Title

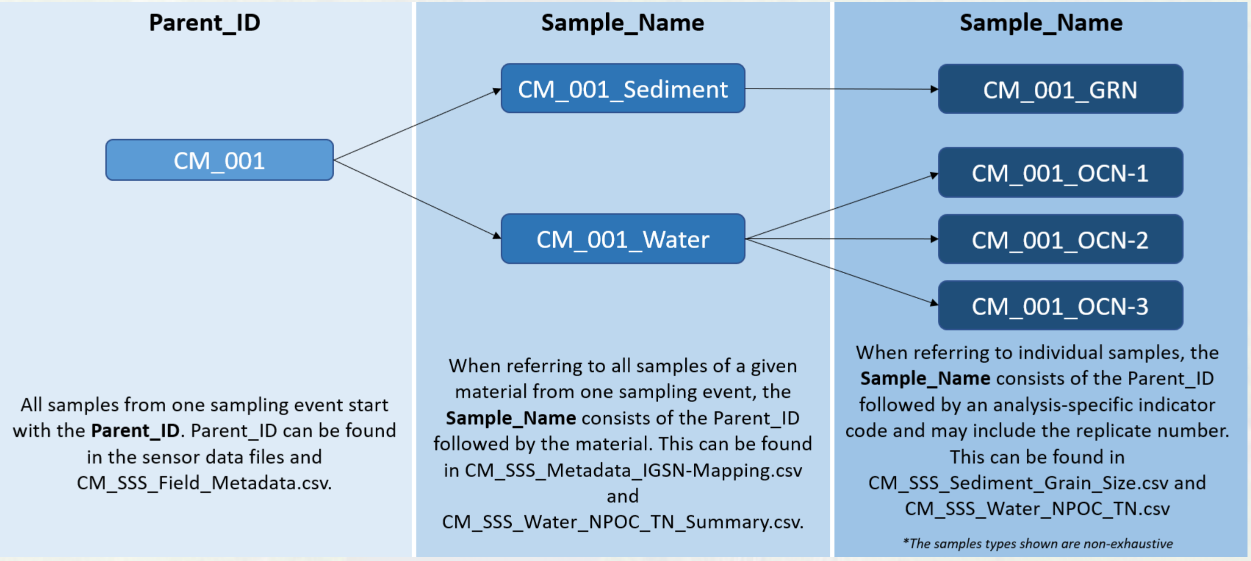
WHONDRS River Corridor Dissolved Oxygen, Temperature, Sediment Aerobic Respiration, Grain Size, and Water Chemistry from Machine-Learning-Informed Sites across the Contiguous United States (v2)

Summary

This dataset supports a broader study examining hyporheic zone respiration rates to improve predictive models at a contiguous United States (CONUS) scale. The CONUS-Scale Model-Sample Study (CM) was designed following ICON (integrated, coordinated, open, and networked) principles to facilitate a model-experiment (ModEx) iteration approach, leveraging crowdsourced sampling across the CONUS. New machine learning models are created every month to guide sampling locations. Data from the resulting samples are used to test and rebuild the machine learning models for the next round of sampling guidance. Sampling began in April 2022, and samples will continue to be collected across the CONUS through 2023 and possibly beyond. This data package will be updated semi-regularly with newly generated data. In addition to widely distributed CONUS sites, a more spatially focused sampling occurred in the Yakima River Basin, WA in summer 2022. Data from this more spatially intensive sampling occurred under the label "Second Spatial Study (SSS)" and were also included in the machine learning models. We acknowledge the Yakama Nation as owners and caretakers of the lands where we collected samples and data for SSS. We thank the Confederated Tribes and Bands of the Yakama Nation Tribal Council and Yakama Nation Fisheries for working with us to facilitate sample collection and optimization of data usage according to their values and worldview. Data from CM and SSS were collected using the same methods. Other data types collected from SSS that were not part of CM were published in a separate data package (https://data.ess-dive.lbl.gov/view/doi:10.15485/1969566).

Critical Details

1 – Each sampling event has a unique Parent_ID in the format CM_# or SSS#. The field metadata and data files all contain these unique IDs and can be mapped across each other accordingly. The Parent_ID may have other indicators appended when referring to samples. See figure below for details.



All samples from one sampling event start with the **Parent_ID**. Parent_ID can be found in the sensor data files and CM_SSS_Field_Metadata.csv.

When referring to all samples of a given material from one sampling event, the **Sample_Name** consists of the Parent_ID followed by the material. This can be found in CM_SSS_Metadata_IGSN-Mapping.csv and CM_SSS_Water_NPOC_TN_Summary.csv.

When referring to individual samples, the **Sample_Name** consists of the Parent_ID followed by an analysis-specific indicator code and may include the replicate number. This can be found in CM_SSS_Sediment_Grain_Size.csv and CM_SSS_Water_NPOC_TN.csv

**The samples types shown are non-exhaustive*

2 – At the end of each month, data collected during the month (dissolved oxygen, temperature, and pH) are used to produce new machine learning models to predict sediment respiration. The models, in turn, provide updated prioritization of potential CONUS sampling sites. Prioritization is based on pushing the bounds of the models (i.e., divergence) (e.g., sampling in environmental conditions not previously

sampled). The first set of models used data from the WHONDRS Summer 2019 Study (S19S) (https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1729719) and the GLObal RIver Chemistry Database (GLORICH; https://doi.org/10.1016/j.proeps.2014.08.005). A map of the top 10 percent of divergent priority sites ("high priority") and top 10% of convergent priority sites ("low priority"). Collaborators signed up to sample, WHONDRS (https://whondrs.pnnl.gov) shipped sampling supplies and protocols to the collaborator, and after sampling, the collaborator shipped the samples to PNNL to be analyzed or stored for future analyses.

3 – The iterative maps generated each month and sent to collaborators can be found at the following link: https://tinyurl.com/CM-Published-Maps.

4 – In addition to the CONUS machine-learning-guided sites and Yakima River Basin sites, three National Ecological Observatory Network (NEON) sites were selected for recurring sampling. These sites will each be sampled bimonthly for one year (six times total) at a minimum by the NEON team to provide time series samples. The Site_IDs of these sites in the field metadata are MART, HOPB, and MAYF.

5 – Experimental Design (Iron, Incubation/Respiration, Dry/Wet amounts): 10 milliliters or 2.5 milliliters of wet sediment is weighed and added to a 40 milliliter glass vial in the laboratory. Unfiltered water from the same sampling site is added until there is no headspace. Partial pressure measurements of oxygen are taken for up to two hours, after which two milliliters of the sediment/water slurry are taken from the vial for iron (II) analysis. The remaining sediment and water in the vial is weighed and then transferred to a 50 milliliter tube to obtain the dry mass of sediment. Dry mass of sediment and mass of water in the vial is used to normalize iron (II) and respiration measurements.

6 – Respiration rate QAQC and normalized respiration rate: Respiration rate is calculated from the change in oxygen over time in the incubation vial. In cases where oxygen consumption was too rapid to measure the oxygen change using the experimental design, two paths were used to calculate rates: (1) a theoretical maximum rate was calculated and used if the experiment could not be repeated with less sediment; (2) the incubation was repeated with a smaller volume (2.5 milliliters of sediment). Given that the ability to measure the respiration rate was dependent in part on the amount of sediment volume in the incubation, a normalized respiration rate file has also been provided in this data package. The normalized respiration rate provides the rate of oxygen consumption per liter of sediment to aid in comparability of measurements. In both the calculation of the theoretical rates and the normalized rates, the project team made assumptions. We encourage any data users to first decide if they are comfortable with these calculations before using the data. See the methods codes file for more detail on calculations.

Data Package Structure

This dataset is comprised of two folders with field photos and videos and one main data folder containing (1) file-level metadata; (2) data dictionary; (3) field metadata; (4) dissolved organic carbon (DOC, measured as non-purgeable organic carbon, NPOC) data and averages; (5) total nitrogen data and averages; (6) sediment grain size data; (7) sediment iron (II) data and averages; (8) miniDOT dissolved oxygen and temperature summary data; (9) sediment incubation respiration rate data and averages; (10) normalized respiration rate data and averages; (11) miniDOT installation methods; (12) field protocols; (13) readme; (14) methods codes; (15) international geo-sample number (IGSN) mapping file; (16) a subfolder with miniDOT dissolved oxygen and temperature data and plots; and (17) a subfolder with sediment incubation respiration data, scripts, and plots. All files are .csv, .pdf, .R, .jpg, .jpeg, .png, .mov, .mp4.

Acknowledgements

A portion of this research was supported by the U.S. Department of Energy (DOE) Biological and Environmental Research (BER) Environmental System Science (ESS) program (https://ess.science.energy.gov/) through the Pacific Northwest National Laboratory River Corridor Science Focus Area (SFA). PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830.

A portion of this research was supported by the U.S. Department of Energy (DOE) Small Business Innovation Research (SBIR) program through Parallel Works (Grant DE-SC0020464: A Platform for Scientific Data Management, Modeling and Analysis with Machine Learning).

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. Data collected/used in this research were obtained through the NEON Assignable Assets program.

WHONDRS consortium members were asked to provide any acknowledgments for their sample collection and the following is a list of acknowledgments that were submitted:

- Research activities were conducted in part on the Wind River Experimental Forest within the Gifford Pinchot National Forest.
- Philadelphia is part of Lenapehoking, the ancestral homelands of the Lenape peoples.
- Land surveyed is the ancestral homelands of the Nookhose'iinenno (Arapaho), Tsis tsis'tas (Cheyenne), and Nuuchu (Ute).
- Georgia Coastal Ecosystem LTER, OCE-1832178.

Contact

James Stegen, james.stegen@pnnl.gov

Change History

Version 1	February 2023	Original data package publication
Version 2	June 2023	• Added data, metadata, and photos for samples collected November 14, 2022 - April 24, 2023. • NPOC limit of detection (LOD) and data were corrected for error in LOD calculation. • Corrected coordinates and/or site IDs in field metadata and/or IGSN mapping file for parent IDs SSS001, SSS004, SSS006, and SSS022. • Added sediment iron (II) data, summary file, and associated methods metadata. • Added water and sediment mass and volume and associated methods metadata. • Removed “00000” from grain size column headers in data file and dd. • Added respiration rate data with updated QAQC. • Added respiration rate data normalized to volume of sediment and associated methods metadata. • Updated flmd, dd, and readme to reflect changes.