Prediction of Distributed River Sediment Respiration Rates using Community-Generated Data and Machine Learning

Stefan F. Gary[2], Timothy D. Scheibe[1], Em Rexer[1],
Alvaro Vidal Torreira[2], Vanessa A. Garayburu-Caruso[1],
Amy Goldman[1], James C. Stegen[1]

[1]Pacific Northwest National Laboratory, Richland, WA, USA
[2]Parallel Works, Inc., Chicago, IL, USA

Key Points:

- Machine learning models can predict river sediment oxygen consumption and explain up to about 65 percent of the variance.
- Sediment organic matter chemistry is one of the most important features for predicting these respiration rates.
- Large scale features like climate are also important factors for these predictions and they can be used to make maps of respiration rates.

Corresponding author: =name=, `=email address=`

Abstract

River sediment microbial respiration is a key indicator of ecosystem functioning and the biogeochemical fluxes across this critical zone link surface and subsurface waters. As such, there is tremendous interest in measuring and mapping these respiration rates. Respiration observations are expensive and labor intensive; there is limited data available to the community. An open science, collaborative initiative is collecting samples for respiration rate analysis and multi-scale metadata; this evolving data set is being used for making machine learning (ML) predictions at unsampled sites to help inform continued community engagement. However, it is a challenge to find an optimum configuration for ML models to work with this feature-rich (i.e. 100+ possible input variables) data set. Here, we present results from a two-tiered approach to managing the analysis of this complex data set: 1) a stacked ensemble of models that automatically optimizes hyperparameters and manages the training of many models and 2) feature permutation importance to detect the most important features in the models. The major elements of this workflow are modular, portable, open, and cloud-based thus making this implementation a potential template for other applications. The models developed here predict that sediment organic matter chemistry is one of the most important features for predicting sediment respiration rate. Other larger-scale, important features fall into the categories of climatic, ecological, geological, and fluvial settings. Leveraging these larger-scale features to generate data-driven estimates of river sediment respiration rates reveals spatially consistent but heterogeneous patterns across the river network of the Columbia River Basin.

Plain Language Summary

We want to determine the environmental factors that impact the amount of oxygen and nutrients that are used by microbes in river sediments. River sediment oxygen and nutrient use are important to river ecosystems but vary a lot between different locations. The number of measurements have been limited but are increasing thanks to volunteers participating in an open science project. Here, we use machine learning with existing data to make predictions of river sediment microbial oxygen consumption. The resulting machine learning models, and their predictions, are then used to estimate which aspects of the environment are the most important for making good predictions. It appears that the presence/absence of different kinds of nutrients for the microbes may be the most important factor in predicting oxygen consumption in sediment. Larger-scale factors, especially the climate, geography, and ecology of the river, have important roles, too. Finally, we use these models to make a map of oxygen consumption in river sediments across the Columbia River Basin. Maps like ours can be combined with river flow models to get a holistic understanding of river systems as well as guide future sampling efforts.

1 Introduction

River channels, together with their surrounding features (riverbed sediments, aquifers, riparian zones and floodplains), form a holistic entity known as the river corridor (Harvey & Gooseff, 2015). Exchanges of water between the actively-flowing surface channel and adjacent more slowly flowing surface and subsurface waters are known as hydrologic exchange flows or HEFs (Harvey et al., 2019). Subsurface HEFs include hyporheic exchanges, bank exchanges, cross-meander flows, and exchanges with in-river structures such as islands and longitudinal bars. These processes expose surface water and its constituents (e.g., nutrients and contaminants) to mineral surfaces and biological agents (e.g., microbes) that catalyze biogeochemical reactions. In particular, aerobic respiration (the mineralization of organic matter to carbon dioxide in the presence of oxygen) is a critical aspect of sediment contributions to riverine biogeochemical function (Butman et al., 2016;

Findlay, 1995; Rode et al., 2015). Biogeochemical reactions in hydrologic exchange zones have been shown to be responsible for up to 96% of respiration within riverine ecosystems (Naegeli & Uehlinger, 1997). As such, these exchange zones have been termed the "river's liver" due to their strong capacity to attenuate contaminants and process significant quantities of organic carbon and other nutrients (Fischer et al., 2005). However, estimates of the influence of exchange zones on river corridor biogeochemistry have varied greatly across systems, ranging from 3% to 96% (Naegeli & Uehlinger, 1997; Battin et al., 2003; Ward et al., 2018; Fuss & Smock, 1996; Kaplan & Newbold, 2000; Jones, 1995), and there is little consensus on the key factors that drive this variation. This leads to significant uncertainty in the parameterization of watershed-scale models that can predict the cumulative impacts of HEFs and river corridor biogeochemistry at large scales. To address this knowledge gap we need quantitative estimates of respiration rates that span diverse ecosystems and environmental conditions.

Recent research by our team has demonstrated that dissolved organic matter (DOM) chemistry is strongly correlated to biogeochemical processes in river corridor sediments, even more so than other variables such as microbial community composition or gene expression (Graham et al., 2018; Stegen et al., 2018; Garayburu-Caruso, Stegen, et al., 2020; Ahamed et al., 2023). Here we define DOM chemistry as thermodynamic properties and elemental composition inferred through molecular formulae assigned to the wide range of water soluble/extractable organic molecules present in samples as characterized by ultra-high-resolution Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS). The central question that this work seeks to explore is: To what extent does DOM chemistry impact river sediment respiration rates relative to other features of a river's environmental context?

We have engaged the global research community through the Worldwide Hydro-biogeochemistry Observation Network for Dynamics River Systems (WHONDRS) consortium (Stegen & Goldman, 2018) to develop consistent distributed data in dynamic river corridors across diverse watershed settings. These data include laboratory measurements of sediment respiration rates and characterization of DOM chemistry by FTICR-MS, as well as many other data and metadata variables of potential interest. This manuscript describes a machine learning (ML) analysis of a North American continental-scale WHONDRS data set to identify key variables that control sediment respiration rates, develop a data-driven predictive model of sediment respiration, and extrapolate observed respiration rates across large domains by cross-referencing to global river physical and chemistry databases, RiverAtlas (Linke et al., 2019) and GLORICH (Hartmann et al., 2014), respectively. These predictions of distributed heterogeneous respiration rates can subsequently be used in process-based watershed models to better understand the impacts of spatial and temporal variability of respiration on cumulative function at watershed and basin scales.

2 Methods and Data

In this section, we document the data sources, and hence the inputs and outputs of the ML models, used in this study. The ML models' inputs and outputs, also known as features and targets, respectively, differ between some of the ML models. First, we describe the observations of river sediment respiration rate that form the core foundation for this study, the WHONDRS data. Then, we outline the larger-scale global databases that are used to supplement the WHONDRS data, thus allowing the ML models to make predictions at rivers which are not in the WHONDRS data. Finally, we outline the details of the ML models.

2.1 River respiration rate observations: WHONDRS

In the summers of 2019 and 2022, the WHONDRS consortium coordinated the collection of surface water and sediment samples from rivers within the contiguous United States using a consistent protocol. Sampling kits were provided to volunteer participants along with video and written instructions for consistent sample collection. Site metadata were also collected in association with each of the physical samples. The resulting samples were sent back to the Pacific Northwest National Laboratory (PNNL) for analysis of a number of biogeochemical variables including DOM chemistry using multiple instrument platforms. Sediment samples were used in laboratory batch experiments to determine aerobic respiration rate coefficients using a standardized protocol. The data from these two sampling campaigns were published openly on the ESS-DIVE public data repository and are referred to herein as S19S (2019 samples, Goldman et al. (2020)) and SSS (2022 samples, Forbes et al. (2023)), respectively. The locations of the sampling sites are shown in Figure 1. The published data packages continue to be updated as additional analyses are completed, and have been used in several published analyses including some open, collaborative papers led by WHONDRS community members (Garayburu-Caruso, Danczak, et al., 2020; Borton et al., 2022; Stegen et al., 2022; Buser-Young et al., 2023; Stadler et al., 2023; Ahamed et al., 2023).

The variables of the WHONDRS observational data set are designated by "WH" in the data source column of Table S1. All of the variables in the WHONDRS data, except respiration rate, can be used as inputs (i.e. "features") for the ML models in this study. The respiration rates in the WHONDRS data are the sole outputs (i.e. "targets") of the ML models. Some sample analyses are ongoing so not all features are available in both data sets (S19S and SSS) and therefore the ML models used here are a progression of models with fewer and fewer features so that it is possible to include respiration rate observations without certain features and make predictions at more river sites. For example, SSS sample analysis using FTICR-MS is not yet complete, while it has been completed for S19S samples. The following subsections detail the collection of field samples and metadata, analysis methods, and postprocessing of the resulting data.

2.1.1 Collection protocol for samples and metadata

Garayburu-Caruso, Danczak, et al. (2020) describes details regarding metadata and sample collection for the S19S study and similar protocols were used for sample collection during 2022. Briefly, shallow sediment samples (1-3 cm depth) were collected at each site from depositional zones (Jensen, n.d.). Sediments were sieved in the field to < 2mm fraction. Surface water samples were collected at each site using a 60 mL syringe. Samples were filtered through a 0.22 μm sterivex filter (EMD Millipore) into vials designated for subsequent analysis (e.g., 40 mL glass vial (I-Chem amber VOA glass vials) for dissolved organic carbon (DOC) analysis, and 4 mL glass vials for isotope analysis). In addition, unfiltered surface water was also collected at each site. Water and sediment samples were stored on ice upon collection and were shipped on blue ice within a day of collection to Pacific Northwest National Laboratory.

Once in the laboratory, filtered water samples were stored at 4C until analysis. Sediment samples were sieved again in the laboratory (< 2mm), homogenized and sub-sampled for C/N analysis, flow cytometry and ultrahigh-resolution mass spectrometry, aliquots were stored at -20C, 4C or -80C respectively. Additionally, homogenized sediments were sub-sampled in triplicate into 40mL clear glass vials (I-Chem amber VOA glass vials) with a 0.5cm diameter factory calibrated oxygen sensor dot (Fibox 3; PreSens GmbH, Regensburg, Germany). Vials with sediments and unfiltered water from each site were kept in the dark inside the environmental chamber at 21C until next-day respiration incubations.

Table 1. Summary of data features used in the ML analyses. For brevity, broad groupings of variables are shown here; the detailed descriptions of the corresponding individual variables (feature IDs in the third column) are provided in Table S1 in the Supplemental Information.

Variable Group	Number of Features	Feature IDs	Description
Ecological Setting	33	0,15-18,23-24,33-36,39-40,46,65-79,112,121,125-126	Vegetation type, land cover/use, watershed features, demographic variables
Geological Setting	13	2-3,6,13-14,19-20,25-30	Includes soil and rock characteristics, and river geomorphology
Fluvial Setting	18	4-5,9-10,37-38,41-45,47,108,114,119-120,124,129	River type and slope, reach geometry, stream order, flow rates, macrophyte cover
Climatic Setting	20	80-84,90-99,102-103,110-111,128	Meteorologic and hydrographic variables over reach and watershed
Riverbed Sediment	9	1,11-12,100-101,105,118,130,132	Grain size, geochemical characteristics
Water Properties	6	21-22,87,106-107,127	Dissolved oxygen, temperature, pH, etc.
Microbiology	3	7-8,123	Cell counts
Sediment Organic Matter Chemistry	26	31-32,48-64,104,109,113,115-117,122	Summary measures of FTICR-MS spectra, e.g. Aromaticity Index, Nominal Oxidation State, Gibbs Free Energy, etc.
Global Databases	5	85-86,88-89,131	Geography of colocated GLO-RICH and RiverAtlas sites
Total Features:	**133**		

Some features of the WHONDRS site metadata consist of ranked and unranked categorical values and special consideration must be made for translating these features to be used by the ML models. The following steps were taken during the preprocessing of the WHONDRS data to prepare it for the ML workflow. The ranked categorical values (e.g. feature ID 129, canopy cover classified as: "no", "partial", or "full" coverage) are mapped to integers for input to the ML models (i.e. a scale from 1 to 3). Unranked categorical features, however, need to be expanded as "one hot" features. For example, feature ID 3 (hydrogeomorphology classes of "braided", "straight", or "meandering") is a feature whose classes should have no intrinsic relative magnitudes and as such they should not be represented on a scale of integers. One hot features replace the original feature with N features, one feature for each class of the unranked category. In the preceding example, a braided river would therefore be represented with the three features as 1, 0, 0. This approach was applied to feature IDs 0, 1, 2, 3, 4, and 5 with 21, 4, 9, 4, 2, and 2 classes, respectively. Technically, the number of inputs to each ML model that uses these one hot features is increased by replacing each unranked categorical feature by its corresponding one hot features. However, for clarity throughout the text and figures, we discuss the one hot features that come from an unranked categorical feature as a single feature and adjust the number of features listed in the tables describing the ML models accordingly. The one hot expanded features are not included in the feature counts; only the original feature is counted (e.g. feature ID 0 instead of its 21 one hot features). Finally, in the feature permutation importance, below, the one hot features that correspond to a single unranked categorical feature are always permuted together as a correlated feature block. Additional features can be added to their block if they are sufficiently correlated.
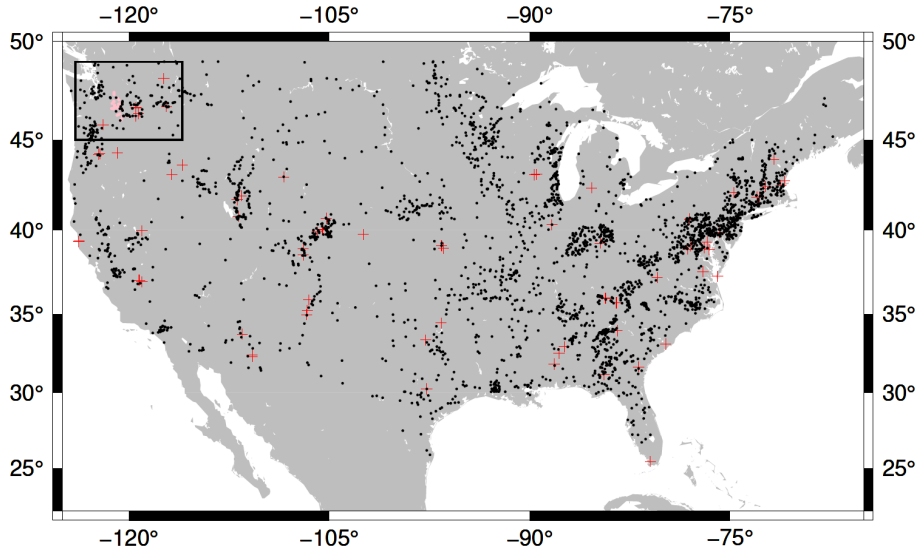

Figure 1. Map of the river sites where we have observations from the WHONDRS S19S (red crosses) and SSS (pink dots) campaigns and where we can make predictions using the merged RiverAtlas and GLORICH databases (black dots). A small number of WHONDRS S19S sites are outside this map but are not shown for simplicity. The black box in the upper left corner corresponds to the domain of the Columbia River Basin map discussed later in the manuscript.

2.1.2 Respiration rates

Respiration rates were determined following methods described by Garayburu-Caruso, Stegen, et al. (2020) and Stegen et al. (2023). Reactors consisted of 10mL of sieved sediments and ∼30–35mL of aerated unfiltered water with no headspace or in some cases also 2.5 mL of sediments and ∼35–40mL of aerated unfiltered water with no headspace. The reactors were shaken at 250 rpm in the dark and 21C for 2h. Dissolved oxygen (DO) concentration was measured noninvasively every 15 min for the first hour and every 30 min during the second hour for S19S and every 5 min for the first 15 min followed by every 15 min and every 30 min for the remainder of the incubation time. These measurements were performed using an oxygen optical meter (Fibox 3; PreSens GmbH, Germany) to read oxygen sensor dots (optodes) on the inside cover of the vials. Respiration rates (mg DO/L water per hour) were calculated as the slope of the linear regression between DO concentration and incubation time for each reactor, these were further normalized to per liter sediment by multiplying the rate by the liters of water in the incubation and dividing it by the liters of sediment in the equation. Therefore, the normalized rates units are milligrams of DO per liter of sediment per hour. Sediment and water volume for each reactor along with normalized and unnormalized rates can be found in the campaign specific data packages published on ESS-DIVE (referenced above).

The river sediment respiration rates reported here are always negative values because they represent the consumption of oxygen over time. Since using the adjectives "larger" and "smaller" can be confusing with negative values, here we use "weaker" and "stronger" to describe respiration rates that correspond to small amounts of oxygen consumed (e.g. -1 mg DO/L/hr) and large amounts oxygen consumed (e.g. -1000 mg DO/L/hr).

2.1.3 FTICR

For this analysis, ultra-high resolution mass spectrometry data was only available for the S19S study. Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) was used to generate mass spectra associated with water-soluble sediment organic matter pools. To access these molecules, field sediments were thawed and extracted with Milli-Q water as described in Garayburu-Caruso, Danczak, et al. (2020). The resulting supernatant was filtered (0.22μm Sterivex) and analyzed for DOC concentrations measured as non-purgeable organic carbon using a Shimadzu combustion carbon analyzer TOC-L CSH/CSN E100V with ASI-L autosampler. To allow comparability across samples and following WHONDRS standard protocol, samples were normalized to 1.5 mg C L^{-1}, acidified to pH 2, and the same volume was extracted with solid phase extraction (SPE) PPL cartridges following procedures described by Dittmar et al. (2008).

FTICR-MS analyses were carried out at the Environmental Molecular Science Laboratory (EMSL) in Richland, WA, using a 12 tesla (12T) Bruker SolariX FTICR mass spectrometer (Bruker, SolariX, Billerica, MA, USA) in negative mode. The instrument details and methodology used to assign molecular formulas are described in detail in Garayburu-Caruso, Danczak, et al. (2020). Briefly, the Compound Identification algorithm within Formularity (Tolic et al., 2017) was used to align mass lists generated using Bruker Data-Analysis V4.2 and assign molecular formulas using S/N > 7, and mass measurement error of < 0.5 ppm as the criteria. Formulas were assigned considering C, H, O, N, S, and P and excluding other elements. Resulting reports were processed using ftmsRanalysis (Bramer et al., 2020). This process removed peaks outside of a high confidence m/z range (200 m/z–900 m/z) and/or with a 13C isotopic signature and calculated the molecular formula properties used for in our analysis listed in Table S1 (Song et al., 2020; Koch & Dittmar, 2006, 2016; LaRowe & Van Cappellen, 2011; Kim et al., 2003).

2.2 Cross-referencing observational data with larger-scale databases

The WHONDRS data contains directly measured river sediment respiration rates, *in situ* chemistry observations, and locally collected site metadata. On the other hand, the global river database RiverAtlas (Linke et al., 2019) contains large-scale data such as climate indices, land use, soil properties, population metrics, and geographic information (e.g. elevation, slope, and stream order) at every river segment detected to 30 m resolution. These features are designated with "RA" in the data source column of Table S1 and are not present in the original WHONDRS data. We merged WHONDRS and RiverAtlas data at each WHONDRS site because both the RiverAtlas large-scale data and the WHONDRS locally collected observations have the potential to inform ML models of respiration rates. In particular, for each WHONDRS site, the closest RiverAtlas river segment was found and that segment's large-scale features were appended to the locally observed WHONDRS features.

The one feature that is available in both WHONDRS and RiverAtlas data sets is stream order, also known as the Horton-Strahler number (Horton, 1945; Strahler, 1957). This overlap provides an opportunity to cross-check the process of co-locating the River-Atlas river segments to WHONDRS sites. Figure 2 shows that the two data sets' stream orders are broadly consistent with each other. Stream order is an integer number with the smallest river segments set to 1 and larger integers representing river segments with more segments feeding into them. Stream order is sensitive to the resolution of the topography data used to find segments and their interconnections because the stream order numbering starts with the smallest streams that are detectable in the topography. The resolution sensitivity is the source of the approximate one-unit systematic offset between the WHONDRS and RiverAtlas stream orders. RiverAtlas is based on the 15 arc-second resolution (about 500 m at the equator) HydroSHEDS data for determining stream order (Linke et al., 2019) while the WHONDRS data were computed with the 30 m resolution National Hydrography Data set Plus (NHDPlus) data set (EPA & USGS, 2022). WHONDRS stream orders in Figure 2 are approximately 1 unit larger than RiverAtlas values because smaller segments are detectable in NHDPlus compared to HydroSHEDS. The 2 outliers in the lower right corner of the plot were manually inspected and determined to be an error in the WHONDRS data and were subsequently corrected; they are retained here as an example of the importance of this type of cross-checking when merging complex data sets. We use the RiverAtlas stream order, and not the WHONDRS stream order, when training the ML models in this study because RiverAtlas stream order is consistent across the sites we use for training the ML models (i.e. the WHONDRS sites where we have respiration rates) and the sites were we will make predictions of respiration rates (e.g. sites were we only have RiverAtlas data).

RiverAtlas does not include *in situ* chemical observations. River water chemistry may be helpful for predicting respiration rates so temperature, pH, oxygen concentrations, and percent saturated oxygen values from the GLObal RIver CHemistry (GLORICH) data set (Hartmann et al., 2014) were downloaded from PANGEA (Hartmann et al., 2019) and merged with the RiverAtlas data in the same way that data at WHONDRS sites were merged with RiverAtlas data described above. In particular, for each site in the GLORICH data with temperature and pH and O_2 or percent saturated O_2, the large-scale data from closest RiverAtlas river segment was appended to the *in situ* chemical observations. For GLORICH sites with time series or multiple samples, only the average values were used. Furthermore, for any GLORICH sites where either the O_2 concentrations or the percent saturated O_2 was missing, whichever available oxygen value that was available was used to reconstruct the corresponding missing oxygen value by estimating the saturated O_2 at a given site based on its temperature, elevation, and an assumed salinity of 0 (Rice et al., 2017; Rajesh & Rehana, 2022). A total of 6,170 GLORICH sites' *in situ* chemistry observations were colocated with larger-scale RiverAtlas data.
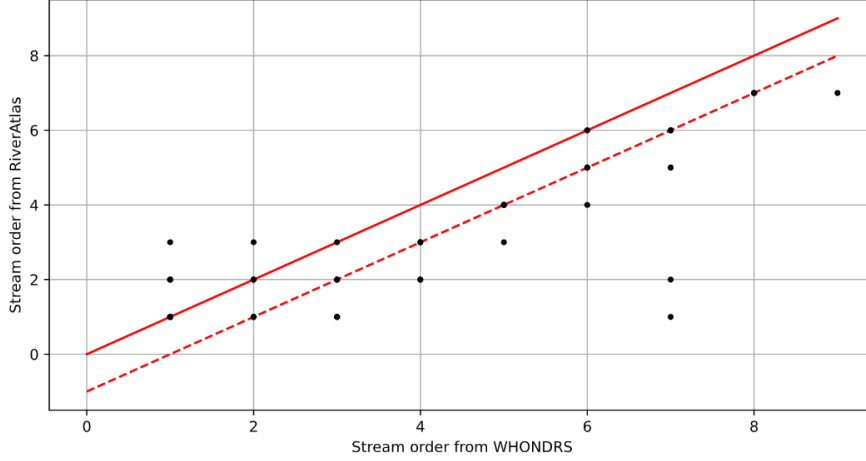
Figure 2. Comparison between stream order from the WHONDRS and RiverAtlas databases at the WHONDRS sites. The lines are the 1:1 line (solid) and the 1:1 line shifted down by exactly one unit (dashed) to highlight the effect of topography resolution differences impacting stream order in the two databases. Since stream order is always an integer, the 97 data points plotted here are often plotted on top of each other. There are only 97 WHONDRS stream order data points since there are multiple samples at some of the S19S WHONDRS sites.

2.3 Machine learning model training

2.3.1 Ensemble of SuperLearner ensembles

The machine learning framework used here (Figure 3) is based on an ensemble of ensembles approach to automatically train and identify the best performing ML models while also accounting for variability due to random train/test data splits and the random initialization of some ML algorithms. Any missing values in the training or testing data were filled with whole data set mean value for that feature. The innermost layer of the ensemble of ensembles is a SuperLearner instance; each SuperLearner is a stack of 15 scikit-learn-compatible ML submodels (Table 2) that are trained in parallel and the outputs (i.e. predictions) of each submodel are weighted by the submodels' relative scores (i.e. the R^2 between the observed and predicted respiration rates in the training data). These submodel weights are determined with a non-negative least squares optimization that is unique to each SuperLearner instance (Owoyele et al., 2021). We use predictions of the stacked ensemble of ML models within each SuperLearner, and not just predictions from the best performing model, because the stacked ensemble usually makes predictions that are as good as or better than the best model within the ensemble (Wolpert, 1992; Owoyele et al., 2022). We use 5-fold cross-validation (CV) for the hyperparameter optimization (HPO) of each submodel independently in each SuperLearner instance. A Bayesan search algorithm is used to explore the best hyperparameters for each submodel as specified by submodel-specific search domains (Owoyele et al., 2021, 2022). This approach to HPO means that all of the data (i.e. training *and* testing sets) are used for selecting the best hyperparameters of each submodel but only the training set is used to train the models themselves once the hyperparameters have been selected.

The second layer of the ML framework consists of an ensemble of 10 SuperLearner instances; each instance is trained and evaluated on a different randomized 75% train and 25% test split of the observational data and each instance runs HPO and trains its own submodels independently of the other SuperLearners. We use an ensemble of SuperLearner instances instead of CV to assess the variability of models because we want

uniform access to all the ensemble members, not just the best model as judged by the model score which is traditionally what is returned by CV. As will be seen below, model score alone is not the sole indicator of model performance for this data set. Automated HPO, training many ML submodels, and ensembles of SuperLearners sidesteps the need to select a single ML model or ML algorithm.

During training and prediction stages, each SuperLearner is initialized with its configuration file that lists the submodels, hyperparameter search ranges, and how the submodels are weighted together. For all the ML models trained here, the majority of the ML models were trained with the respiration rates first transformed by a log10 filter and then scaled by a min-max filter. These transformations were applied to the targets because the respiration rates span many orders of magnitude and the log10 filter helps to keep the data closer to a Gaussian distribution which is better for most ML regression models. One ML model training case was reserved for training a model without the log10 transformation as a way to check for its impact.

All the ML model ensemble members were trained on a cloud-based SLURM cluster in parallel. Each SuperLearner instance's compute needs were relatively modest (2 CPU and 16 GB RAM) with training scaling from 15 minutes to up to 3 hours based on the size of the training set and the number of features in the ML models. We set up a GitHub Action to automate the launch of the ML model training workflow on cloud clusters via API calls to the Parallel Works platform. This workflow automatically pushed data back to a public repository (see the Open Research section) so that all ML models and their corresponding code, training data, predictions, and logs are stored on separate branches for reproducibility and traceability of each ML model.

Table 2: Submodels and corresponding hyperparameter (HP) search spaces for each SuperLearner instance. All submodels are from scikit-learn except for XGBoost. HP search spaces are lists of specific values to be used unless a search method is indicated with a comma after the search range. Search methods are uniform ("uni") and log-uniform ("log-uni").

Submodel and source	Hyperparameter search space
Nu support vector regression with RBF kernel	$C = [10^{-6}, 10^{2.5}]$, log-uni $\nu = [10^{-10}, 0.99]$, uni $\gamma = [10-6, 0.99]$, log-uni
Nu support vector regression with linear kernel	$C = [10^{-6}, 10^{2.5}]$, log-uni $\nu = [10^{-10}, 0.99]$, uni
Nu support vector regression with polynomial kernel	$C = [10^{-6}, 10^{2.5}]$, log-uni $\nu = [10^{-10}, 0.99]$, uni Polynomial degree $= [1, 2, 3]$
Nu support vector regression with sigmoid kernel	$C = [10^{-6}, 10^{2.5}]$, log-uni $\nu = [10^{-10}, 0.99]$, uni coef0 $= [-0.99, 0.99]$, uni
k-nearest neighbors regression with uniform weights	N neighbors $= [1, 10]$, uni
k-nearest neighbors regression distance weighted	N neighbors $= [1, 10]$, uni
Partial least squares regression	N components $= [1, 10]$, uni
Multi-layer perceptron regression	Hidden layer sizes $= [10, 250]$ Solver $= [\text{lbfgs}, \text{sgd}, \text{adam}]$ $\alpha = [10^{-6}, 0.99]$, log-uni Tolerance $= [10^{-6}, 10^{-2}]$, log-uni

Continuation of Table 2	
Submodel and source	**Hyperparameter search space**
Ridge linear regression	Polynomial degree $= [1, 2, 3]$
	$\alpha = [10^{-6}, 0.99]$, log-uni
Lasso linear regression	Polynomial degree $= [1, 2, 3]$
	$\alpha = [10^{-6}, 0.99]$, log-uni
Linear regression	Polynomial degree $= [1, 2, 3]$
Elastic net linear regression	Polynomial degree $= [1, 2, 3]$
	$\alpha = [10^{-6}, 0.99]$, log-uni
	$l1$ ratio $= [10^{-6}, 0.99]$, log-uni
Huber linear regression	Polynomial degree $= [1, 2, 3]$
	$\alpha = [10^{-6}, 0.99]$, log-uni
	$\epsilon = [1.35, 1.9]$, uni
XGBoost regression	N estimators $= [100, 10000]$
	Learn rate $= [10^{-4}, 0.99]$, log-uni
	Maximum depth $= [2, 3, 4, 5, 6, 7, 8]$
Extra trees regression	N estimators $= [100, 10000]$
	$\alpha = [0, 0.001, 0.01, 0.1]$
	Max feature $= [0.1, 0.3, 0.5, 0.8, 1.0]$
	Max depth $= [2, 3, 4, 5, 6, 7, 8]$
	Min samples split $= [0.1, 0.2, 0.3]$
	Min samples leaf $= [0.1, 0.2, 0.3]$
	$C = [\text{err.}^2, |\text{err.}|, \text{FMSE}, \text{Poisson}]$

End of Table 2

2.3.2 Feature permutation importance

We use feature permutation importance (FPI) (Ojala & Garriga, 2010) to determine which features have the greatest impact on the predictive power of an ML model. FPI is a brute-force method that checks the sensitivity of ML model predictions by selectively mixing up specific subsets of features. To quantify the importance of a feature, the values of just that feature are randomly mixed up, leaving the other features in the training and testing data sets the same. In all cases, we use the combined training and testing data for FPI analysis because there are a relatively small number of data points. Then, the relative change in the skill of the ML model's predictions using data with one feature scrambled are represented as the model score R^2 without scrambling divided by the model score with scrambling; values greater than 1 indicate the ML model's score was reduced by scrambling the feature (i.e., the feature is important). The greater the ratio, the more important the feature. FPI is frequently used in ML analyses because it can be used with nearly any ML model, it is relatively computationally efficient, and random permutations allow for multiple FPI runs to assess the mean and variance of the importance results.

In practice, subsets of features, instead of individual features, are permuted together. FPI works best if highly correlated features are permuted together as a block. Otherwise, if only one feature is permuted at a time, information encoded in a single permuted feature could still be available to the ML model via a different, unpermuted feature that is correlated with the permuted feature. Here, we use correlation cutoffs of $R = 0.7$ and 0.9 to identify features that may be sufficiently correlated to impact FPI analysis. Our approach to grouping features is agglomerative hierarchical clustering where the highest correlated features, with correlations above the cutoff, are first grouped together. The next highest correlated features are grouped and so-on as the algorithm iterates over all correlations between all features. Groups whose features are correlated with features in another group are merged iteratively until we reach the correlation cutoff at which point
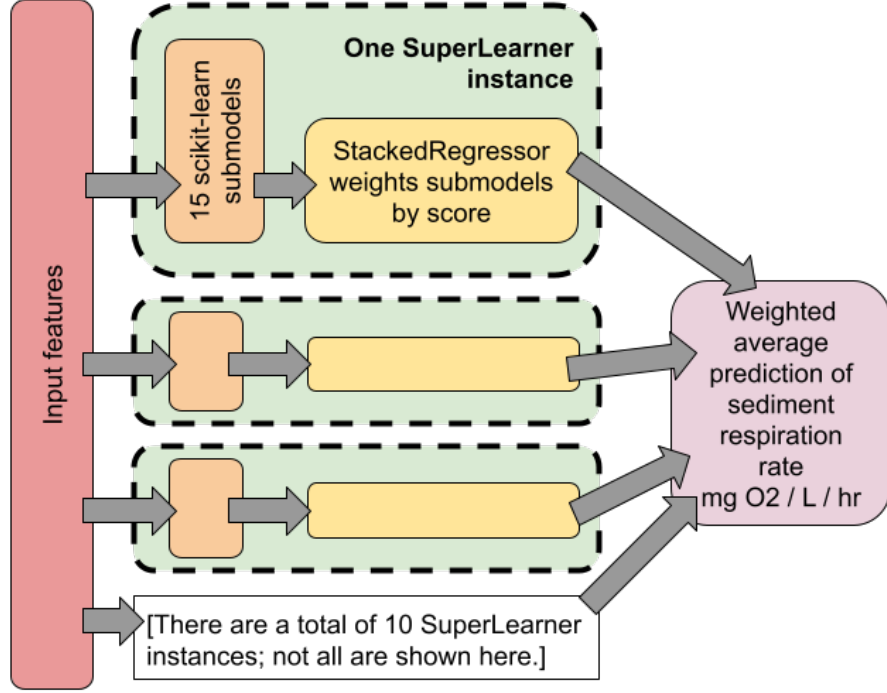
Figure 3. Schematic of the architecture of the ensemble of ensembles machine learning framework. 15 submodels are weighted in each SuperLearner instance based on their relative performance as described in the Methods section. The predictions of 10 SuperLearner instances are averaged together to account for random variability due to train/test data splits and algorithm initialization in some of the submodels in each SuperLearner instance. Hyperparameter optimization is performed as part of the training within each SuperLearner instance; hyperparameter values can be different across SuperLearner instances.

no additional groups are defined or merged. Any features that are not in any groups are perturbed independently of all the other remaining individual features and grouped features.

2.3.3 Machine learning model scenarios

Based on the many types and scales of data that are available for predicting river sediment respiration rates, there are a plethora of possible ML models that could be used in this project. A key goal in this work is to explore which features are most important for predicting river sediment respiration rates. Here, we take an approach of successively reducing the number of features of ML models. This progression is outlined in Table 3. The rationale for this progression is as follows for each ML training run ID:

1. **IDs 1 and 2 in Table 3:** ML models were trained on a data set with the largest possible number of features (from WHONDRS S19S and RiverAtlas co-located data). These runs were our main starting point.
2. **ID 3 in Table 3:** ML models were retrained using only the top 20 features identified by the FPI analysis from IDs 1 and 2. This step is designed to evaluate the change in ML models' predictive power associated with reduction in the number of training features.

3. **ID 4 in Table 3:** ML models were trained with an expanded data set including WHONDRS S19S and SSS observations and co-located RiverAtlas data. Because not all features were available in both S19S and SSS data sets, this step resulted in a decreased number of features but an increased number of data points for training. Comparison to IDs 1 and 2 provides information on changes in model performance associated with the reduced feature set but increased training data.

4. **IDs 5-7 in Table 3:** In order to predict respiration rates at unsampled locations, we retrained the ML models on an even further reduced set of features from the merged RiverAtlas (large-scale data) and WHONDRS *in situ* chemistry (temperature, pH, O_2, and percent saturated O_2) and river sediment respiration rates. Furthermore, some ML models (ID 6) also drop the *in situ* chemistry data from WHONDRS and only use RiverAtlas large-scale data and WHONDRS respiration rates to train the ML models. The ML models from this final stage could then be used to predict respiration at many unsampled sites across the Columbia River Basin as a concrete demonstration of data-driven extrapolation of respiration rate estimates at nearly any river segment for use in large-scale watershed modeling studies.

3 Results and Discussion

3.1 Model Predictive Skill

ML models are often evaluated based on the extent to which the predictions they make using the testing data set match the original target values. Figure 4 summarizes this model skill test for the S19S-SSS-log10-extrap-r01 (ID 5) and S19S-SSS-nolog-extrap-r01 (ID 7) ML model runs. A highly skilled ML model would show a very tight correspondence between the observed respiration rate (i.e. the target, horizontal axis) and the respiration rate predicted by the ML model using the corresponding features at that site. The ML models in this work exhibit moderate skill since there is a significant amount of scatter in the points. Furthermore, the scatter of the two ML model ensembles seem comparable by eye - neither model seems to be better than the other.

Model skill is quantified by the R^2 correlation coefficient "score" computed from the scatter of points in Figure 4.Consistent with the observations above, the model scores are not significantly different from each other in Figure 5. Both Figures 4 and 5 show that the river sediment respiration rates span several orders of magnitude and they have bimodal distributions with many observations of relatively weak respiration rates on the left side of each plot and a smaller number of relatively strong respiration rates on the right side of each plot.There is a breakpoint between the weak and strong respiration rates at about -500 mg DO/L/hr. The ML models generally recover this bimodal distribution when evaluated with the testing and training data (Fig. 5 bottom panel). However, all the ML models are significantly biased with the predicted, weak respiration rates being significantly stronger (i.e. shifted right) relative to the observed respiration rates. This bias is an example of the classic tendency for ML models to make predictions near the mean value of the training data.

The S19S-SSS-log10-extrap-r01 (ID 5) and S19S-SSS-nolog-extrap-r01 (ID 7) ML training runs were presented first in Figure 4 because their contrast highlights an ML model design decision that applies to the rest of the ML models discussed in this work. The only difference between these two ML model ensembles is the presence or absence of a log10 transform on the river sediment respiration rates as described in the Methods section. A log10 transformation was considered because the respiration rates span many orders of magnitude and such a transformation could help resolve the biases in ML model predictions noted above. Figures 4 and 5 (top panel) show that run ID 7 has only a slightly higher model score than run ID 5. On the other hand, the histogram of respiration rate predictions for run ID 5 and all the other run IDs with the log10 transformer

Table 3. Summary of the ML training runs discussed in this manuscript. The number of data points available for the train/test data split is listed in the N pts. column and the number of features used to train each model is in the N feat. column. In all cases, 25% of the data are used for the testing set for evaluating the ML models while 75% of the data are used for training the ML models. The exception to this general rule is that *all* data points are used for cross-validated hyperparameter optimization as described in the Methods. The ID for each model is used to locate a particular model's results in Figures 5 and 6.

ID	ML run branch name	N pts.	N feat.	Description
1	Summer-2019-log10-r07	227	133	S19S WHONDRS observations with RiverAtlas co-located data, all possible features, FPI corr. cutoff 0.7.
2	Summer-2019-log10-r08	227	133	Same as ID 1, above, except FPI corr. cutoff set to 0.9.
3	Summer-2019-log10-r09	227	29	Same as ID 1, above, but trained with only the top 20 features identified by FPI analysis in ID 1 and 2 (the top 20 features in each run were merged).
4	S19S-SSS-log10-r02	367	94	Merged S19S and SSS WHONDRS observations and using all features present in both data sets, FPI corr. cutoff 0.7.
5	S19S-SSS-log10-extrap-r01	367	78	Merged S19S and SSS WHONDRS observations and using only features available from RiverAtlas and WHONDRS temp., pH, O_2, and % sat. O_2 for use with making predictions at sites with GLORICH and RiverAtlas data, FPI corr. cutoff 0.7.
6	S19S-SSS-log10-extrap-r03	367	74	Same as ID 5, above, but using only features available from RiverAtlas for use with making predictions at any RiverAtlas river segment, FPI corr. cutoff 0.7.
7	S19S-SSS-nolog-extrap-r01	367	78	Same as ID 5, above, but without a log10 filter on the sediment respiration rate.

(i.e. 1-6) exhibit slightly stronger bimodality with a local minima in their histograms that are smaller than the local minimum of the histogram of run ID 7 (Fig. 5 bottom panel). Furthermore, if we consider -500 mg DO/L/hr as the boundary between weak and strong respiration rates,run ID 7 is slightly more likely (5.2%) than run ID 5 (4.4%) to make a classification error by predicting weak respiration rates for sites that are actually observed with strong respiration rates or vice-versa. While there is only a slight difference in the performance of ML models with or without the log10 transformation of the respiration rates, we ultimately chose to use ML models with the log10 transformer due to the slightly better representation of the bimodal distribution of respiration rates. This weak-strong bimodality is a key aspect of the observed river sediment respiration rates so the extent to which an ML model can recover bimodality is a more important criterion than the model score alone.

When comparing the other ML model runs, ML model ensemble skill is most improved by increasing the number of data points in the training data (e.g. compare run IDs 1-3 to 4-7 of Figure 5, top panel). This observation is consistent with the expectation of ML model skill improving with more data and the fact that our training data is relatively small at a few hundred data points while ML generally performs better with thousands or more data points. Adding training data points, however, does not appear to have as large an impact on improving the biases of the ML models; the peaks of the histograms of model predictions of respiration rates in Figure 5 (bottom panel) are all generally centered at the same two locations. The predicted weak respiration rates exhibit a consistent bias compared to the observations. However, the broad weak-strong respiration rate bimodality is retained across all models. Finally, changes to the number of features in each ML model does not have a significant impact on either the model skill or model biases (e.g. comparing run IDs 1-2 to 3 or run ID 4 to 5-6 in Figure 5).

3.2 Feature Importance

We have been building up to Figure 6 which can now help answer the central question posed in this work, namely: Which features are most important for predicting river sediment respiration rates? When interpreting Figure 6 it is important to remember that the progression of reducing features of the ML model ensembles outlined in Table 3 means that not all features are present in all cases. Comparing run IDs 1 and 2 shows that the choice of the FPI correlation cutoff has a minor impact on the results. Larger groups of features, which are visible in Figure 6 as blocks of adjacent features with exactly the same feature importance mean and standard deviation, tend to have greater importances. Lower FPI correlation cutoffs will result in larger groupings of features and vice-versa. This effect can be explained by the fact that a greater proportion of the total amount of information available to the ML models is contained in a group with more features than a group with fewer features. As the number of features is reduced in the ML model ensembles, the effect of feature grouping is intensified (e.g. the elevated feature importances of run IDs 3-7 relative to 1-2) because the proportion of information in a group of features relative to the total amount of information available to the model increases. Despite this effect, feature importances across all run IDs are broadly consistent with each other so the FPI correlation cutoff does not have a significant impact on the overall results as long as the features we want to compare relative to each other are resolved in different groups. Since feature grouping depends on the correlations between features for a specific input data set, no single correlation cutoff can be specified for all data. For the merged WHONDRS and RiverAtlas data, we found that an FPI correlation cutoff of about 0.7 suited the needs of this analysis.

The features that are consistently considered important across different ML model ensembles in Figure 6 can be organized into the broad categories of Table 1:

- ecological setting with feature IDs 15-18, 23-24, 33-36, 39-40, 65-75, 76-77, 112, 121, and 125-126;
- geological setting with feature IDs 13-14, 19-20, 25-30;
- fluvial setting with feature IDs 5, 10, 37-38, 41-45, 47, and 119-120;
- climatic setting with feature IDs 80-84, 90-99, 102-103, 110-111, and 128.
- riverbed sediment with feature IDs 11-12;
- water properties with feature IDs 21, 22, 87, and 107; and
- sediment organic matter chemistry with feature IDs 50-64 and 115.

This list of important features is long and hard to interpret, so we will leverage the progressive reduction of features in ML model IDs 1-7 in Table 3 as a context for highlighting particular categories of important features. For the cases where we can train ML models with all available features, ML model IDs 1 and 2, it is the FTICR data, especially feature IDs 50-58 (metrics of average thermodynamic efficiency, Gibbs free energy, oxidation state, and aromaticity of the chemicals detected in the sediment sample), that are consistently the most important features. This result is robust to changes in the FPI correlation cutoff, the only difference between ML model IDs 1 and 2 (aside from the randomized differences in train/test splitting during training). FTICR features are also available in ML model ID 3 and the FPI analysis suggests that these features are extremely important. However, the highly elevated FTICR feature importance in ML model ID 3 is also amplified by the fact that most of the FTICR features are grouped together as a single block and they account for the majority of the information used to train the ML models in this ensemble.

Unfortunately, the FPI analysis for ML model IDs 1-3 also exhibits a very large uncertainty. The FTICR features are only available in the S91S subset of the WHONDRS data so ML models 1-3 were trained with only 227 data points. As such, there is only one feature, 123 (the count of phototropic cells), in only one ML model ensemble, 1, whose feature importance is greater than $1 + s$ where s is the sample standard deviation of the feature importance scores over the ML model ensemble members for a particular feature. A feature importance score of 1 is a totally neutral feature (i.e. the ML models make similar predictions with or without the feature scrambled). The WHONDRS S19S subset used to train ML model IDs 1-3 is 62% of the merged S19S and SSS WHONDRS data that was used to train ML models 4-7. As noted above, model performance increases significantly with more training data (Fig. 5) but the uncertainties of the feature importances remain large, even for ML models 4-7. Despite the large uncertainties, there are substantial numbers of features for ML model IDs 4-7 whose feature importances are greater than $1+s$: 94, 44, 64, and 78 total features for ML model IDs 4, 5, 6, and 7, respectively.

The features with the greatest importances are generally consistent across ML models with IDs 4-7. Broadly, it appears that features that fall into the category of climatic setting may have the greatest importance followed by ecological, geological, and fluvial settings. However, although these blocks of features may have importances substantially different from a neutral importance of 1, their uncertainties remain sufficiently high so that the uncertainty ranges of feature importance often overlap and it is not possible to definitively say which categories of features are more important than the others. Special consideration should be given to the features that fall into the category of readily available *in situ* chemistry observations, principally water temperature, pH, DO, and percent saturated DO with feature IDs 87, 107, 21, and 22, respectively. These features are available in the WHONDRS and GLORICH data and are relatively easy to collect in the field with well-established methods and sensors. We initially hypothesized that having local, *in situ* observations would provide an important environmental context for making predictions of river sediment respiration rate when paired with the large-scale data in RiverAtlas. In fact, including or excluding these features is the distinction between training ML model IDs 5 and 7 (including temperature, pH, DO, and percent saturated

DO) and ML model ID 6 (no *in situ* data at all). We can see that these features are significantly important (i.e. importances greater than $1 + s$) for ML run IDs 4, 5 and 7, but they only have modest importance (Fig. 6). This observation is consistent with the fact that ML model ID 6 has comparable performance to ML models with IDs 4, 5, and 7 (Fig. 5). ML model IDs 1 and 2 are our only mechanism for comparing the feature importances of temperature, pH, DO, and percent saturated DO compared to the additional chemistry features available in the WHONDRS data. In both cases of ML models, temperature, pH, DO, and percent saturated DO do not appear to be critically important. This observation is consistent with the lower importances of these features in the other ML models (e.g. IDs 4-7).

3.3 Extrapolation to Unsampled Locations

The final step in our analysis is to use the ML models trained with the merged WHON-DRS and RiverAtlas data to make predictions of river sediment respiration rate across the Columbia River Basin. We chose the Columbia River because this river basin has the greatest concentration of WHONDRS data; the 2022 SSS sampling campaign was focused there (Fig. 1). Since the RiverAtlas data set is global, we could in principle use the ML models to make predictions anywhere there are temperature, pH, DO, and percent saturated DO observations (ML model ID 5) or simply anywhere (ML model ID 6). Figure 7 shows river sediment respiration rates across the Columbia River Basin. Respiration rates exhibit patterns that are spatially consistent; most adjacent river segments have similar respiration rates with gradual changes along streams. Trends in these patterns deviate from the underlying topography and relative size of river segments in the overall network suggesting that there are many significant factors governing the predictions of river sediment respiration rates. This observation is consistent with the FPI analysis, above, whose results show that many features are important in the ML models. Finally, the magnitudes and spatial patterns of the predictions made at every river segment in RiverAtlas are broadly consistent with the predictions at the GLORICH sites.

Histograms of the ML models' predictions over the Columbia River Basin (Fig. 8), however, reveal an important discrepancy between the predictions made by ML models including (ML model ID 5) and excluding (ML model ID 6) readily available *in situ* chemistry observations. The ML models with ID 5 and 6 are able to both recreate the observed bimodal distribution in river sediment respiration rates (gray lines in Fig. 8) albeit with the bias in weak respiration rates discussed above. However, when these models are used to predict respiration rates across the Columbia River Basin (i.e. extrapolate), the results from ML model ID 5 exhibit a pronounced bias toward strong respiration rates while the results from ML model ID 6 are distributed more closely to the predictions based on the training data. We explored several avenues for uncovering the source of the bias of ML model ID 5. For example, since the WHONDRS data are all collected in the summers of 2019 and 2022, they are $\sim 2^oC$ warmer compared to the mean temperature across all GLORICH sites. The WHONDRS data also have a corresponding ~ 1.3 mg DO/L/hr lower DO bias due to the temperature dependence of DO in water. Shifting all the temperature, pH, DO, and percent saturated DO values in GLORICH so they have the same mean as the WHONDRS data did not change the bias of the predictions at the GLORICH sites (not shown), nor did setting temperature, pH, DO, and percent saturated DO to the constant, mean values of the WHONDRS data (Fig. 8).

The high bias of ML model ID 5 is perplexing given the relatively low importances associated with temperature, pH, DO, and percent saturated DO and the fact that ML model ID 6 is identical except that its training set excluded temperature, pH, DO, and percent saturated DO. The models are likely overfit (i.e. they have trouble generalizing to new, unseen data) due to the small number of points in the training data set. Supporting this conjecture is the observation that for both ML model ID 5 and 6 predictions across the Columbia River Basin (red and pink lines) do not exhibit the same strong bi-

modality as their predictions made with the training data (gray lines, Fig. 8). There-
fore, when using the new, unseen, and variable data from 6,171 GLORICH sites and 86,054
RiverAtlas segments, these ML models have difficulty making clear, decisive predictions
of which sites have weak or strong respiration rates. For ML model ID 5, in particular,
it is possible that weak learned correlations between temperature, pH, DO and percent
saturatd DO and the other features are amplified into bias when using data outside the
training set. The predictions of river sediment sediment respiration rates in Figure 7 by
ML model ID 6 are likely "blurred" or "smoothed out" with more overlap between the
segments with weak and strong respiration rates. Despite this overlap, there is still an
asymmetric distribution in ML model ID 6's predictions of respiration rates across the
Columbia River Basin that are broadly consistent with observations.

Figure 4.   Observed respiration rates (i.e. original respiration rates provided in the WHON-
DRS data which are used as "targets" in training) versus the respiration rates predicted by an
ensemble of machine learning models using most of the other WHONDRS and RiverAtlas data as
inputs. Black (red) symbols are for the S19S-SSS-log10-extrap-r01 (S19S-SSS-nolog-extrap-r01)
ML-runs defined in Table 3. The same data points are plotted in regular units (left) and log10
space (right). Model scores (i.e. the correlation between targets and predicted values in regular
units) for these ML-runs and all other runs used here are in Figure 5. Pale plus signs correspond
to predictions made by each individual SuperLearner instance while the solid circles are the aver-
age prediction across all 10 of the SuperLearner instances that make up the ensemble of models
for each ML workflow run.

4 Conclusions

This study demonstrates the application of machine learning (ML) methods to de-
velop predictive models of river sediment respiration based on community-generated dis-
tributed data. The data-driven models provide new insights into the spatial variability
of respiration and the relative importance of various input variables (system features)
to those predictions.

We analyzed a data set, compiled from community-generated data combined with
public river databases, containing as many as 367 samples and 133 features. Not all sam-
ples contained all features, so a progressive ML training process was followed to test the
impact of numbers of samples and features on ML model performance. The various runs
also tested impacts of model decisions such as FPI correlation cutoffs and data trans-
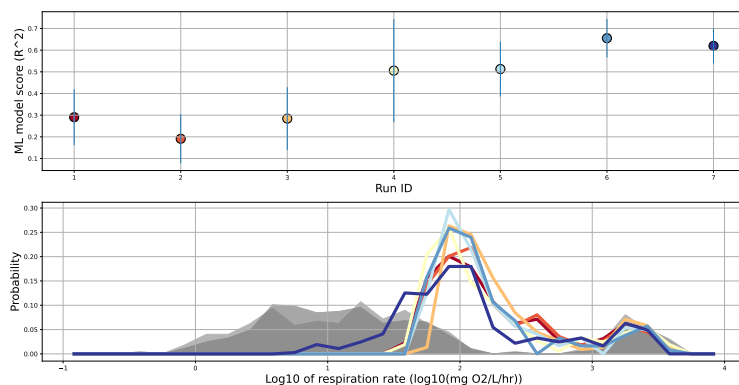formations.

Figure 5. Model scores computed with the testing data (i.e. hold-out data) only (top) and histograms of predictions made with both the training and testing data (bottom) for the ML-runs as numbered in Table 3. The color coding with run ID is the same in both panels of this figure as well as in Figure 6.
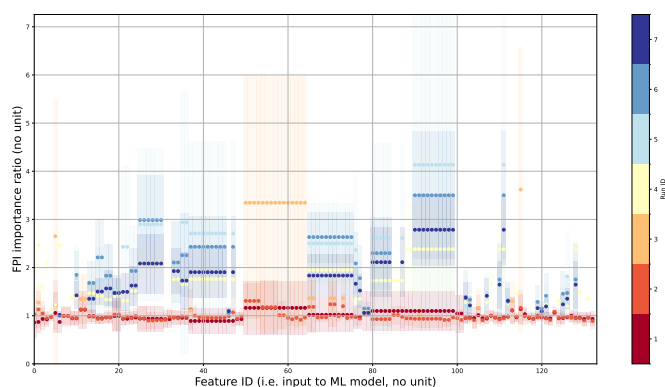

Figure 6. Summary of highly ranked features in the merged WHONDRS and RiverAtlas data for predicting respiration rates. Dots represent the FPI improvement ratio for each feature listed in Tables 1 and S1 averaged over all the SuperLearner ensemble members for a particular ML model ID listed in Table 3. Higher FPI ratios imply greater feature importance. The translucent rectangles represent the one standard deviation of the FPI improvement ratio among the Super-Learner instances. Feature ID's can be looked up in Table 1 and SI Table S1.

4.1 Most Important System Features

We had hypothesized that organic matter chemistry, as reflected in summary measures of FTICR-MS spectra on extractable sediment organic matter, could be a major determinant of sediment respiration. The results of the feature permutation importance (FPI) analysis support this hypothesis, and indicate that the FTICR data may be the most important features in predicting respiration rates. However, because of the relatively small number of samples with FTICR data, significant uncertainty exists regarding this conclusion. We are currently analyzing additional samples using FTICR and expect that future analyses will show reduced uncertainty.
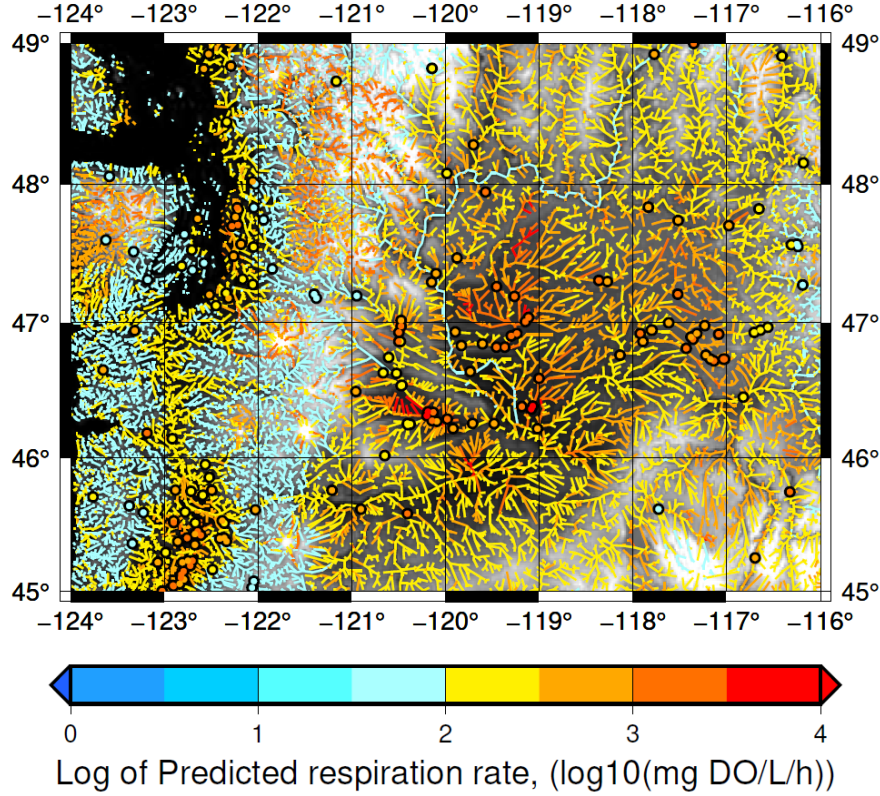
Figure 7. Predicted river sediment respiration rates across the northwest corner of the continental US (inset of Fig. 1). The network of all RiverAtlas segments in this map domain is color coded by predictions of river sediment respiration rate from ML model ID 6 at each segment. Predictions of respiration rate at all sites where there are GLORICH *in situ* temperature, pH, and DO or percent saturated DO are made by ML model ID 5 and are color coded dots using the same colorbar as the river segments. Predictions at GLORICH sites all have a larger solid black circle behind them so that the color coded dots do not blend into the river network behind. Relative topography is shaded in the background on a grayscale (sea level is black, darker lowlands, lighter highlands).

For a set of different ML models trained on data that did not contain FTICR features, the results suggest that large-scale features of climate zone, population, and physical geography are important for prediction river sediment respiration rates. These features can be extracted from widely-available public databases. Supplementing these data with site-specific (*in situ*) measurements of river temperature, pH, DO, and percent DO saturation showed some benefit to ML model performance, but did not appear to be critically important.

4.2 Transferability and Watershed Model Parameterization

By selecting those features that are deemed both important to the prediction and available in data sets with high spatial coverage, we can extrapolate the understanding gained from a relatively small set of sampling locations to large set of locations across broad domains of interest. We demonstrated this potential capability by predicting respiration rates in river reaches spanning the Columbia River Basin.
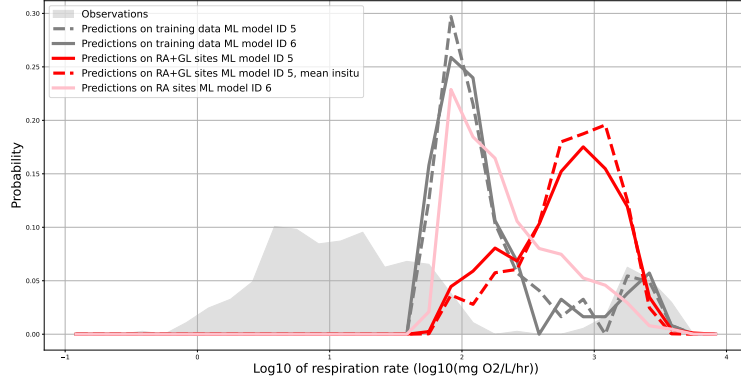
Figure 8. Histograms of observed and predicted river sediment respiration rates extrapolated at a large-scale. Observed respiration rates from the WHONDRS sites (filled gray) and the predicted respiration rates at the WHONDRS sites by ML model IDs 5 and 6 (solid and dashed gray lines) have identical counterparts in the bottom panel of Figure 5. The distribution of respiration rate predictions by ML model ID 5 at sites where there are GLORICH *in situ* temperature, pH, DO, and percent saturated DO (6,171 sites) are in red while predictions by ML model ID 6 at each RiverAtlas river segment in the northwest corner of the US (86,054 segments) are in pink. The dashed red line are predictions by ML model ID 5 similar to the red line except that all temperature, pH, DO, and percent saturated DO are set to their respective constant mean values from the WHONDRS data while the other larger-scale features are allowed to vary as in the predictions with the solid red line.

This information can potentially be used in the parameterization of process-based watershed models to assess large-scale (cumulative) watershed functions such as total carbon dioxide efflux from rivers across watersheds or basins.

4.3 Future Research

Although the ML models developed and presented here demonstrated significant predictive ability, a high level of uncertainty remains both in the predictions of respiration rates and the assessments of which features are most important. Ongoing research is expanding on this study to both 1) incorporate additional FTICR-MS data into the existing models as these analyses are completed for existing samples, and 2) collect additional samples to increase the sample size and spatial coverage, and ostensibly to improve ML model predictive ability.

5 Open Research

WHONDRS data used here were published previously by Goldman et al. (2020) and Forbes et al. (2023) in the ESS-DIVE repository (https://ess-dive.lbl.gov/) and are licensed for reuse under the Creative Commons Attribution 4.0 International License. The data associated specifically with this manuscript are also available on the ESS-DIVE repository (Gary et al., 2024). All the ML models (code, model files, preprocessed training data, predictions, and Conda environment definition files) are available in the public GitHub repository https://github.com/parallelworks/sl-archive-whondrs which is also snapshotted in the ESS-DIVE data package (Gary et al., 2024) together with the scripts

that generated the figures in this manuscript. The original RiverAtlas (Linke et al., 2019) and GLORICH (Hartmann et al., 2019) data sets are publicly available and their pre-processed versions (along with all preprocessing scripts) for use with the ML workflow used in this work are available in the public GitHub repository https://github.com/parallelworks/global-river-databases.

Acknowledgments

References

Ahamed, F., You, Y., Burgin, A., Stegen, J. C., Scheibe, T. D., & Song, H.-S. (2023). Exploring the determinants of organic matter bioavailability through substrate-explicit thermodynamic modeling. *Frontiers in Water*, *5*. doi: 10.3389/frwa.2023.1169701

Battin, T. J., Kaplan, L. A., Newbold, J. D., & Hendricks, S. P. (2003). A mixing model analysis of stream solute dynamics and the contribution of a hyporheic zone to ecosystem function [Journal Article]. *Freshwater Biology*, *48*(6), 995-1014. Retrieved from `<GotoISI>://WOS:000182948500005https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-2427.2003.01062.x` doi: DOI10.1046/j.1365-2427.2003.01062.x

Borton, M., Collins, S., Graham, E., Garayburu-Caruso, V. A., Goldman, A., de Melo, M., ... Consortium, W. C. (2022). It takes a village: Using a crowdsourced approach to investigate organic matter composition in global rivers through the lens of ecological theory. *Frontiers in Water*, *4*. doi: 10.3389/frwa.2022.870453

Bramer, L. M., White, A. M., Stratton, K. G., Thompson, A. M., Claborne, D., Hofmockel, K., & McCue, L. A. (2020). ftmsranalsysi: An r package for exploratory data analysis and interactive visualization of ft-ms data. *PLoS Computational Biology*, *16*(3), e1007654. doi: 10.1371/journal.pcbi.1007654

Buser-Young, J., Garcia, P., Schrenk, M., Regier, P., Ward, N., Biçe, K., ... Lønborg, C. (2023). Determining the biogeochemical transformations of organic matter composition in rivers using molecular signatures. *Frontiers in Water*, *5*. doi: 10.3389/frwa.2023.1005792

Butman, D., Stackpoole, S., Stets, E., McDonald, C. P., Clow, D. W., & Striegl, R. G. (2016). Aquatic carbon cycling in the conterminous united states and implications for terrestrial carbon accounting. *PNAS*, *113*(5), 58-63. doi: 10.1073/pnas.1512651112

Dittmar, T., Boris, K., Hertkorn, N., & Kattner, G. (2008). A simple and efficient method for the solid-phase extraction of dissolved organic matter (spe-dom) from seawater. *Limnology and Oceanography Methods*, *6*(6), 230-235. doi: 10.4319/lom.2008.6.230

EPA, & USGS. (2022). *National hydrography dataset plus, 2022 snapshot* [data set]. US EPA. Retrieved from `https://www.epa.gov/waterdata/get-nhdplus-national-hydrography-dataset-plus-data`

Findlay, S. (1995). Importance of surface-subsurface exchange in stream ecosystems: The hyporheic zone. *Limnology and Oceanography*, *40*(1), 159-164. doi: 10

.4319/lo.1995.40.1.0159

Fischer, H., Kloep, F., Wilzcek, S., & Pusch, M. T. (2005). A river's liver - microbial processes within the hyporheic zone of a large lowland river [Journal Article]. *Biogeochemistry*, *76*(2), 349-371. Retrieved from `<GotoISI>://WOS: 000233242400008`

Forbes, B., Barnes, M., Boehnke, B. T., Chen, X., Cornwell, K., Delgado, D., ... Consortium, T. W. (2023). Whondrs river corridor dissolved oxygen, temperature, sediment aerobic respiration, grain size, and water chemistry from machine-learning-informed sites across the contiguous united states [dataset]. *ESS-DIVE repository*. doi: 10.15485/1923689

Fuss, C. L., & Smock, L. A. (1996). Spatial and temporal variation of microbial respiration rates in a blackwater stream [Journal Article]. *Freshwater Biology*, *36*(2), 339-349. Retrieved from `<GotoISI>://WOS:A1996VM61200010https:// onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2427.1996.00095.x` doi: DOI10.1046/j.1365-2427.1996.00095.x

Garayburu-Caruso, V. A., Danczak, R. E., Stegen, J. C., Renteria, L., Mccall, M., Goldman, A. E., ... Graham, E. B. (2020). Using community science to reveal the global chemogeography of river metabolomes. *Metabolites*, *10*(12). doi: 10.3390/metabo10120518

Garayburu-Caruso, V. A., Stegen, J. C., Song, H. S., Renteria, L., Wells, J., Garcia, W., ... Graham, E. B. (2020). Carbon limitation leads to thermodynamic regulation of aerobic metabolism [Journal Article]. *Environmental Science & Technology Letters*, *7*(7), 517-524. doi: 10.1021/acs.estlett.0c00258

Gary, S., Scheibe, T. D., Rexer, E., Wilde, M., Vidal Torreira, A., Garayburu-Caruso, V. A., ... Stegen, J. C. (2024). Models, data, and scripts associated with "prediction of distributed river sediment respiration rates using community-generated data and machine learning.". *ESS-DIVE repository*. doi: 10.15485/2318723

Goldman, A. E., Arnon, S., Bar-Zeev, E., Chu, R. K., Danczak, R. E., Daly, R. A., ... Consortium, T. W. (2020). Whondrs summer 2019 sampling campaign: Global river corridor sediment fticr-ms, dissolved organic carbon, aerobic respiration, elemental composition, grain size, total nitrogen and organic carbon content, bacterial abundance, and stable isotopes. *ESS-DIVE repository*. doi: 10.15485/1729719

Graham, E. B., Crump, A. R., Kennedy, D. W., Arntzen, E., Fansler, S., Purvine, S. O., ... Stege, J. C. (2018). Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone [Journal Article]. *Science of the Total Environment*, *642*, 742-753. Retrieved from `<GotoISI>://WOS:000439405600074`

Hartmann, J., Lauerwald, R., & Moosdorf, N. (2014). A brief overview of the global river chemistry database, glorich. *Procedia Earth and Planetary Science*, *10*, 23-27. Retrieved from `https://www.sciencedirect.com/science/ article/pii/S1878522014000678` (Geochemistry of the Earth's surface GES-10 Paris France, 18-23 August, 2014.) doi: https://doi.org/10.1016/ j.proeps.2014.08.005

Hartmann, J., Lauerwald, R., & Moosdorf, N. (2019). *GLORICH - Global river chemistry database* [data set]. PANGAEA. Retrieved from `https://doi.org/10.1594/PANGAEA.902360` (Supplement to: Hartmann, J et al. (2014): A Brief Overview of the GLObal RIver Chemistry Database, GLORICH. Procedia Earth and Planetary Science, 10, 23-27, https://doi.org/10.1016/j.proeps.2014.08.005) doi: 10.1594/PANGAEA .902360

Harvey, J., Gomez-Velez, J., Schmadel, N., Scott, D., Boyer, E., Alexander, R., ... Choi, J. (2019). How hydrologic connectivity regulates water quality in river

corridors. *JAWRA Journal of the American Water Resources Association*, *55*(2), 369-381. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/1752-1688.12691` doi: 10.1111/1752-1688.12691

Harvey, J., & Gooseff, M. (2015, sep). River corridor science: Hydrologic exchange and ecological consequences from bedforms to basins. *Water Resources Research*, *51*(9), 6893–6922. Retrieved from `http://doi.wiley.com/10.1002/2015WR017617` doi: 10.1002/2015WR017617

Horton, R. E. (1945). Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Geological society of America bulletin*, *56*(3), 275–370.

Jensen, B. (n.d.). *Aos protocol and procedure: Sediment chemistry sampling in wadeable streams.* `http://data.neonscience.org/documents`. Retrieved from `http://data.neonscience.org/documents` (Accessed: 12 June 2019)

Jones, J. B. (1995). Factors controlling hyporheic respiration in a desert stream [Journal Article]. *Freshwater Biology*, *34*(1), 91-99. Retrieved from `<GotoISI>://WOS:A1995RP65400010https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2427.1995.tb00426.x` doi: DOI10.1111/j.1365-2427.1995.tb00426.x

Kaplan, L., & Newbold, J. (2000). Surface and subsurface dissolved organic carbon [Book Section]. In J. Jones & P. Mulholland (Eds.), *Stream and ground waters* (p. 237-258). San-Diego, CA: Academic Press.

Kim, S., Kramer, R. W., & Hatcher, P. G. (2003). Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van krevelen diagram. *Analytical chemistry*, *75*(20), 5336–5344.

Koch, B. P., & Dittmar, T. (2006). From mass to structure: An aromaticity index for high-resolution mass data of natural organic matter. *Rapid communications in mass spectrometry*, *20*(5), 926–932.

Koch, B. P., & Dittmar, T. (2016). From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry*, *30*(1), 250-250. Retrieved from `https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.7433` doi: https://doi.org/10.1002/rcm.7433

LaRowe, D. E., & Van Cappellen, P. (2011). Degradation of natural organic matter: a thermodynamic analysis. *Geochimica et Cosmochimica Acta*, *75*(8), 2030–2042.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., … others (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific data*, *6*(1), 283. doi: https://doi.org/10.1038/s41597-019-0300-6

Naegeli, M. W., & Uehlinger, U. (1997). Contribution of the hyporheic zone to ecosystem metabolism in a prealpine gravel-bed river [Journal Article]. *Journal of the North American Benthological Society*, *16*(4), 794-804. Retrieved from `<GotoISI>://WOS:000071795200006`

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*(6), 1833-1863.

Owoyele, O., Pal, P., Torreira, A. V., Probst, D., Shaxted, M., Wilde, M., & Senecal, P. K. (2022). Application of an automated machine learning-genetic algorithm (automl-ga) coupled with computational fluid dynamics simulations for rapid engine design optimization. *International Journal of Engine Research*, *23*(9), 1586-1601. Retrieved from `https://doi.org/10.1177/14680874211023466` doi: 10.1177/14680874211023466

Owoyele, O., Pal, P., & Vidal Torreira, A. (2021). An automated machine learning-genetic algorithm framework with active learning for design optimization. *Journal of Energy Resources Technology*, *143*(8), 082305.

Rajesh, M., & Rehana, S. (2022). Impact of climate change on river water tempera-

ture and dissolved oxygen: Indian riverine thermal regimes. *Scientific Reports*, *12*(1), 9222.

Rice, E. W., Baird, R. B., & Eaton, A. D. (2017). *Standard methods for the examination of water and wastewater 23rd ed.* Washington, DC, USA: American Public Health Association (APHA).

Rode, M., Hartwig, M., Wagenschein, D., Kebede, T., & Borchardt, D. (2015). The importance of hyporheic zone processes on ecological functioning and solute transport of streams and rivers. In L. Chicharo, F. Müller, & N. Fohrer (Eds.), *Ecosystem services and river basin ecohydrology.* Springer. doi: 10.1007/978-94-017-9846-4_4

Song, H.-S., Stegen, J. C., Graham, E. B., Lee, J.-Y., Garayburu-Caruso, V. A., Nelson, W. C., . . . Scheibe, T. D. (2020). Representing organic matter thermodynamics in biogeochemical reactions via substrate-explicit modeling. *Frontiers in Microbiology*, *11*, 531756.

Stadler, M., Barnard, M., Bice, K., de Melo, M., Dwivedi, D., Freeman, E., . . . Meile, C. (2023). Applying the core-satellite species concept: Characteristics of rare and common riverine dissolved organic matter. *Frontiers in Water*, *5*. doi: 10.3389/frwa.2023.1156042

Stegen, J. C., Fansler, S. J., Tfaily, M. M., Garayburu-Caruso, V. A., Goldman, A. E., Danczak, R. E., . . . Toyoda, J. (2022). Organic matter transformations are disconnected between surface water and the hyporheic zone. *Biogeosciences*, *19*, 3099-3110. doi: 10.5194/bg-19-3099-2022

Stegen, J. C., Garayburu-Caruso, V. A., Danczak, R. E., Goldman, A. E., Renteria, L., Torgeson, J. M., & Hager, J. (2023). Maximum respiration rates in hyporheic zone sediments are primarily constrained by organic carbon concentration and secondarily by organic matter chemistry. *Biogeosciences*, *20*(14), 2857-2867. doi: 10.5194/bg-20-2857-2023

Stegen, J. C., & Goldman, A. E. (2018). Whondrs: a community resource for studying dynamic river corridors [Journal Article]. *Msystems*, *3*(5). Retrieved from `<GotoISI>://WOS:000449523700017` doi: ARTNe00151-1810.1128/mSystems .00151-18

Stegen, J. C., Johnson, T., Fredrickson, J. K., Wilkins, M. J., Konopka, A. E., Nelson, W. C., . . . Zachara, J. (2018, dec). Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. *Nature Communications*, *9*(1), 585. Retrieved from `http://www.nature.com/articles/ s41467-018-02922-9` doi: 10.1038/s41467-018-02922-9

Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, *38*(6), 913–920.

Tolic, N., Liu, Y., Liyu, A., Shen, Y., Tfaily, M. M., Kujawinski, E. B., . . . Hess, N. (2017). Formularity: Software for automated formula assignment of natural and other organic matter from ultrahigh-resolution mass spectra. *Analytical Chemistry*, *89*(23), 12659-12665. doi: 10.1021/acs.analchem.7b03318

Ward, N. D., Sawakuchi, H. O., Neu, V., Less, D. F. S., Valerio, A. M., Cunha, A. C., . . . Keil, R. G. (2018). Velocity-amplified microbial respiration rates in the lower amazon river [Journal Article]. *Limnology and Oceanography Letters*, *3*(3), 265-274. doi: 10.1002/lol2.10062

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.