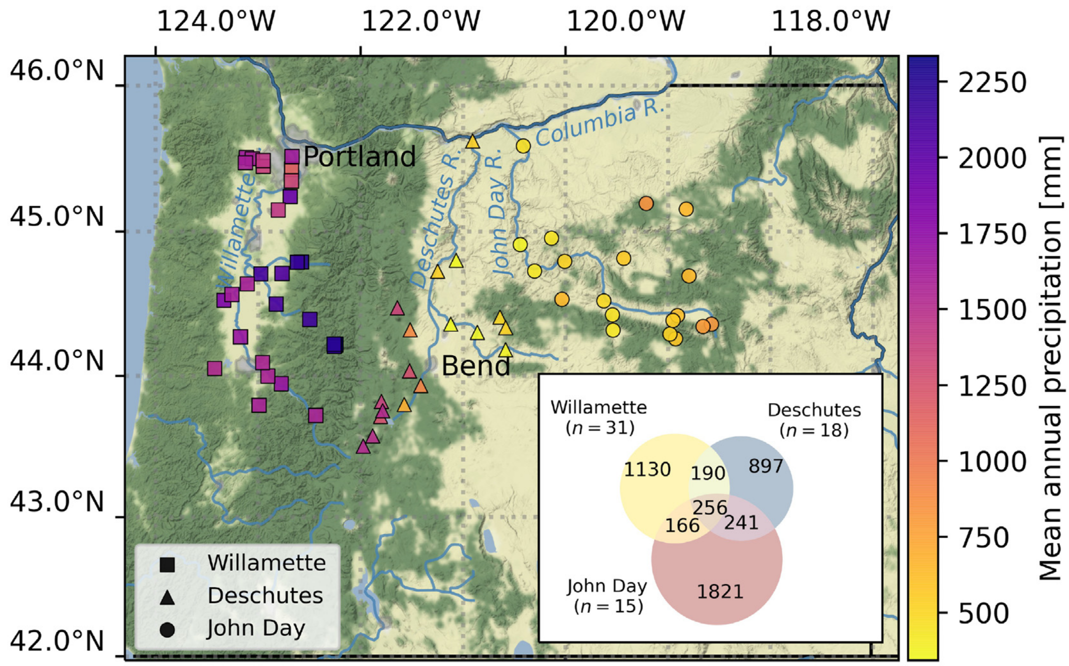


Fig. 1. Map of co-located stream gages and streamwater DNA sampling sites across Willamette (2017), Deschutes (2017), and John Day (2018) basins in Oregon, USA. Marker colors indicate mean annual precipitation in sample catchments (United States Geological Survey, 2017). Inset indicates number of unique and common microbial amplified sequence variants detected in *n* samples across each basin.

of the day and sample-rinsed between sites. We filtered and extracted DNA samples from collected streamwater as described in Crump et al. (2003). Although some DNA sampling protocols employ pre-filtering in an effort to exclude particle-associated microbial DNA, our sampling protocol includes no such pre-filtering. Consequently, samples we collected include a combination of aquatic bacterioplankton as well as microbes originating from soil-water, hyporheic, and benthic habitats (Crump et al., 2007). After extracting and isolating the DNA following common accepted protocols, we PCR-amplified 16S rRNA genes were with dual-barcoded primers targeting the V4 region (515f GTGCCAGCMGCCGCGGTAA, 806r GGACTACHVGGGTWTCTAAT; Caporaso et al., 2011) that were linked to barcodes and Illumina adapters following Kozich et al. (2013). We sequenced PCR products with Illumina Miseq V.2 paired end 250 bp sequencing. 16S rRNA gene amplicon sequences have been deposited in the NCBI Sequence Read Archive (SRA) under the bioproject accession number PRJNA642636 (https:// www.ncbi.nlm.nih.gov); experiment accessions SRX8627679 - SRX8627772. A detailed description of collection and processing of this dataset is found in URycki et al. (2020). To approximate even sampling depth, we rarefied the sequence dataset to 1450 sequences per sample. We selected a rarefaction threshold of 1450 sequences because it represented the largest tolerable loss of data while retaining as many samples as possible. This rarefaction cut 17 of 81 samples that either did not PCR amplify or produced fewer than 1450 sequences. Sequences were taxonomically classified with the SILVA 16S rRNA gene database v.132 (Quast et al., 2013). We analyzed these data as a matrix of the relative abundance (i.e., sequence counts) of each unique amplified sequence variant (ASV) detected at each site. In microbiome research, individual ASVs are considered distinct taxonomic groups, usually at the genus level; as such, we use the terms *ASV* and *taxon* interchangeably throughout this paper.

After rarefying the raw sequences, our dataset consisted of 4701 unique ASVs from 64 sample sites across the three watersheds. We detected 1742 unique ASVs across 31 samples in the Willamette Basin, 1584 unique ASVs across 18 samples in the Deschutes Basin, and 2484 unique ASVs across 15 samples in the John Day Basin (Fig. 1). To reduce computational and analytical complexity and explore more generalizable patterns, we focus our analysis on the subset of microbial taxa that were detected across all three river basins. A total of 256 ASVs were detected at least once in all three watersheds (Fig. 1), and within this subset we identify these unique ASVs by their abundance rank, with the most abundant ASV across all sites identified as *ASV 1* and the least abundant as *ASV 256* (Table S2).

2.3. Hydrologic metrics

We calculated hydrologic metrics from daily mean discharge records at 64 stream gages (Fig. 1). Study site stream gages spanned headwaters, tributaries, and outlets of the three major rivers and with the intention to sample the range of land use, landcover, and disturbance in each watershed. Stream gages were managed by the United States Geological Survey (USGS; U.S. Geological Survey, 2016), Oregon Water Resources Department (OWRD; Oregon Water Resources Department, 2021), or H.J. Andrews Experimental Forest (HJA; Johnson et al., 2020). We obtained available records of daily mean discharge (ft^3/s; converted to m^3/s) for each stream gage for the 10-year period preceding DNA sample collection (see *2.2 DNA collection and sequencing*). We obtained 10 years of data for 54 sites; 6.9–9.9 years of data for six sites, and <5 years of data for four sites (Table S1).

We sought to analyze a set of hydrologic metrics that would characterize discharge dynamics in our study area at different time and flow scales. We therefore calculated 76 hydrologic metrics across three categories: daily discharge, mean monthly discharge, and seasonal high and low flow durations. To describe current hydrologic conditions, we used daily mean absolute discharge [m^3/s] on to the date (t) of DNA sample collections and at lags of even numbers of days, up to 30 days, prior to sampling (Q_t, $Q_{(t-2\ days)}$, $Q_{(t-4\ days)}$, …, $Q_{(t-30\ days)}$) at all 64 stream gages. To characterize typical catchment conditions, we calculated mean monthly absolute discharge ($\overline{Q}_{mon}$) over the period of study for each month of the water year from October to September for the 60 sites for which we had >6.9 years of data. To capture more extreme hydrologic responses, we calculated seasonal high and low flow durations over the period of study for each of four seasons and annually beginning at the start of the water year ($Q_{P,\ s}$ for P = *5*- and *95* % exceedance probability [95th and 5th percentile, respectively] for seasons s = *OND*, *JFM*, *AMJ*, *JAS*, and *Ann* [annual]), again for the 60 sites for which we had >6.9 years of data. We normalized absolute discharge values by the sub-catchment area, derived from the StreamStats web application developed by the USGS (https://streamstats.usgs.gov/ss/; USGS, 2017), to obtain daily specific discharge ($q_{(t-n\ days)}$ [$m^3 km^{-2} s^{-1}$]), mean monthly specific discharge ($\overline{q}_{mon}$), and seasonal specific discharge flow durations ($q_{P,\ s}$).

Across sites, daily discharge generally decreased in the 30 days prior to sampling, ranging $Q_{(t-2\ days)} = 0 - 297\ m^3/s$, $Q_{(t-10\ days)} = 0 - 340\ m^3/s$, and $Q_{(t-30\ days)} = 0 - 521\ m^3/s$ (Table S1). Mean monthly discharge across sites in January and July ranges $\overline{Q}_{Jan} = 0 - 1,755\ m^3/s$ and $\overline{Q}_{Jul} = 0 - 343\ m^3/s$. Summer and winter low flows across sites range $Q_{95,\ JAS} = 0 - 208\ m^3/s$ and $Q_{95,\ JFM} = 0 - 479\ m^3/s$. Summer and winter high flows range $Q_{5,\ JAS} = 0 - 456\ m^3/s$ and $Q_{5,\ JFM} = 0 - 3512\ m^3/s$.

Median catchment drainage area ranges 0.03 − 29,007.89 km^2 (Table S1). Daily specific discharge across sites ranged from zero to 0.05 $m^3 km^{-2} s^{-1}$, 0.06 $m^3 km^{-2} s^{-1}$, and 0.09 $m^3 km^{-2} s^{-1}$ for $q_{(t-2\ days)}$, $q_{(t-10\ days)}$, and $q_{(t-30\ days)}$ respectively. Mean monthly specific discharge in January and July ranges $\overline{q}_{Jan} = 0.0 - 0.22\ m^3 km^{-2} s^{-1}$ and $\overline{q}_{Jul} = 0.0 - 0.05\ m^3 km^{-2} s^{-1}$. Summer and winter low flows across sites range $q_{95,\ JAS} = 0.0 - 0.04\ m^3 km^{-2} s^{-1}$ and $q_{95,\ JFM} = 0.0 - 0.03\ m^3 km^{-2} s^{-1}$. Summer and winter high flows range $q_{5,\ JAS} = 0.0 - 0.06\ m^3 km^{-2} s^{-1}$ and $q_{5,\ JFM}$ $0.0 - 0.66\ m^3 km^{-2} s^{-1}$.

2.4. Information metrics

We leveraged information theory to analyze the relationships between microbial communities and catchment hydrology. Information theory is useful for understanding complex systems in a range of research domains, including in hydrology and in the geosciences more broadly (Ruddell and Kumar, 2009; Ehret et al., 2014; Olds et al., 2016; Franzen et al., 2020; Goodwell et al., 2020; Li et al., 2021). Information theory is based on Shannon's entropy ($H(X)$ [*bits*]), which quantifies the amount of uncertainty in a discrete random variable X (Shannon, 1948) and is defined as:

$$H(X) = -\Sigma p(x)\ log_2 p(x), \tag{1}$$

where $p(x)$ is the probability distribution function of random variable X and summation is over all possible states of $X = x$. Likewise, the joint entropy ($H(X,Y)$[*bits*]), the total amount of information necessary to describe two random variables X and Y, is defined as:

$$H(X,Y) = -\Sigma p(x,y)\ log_2 p(x,y), \tag{2}$$

where ($p(x,y)$) is the joint probability distribution of X and Y and summation is over all possible states of $X = x$ and $Y = y$. Mutual information ($I(X;Y)$ [*bits*]) is the reduction in uncertainty in one random variable X as a result of knowledge of a second random variable Y and is calculated from the joint and marginal probability distributions of X and Y as:

$$I(X;Y) = \Sigma p(x)\ log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

Mutual information can be equivalently expressed (Cover and Thomas, 2005) in terms of the entropy of X and Y as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

We calculated the mutual information between each target hydrologic metric Y and abundance of microbial taxon X in terms of their marginal and joint entropies across study sites. To compute the information metrics, we (i) log-transformed hydrologic metrics and ASV abundances, adding 0.001 to each hydrologic metric and unity to each ASV abundance value to avoid taking the log of zero; (ii) standardized the log-transformed data between zero and unity using the range of each variable, (iii) discretized the standardized data for each X ASV abundance and Y hydrologic metric into five evenly spaced bins. We estimated the marginal and joint probability density functions (*pdfs*) with a fixed binning method to ensure that information metrics for each variable are comparable and transparent. We selected this fixed number of bins based on sample size and standard deviation according to a commonly used rule of thumb (Scott, 1979). We computed the optimal bin size for all variables and found that it ranged between 4 and 9 bins and was ≥5 bins for 95 % of the variables. For comparison, we computed information metrics with up to 20 bins and using gaussian kernel densities and verified that prescribing a fixed bin size of 5 provided the most robust results for this dataset. We also note that although the bin size and *pdf* estimation methods do influence entropy values, this choice does not typically affect comparison patterns (Loritz et al., 2019) which is the focus of this analysis. We then employed the marginal and joint fixed-bin *pdfs* in Eqs. (1) and (2), respectively, to obtain $H(X)$, $H(Y)$, and $H(X,Y)$. Finally, we applied eq. (4) to compute the mutual information shared between each microbial taxon and each hydrologic metric.

To identify statistically significant relationships between microbial taxa and hydrologic metrics, we used a shuffled surrogates method (Ruddell and Kumar, 2009). For 1000 iterations of each pair of X and Y, we randomly shuffled X to destroy any structure while maintaining the distribution of the data and then computed mutual information of the shuffled data. We considered $I(X;Y)$ statistically significant if the value was greater than the 99th percentile of shuffled iterations.

Finally, to enable comparison of mutual information values across microbial taxa and hydrologic metrics, we normalized mutual information scores by the entropy of the target hydrologic metrics as:

$$I(X;Y)_{norm} = \frac{I(X;Y)}{H(Y)} \tag{5}$$

The normalized mutual information value is thus the fraction reduction in uncertainty of the hydrologic metric that comes from observing the abundance of a particular microbial taxon.

To analyze patterns in hydrologic information shared with the microbial community, we summarized mutual information in two ways: 1) median value of normalized information across all informative microbial taxa for each hydrologic metric, and 2) the number of informative microbial taxa for each hydrologic metric. We define informative taxa as those with $I(X;Y)_{norm} > 0$.

We first sought to compare the value of information shared between the streamwater microbial community and absolute versus specific discharge and among the three major categories of hydrologic metrics: daily discharge, mean monthly discharge, and seasonal high and low flow durations. To compare information shared with absolute versus specific discharge, we applied a non-parametric Mann-Whitney U test (Mann and Whitney, 1947) to test for differences in median value of information shared between microbial taxa and absolute versus specific discharge metrics. We then applied a paired sample t-test to compare the number of informative taxa for absolute versus specific discharge for each hydrologic metric. To satisfy the parameters of a paired sample test, we assigned a value of zero to all ASVs that did not share significant information with hydrologic metrics. (Throughout the rest of the analysis, non-significant values of mutual information were excluded from calculations). To compare the strength of the relationships among the three hydrologic categories, we applied Mann-Whitney U tests to test for differences in median value of shared information, and one-way analysis of variance (ANOVA) F-tests to test for differences in the mean number of informative ASVs, across all metrics for each category (e.g., across all time lags for daily discharge).

We performed regression analyses to analyze patterns in mutual information between the microbial community and discharge through time for daily discharge and mean monthly discharge. For daily discharge, we fit regression functions relating the median value of shared information to the number of informative ASVs across time lags. For mean monthly discharge, we fit regression functions relating the median value of shared information to the number of informative ASVs across months. We assessed model fit with Pearson's correlation (ρ).

To assess how the strength of relationships with streamwater microbial taxa may be related to discharge magnitude, we used non-parametric Kruskal-Wallis H-tests (i.e., one-way ANOVA on ranks; Kruskal and Wallis, 1952) to test for differences in median value of mutual information between high and low flow durations. We conducted a one-way ANOVA *F*-test to test for differences in the mean number of informative ASVs between high and low flow durations.

Finally, we sought to identify patterns in mutual information related to abundance of microbial taxa across sites. We applied a non-parametric Spearman's rank correlation (r_s) to test for relationships between the value of mutual information for each hydrologic metric and 1) total abundance of an ASV across all sites and 2) the number of sites at which an ASV was detected. We conducted all statistical tests at significance level $\alpha = 0.05$ using *SciPy* (v 1.6.2) for Python.

3. Results

3.1. Stream microbial community composition

Overall, gammaproteobacteria was the most abundant taxonomic group, comprising 20–40 % of most stream communities we sampled (Fig. 2). Phylum Bacteroidota was also common across all sites. The proportion of unclassified Other microbes, often consisting of more rare taxa, was generally greater in sites with lower daily stream discharge on the date of DNA sample collection (Q_t). Phyla Verrucomicrobiota and Actinobacteriota were, very broadly, more common at sites with higher discharge. Alphaproteobacteria, Cyanobacteria, and Planctomycetota were also detected in smaller numbers across most sites.

3.2. Mutual information between microbial communities and hydrologic metrics

We calculated normalized mutual information for each ASV common to all three watersheds for each hydrologic metric, resulting in a matrix of 256 ASVs × 76 hydrologic metrics (and area; Fig. 3). Of the 256 common ASVs, 102 had statistically significant mutual information with at least one hydrologic metric, and each hydrologic metric shared information with at least one ASV. Mutual information is generally concentrated among more abundant taxa with a notable exception. *ASV 69,* a Bacteroidota classified to the lake-inhabiting genus *Lacihabitans* (Joung et al., 2014) shares information with all of the absolute and nearly all of the specific hydrologic metrics. *ASV 69* demonstrates an especially strong relationship with daily discharge (Fig. 3; Table S2). Across all absolute and specific hydrologic metrics, 9.7 % of taxa, on average, share information with a given hydrologic metric, reducing uncertainty of that metric by a median of 15.6 %. The maximum value of shared information is $I\left(ASV\ 4; Q_{-Feb}\right)_{norm} = 0.447\ bits/bit$, corresponding to a 44.7 % reduction in uncertainty of mean February absolute discharge by observing the abundance of *ASV 4*, a Gammaproteobacteria classified to the planktonic genus *Limnohabitans,* the fourth most abundant microbial taxon detected across all sites; the minimum nonzero value is $I(ASV\ 3; q_{(t)})_{norm} = 0.124\ bits/bit$. Note that this taxon and many other informative taxa classified as Gammaproteobacteria by the Silva 3.24 database are often classified as Betaproteobacteria by other databases.

Microbial taxa share more information with absolute than with specific discharge metrics (Fig. 3). The median value of mutual information is significantly higher for absolute versus specific discharge ($I(X;Q)_{norm} = 0.163\ bits/bit$ vs $I(X;q)_{norm} = 0.150\ bits/bit$; $U = 126{,}985$, $p < 0.001$). Furthermore, a greater number of taxa share information with absolute (28.2 ± 4.1 SD) than with specific (21.4 ± 5.5 SD) hydrologic metrics (two-sample one-sided $t = 6.13$, $p < 0.001$). Because relationships with the microbial community are so much stronger, in terms of both number of informative taxa and median value of shared information, we focus the remainder of the analysis on absolute discharge metrics only. (Results of the complementary analyses of specific discharge metrics are presented in supplementary figures S1-S3.)

In comparing mutual information between the streamwater microbial community and the three hydrologic categories, we found that the microbial community is, on average, more strongly related to mean monthly discharge and typical seasonal extreme flows, in terms of median value of shared information but not necessarily in terms of the number of informative taxa. Median value of shared information between microbial taxa and daily discharge ($I(X;Q_{(t-n\ days)})_{norm} = 0.159\ bits/bit$) is significantly lower than median value of shared information for mean monthly discharge ($I\left(X;\overline{Q}_{mon}\right)_{norm} = 0.165\ \text{bits/bit}$; $U = 473{,}779$, $p < 0.001$) and seasonal flow durations (median $I(X;Q_{P,\ s})_{norm} = 0.165\ bits/bit$; $U = 619{,}418$, $p < 0.001$). Mean number of informative taxa for each category is similar: 27.4 (±2.6 SD) taxa, 29.3 (±4.7 SD) taxa, and 28.1 (±5.2 SD) taxa for daily discharge, monthly mean discharge, and seasonal flow durations, respectively.

For daily absolute discharge, information shared with the microbial community across all time lags up to 30 days before sampling ranges $I(ASV\ 47; Q_{(t-20\ days)})_{norm} = 0.127\ bits/bit$ to $I(ASV\ 4; Q_{(t-2\ days)})_{norm} = 0.250\ bits/bit$ (Table 1, Fig. 4). Median value of mutual information over time follows a second-order polynomial function with a peak in shared information with discharge at 16 days prior to sampling (Pearson's $\rho = 0.73$, $p = 0.001$; Fig. 4). The number of informative ASVs through time also fits a second-order polynomial with a maximum number of informative taxa for discharge at 22 days prior to sampling ($\rho = 0.68$, $p = 0.004$; Fig. 4).

For mean monthly discharge, shared information across all months ranges $I\left(ASV\ 4; \overline{Q}_{Feb}\right)_{norm} = 0.447\ \text{bits/bit}$ to $I\left(ASV\ 24; \overline{Q}_{Jan}\right)_{norm} = 0.133\ \text{bits/bit}$ (Table 1, Fig. 5). Unlike for daily discharge, we observed no trend in

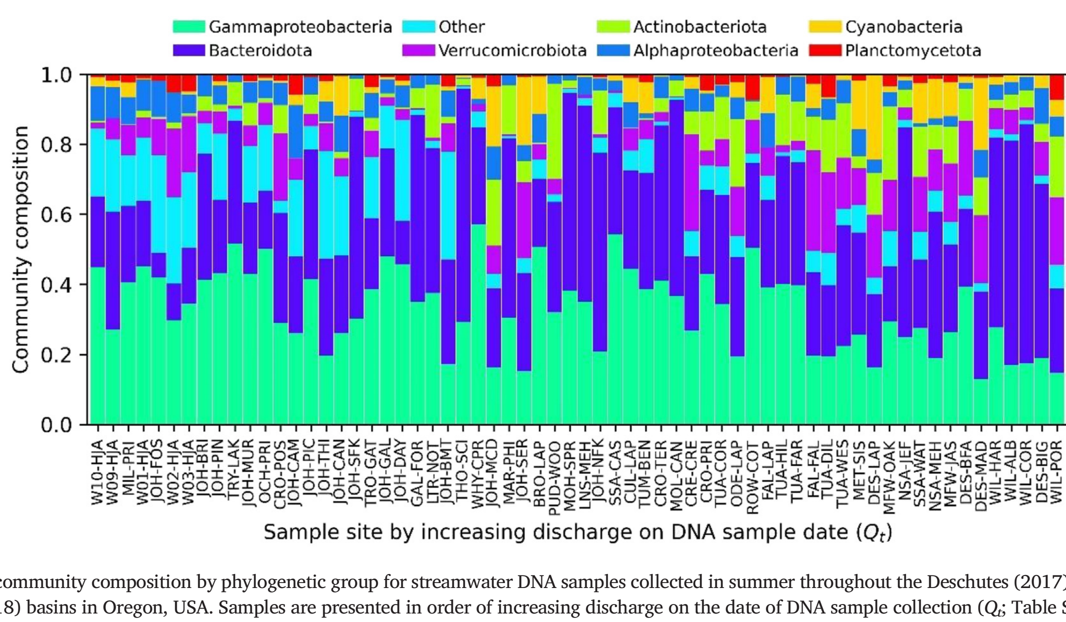


Fig. 2. Microbial community composition by phylogenetic group for streamwater DNA samples collected in summer throughout the Deschutes (2017), Willamette (2017), and John Day (2018) basins in Oregon, USA. Samples are presented in order of increasing discharge on the date of DNA sample collection (Q_t; Table S1).

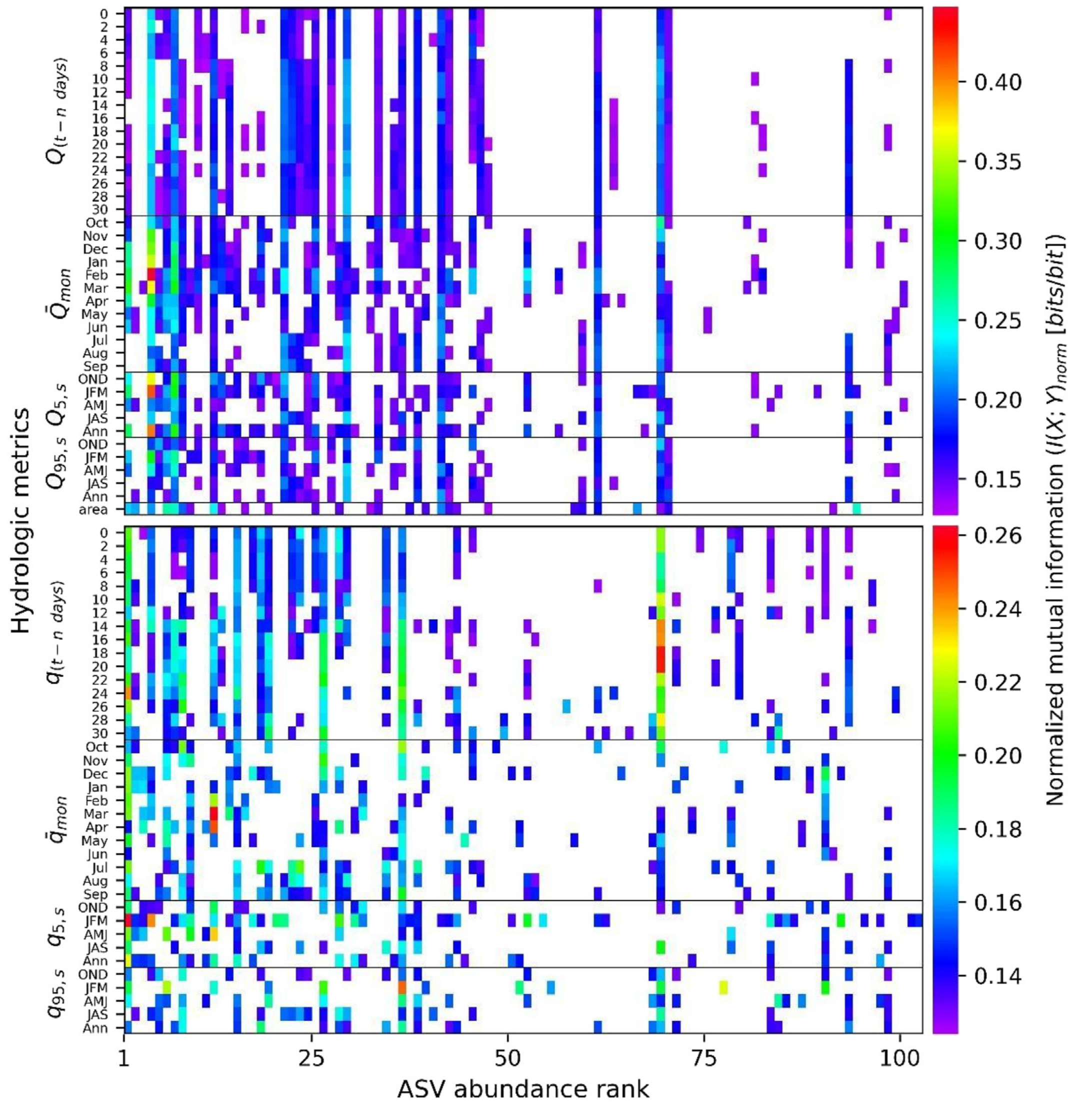


Fig. 3. Heatmap illustrating the normalized mutual information ($I(X;Y)_{norm}$ [bits/bit]) between streamwater microbial amplified sequence variants (X = ASVs) and absolute discharge ($Y = Q$; top), specific discharge ($Y = q$; bottom) hydrologic metrics, as well as basin drainage area (top, bottom line), for study streams in Oregon, USA. We collected microbial DNA samples in summer in the Willamette (2017), Deschutes (2017), and John Day (2018) basins. Hydrologic metrics include daily discharge at time lags n days prior to DNA sample day t ($Q_{(t-n\ days)}$, $q_{(t-n\ days)}$), mean monthly discharge ($\overline{Q}_{mon}$, $\overline{q}_{mon}$) for months October to September, and seasonal high and low flow durations ($Q_{P,\ s}$, $q_{P,\ s}$ for P = 5- and 95 % exceedance probability for seasons s = fall [*OND*], winter [*JFM*], spring [*AMJ*], summer [*JAS*], and annually [*Ann*]).

mutual information by month (Fig. 5). Number of informative ASVs fits a second-order polynomial with the greatest number of taxa sharing information with January mean discharge (ρ = 0.70, p = 0.01; Fig. 5). Across seasons and high and low flow durations, mutual information shared with microbial taxa ranges $I(ASV\ 20; Q_{5,\ OND})_{norm} = 0.130$ *bits/bit* to $I(ASV\ 4; Q_{5,\ JFM})_{norm} = 0.414$ *bits/bit* (Fig. 6). We did not detect a difference in shared information between high versus low flows, neither in terms of median value of shared information nor in the number of informative taxa.

Mutual information increases with increasing detection of microbial taxa for all three categories of hydrologic metrics, although the relationship is not as strong for daily discharge. For all three categories, mutual information increases linearly with the log of abundance of a microbial taxon across all sites for mean monthly discharge (ρ = 0.42, p < 0.001), seasonal flow durations (ρ = 0.37, p < 0.001), and daily discharge (ρ = 0.12, p = 0.013; Fig. 7). Mutual information is most strongly correlated with the number of sites at which a microbial taxon is detected (ρ = 0.49, 0.48, 0.26) for mean monthly discharge, seasonal flow durations, and daily discharge, respectively; all p < 0.001).

4. Discussion

We detected significant relationships between summer streamwater microbial communities and both absolute and specific (per unit area) hydrologic metrics. Drainage area, the upslope area that drains to a point in a stream, is one of the strongest drivers of stream discharge across timescales, and as such, discharge normalized by area reflects a different set of controls than absolute discharge. Streams draining large watersheds generally respond more slowly to precipitation events because of greater infiltration rates, longer transit times, and diminishing peak event flows, usually resulting in lower rates of discharge per unit area than might be observed in an otherwise similar smaller watershed (Dingman, 2015). The strength of relationships with microbial taxa was greater for absolute than for specific discharge, as indicated by both a higher median value of mutual information and a greater number of informative taxa across absolute discharge metrics. However, patterns in the information shared between microbial taxa and absolute discharge appear different than patterns in information shared with specific discharge (Fig. 3); that is, information shared with discharge per unit area across the microbial community is not simply a

Table 1

Medians and ranges of entropy ($H(Y)$ [bits]) and normalized mutual information ($I(X;Y)/H(Y)$ [bits/bit]) between streamwater microbial amplified sequence variants (ASVs) and hydrologic metrics: absolute discharge (Q [m^3s^{-1}]) and specific (per unit area) discharge (q [$m^3km^{-2}s^{-1}$]). We collected microbial DNA samples in summer in the Willamette (2017), Deschutes (2017), and John Day (2018) basins in Oregon, USA. Hydrologic metrics include daily discharge at time lags up to $n = 30$ days prior to DNA sample collection, mean monthly discharge for mon = all months October to September, and seasonal high and low flow durations for P = 5- and 95-perecent exceedance probability and seasons s = fall [OND], winter [JFM], spring [AMJ], summer [JAS], and annually [Ann]. Catchment area is derived from the StreamStats web application developed by the USGS (USGS, 2017).

Hydrologic metric		Information metric					
		Absolute discharge (Q)			Specific discharge (q)		
		$H(Y)$	$\frac{I(X;Y)}{H(Y)}$		$H(Y)$	$\frac{I(X;Y)}{H(Y)}$	
			Median	(Range)		Median	(Range)
Daily discharge ($Q_{(t-n\ days)}$, $q_{(t-n\ days)}$)	0	2.107	0.151	(0.100)	2.181	0.145	(0.092)
	2	2.098	0.161	(0.120)	2.179	0.147	(0.088)
	4	2.082	0.155	(0.092)	2.185	0.148	(0.068)
	6	2.078	0.158	(0.082)	2.185	0.147	(0.069)
	8	2.064	0.157	(0.108)	2.150	0.147	(0.066)
	10	2.052	0.163	(0.105)	2.148	0.146	(0.097)
	12	2.070	0.169	(0.116)	2.136	0.143	(0.085)
	14	2.093	0.166	(0.115)	2.116	0.146	(0.117)
	16	2.079	0.166	(0.114)	2.073	0.148	(0.115)
	18	2.091	0.157	(0.105)	2.069	0.147	(0.126)
	20	2.100	0.162	(0.107)	2.062	0.158	(0.130)
	22	2.065	0.164	(0.118)	2.032	0.162	(0.089)
	24	2.060	0.164	(0.087)	1.960	0.155	(0.114)
	26	2.079	0.163	(0.091)	2.047	0.151	(0.083)
	28	2.055	0.157	(0.096)	2.041	0.151	(0.094)
	30	2.055	0.157	(0.095)	2.054	0.147	(0.082)
Mean monthly discharge ($\overline{Q}_{mon}$, $\overline{q}_{mon}$)	Oct	2.100	0.167	(0.129)	1.995	0.155	(0.078)
	Nov	2.208	0.164	(0.186)	1.972	0.172	(0.064)
	Dec	2.174	0.161	(0.197)	1.992	0.152	(0.077)
	Jan	2.150	0.158	(0.223)	2.115	0.153	(0.080)
	Feb	1.925	0.185	(0.304)	2.186	0.160	(0.081)
	Mar	2.003	0.164	(0.236)	2.121	0.159	(0.126)
	Apr	2.024	0.164	(0.160)	2.054	0.149	(0.114)
	May	2.000	0.165	(0.100)	2.103	0.145	(0.062)
	Jun	2.015	0.155	(0.103)	2.100	0.144	(0.041)
	Jul	1.964	0.174	(0.101)	2.142	0.151	(0.079)
	Aug	2.061	0.167	(0.091)	2.122	0.149	(0.051)
	Sep	2.092	0.172	(0.097)	2.114	0.146	(0.064)
Seasonal flow durations ($Q_{P,\ s}$, $q_{P,\ s}$)	5, OND	2.160	0.163	(0.236)	2.186	0.145	(0.060)
	5, JFM	2.041	0.165	(0.276)	2.100	0.150	(0.127)
	5, AMJ	2.056	0.154	(0.093)	2.074	0.165	(0.093)
	5, JAS	1.949	0.179	(0.086)	2.076	0.150	(0.066)
	5, Ann	2.012	0.171	(0.267)	2.168	0.152	(0.094)
	95, OND	2.006	0.161	(0.088)	2.205	0.142	(0.047)
	95, JFM	2.038	0.167	(0.166)	1.703	0.193	(0.095)
	95, AMJ	2.015	0.163	(0.118)	1.946	0.155	(0.049)
	95, JAS	2.067	0.165	(0.087)	2.236	0.149	(0.051)
	95, Ann	2.024	0.154	(0.088)	2.169	0.151	(0.055)
Catchment area		1.81	0.157	(0.131)			

characteristically lesser value of the information shared with absolute discharge. Furthermore, patterns in information shared between both absolute discharge and discharge per unit area appear independent of the patterns in information shared with catchment drainage area (Fig. 3). The different strengths and patterns of mutual information between microbial taxa and absolute versus specific discharge, and that these relationships do not appear to be driven by catchment area, suggests that distinct aspects of the streamwater microbial communities integrate information from a broad range of complex processes and interactions.

We furthermore found that summer microbial taxa were, on average, more informative of long-term mean monthly discharge and seasonal flow durations than of daily discharge concurrent or up to 30-days antecedent to the microbial sampling dates. The hydrologic regime of a catchment, including typical extremes as well as average monthly or seasonal stream discharge, is shaped by larger scale time and space characteristics, such as climate and topographic organization of the catchment (McGuire et al., 2005). On the other hand, daily discharge is influenced to varying degrees by short- and moderate-timescale localized variable conditions, including duration and intensity of recent precipitation events and antecedent soil moisture and storage, which impact infiltration rates and water transit times (Dingman, 2015). That the summer communities we sampled were more informative of metrics of general hydrologic regime suggests that these communities are shaped by broader catchment characteristics, further supporting earlier findings connecting streamwater microbial diversity to geomorphic and climatic characteristics of the catchment (URycki et al., 2020). However, the summer microbial community was also informative of recent daily discharge, suggesting that, although seasonality and static properties of the catchment contribute to the development of characteristic microbial communities (Crump et al., 2009), these communities also respond to conditions at shorter timescales. Additional research on microbial community dynamics at higher temporal resolution would be useful to further test these responses.

The number of informative microbial taxa is greater for discharge metrics in the days or months prior to DNA sampling than the day or season of sampling. For daily discharge, this pattern suggests that informative taxa may accumulate over time (~2 weeks in this case) as microbial communities develop in response to environmental conditions that also control discharge. For monthly discharge, an explanation for this lag in microbial community response is less straightforward. Given that freshwater microbial communities experience seasonal shifts but return to characteristic

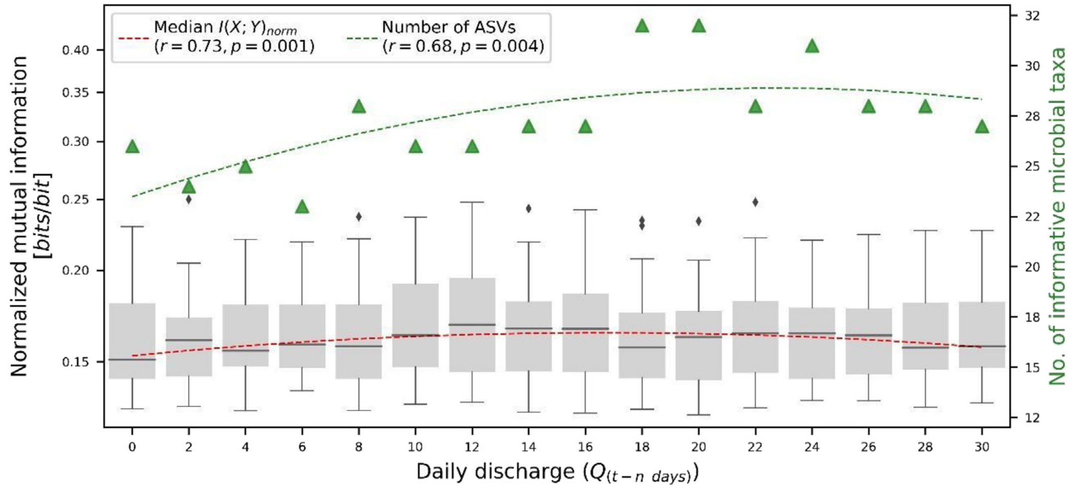


Fig. 4. Median value of normalized mutual information ($I(ASV; Q_{(t-n\ days)})_{norm}$ [bits/bit]) between informative streamwater microbial amplified sequence variants (ASVs) and daily discharge at different time lags ($Q_{(t-n\ days)}$ for time lags up to $n = 30$ days before sample date) for study streams in Oregon, USA. Boxes show medians and interquartile ranges; whiskers show values within 1.5 times the interquartile range. Green triangles indicate the number of ASVs with statistically significant mutual information (99 % confidence). Dashed lines show best fit curves between time lags and median value of normalized information (red) and number of informative ASVs (green), with Pearson correlation (r) and p-value indicated in legend. We collected microbial DNA samples in summer in the Willamette (2017), Deschutes (2017), and John Day (2018) basins.

core summer and winter communities (Crump and Hobbie, 2005; Crump et al., 2009), we might expect the strongest relationships between microbial taxa and hydrologic measures to be observed during periods when conditions are similar to those when microbial communities are sampled. Our results suggest the opposite pattern may be true. For monthly discharge, the number of informative ASVs is greater in months more distant (e.g., March and November) than those in which we sampled (i.e., July and August; Fig. 5). Furthermore, although we did not observe a clear trend in the median value of mutual information across months, the highest values of mutual information shared between microbial taxa and monthly discharge occur with a time lag of 4–6 months, as observed by the outlying high values of mutual information in fall, winter, and spring months, but not in summer (Fig. 5). One explanation for this phenomenon is that patterns in shared information might be more related to discharge dynamics (e.g., flow volume) than to temporal lags. The peak in shared information we observed between summer microbes and winter flows might indicate that the microbial community is more informative of higher flow conditions, which typically occur in winter in the Mediterranean climate of the Pacific Northwest where our study is located, than the lower flows observed in summer when we collected DNA samples. However, the patterns in shared information we identified between seasonal high and low flows (i.e., Q_5, Q_{95}) would appear to contradict this explanation; if summer microbes are more informative of high flows (typically winter) than low flows (typically summer), we would expect to see more information shared between microbial taxa and seasonal high flow durations (Q_5), but we observed no difference in shared information between high and low seasonal flow durations (Fig. 6). Future research might examine whether this pattern of greater shared information with trends in the opposite season holds when winter microbial communities are sampled. Furthermore, considering that our results indicate that microbial communities also respond to local environmental conditions at shorter timescales (days and weeks), it may be that the relationships between microbial taxa and typical flow regime are confounded by short-term perturbations to the community. Additional research on microbial community diversity at higher temporal

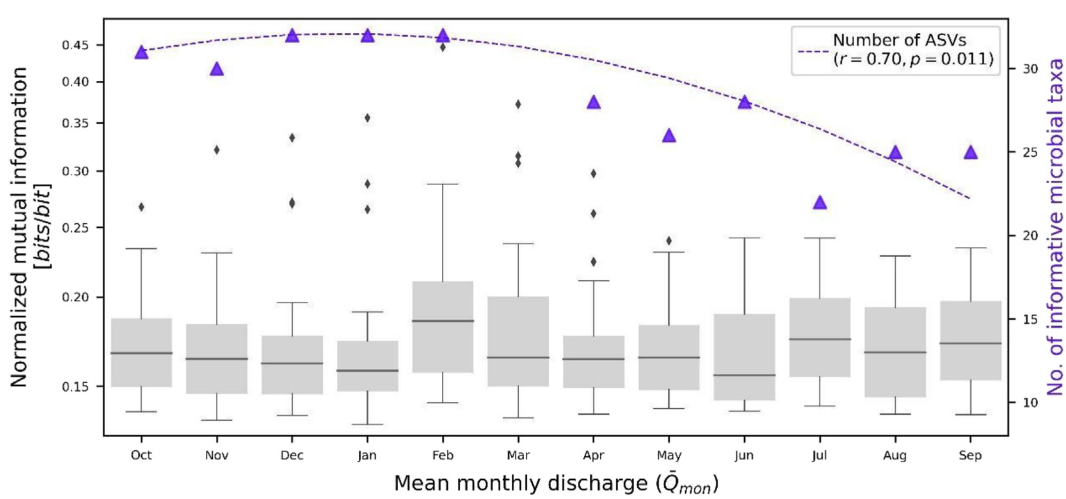


Fig. 5. Median value of normalized mutual information ($I(ASV; \overline{Q}_{mon})_{norm}$ [bits/bit]) between informative streamwater microbial amplified sequence variants (ASVs) and mean monthly discharge ($\overline{Q}_{mon}$) of study streams in Oregon, USA. Boxes show medians and interquartile ranges; whiskers show values within 1.5 times the interquartile range. Purple triangles indicate the number of ASVs with statistically significant mutual information (99 % confidence). Dashed purple line shows best fit curve between months and number of informative ASVs, with Pearson correlation (r) and p-value indicated in legend. We observed no significant relationship between month and median value of mutual information. We collected microbial DNA samples in July and August in the Willamette (2017), Deschutes (2017), and John Day (2018) basins.

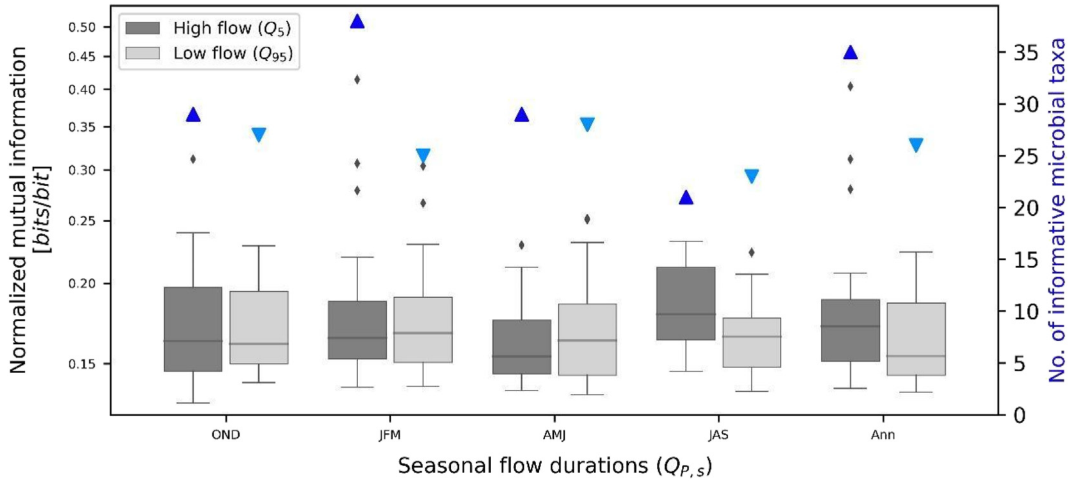


Fig. 6. Normalized mutual information ($I(ASV;Q_{P,\,s})_{norm}$ [bits/bit]) between unique streamwater microbial amplified sequence variants (ASVs) and seasonal high (dark) and low (light) flow durations ($Q_{P,\,s}$ for P = 5- and 95 % exceedance probability for seasons s = fall [*OND*], winter [*JFM*], spring [*AMJ*], summer [*JAS*], and annually [*Ann*]) at study streams in Oregon, USA. Boxes show medians and interquartile ranges; whiskers show values within 1.5 times the interquartile range. Blue triangles indicate the number of ASVs with statistically significant mutual information (99 % confidence) for high flows (dark) and low flows (light). We collected microbial DNA samples in summer in the Willamette (2017), Deschutes (2017), and John Day (2018) basin.

resolution is necessary to determine the timescale and magnitude at which the streamwater microbial community responds to variable local conditions.

Microbial community composition likely reflects different components of hydrologic and ecosystem function, even within the relatively small suite of discharge metrics we analyzed. Future work might also analyze more closely the sets of taxa that share information with one metric versus another, for instance daily discharge versus mean monthly discharge, and whether and how those sets of informative taxa overlap. Investigation of the taxonomy and phylogeny of informative taxa, coupled with analysis of the functional roles of these taxa within the microbial community and within the larger ecosystem, will improve our understanding of the dynamics and interactions between the streamwater microbial community and hydrology, as well as contribute new knowledge to the fields of microbiology and ecology, among others.

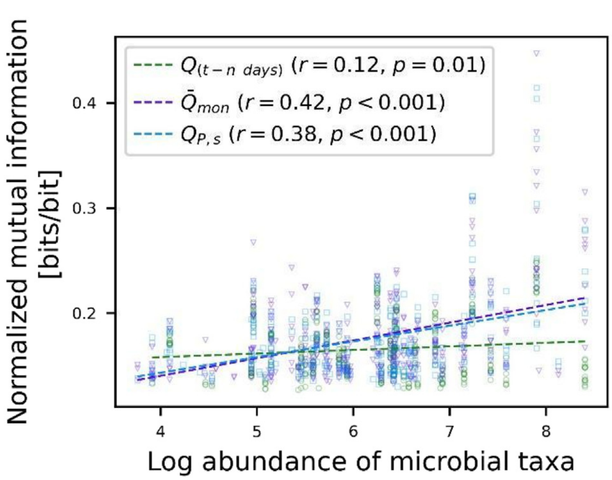


Fig. 7. Normalized mutual information ($I(ASV;Y)_{norm}$ [bits/bit]) between hydrologic metrics and streamwater microbial taxa versus the log of abundance of taxa in streams across Oregon, USA. We collected microbial DNA samples in summer in the Willamette (2017), Deschutes (2017), and John Day (2018) basins. Hydrologic metrics include daily discharge at time lags up to n = 30 days prior to DNA sample collection ($Q_{(t-n\ days)}$; green circles), mean monthly discharge ($\overline{Q}_{mon}$; purple triangles), and seasonal high and low flow durations ($Q_{P,\,s}$; 5- and 95 % exceedence probability for all seasons and annually; blue squares). Legend shows Pearson's correlation (r) and p-value of the linear regression.

While we identified significant shared information between a number of microbial taxa and catchment hydrology, it is impossible from this analysis to quantify the sum total informational value of the community overall. At least some of the hydrologic information encoded within individual taxa likely overlaps with that of other taxa (i.e., is redundant); on the other hand it is also plausible that information encoded within different taxa is more valuable when considered jointly (i.e., is synergistic). Future work might enlist additional information theory metrics, such as mutual information conditioned on a third microbial or hydrologic variable and information decomposition (Goodwell et al., 2020; Gutknecht et al., 2021). Such techniques can quantify joint relations and parse information types (redundant, synergistic and unique) to better explore causality and broaden our understanding of the dynamics of microbial community development and catchment hydrology.

We evaluated possible sources of uncertainty in our results and used our best judgment in the analysis and parameter choices required for investigating this novel type of hydrologic data. For instance, we selected only those microbial taxa that appeared in all three major study watersheds, which reduced our dataset from 4701 taxa to 265 taxa. We reasoned that greater detection over a wider range of conditions would likely result in identifying stronger, more broadly meaningful relationships between microbial taxa and hydrologic metrics. To test this assumption, we ran the analysis over the complete set of 4701 taxa. Analysis of the full set of ASVs resulted in identification of many additional taxa containing very small amounts of information, but maximum and median values of shared information were greatest among taxa that were identified in all three watersheds (Fig. S4). Therefore, although our selection criteria may have precluded identification of some potentially meaningful relationships, our analysis captured the strongest relationships while also considerably reducing analytical and computational complexity. Additionally, we applied a strict significance threshold (α = 0.01) for mutual information values, which resulted in more robust results. Although this conservative threshold likely resulted in the loss of some data, it also contributes to stronger confidence in the patterns we identify. On the other hand, we opted to include some study sites with less than a 10-year period of hydrologic record (minimum = 6.9 years; Table S1) in calculations of mutual information for monthly discharge and seasonal flow durations. We determined that small differences in hydrologic statistics for a fraction of sites would not bias our overall results and that including a greater number of co-located hydrological and microbial observations would strengthen the overarching insights gained from the analysis.

An important caveat of this analysis is that interpretation is confined, both spatially and temporally, to the microbial communities we sampled.

Collecting microbial DNA during the dry summer season offered the important advantage of a reasonable assumption of uniform hydrologic conditions across sites, given that precipitation inputs in the summer are typically negligible. The assumption of baseflow conditions across sites is important particularly for the information-theoretic approach we applied here, as the information metrics we computed rely on joint probabilities of hydrologic characteristics and abundance of microbial taxa across sites. Minimizing the influence of local perturbations to hydrologic measurements and microbial community assemblages thus lends additional strength to these results. On the other hand, for these same reasons, it is not possible to extrapolate these results beyond summer microbial communities in the watersheds we sampled. Replicating this summer microbial community sampling and analysis, potentially across a broader area, would increase the interpretive power of our results. Additionally, a complementary analysis on winter (or wet season) communities could potentially expand the interpretation of these results, though careful study design would be necessary to meet the assumptions of the approach we applied here.

5. Conclusions

Microbial diversity and hydrologic metrics are driven by common processes, and we found that the relative abundance (i.e., counts) of many microbial taxonomic groups are statistically related to all 76 hydrologic metrics we analyzed. An average of 9.6 % of the summer streamwater microbial taxa we analyzed shared information with a given hydrologic metric, including mean monthly discharge, seasonal high and low flow durations, and daily discharge, in some cases reducing the uncertainty of a hydrologic metric by >40 %. The summer microbial community shared the most information with winter mean flows, which also coincide with high flow periods in our study area. The microbial community also shared information with daily discharge, most strongly at an approximately two-week lag from sampling dates. Considering that a single streamwater DNA sample can yield thousands of data points, many more than traditional hydrologic observations, our study lends further support to the value of microbial community composition as a hydrologic observation at multiple timescales. Microbiome analyses therefore have the potential to contribute new and complex insights to outstanding questions in the field of hydrology, and the value and application of these data should continue to be explored.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.157911.

Funding

This work was supported by the National Science Foundation grant EAR 1836768. DRU would like to acknowledge STEM Scholarship support from NSF grant 1153490. MB received funding from the European Commission and Swedish Research Council for Sustainable Development (FORMAS) (grant 2018-02787) in the frame of the international consortium iAqueduct financed under the 2018 Joint call of the WaterWorks2017 ERA-NET Cofund. Data and facilities for a portion of this research were provided by the HJ Andrews Experimental Forest and Long Term Ecological Research (LTER) program, administered cooperatively by the USDA Forest Service Pacific Northwest Research Station, Oregon State University, and the Willamette National Forest. This material is based upon work supported by the National Science Foundation under the LTER Grants: LTER8 DEB-2025755 (2020-2026) and LTER7 DEB-1440409 (2012-2020).

Data availability

Data and code are available at links provided in article text.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Author contributions are as follows: DRU, conceptualization, data curation, formal analysis, writing – original draft; MB: formal analysis, methodology, writing – review and editing; SPG, conceptualization, funding acquisition, resources, writing – review and editing; BCC, funding acquisition, data curation, resources, writing – review and editing; BL, methodology, writing – review and editing. We gratefully acknowledge the graduate and undergraduate students who contributed many long hours of lab and field work. The authors further acknowledge that Oregon State University in Corvallis, Oregon is located within the traditional homelands of the Ampinefu Band of Kalapuya. Following the Willamette Valley Treaty of 1855, Kalapuya people were forcibly removed to reservations in western Oregon. Today, living descendants of these people are a part of the Confederated Tribes of Grand Ronde Community of Oregon and the Confederated Tribes of the Siletz Indians.

Data availability

DNA sequence data is archived under BioProject PRJNA642636 under run accessions SRX8627687-SRX8627763 at the National Center for Biotechnology Information (NCBI). Code for the complete analysis can be found at zenodo.org (URycki and Bassiouni, 2022).

References

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., et al., 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82, 2415–2434. https://doi.org/10.1175/1520-0477.

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al., 2019. Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrol. Sci. J. 64, 1141–1158. https://doi.org/10.1080/02626667.2019.1620507.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., et al., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. 108, 4516–4522. https://doi.org/10.1073/pnas.1000080107.

Cover, T.M., Thomas, J.A., 2005. Elements of information theory. Elem. Inf. Theory 1–748. https://doi.org/10.1002/047174882X.

Crump, B.C., Hobbie, J.E., 2005. Synchrony and seasonality in bacterioplankton communities of two temperate rivers. Limnol. Oceanogr. 50, 1718–1729. https://doi.org/10.4319/lo.2005.50.6.1718.

Crump, B.C., Kling, G.W., Bahr, M., Hobbie, J.E., 2003. Bacterioplankton community shifts in an Arctic lake correlate with seasonal changes in organic matter source. Appl. Environ. Microbiol. 69, 2253–2268. https://doi.org/10.1128/AEM.69.4.2253-2268.2003.

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, 1365–1378. https://doi.org/10.1890/06-0387.

Crump, B.C., Peterson, B.J., Raymond, P.A., Amon, R.M.W., Rinehart, A., McClelland, J.W., et al., 2009. Circumpolar synchrony in big river bacterioplankton. Proc. Natl. Acad. Sci. U. S. A. 106, 21208–21212. https://doi.org/10.1073/pnas.0906149106.

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

Dingman, S.L., 2015. Physical hydrology. 3rd ed. Prentice Hall.

Droppo, I.G., Liss, S.N., Williams, D., Nelson, T., Jaskot, C., Trapp, B., 2009. Dynamic existence of waterborne pathogens within river sediment compartments. Implications for water quality regulatory affairs. Environ. Sci. Technol. 43, 1737–1743. https://doi.org/10.1021/es802321w.

Ehret, U., Gupta, H.V., Sivapalan, M., Weijs, S.V., Schymanski, S.J., Blöschl, G., et al., 2014. Advancing catchment hydrology to deal with predictions under change. Hydrol. Earth Syst. Sci. 18, 649–671. https://doi.org/10.5194/HESS-18-649-2014.

Franzen, S.E., Farahani, M.A., Goodwell, A.E., 2020. Information flows: characterizing precipitation-streamflow dependencies in the Colorado headwaters with an information theory approach. Water Resour. Res. 56, e2019WR026133. https://doi.org/10.1029/2019WR026133.

Good, S.P., Urycki, D.R., Crump, B.C., 2018. Predicting hydrologic function with aquatic gene fragments. Water Resour. Res. 54, 2424–2435. https://doi.org/10.1002/2017WR021974.

Goodwell, A.E., Jiang, P., Ruddell, B.L., Kumar, P., 2020. Debates—does information theory provide a new paradigm for earth science? Causality, interaction, and feedback. Water Resour. Res. 56, e2019WR024940. https://doi.org/10.1029/2019WR024940.

Gutknecht, A.J., Wibral, M., Makkeh, A., 2021. Bits and pieces: understanding information decomposition from part-whole relationships and formal logic. Proc. R. Soc. A 477. https://doi.org/10.1098/RSPA.2021.0110.

Hermans, S.M., Buckley, H.L., Case, B.S., Lear, G., 2019. Connecting through space and time: catchment-scale distributions of bacteria in soil, stream water and sediment. Environ. Microbiol. 1462–2920, 14792. https://doi.org/10.1111/1462-2920.14792.

Hrachowitz, M., Benettin, P., van Breukelen, B.M., Fovet, O., Howden, N.J.K., Ruiz, L., et al., 2016. Transit times-the link between hydrology and water quality at the catchment scale. Wiley Interdiscip. Rev. Water 3, 629–657. https://doi.org/10.1002/wat2.1155.

Johnson, S.L., Rothacher, J.S., Wondzell, S.M., 2020. Stream discharge in gaged watersheds at the HJ Andrews experimental Forest, 1949 to present ver 33. Environmental Data Initiative. https://doi.org/10.6073/pasta/0066d6b04e736af5f234d95d97ee84f3.

Joung, Y., Kim, H., Kang, H., Lee, B.Il, Ahn, T.S., Joh, K., 2014. Lacihabitans soyangensis gen. nov., sp. nov., a new member of the family cytophagaceae, isolated from a freshwater reservoir. Int. J. Syst. Evol. Microbiol. 64, 3188–3194. https://doi.org/10.1099/IJS.0.058511-0/CITE/REFWORKS.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Appl. Environ. Microbiol. 79, 5112–5120. https://doi.org/10.1128/AEM.01043-13.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583. https://doi.org/10.2307/2280779.

Li, B., Good, S.P., URycki, D.R., 2021. The value of L-band soil moisture and vegetation optical depth estimates in the prediction of vegetation phenology. Remote Sens. 13. https://doi.org/10.3390/rs13071343.

Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H., et al., 2019. A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation. Hydrol. Earth Syst. Sci. 23, 3807–3821. https://doi.org/10.5194/HESS-23-3807-2019.

Mann, H.B., Whitney, D.R., 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. 18, pp. 50–60. https://doi.org/10.1214/aoms/1177730491.

McGuire, K.J., McDonnell, J.J., Weiler, M., Kendall, C., McGlynn, B.L., Welker, J.M., et al., 2005. The role of topography on catchment-scale water residence time. Water Resour. Res. 41, 1–14. https://doi.org/10.1029/2004WR003657.

Miller, J.B., Frisbee, M.D., Hamilton, T.L., Murugapiran, S.K., 2021. Recharge from glacial meltwater is critical for alpine springs and their microbiomes. Environ. Res. Lett. 16, 64012. https://doi.org/10.1088/1748-9326/abf06b.

Newby, D.T., Pepper, I.L., Maier, R.M., 2009. Microbial transport. Environmental Microbiology. Elsevier Inc., pp. 365–383. https://doi.org/10.1016/B978-0-12-370519-8.00019-5.

Olds, B.P., Jerde, C.L., Renshaw, M.A., Li, Y., Evans, N.T., Turner, C.R., et al., 2016. Estimating species richness using environmental DNA. Ecol. Evol. 6, 4214–4226. https://doi.org/10.1002/ece3.2186.

Oregon Water Resources Department, 2021. Near Real Time Hydrographics Data accessed 3 September 2021.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590. https://doi.org/10.1093/nar/gks1219.

Read, D.S., Gweon, H.S., Bowes, M.J., Newbold, L.K., Field, D., Bailey, M.J., et al., 2015. Catchment-scale biogeography of riverine bacterioplankton. ISME J. 9, 516–526. https://doi.org/10.1038/ismej.2014.166.

Risley, J., Stonewall, A., Haluska, T., 2008. Estimating Flow-duration and Low-flow Frequency Statistics for Unregulated Streams in Oregon: U.S. Geological Survey Scientific Investigations Report 2008-5126 22 p.

Ruddell, B.L., Kumar, P., 2009. Ecohydrologic process networks: 1. Identification. Water Resour. Res. 45. https://doi.org/10.1029/2008WR007279.

Savio, D., Sinclair, L., Ijaz, U.Z., Parajka, J., Reischer, G.H., Stadler, P., et al., 2015. Bacterial diversity along a 2600?km river continuum. Environ. Microbiol. 17, 4994–5007. https://doi.org/10.1111/1462-2920.12886.

Schimel, D., Hargrove, W., Hoffman, F., MacMahon, J., 2007. NEON: a hierarchically designed national ecological network. Front. Ecol. Environ. 5, 59. https://doi.org/10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2.

Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610. https://doi.org/10.1093/BIOMET/66.3.605.

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. Water Resour. Res. 38. https://doi.org/10.1029/2001WR000978 23–1.

Shannon, C., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sorensen, J.P.R., Maurice, L., Edwards, F.K., Lapworth, D.J., Read, D.S., Allen, D., et al., 2013. Using boreholes as windows into groundwater ecosystems. PLoS One 8, e70264. https://doi.org/10.1371/journal.pone.0070264.

Sugiyama, A., Masuda, S., Nagaosa, K., Tsujimura, M., Kato, K., 2018. Tracking the direct impact of rainfall on groundwater at mt. Fuji by multiple analyses including microbial DNA. Biogeosciences 15, 721–732. https://doi.org/10.5194/bg-15-721-2018.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., et al., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. https://doi.org/10.1038/nature24621.

U.S. Geological Survey, 2016. National Water Information System Data Available on the World Wide Web (USGS Water Data for the Nation). https://doi.org/10.5066/F7P55KJN accessed 9 May 2020.

U.S. Geological Survey, 2017. StreamStats, Version 4. https://doi.org/10.3133/fs20173046.

URycki, D.R., Bassiouni, M., 2022. Mutual Information Between Streamwater Microbiome and Hydrology (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.6835083.

URycki, D.R., Good, S.P., Crump, B.C., Chadwick, J., Jones, G.D., 2020. River microbiome composition reflects macroscale climatic and geomorphic differences in headwater streams. Front. Water 2, 574728. https://doi.org/10.3389/frwa.2020.574728.