Supplementary information: Identifying reactive components of microbes and organic matter reveals unique links along aquatic networks

Masumi Stadler[1,*], François Guillemette[2], Trista J. Vick-Majors[3], and Paul A. del Giorgio[1]

[1]Groupe de Recherche Interuniversitaire en Limnologie, Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, QC, Canada
[3]Department of Biological Sciences, Michigan Technological University, Houghton, Michigan, USA
[2]Département des Sciences de l'Environnement, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

Contents

List of Figures

List of Tables

1 Supplementary Methods

1.1 Microbial abundance

To determine bacterial abundance, samples were preserved on the same day of collection by adding para-formaldehyde (PFA) to a final concentration of 1% and glutaraldehyde (G) to 0.05%, then stored at -80℃ until they were ready for analysis [del Giorgio et al., 1996]. Before analysis, the samples were thawed and stained with SYTO 13 (diluted in dimethyl sulfoxide (DMSO), 2.5 μM; Invitrogen, Waltham, MA, USA) at 0.025% of the sample volume. The stained samples were then analysed using an Accuri C6 flow cytometer (BD Bioscience, San José, CA, USA) at a flow rate of 14 L min^{-1}, utilising side scatter and green fluorescence (FL1-H) detection.

1.2 Hydrological estimates

To visualise and study the spatial patterns of microbes and molecules along a true hydrological continuum, flow-weighted water age (FWWA) was estimated for the studied watershed. FWWA represents the average time water takes to arrive at any given point in the hydrological network. A digital elevation model (18 x 18 m) was obtained from GeoGratis Canada [Natural Resources Canada, 2017] to delineate the watershed and calculate metrics such as flow accumulation, flow length, and pixel area using the Spatial Analyst Toolbox in ArcMap (v10.7.1, ESRI Inc., Redland, CA, USA). A flow accumulation threshold of 3,000 pixels was used to identify the stream network. Flow accumulation was converted to catchment area by multiplying it by the pixel area.

1.2.1 Reconstruction of a true hydrological network

To our knowledge there are no attempts to estimate FWWA within a watershed by considering the various ecosystems (i.e., lentic versus lotic systems) within a hydrological network, and they commonly assume that there are no lentic systems. To reconstruct a 'true' hydrological network, we first identified which pixels were lentic or lotic systems by overlaying the HydroLAKES [Messager et al., 2016] and reservoir (Hydro-Québec) polygons to the stream network and applied a buffer of 0.004 degrees. To identify the downstream pixel for each pixel in the watershed, we sorted the stream network by flow length (*Flow length* tool in ArcMap) within unique reach IDs assigned to stream vertices (*Stream to Feature* tool in ArcMap). To identify downstream pixels at confluences, unique node connection IDs that connect two stream reaches were utilised (extracted from *Stream to Feature* tool in ArcMap). Once the ecosystems along the stream network and the flow path were identified, tributaries into lakes and reservoirs were identified by finding pixels of ecosystem shifts (e.g., fluvial to lake) along the stream network. The lake/reservoir main inlet was identified as the tributary with the longest flow length. Similarly, the lake/reservoir outlet was identified as the pixel within the lake/reservoir with the longest flow length. The main channel for each lake and reservoir was subsequently identified by tracking the inlet reach downstream until the outlet. Other reaches that flow within lentic systems were not considered for subsequent water age calculations, and hence the identified tributaries' confluence was moved onto the main channel of the lentic system (Fig. S7). Hydrological network sorting, and identifying the flow paths, tributaries, lake inlets and outlets were programmed in R (v.4.3.1, R Core Team [2024]).

1.2.2 Fluvial discharge, velocity and residence time

Discharge and velocity were measured in a few streams ranging Strahler orders 1-4 within two headwater sub-watersheds (Petite Romaine and Bernard) in 2015/2016 and 2021/2022, respectively, using a 2-D Acoustic Doppler Velocimeter (FlowTracker, Sontek, San Diego, CA, USA). Additionally, a few hydrological stations located at Strahler orders 5-7 were periodically measured for discharge and velocity using a vessel mounted Acoustic Doppler current profiler (data provided by Hydro-Québec, Montréal, QC, Canada). All May/June measurements were categorised as high flow, while all other months were identified as low flow conditions. Combining these datasets spanning over the years 2010-2022 and Strahler orders 1-7 (n = 103), we identified a model to estimate discharge (Q) using catchment area (km^2) by flow condition (Fig. S8a):

$$log\ Q_{high} = 0.26 + 14.03 \times log\ CA - 1.96 \times log\ CA^2$$

$$log\ Q_{low} = 0.26 + 14.03 - 0.28 \times log\ CA - 1.96 \times log\ CA^2$$

where *CA* is catchment area in km^2 and *Q* is discharge in m^3 s^{-1} (R^2 = 0.88). Discharge and catchment area were log-transformed to fulfil model assumptions. Velocity was also estimated using catchment area and flow condition as follows (Fig. S8b):

$$log\ v_{high} = -0.49 + 3.22 \times log\ CA - 2.17 \times log\ CA^2$$

$$log\ v_{low} = -0.49 + 3.22 - 0.29 \times log\ CA - 2.17 \times log\ CA^2$$

where *CA* is catchment area in km^2 and *v* is velocity in m s^{-1}. Although the R^2 was not particularly high (R^2 = 0.46), the magnitude of the measured velocity as well as the relationship of velocity among Strahler orders found in the empirical data was captured with this model (Fig. S9). These two models were the most accurate and parsimonious solutions in comparison to directly calculating velocity from estimated channel cross-sectional area and discharge. WRT in each fluvial pixel (WRT_{px}) was estimated by dividing the pixel length by the estimated local velocity.

1.2.3 Reservoir water residence time

To estimate WRT_{px} within the reservoir, the overall water residence time within the reservoir was calculated for each month. Daily water level measurements for the reservoir were transformed into volume estimates using a conversion key provided by Hydro-Québec. The conversion key was implemented using a general additive model (GAM) approach as the relationship between water level and volume was not linear (data not provided). Discharge stations located at the turbine and overflow (water not used for electricity production) of the reservoir were used to calculate the daily outflow from the reservoir. Daily estimates were averaged by month and monthly water residence time was computed as:

$$WRT_{Reservoir} = \frac{V}{Q_{out}}$$

where *V* represents reservoir volume in m^3 and Q_{out} the discharge at the outflow in m^3 s^{-1}. Subsequently, the monthly reservoir WRT was divided by the number of pixels along the main channel of the reservoir to estimate the WRT_{px} along the main channel.

1.2.4 Lake water residence time

Due to the lack of data for remote lakes sampled in the watershed, WRT estimates were taken from the HydroLAKES dataset [Messager et al., 2016]. Similarly to reservoir estimates, the overall WRT in the lakes was divided by the number of pixels along the main channel flowing through the lake to estimate WRT_{px}.

1.2.5 Flow-weighted water age

A schematic representation of the FWWA calculation can be found in figure S10. The calculated FWWA ranged between 0.002 and 1,355 and 0.004 and 2,114 days for the sampled sites within the watershed in spring and summer, respectively. Stream FWWA increased along the Strahler order gradient with lowest FWWA in orders 1 (0.002 - 35.8 d), 2 (0.06 - 287.0 d), 3 (71.6 - 172.9 d), and higher values found in orders 4 (166.7 - 1,104.8 d), 5 (457.4 - 1,355.9 d), 6 (La Romaine upstream of reservoir cascade, 518.4 - 978.8 d) and 7 (La Romaine downstream of reservoirs, 523.0 - 820.1 d). Systems with long water residence time, such as lakes and reservoirs, led to increased FWWA, with reservoir FWWA ranging between 522.5 and 726.3 days, and lake FWWA between 3.0 and 2,114.3 days. These patterns emerge due to the lower water velocities within lakes and reservoirs. The outlet of a lake and reservoir always reflects this longer water age through the system, yet as streams and rivers with low FWWA merge into these outlets further downstream, the FWWA starts to decrease once again. Hence, depending on the upstream history of each stream and river, streams of the same order can have very different FWWA.

1.3 Modelling and classifying spatial patterns

Before modelling spatial patterns, the dataset was filtered by each sampling campaign. Extreme outlier observations of an operational taxonomic unit (OTU) or molecular formula (MF) were removed (values above or below 3 times the interquartile range), MF and OTUs with less than 8 and 7 observations, respectively, within a campaign were removed (based on histogram observations). Observations were z-scaled for each OTU and MF per campaign.

Prior to the modelling exercise, observations were binned and averaged at a 50-day interval along the water age gradient. A MF and OTU was only considered for modelling if more than 3 bins recorded an actual observation. To allow various dynamic patterns to be modelled, we established a decision tree that selects the best model for individual MF and OTUs along the FWWA gradient (illustrated in Fig. S11). The decision tree starts with first fitting a linear model and then comparing it to polynomial (2nd and 3rd order) and GAM models using maximum likelihood estimation (*gam()* function; *mgcv* package; Wood [2011]). Each bin was weighed by its corresponding number of observations. The best model was selected based on the smallest Akaike Information Criterion (AIC). If the model with the smallest AIC was a non-linear model, non-linearity was justified by testing it against the linear model using a χ^2-test (*anova()*; *mgcv* package; Wood [2011]). Once the best model was selected, model statistics such as *p*-value and R^2 were extracted. The slope of the initially fit linear model was extracted regardless of which model type was selected to aid in spatial pattern classification described below. For all non-linear models, the 2nd derivative was utilised to find peak locations along the FWWA gradient.

All linear models that had a positive slope were classified as 'increase', while all linear models with negative slopes were identified as 'decrease'. All non-linear models were first classified by their number of peaks. Models with a single peak were classified by where their peak was located. For each model, the centre along their respective water age range was identified. Subsequently, a buffer area was defined around the centre by adding and subtracting 1/6 of the FWWA range to the centre. If a peak was located within the buffer area, the model was classified as 'unimodal'. All other one-peak models were classified as 'non-linear decrease' if their peak was located below the FWWA range centre and models with their peak located higher than the FWWA range centre was classified as 'non-linear increase'. Any non-linear models with more than one peak were classified as multimodal increase and decrease when their linear model slope was positive and negative, respectively. Non-linear and multimodal spatial patterns were merged depending on the analysis. All models that had a slope of 0 or did not return a *p*-value were removed from downstream analyses.

2 Supplementary results

2.1 Environmental conditions

Dissolved organic carbon (DOC) concentrations in La Romane river, Romaine 1 and 2 reservoirs and in the lakes within the watershed ranged from 5.5 to 7 mg C L^{-1}, and where higher in some of the tributaries, up to 10 mg C L^{-1} [Barbosa et al., 2023]. Bacterial abundance (ml^{-1}) was highest in soilwaters (4.7 x10^7 $\pm$ 8.6 x10^7; means $\pm$ standard deviation), followed by streams Strahler order 2 (1.6 x10^6 $\pm$ 1.1 x10^6). Rivers of orders 6 and 7, lakes and reservoirs had similar bacterial abundances between 1.4 x10^6 and 1.5 x10^6. Abundance in streams of orders 1 to 5 ranged between 1.0 and 1.2 x10^6. Bacterial density in aquatic samples were generally higher in summer (1.6 x10^6 $\pm$ 9.4 x10^5) over spring (1.3 x10^6 $\pm$ 5.5 x10^5).

2.2 Richness of DOM and microbial assemblages

We did not observe a clear seasonal pattern in the number of examined entities (richness) for each assemblage. For the microbial community, there were in general more OTUs in 2015 (2,135 $\pm$ 35) than 2016 (1,532 $\pm$ 25). Across sampling campaigns, the number of OTUs was mostly stable within years with no seasonal differences. In contrast, the DOM assemblage exhibited seasonal and inter-annual differences in the number of MF. While the MF count was stable within the year 2015 (7,375 $\pm$ 21), 2016 exhibited in general a higher number of MF (10,616 $\pm$ 3,469) and an exceptionally high number of MF in summer ($\sim$13,000). The number of unique OTUs and MF within each campaign followed in general the same trend as the overall number and ranged between 176 - 482 for OTUs, and 0 - 4,030 for MF.

2.3 Spatial distributions of reactive MF and OTUs

We examined whether the two studied years were different in their spatial range along the continuum (i.e., whether the entire habitat range was sampled) to explore reasons underlying the observed differences in inter-annual patterns of MF spatial patterns. The spatial patterns in OTUs and MF that we identified were not constrained spatially and could occur in portions anywhere along the hydrological continuum. Hence, it is important to understand when and where along the network the spatial patterns preferentially occurred, and the network source of the moieties involved. We identified where the OTUs and MF first emerged (i.e., soil versus stream), and where they were last detected along the FWWA continuum (i.e., La Romaine river mouth versus lakes) by season and year. Larger lakes represent the endpoint of our hydrological

water age continuum, since water age is reduced in fluvial systems beyond the lake outlet due to influx of smaller streams and rivers. Most spatial patterns of MF spanned the entire continuum from soils to lakes, except for Summer 2015 and Spring 2016 as no soil and soil water samples were retained after initial quality filtration (Fig. S1). Given that the spatial pattern distribution is very dissimilar between Summer 2015 and Spring 2016, with higher proportions of decreasing MF and increasing MF, respectively (Fig. 2c), it is unlikely that the absence of terrestrial samples in both campaigns is contributing to the observed trends. Overall, the majority (>90%) of MF could be retraced to soils in both spring and summer, and the majority of these MF were detected along the entire range of FWWA (95% in summer and 77% in spring), such that most of the spatial patterns ranged from soils to larger lakes (Fig. S1). For the microbial assemblage the spatial range was very consistent between years, with 75% - 97% of DNA and RNA spatial ranges originating in soils regardless of season and year, yet the endpoint of detection within the network varied greatly. A greater proportion of OTUs ($\sim$60%) were only found until the river mouth especially in summer, indicating that not all OTUs are present in lakes. These results collectively indicate that microbes are patchier in their spatial distribution, while molecular formulae are more continuous along the network. Regardless, both assemblages could be overwhelmingly traced back to soils.

2.4 Phylogenetic underpinning of microbial spatial patterns

To evaluate whether there was a phylogenetic signal underlying the spatial patterns, a heatmap representing the reactive spatial patterns (i.e., increase, unimodal, and decrease) was plotted onto the microbial phylogenetic tree (Fig. S2). It was visually evident that many microbial OTUs had very few reactive representatives (empty spaces), however, those that were identified as reactive, generally seemed to have the same spatial pattern across campaigns and between DNA and RNA. To directly test whether there is phylogenetic signal in the distribution of spatial patterns, UniFrac distances were computed within two OTU matrices: bulk relatedness and the relatedness among reactivity groups (i.e., reactive versus non-reactive). Within the bulk pool, a strong phylogenetic signal among spatial patterns was found (Fig. S3a), which was statistically supported by a PERMANOVA analysis (F_6 = 2.03, R^2 = 0.37, $p < 0.0001$). However, there was no phylogenetic signal by season or year ($p > 0.8$). Decreasing and increasing linear OTUs were closely related, while non-linear spatial patterns started to exhibit phylogenetic divergence with unimodal OTUs being most phylogenetically dispersed. When we considered the reactive and non-reactive fractions in a separate analysis, it was also found that there was a phylogenetic signal between spatial patterns (F_6 = 1.75, R^2 = 0.16, $p < 0.0001$) and that they also differed by reactivity groups (F_1 = 5.53, R^2 = 0.09, $p < 0.0001$). Similarly to the bulk pool, no seasonal or annual pattern was found ($p > 0.7$) (visually represented in a NMDS and PCoA in Fig. S3b-c, respectively). Within the phylogenetic multivariate space, linear patterns of the unreactive fraction were clustered, and the unreactive non-linear patterns were phylogenetically similar to the linear reactive spatial patterns. There was a general trend across the reactive fraction, where increasing spatial patterns were more phylogenetically similar to each other, while decreasing patterns were more dispersed in multivariate space (Fig. S3b). Phylogenetic clustering of unreactive non-linear patterns with reactive patterns may indicate that our modelling and filtering approach used to determine the reactive fractions may be conservative. Non-linear models may not be statistically significant due to their highly dynamic nature, and hence, a greater sample size and deeper sequencing depth may be needed to detect more reads for these OTUs to achieve 'reactiveness' as defined in our approach.

We computed indices of phylogenetic relatedness by spatial pattern to assess whether some spatial groups are more phylogenetically constrained than others. Tests were carried on the nearest taxon index (NTI) and net relatedness index (NRI) for both the bulk and reactive pool separately, to examine whether parsing out the reactive portion helps in identifying a phylogenetic signal (Kruskal-Wallis Rank Sum Test). Only NRI was significantly different among spatial patterns within the reactive pool (NRI: *d.f.* = 2, χ^2 = 7.42, $p < 0.05$; NTI: *d.f.* = 2, χ^2 = 5.69, p = 0.058). Dunn's test revealed that decreasers were statistically more over-dispersed than increasers and unimodals ($p < 0.05$, adjusted via Benjamini-Hochberg). In contrast, there were no phylogenetic differences among spatial patterns within the bulk pool (NTI: *d.f.* = 2, χ^2 = 2, p = 0.37; NRI: *d.f.* = 2, χ^2 = 3.5, p = 0.17) (Fig. 3a, data of NTI not shown). These results suggest that identification of the reactive pool on the basis of spatial patterns enhances our insight or understanding of the ecological and phylogenetic basis of microbial community assembly, which is not forthcoming from the analysis of the bulk community.

2.5 Chemical and functional similarity underlying molecular spatial patterns

We used chemical similarity of DOM formulae to build a hierarchical clustering of MF, as an analogous analysis of microbial phylogenetic similarity. The unsupervised clustering technique revealed five clusters that were characterised by different mass, aromaticity (AI_{mod}) and metabolic potential (NOSC). In essence, the clustering order (1-5) represents

an ascending gradient for H/C and descending gradient for O/C, AI_{mod} and NOSC (Fig. S4, Table S2). All clusters were significantly different from each other in median H/C, O/C, C/N, AI_{mod}, NOSC and in the number of C, H, O atoms. Clusters 1 and 5, and cluster 2 and 4 were not different from each other in mass. Clusters 1 and 2 had no statistical difference in the number of N. The retrieved chemical dendrogram from the hierarchical clustering approach was plotted together with the reactive spatial pattern heatmap (Fig. S2) to visualise the distribution of MF spatial patterns by chemical similarity. In contrast to the microbial heatmap, many more MF exhibited a reactive spatial pattern, yet the direction of the pattern was often not consistent for a given MF and varied mostly by year.

To similarly test whether spatial patterns in MF were associated to chemical similarity, we applied the same phylogenetic indices (NTI and NRI) to the chemical dendrogram and tested whether there was a difference by spatial pattern within bulk and reactive pools (Kruskal-Wallis Rank Sum test). No difference was found in how the spatial patterns were spread across the dendrogram within the bulk pool (NTI: *d.f.* = 2, χ^2 = 4.9, p = 0.09; NRI: *d.f.* = 2, χ^2 = 3.5, p = 0.17). However, within the reactive pool a statistical difference was found for NTI (*d.f.* = 2, χ^2 = 8.0, $p < 0.05$) but not for NRI (*d.f.* = 2, χ^2 = 1.0, p = 0.59). For NTI, Dunn's test revealed that increasers were more clustered than unimodal patterns ($p < 0.05$, adjusted via Benjamini-Hochberg) (Fig. 3b). It is noteworthy that in general all spatial groups were clustered (NTI > 0) rather than overdispersed, with the least clustering observed in unimodal MF.

We further tested whether the spatial patterns were associated with functional differences in their nominal oxidation state of carbon (NOSC) and aromaticity index. There were statistical differences in both metrics among spatial patterns (KW-test, bulk: *d.f.* = 2, χ^2 = 355.58, $p < 0.0001$; reactive: *d.f.* = 2, χ^2 = 114.1, $p < 0.0001$) regardless of the examined pool ($p < 0.0001$, Fig. S5). NOSC increased from decreasing (0.02 ± 0.48, mean $\pm$ SD), to increasing (0.08 ± 0.48) to unimodal MF (0.23 ± 0.39). The spatial patterns were also significantly different in their aromaticity regardless of the reactivity pool ($p < 0.0001$), with AI_{mod} increasing from decreasing (0.28 ± 0.18), to increasing (0.32 ± 0.20) to unimodal (0.39 ± 0.20) MF. These results indicate that there are strong functional differences among spatial patterns even when the bulk pool is considered.

2.6 DOM properties underlying the spatial correlation between DOM formulae and microbial OTUs

In the previous section we showed that defining and extracting the most reactive microbial and molecular moieties results in the reduction of spurious patterns and in an improvement in our capacity to distinguish coherent relationships between DOM MF and microbial assemblages. Four correlation categories (hereafter, CCs) were identified: increasing OTU x increasing MF (CC1), increasing OTU x decreasing MF (CC2), decreasing OTU x increasing MF (CC3), decreasing OTU x decreasing MF (CC4). We further explored if the correlations among reactive moieties showed any biogeochemically coherent patterns in terms of distribution of DOM properties (Fig. 5). We calculated the proportion of significant correlations within each correlation category that involved MF in each of the 5 DOM clusters that we had previously defined. To simplify this analysis, we removed any correlations involving unimodal patterns, and those correlations that did not match our expected correlation signs, and we only focused on statistically significant correlations.

Unlike the proportion in spatial patterns, which did not show any clear seasonality (Fig. 2c), the relative contribution of DOM clusters to the total number of significant correlations by correlation category did (Fig. 5). The four CCs between MF and OTUs appear to be preferentially associated to specific DOM clusters, but these associations varied seasonally and inter-annually. Overall, the patterns of association were different between spring and summer, and inter-annual differences were especially evident in spring. In spring 2015, an overwhelming dominance of Cluster 1 on all four CCs was observed ($\sim$66%), however, overall dominance shifted to Cluster 2 in 2016 ($\sim$42%). This pattern of dominance of a few clusters in spring was systematically different from the pattern observed in summer. There is remarkable consistency in the summer patterns across years, where contributions of Clusters 2, 3 and 4 are much higher in CCs involving declining MF. In contrast, CCs involving increasing MF showed larger contributions of Cluster 1, 2 and depending on the year, cluster 4. Despite the observed consistency in the patterns of contribution of specific clusters across seasons, it is also clear that the microbial-DOM links were very dynamic – and there were inter-annual shifts in the contribution of certain clusters. For example, cluster 2 and 4 are almost non-existent in spring 2015, however, they contribute largely to the CCs in spring 2016 (42 and 36%, respectively). These results indicate that there are seasonal consistencies as well as inter-annual differences in the pools involved in microbial-DOM interactions that may be linked to hydrology, climate and watershed shifts (i.e., number/age of reservoirs in watershed).

3 Supplementary figures

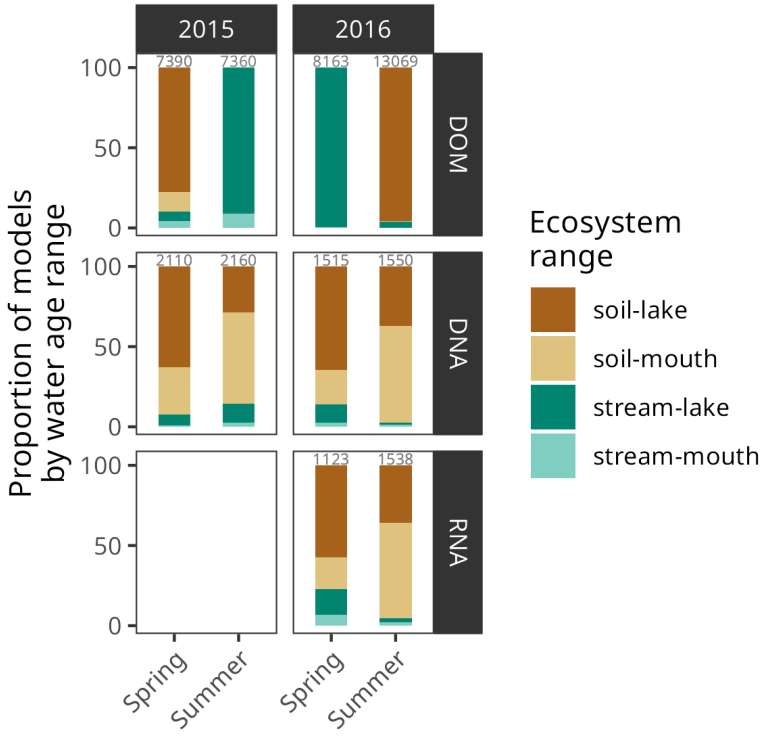

Figure S1: Distribution by ecosystem range for all spatial patterns. Soil includes soilwater and groundwater sites. Lake indicates any MF/OTU observed beyond the flow-weighted water age at the river mouth.

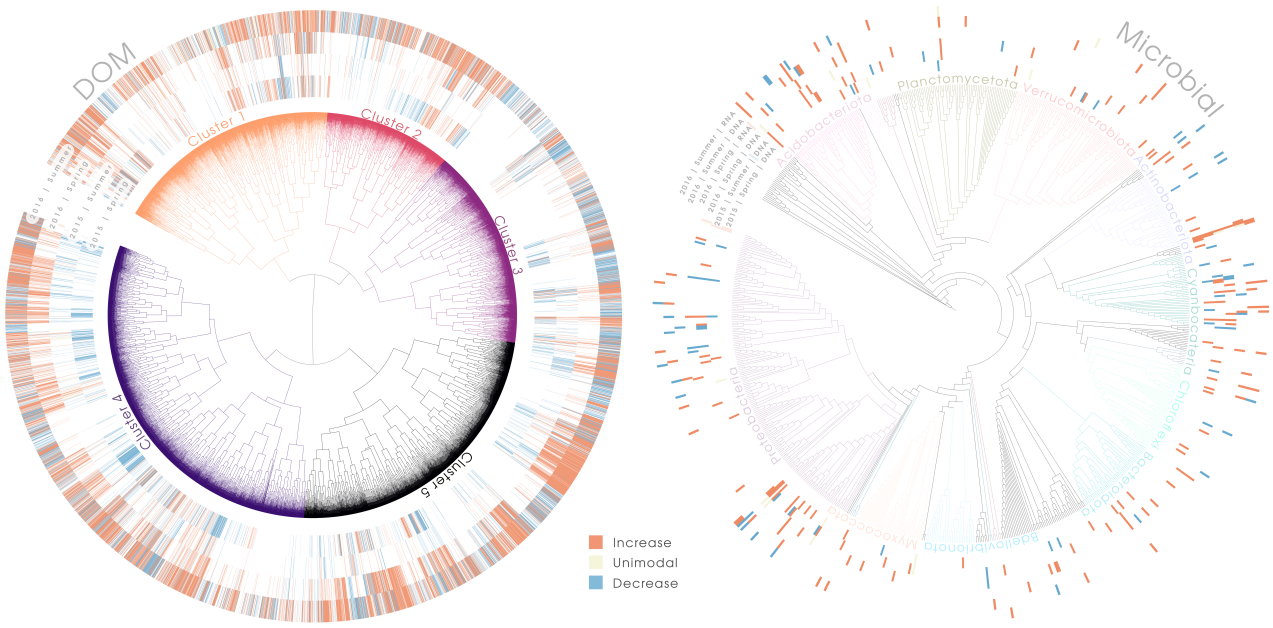

Figure S2: Heatmap of reactive spatial patterns on molecular and microbial tree. Left) Molecular tree derived from hierarchical clustering coloured by the identified 4 clusters. Heatmap rings represent 2015 Spring, 2015 Summer, 2016 Spring to 2016 Summer from inner to outer most ring. Right) Microbial phylogenetic tree coloured by phyla. Heatmap rings indicate the same temporal order as DOM tree, however, additional rings representing RNA for each campaign are added after DNA for 2016. Any spatial pattern that was identified as unreactive in a specific campaign is left blank.

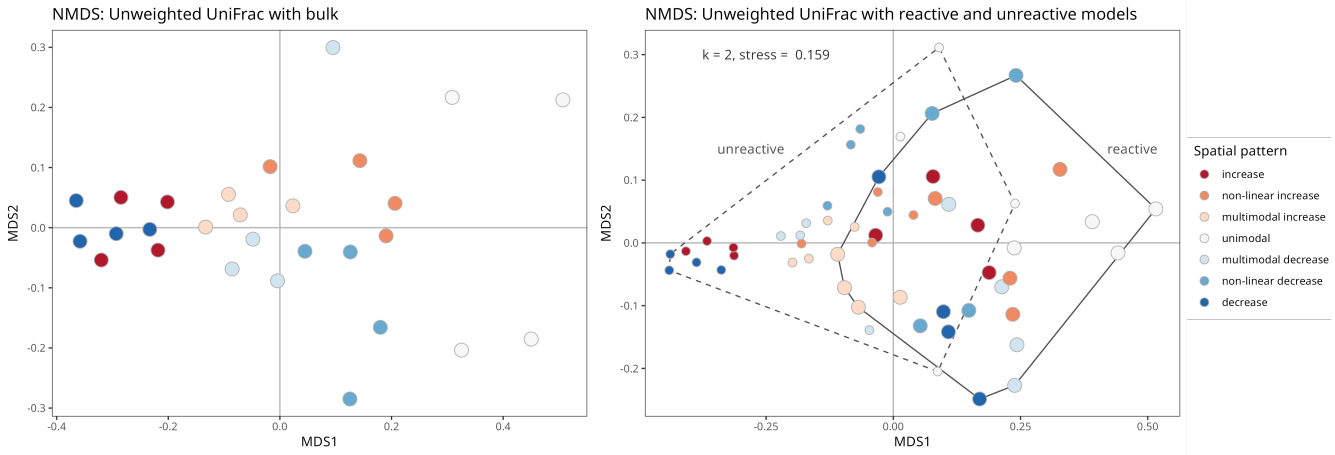

Figure S3: Unweighted UniFrac phylogenetic distance among spatial patterns in bulk and reactivity pools of microbial dataset. Non-metric multidimensional scaling (NMDS) of unweighted UniFrac distance on presence-absence transformed community matrix of microbial OTUs. Spatial patterns are distinguished by colour, reactivity pools (unreactive versus reactive) are depicted as different sizes in points as well as surrounded by polygons.

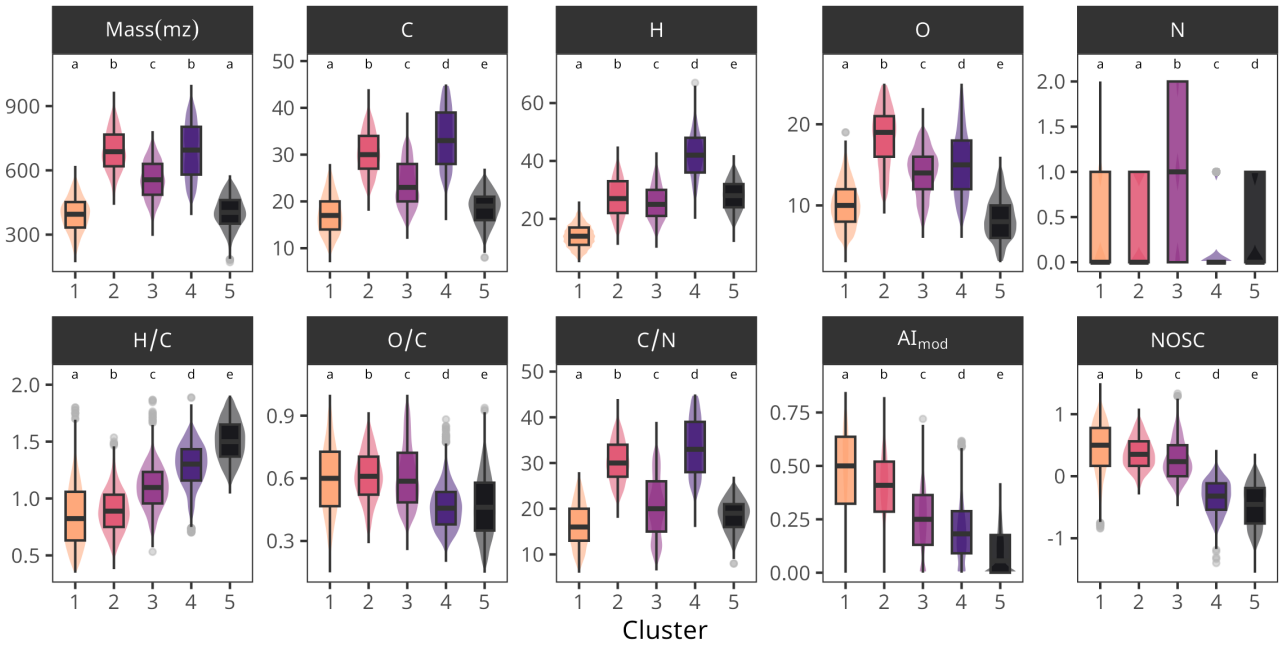

Figure S4: Chemical metrics used in hierarchical clustering analysis and their distribution among identified clusters. Given are chemical metrics such as the number of elements within a molecular formula (C = carbon, H = hydrogen, O = oxygen, N = nitrogen), mass (in *mz*), elemental ratios (H/C, O/C, C/N) and indicators of aromaticity (AI_{mod}) as well as nominal oxidation state of carbon (NOSC). Middle lines of boxplots represent the median, while the upper and lower hinges represent the 25^{th} and 75^{th} percentiles. Upper and lower whiskers expand to the largest and smallest value, respectively, no further than 1.5 times the inter-quartile range (IQR) from the hinge. Outliers are depicted as points. Clusters are identified as colours. The distribution of data is additionally depicted in the cluster colours around the boxplots.

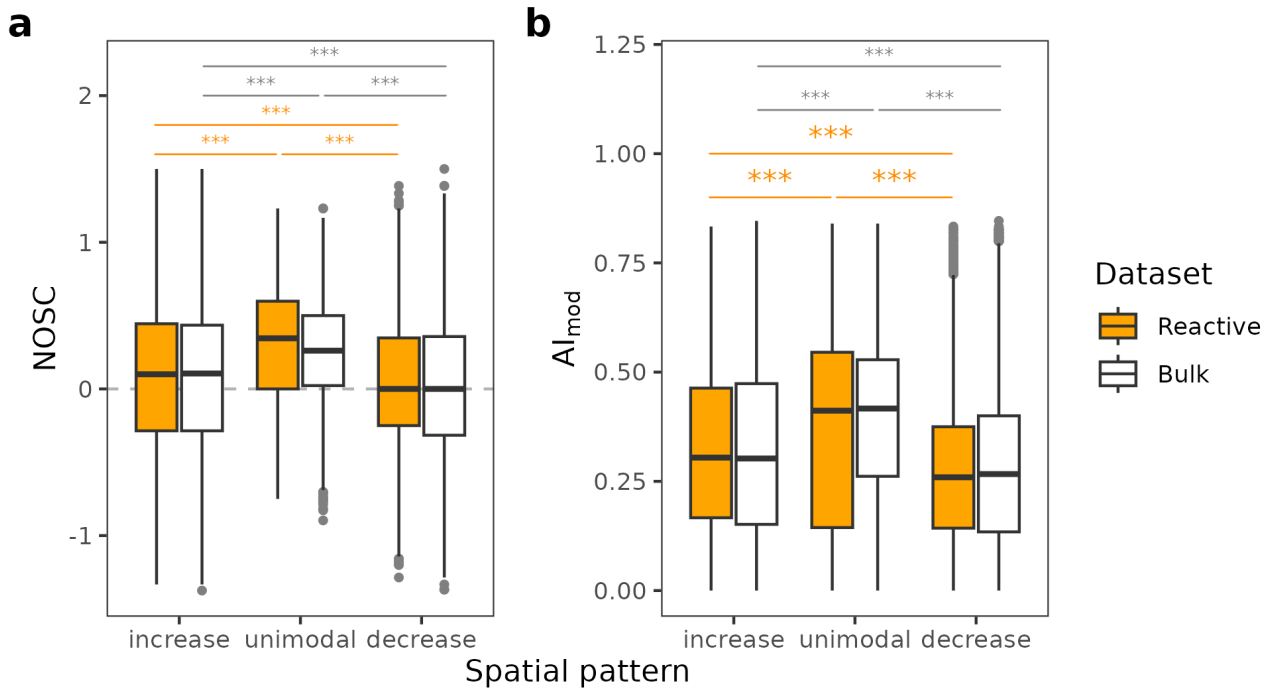

Figure S5: Differences in DOM intrinsic molecular properties between spatial categories. Differences in a) nominal oxidation state of carbon (NOSC) and b) in aromaticity (AI_{mod}) between spatial categories. The number of asterisks increase with lower *p*-values (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$) according to pair-wise comparisons with Dunn's tests. Statistical tests were only conducted to test differences between spatial patterns within the same pool (i.e. reactive and bulk).

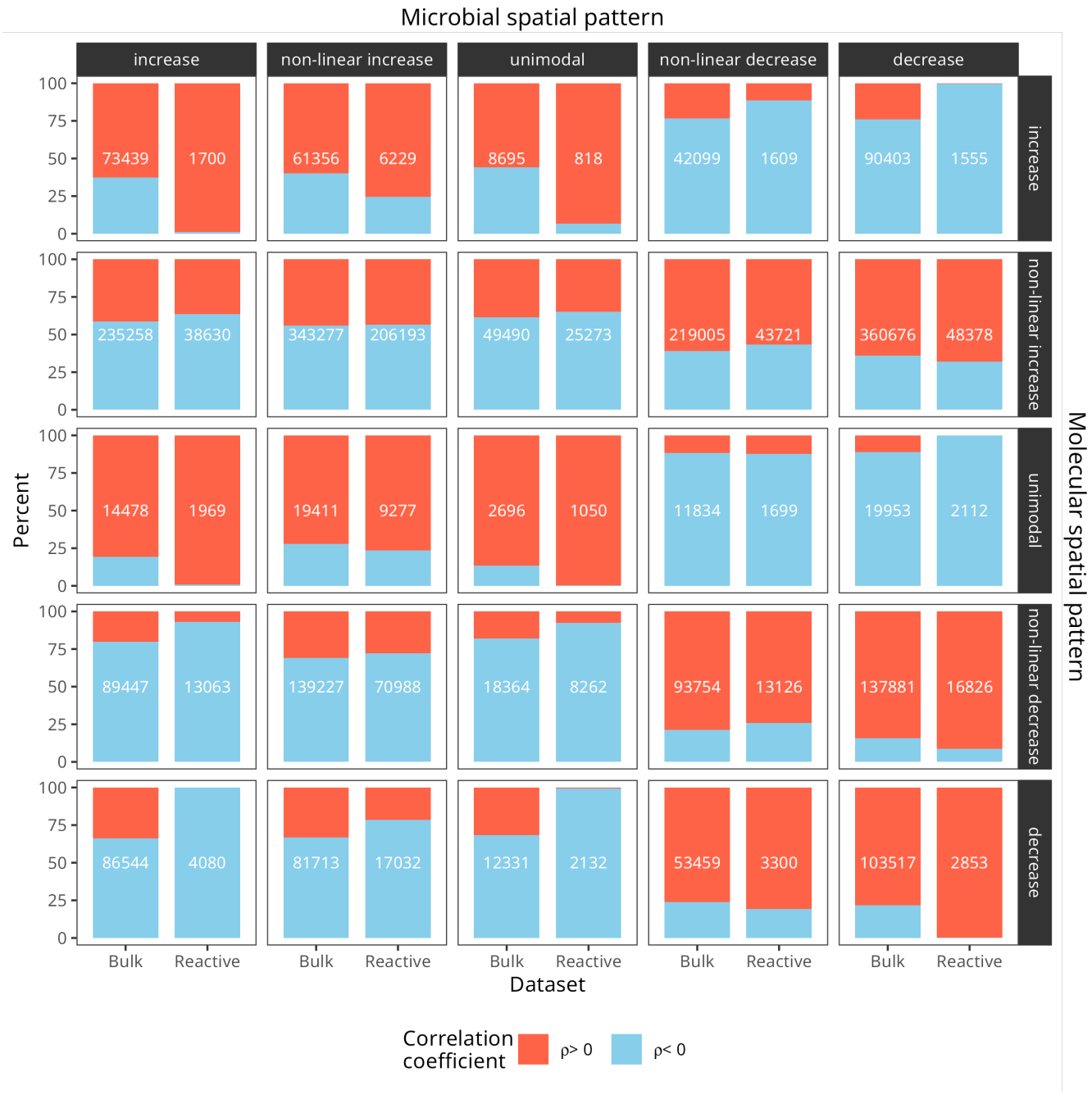

Figure S6: Proportion of significant positive and negative relationships between microbial and molecular spatial patterns. Percentages are given for the total number of correlations by pool (bulk versus reactive) and spatial pattern combinations.

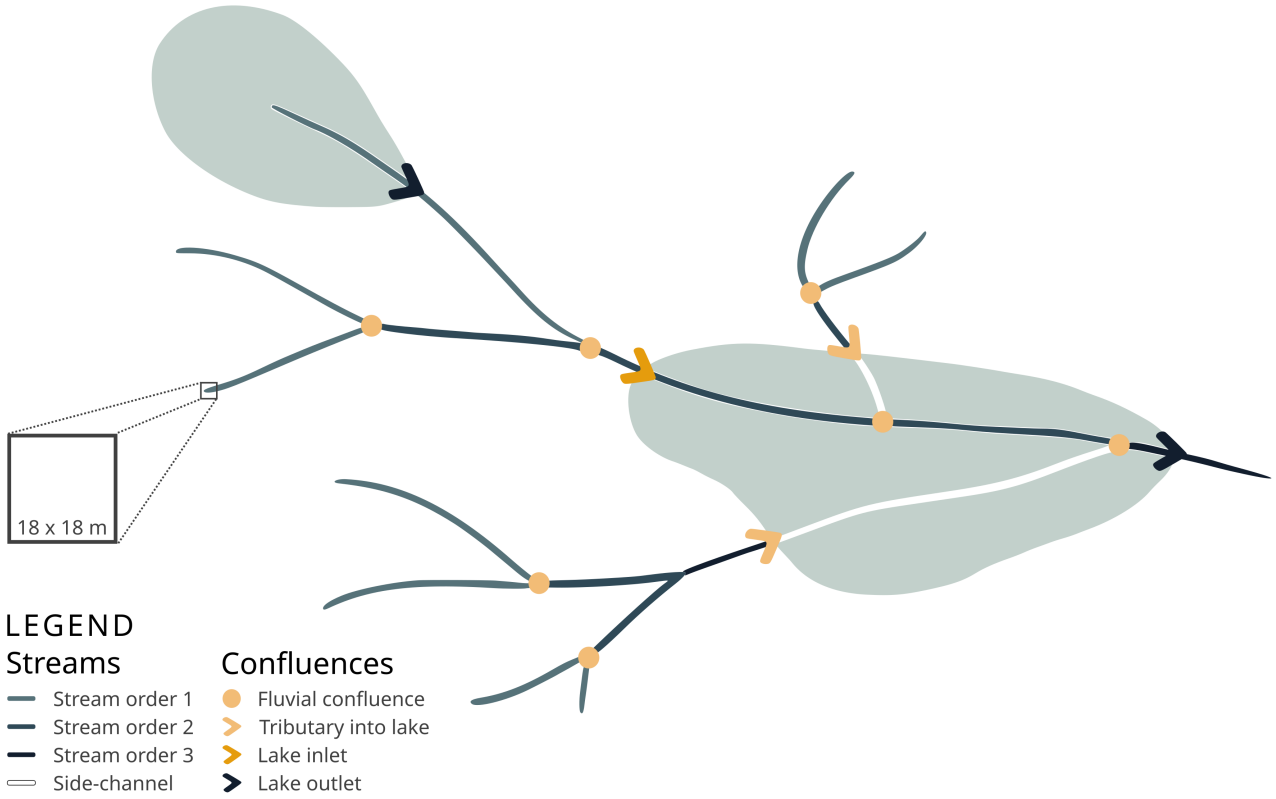

Figure S7: Schematic representation of flow-weighted water age calculation. Stream colours represent Strahler order. Within lakes only the coloured channel was considered when calculating flow-weighted water age (= main channel). The white channels' pixels were skipped during the cumulative calculation. Hence, the FWWA until the confluence to the lake (yellow arrow) was summed to the main channel where the side channel merges into the main channel (yellow point).

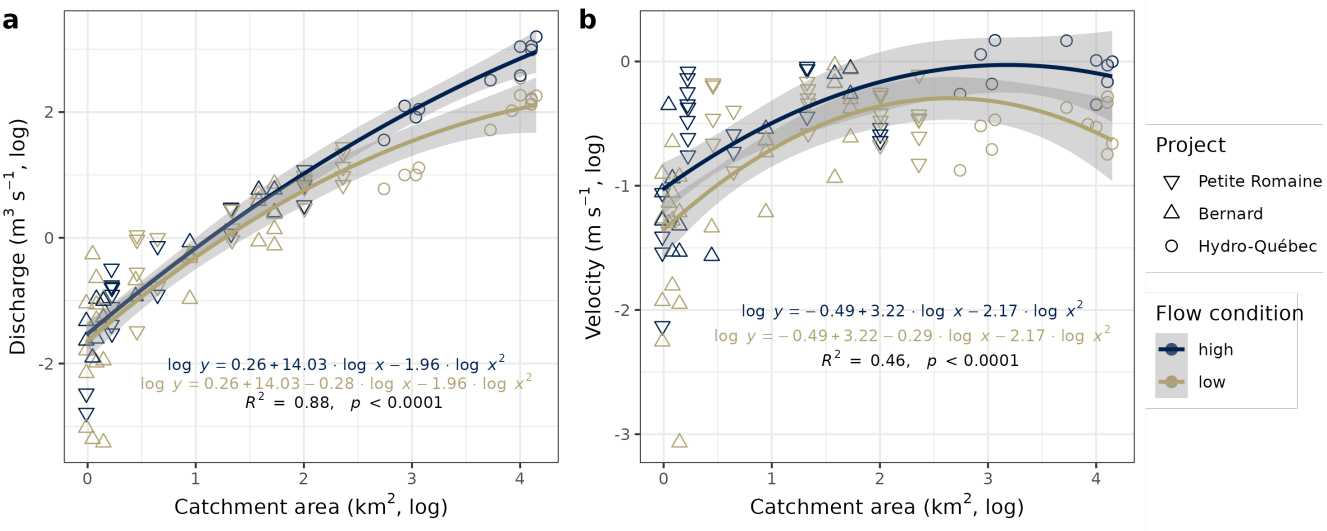

Figure S8: Hydrological models used to estimate water age in the watershed. a) Log-transformed discharge as a function of log catchment area in km^2. Model equations by flow condition are given in blue for high and brown for low flow. b) Log-transformed velocity as a function of log catchment area. Model equations are likewise given by flow condition. Various point shapes indicate the source of empirical measurements used to construct models. Petite Romaine and Bernard are sub-watersheds of La Romaine watershed, representing small headwater watersheds. Hydro-Québec data capture larger rivers within the watershed.

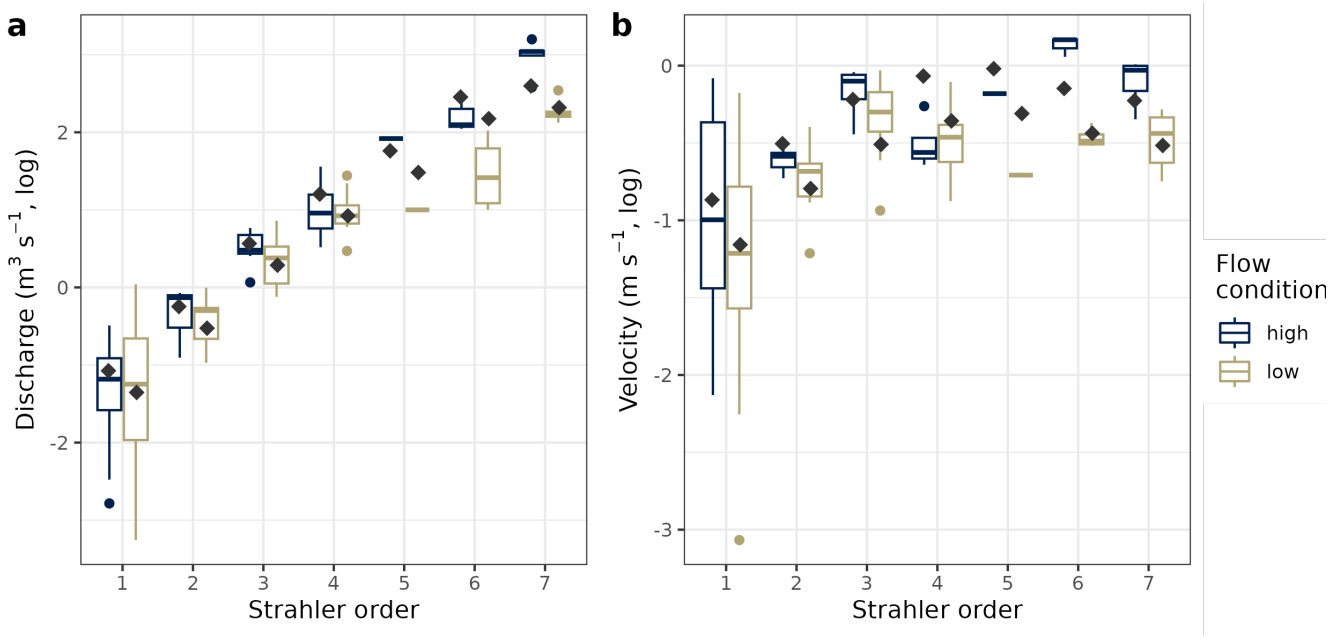

Figure S9: Measured versus estimated discharge and velocity by Strahler order. Diamonds indicate the median estimated value for each Strahler order and flow condition using the models presented in figure S8. Boxplots represent the measured discharge (a) and velocity (b) within the watershed. Blue and brown colours indicate high and low flow, respectively. The boxplot middle line represents the median, lower and upper hinges correspond to the 25th and 75th percentiles. Upper and lower whiskers expand to the largest and smallest value, respectively, no further than 1.5 times the inter-quartile range (IQR) from the hinge. Outliers are depicted as points.

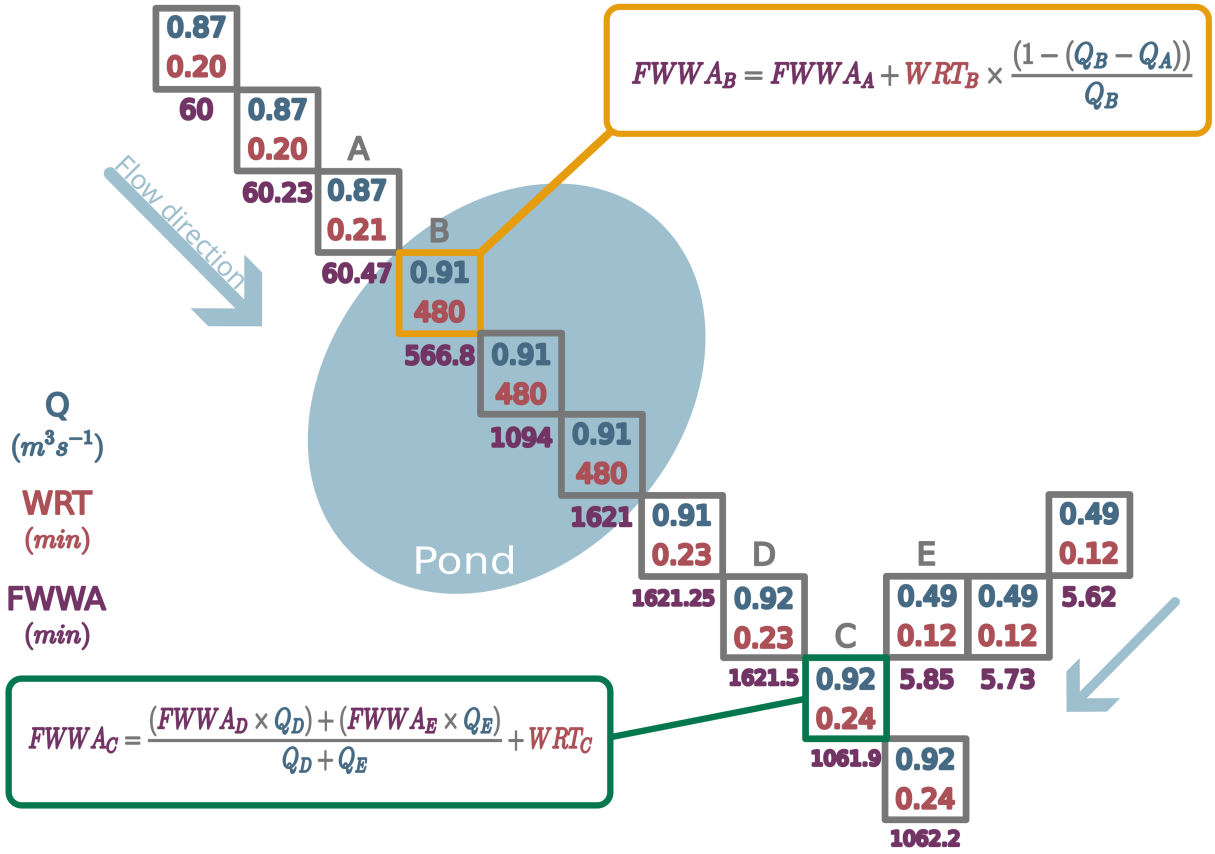

Figure S10: Example of flow-weighted water age calculation along a reach. Each square represents a pixel (GIS) along the aquatic network. For each pixel, colours distinguish between discharge (Q), water residence time (WRT) and flow-weighted water age (FWWA). Examples of mathematical formulae are given along a reach (yellow) and at a confluence (green).

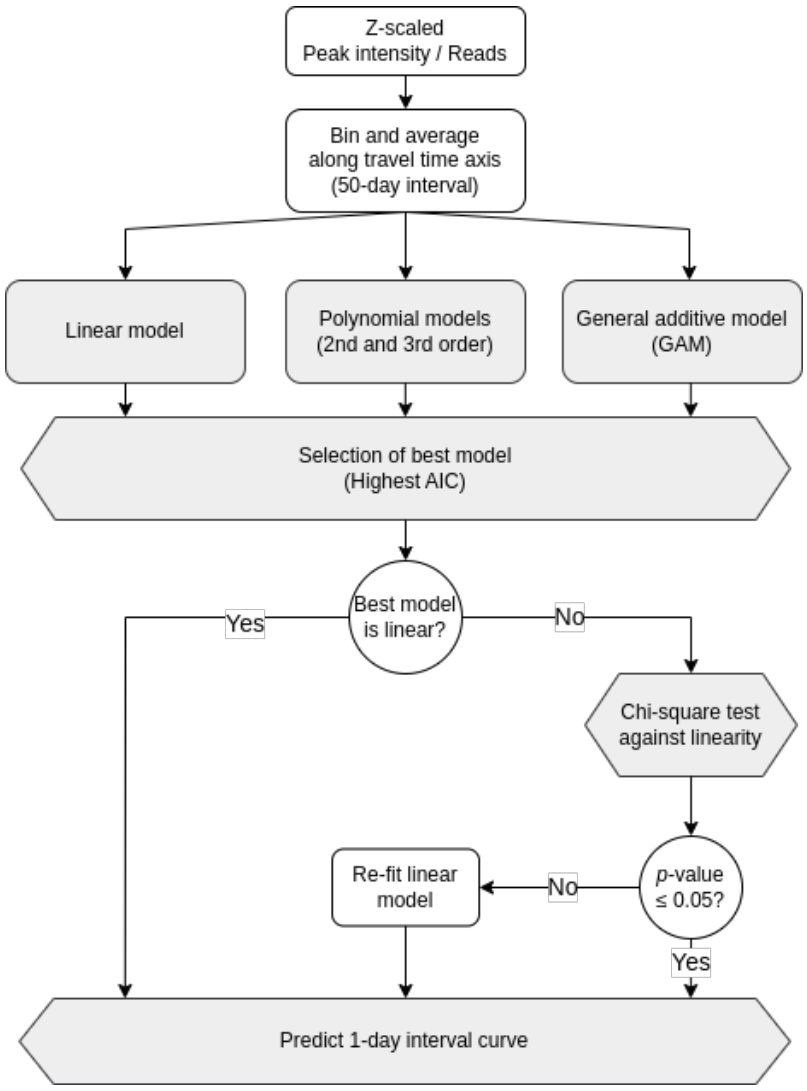

Figure S11: Flow chart illustrating decision tree utilised in modelling framework. Steps in the depicted decision tree were followed to find the best fitting model to characterise the spatial pattern for each operational taxonomic unit (OTU) and molecular formulae (MF) for each season and year. Grey boxes indicate steps involving decisions based on statistical information.

4 Supplementary tables

Table S1: Number of overall (n) and unique microbial OTUs and DOM molecular formulae per campaign.

		Microbial		DOM	
Year	Season	n	unique	n	unique
2015	Spring	2,110	447	7,390	66
2015	Summer	2,160	482	7,360	92
2016	Spring	1,515	176	8,163	0
2016	Summer	1,550	238	13,069	4,030

Table S2: Averages and standard deviations of chemical indices for identified molecular formulae clusters.

Cluster	Mass (m/z)	H/C	O/C	C/N	AI_{mod}	NOSC
1	392.6 $\pm$ 86.5	0.9 $\pm$ 0.3	0.6 $\pm$ 0.2	16.5 $\pm$ 5.1	0.5 $\pm$ 0.2	0.5 $\pm$ 0.4
2	691.9 $\pm$ 102.7	0.9 $\pm$ 0.2	0.6 $\pm$ 0.1	30.5 $\pm$ 4.9	0.4 $\pm$ 0.2	0.4 $\pm$ 0.3
3	556.2 $\pm$ 95.3	1.1 $\pm$ 0.2	0.6 $\pm$ 0.2	20.5 $\pm$ 7.1	0.2 $\pm$ 0.2	0.3 $\pm$ 0.3
4	693.0 $\pm$ 136.8	1.3 $\pm$ 0.2	0.5 $\pm$ 0.1	33.4 $\pm$ 6.6	0.2 $\pm$ 0.1	-0.5 $\pm$ 0.4
5	402.7 $\pm$ 79.7	1.5 $\pm$ 0.2	0.5 $\pm$ 0.2	18.6 $\pm$ 3.6	0.1 $\pm$ 0.1	-0.5 $\pm$ 0.4

References

Paul A. del Giorgio, David F. Bird, Yves T. Prairie, and Dolors Planas. Flow cytometric determination of bacterial abundance in lake plankton with the green nucleic acid stain syto 13. *Limnology and Oceanography*, 41:783–789, 1996. ISSN 0024-3590. doi: 10.4319/lo.1996.41.4.0783.

Natural Resources Canada. Canadian digital elevation model, 2017. URL `https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333`.

Mathis Loïc Messager, Bernhard Lehner, Günther Grill, Irena Nedeva, and Oliver Schmitt. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, 7:1–11, 2016. ISSN 20411723. doi: 10.1038/ncomms13603.

R Core Team. R: A language and environment for statistical computing, 2024. URL `https://www.r-project.org/`.

S.N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73:3–36, 2011.

Pedro M Barbosa, Pascal Bodmer, Masumi Stadler, Felipe Rust, Alain Tremblay, and Paul A. del Giorgio. Ecosystem metabolism is the dominant source of carbon dioxide in three young boreal cascade-reservoirs (la romaine complex, québec). *Journal of Geophysical Research: Biogeosciences*, 128:1–21, 2023. ISSN 2169-8953. doi: 10.1029/2022JG007253.