Title

Models, data, and scripts associated with “Prediction of Distributed River Sediment Respiration Rates using Community-Generated Data and Machine Learning”

Summary

This data package is associated with the publication “Prediction of Distributed River Sediment Respiration Rates using Community-Generated Data and Machine Learning’’ submitted to the Journal of Geophysical Research: Machine Learning and Computation (Scheibe et al. 2024). River sediment respiration observations are expensive and labor intensive to obtain and there is no physical model for predicting this quantity. The Worldwide Hydrobiogeochemisty Observation Network for Dynamic River Systems (WHONDRS) observational data set (Goldman et al.; 2020) is used to train machine learning (ML) models to predict respiration rates at unsampled sites. This repository archives training data, ML models, predictions, and model evaluation results for the purposes of reproducibility of the results in the associated manuscript and community reuse of the ML models trained in this project. One of the key challenges in this work was to find an optimum configuration for machine learning models to work with this feature-rich (i.e. 100+ possible input variables) data set. Here, we used a two-tiered approach to managing the analysis of this complex data set: 1) a stacked ensemble of ML models that can automatically optimize hyperparameters to accelerate the process of model selection and tuning and 2) feature permutation importance to iteratively select the most important features (i.e. inputs) to the ML models. The major elements of this ML workflow are modular, portable, open, and cloud-based, thus making this implementation a potential template for other applications. This data package is associated with the GitHub repository found at https://github.com/parallelworks/sl-archive-whondrs. A static copy of the GitHub repository is included in this data package as an archived version at the time of publishing this data package (March 2023). However, we recommend accessing these files via GitHub for full functionality.

Brief Overview of Methods

Machine learning models were trained on the WHONDRS observations of river sediment respiration rates corresponding site metadata (Goldman et al., 2020), and Fourier-transform ion cyclotron resonance mass spectrometry (FTICR-MS) analysis (Garayburu-Caruso, 2022). Additional data from large-scale, global databases RiverAtlas (Linke et al., 2019) and GLORICH (Hartmann et al., 2014) were used to supplement the local, site-based metadata of the WHONDRS data. The most important input data to the ML models were quantified with a feature permutation importance analysis (FPI). All ML models were trained and assessed via a workflow that was automatically launched through a GitHub Action triggered by each new release published in the GitHub repository. Each ML model/release is tagged and on a unique git branch. ML models, the training data, and all training and model assessment results were pushed back to the GitHub repository to their respective unique branches. Extensive documentation on the methods is available in the manuscript associated with this data package.

Critical Details

1 – The scripts provided in this repository are intended to run in a Linux Bash shell, Python in Jupyter notebooks, or in a Conda environment which has installed Python packages as described in the .yaml and .txt files in the ./examples directory of the repository. Some plotting scripts use the Generic

Mapping Tools (GMT). For portability, GMT is delivered as a Docker container; to run the scripts as they are in this repository, you will need to install Docker as well.
2 – The main branch of this repository contains the necessary starting points for training machine learning models, but it does not contain any of the trained ML models and their results. Instead, ML models, along with all their supporting data, are stored in the Summer-2019-* and S19S-SSS-* branches of this repository. The main branch also contains the most up-to-date version of the plotting scripts and notebooks used for preparing the manuscript associated with this repository. As such, the main branch has the most recent plotting routines while other branches contain the results of the machine learning workflow. No single branch has all the files listed in the file level metadata (FLMD); you will need to checkout ML-result branches if you want access to the ML models and you will need to checkout the main branch if you want access to the up-to-date plotting scripts. More details about branches are provided in the top level README.md of this repository. Examples for managing branches and comparing results from multiple ML models on different branches are in the Jupyter notebooks in the ./examples directory in this repository.

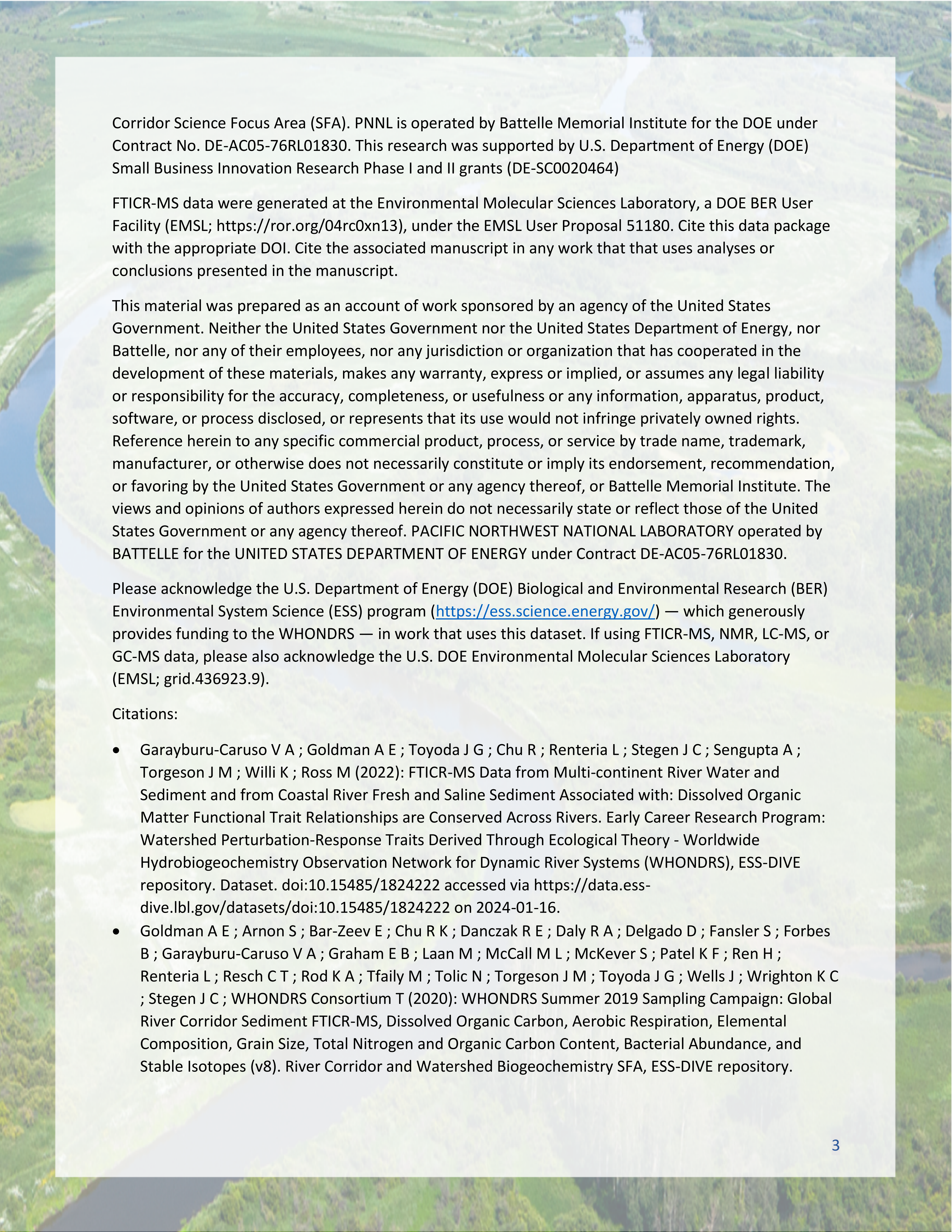
3 – The dates and times of the observations were removed during preprocessing and the ML models in this project were not trained with any awareness of time (either dates or seasons). The results here are meant to be interpreted in a mean, climatological sense.

Data Package Structure

Please see the file level metadata (flmd; “sl_archive_whondrs_flmd.csv”) for a list of all files contained in this data package and descriptions for each. Please see the data dictionary (dd; “sl_archive_whondrs_dd.csv”) for a list of all column headers contained within comma separated value (csv) files in this data package and descriptions for each. The GitHub repository is organized into five top-level directories: (1) “input_data” holds the training data for the ML models; (2) “ml_models” holds machine learning models trained on the data in “input_data”; (3) “scripts” contains data preprocessing and postprocessing scripts and intermediate results specific to this data set that bookend the ML workflow; (4) “examples” contains the visualization of the results in this repository including plotting scripts for the manuscript (e.g., model evaluation, FPI results) and scripts for running predictions with the ML models (i.e., reusing the trained ML models); (5) “output_data” holds the overall results of the ML model on that branch. Each trained ML model resides on its own branch in the repository; this means that inputs and outputs can be different branch-to-branch. Furthermore, depending on the number of features used to train the ML models, the preprocessing and postprocessing scripts, and their intermediate results, can also be different branch-to-branch. The “main-*” branches are meant to be starting points (i.e. trunks) for each model branch (i.e. sprouts). Please see the Branch Navigation section in the top-level README.md in the GitHub repository for more details. There is also one hidden directory “.github/workflows”. This hidden directory contains information for how to run the ML workflow as an end-to-end automated GitHub Action but it is not needed for reusing the ML models archived here. Please the top-level README.md in the GitHub repository for more details on the automation.

Citations and Acknowledgements

This research was supported by the U.S. Department of Energy (DOE) Office of Science, Biological and Environmental Research (BER) Program, Environmental System Science (ESS) Program (https://ess.science.energy.gov/) through the Pacific Northwest National Laboratory (PNNL) River



Corridor Science Focus Area (SFA). PNNL is operated by Battelle Memorial Institute for the DOE under Contract No. DE-AC05-76RL01830. This research was supported by U.S. Department of Energy (DOE) Small Business Innovation Research Phase I and II grants (DE-SC0020464)

FTICR-MS data were generated at the Environmental Molecular Sciences Laboratory, a DOE BER User Facility (EMSL; https://ror.org/04rc0xn13), under the EMSL User Proposal 51180. Cite this data package with the appropriate DOI. Cite the associated manuscript in any work that that uses analyses or conclusions presented in the manuscript.

This material was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the United States Department of Energy, nor Battelle, nor any of their employees, nor any jurisdiction or organization that has cooperated in the development of these materials, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness or any information, apparatus, product, software, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. PACIFIC NORTHWEST NATIONAL LABORATORY operated by BATTELLE for the UNITED STATES DEPARTMENT OF ENERGY under Contract DE-AC05-76RL01830.

Please acknowledge the U.S. Department of Energy (DOE) Biological and Environmental Research (BER) Environmental System Science (ESS) program (https://ess.science.energy.gov/) — which generously provides funding to the WHONDRS — in work that uses this dataset. If using FTICR-MS, NMR, LC-MS, or GC-MS data, please also acknowledge the U.S. DOE Environmental Molecular Sciences Laboratory (EMSL; grid.436923.9).

Citations:

- Garayburu-Caruso V A ; Goldman A E ; Toyoda J G ; Chu R ; Renteria L ; Stegen J C ; Sengupta A ; Torgeson J M ; Willi K ; Ross M (2022): FTICR-MS Data from Multi-continent River Water and Sediment and from Coastal River Fresh and Saline Sediment Associated with: Dissolved Organic Matter Functional Trait Relationships are Conserved Across Rivers. Early Career Research Program: Watershed Perturbation-Response Traits Derived Through Ecological Theory - Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS), ESS-DIVE repository. Dataset. doi:10.15485/1824222 accessed via https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1824222 on 2024-01-16.
- Goldman A E ; Arnon S ; Bar-Zeev E ; Chu R K ; Danczak R E ; Daly R A ; Delgado D ; Fansler S ; Forbes B ; Garayburu-Caruso V A ; Graham E B ; Laan M ; McCall M L ; McKever S ; Patel K F ; Ren H ; Renteria L ; Resch C T ; Rod K A ; Tfaily M ; Tolic N ; Torgeson J M ; Toyoda J G ; Wells J ; Wrighton K C ; Stegen J C ; WHONDRS Consortium T (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 (v8). River Corridor and Watershed Biogeochemistry SFA, ESS-DIVE repository.

Dataset. doi:10.15485/1729719 accessed via https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1729719 on 2024-01-16.

- Hartmann 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.
- Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Scientific Data 6: 283. doi: https://doi.org/10.1038/s41597-019-0300-6.

Contact

Stefan Gary, sfgary@parallelworks.com

Timothy Scheibe, tim.scheibe@pnnl.gov

Change History

Data Package Version	Changes
Version 1 *March 2024*	Original data package publication