## Title

WHONDRS River Corridor Sediment and Water Geochemistry and In Situ Sensor Data from Machine-Learning-Informed Sites across the Contiguous United States (v3)

## 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 ended in October 2023 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).

This data package was originally published in February 2023. It was updated in June 2023 (v2; new and modified files) and again in December 2023 (v3; new and modified files). See the change history section below for more details.

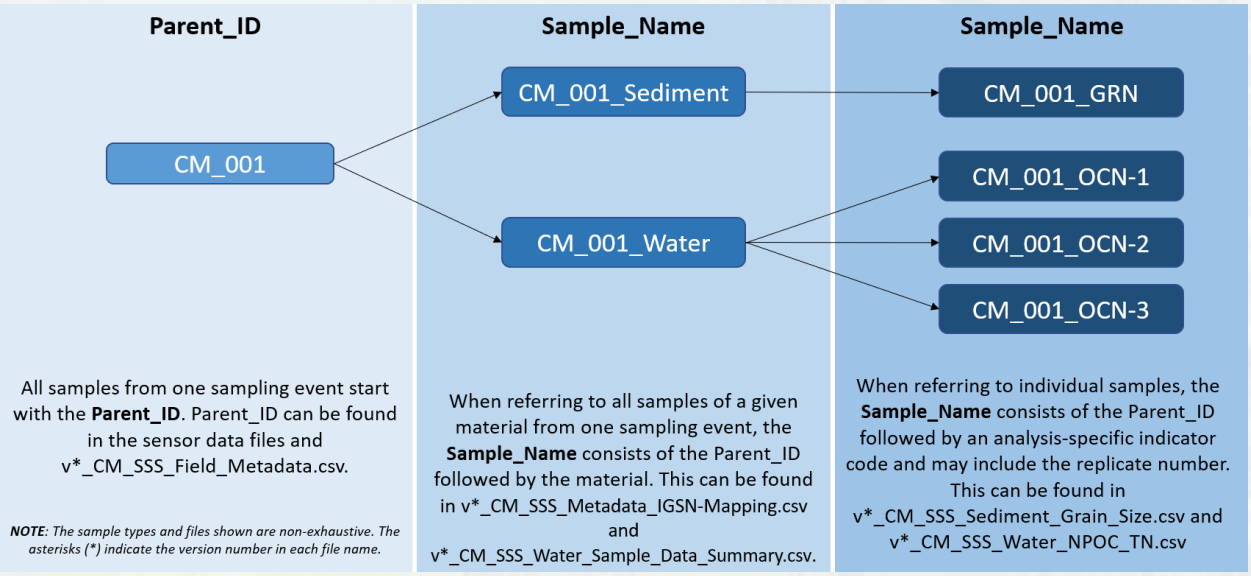
## Brief Overview of Methods

Samples were collected across the contiguous United States via WHONDRS (https://whondrs.pnnl.gov) crowdsourced sampling. Site metadata, including general environmental information, were manually recorded in the field. See CM_Sampling_Protocol.pdf and SSS_Field_Protocol.pdf for details on sample collection, sensor deployment, and site metadata collection. Note that the SSS protocol includes data types published in a separate data package, because they are not included as part of CM. A miniDOT logger was deployed underwater upstream of the sampling site to measure dissolved oxygen and temperature during sampling. Unfiltered surface water and filtered surface water were collected at 50 percent water column depth. Sediment samples were collected underwater at 1 to 3 centimeters depth below the streambed and sieved to less than 2 millimeters. The samples were refrigerated after sampling and shipped or transported on blue ice to Pacific Northwest National Laboratory. Sediment aerobic respiration rate was calculated from time series dissolved oxygen data measured from laboratory sediment/water incubations using a non-invasive fiber-optic approach. For details regarding laboratory and instrument methods for this analysis and others, see the alphanumeric methods codes located in the header rows of the chemistry data csv file and their associated definitions in v*_CM_SSS_Methods_Codes.csv. Data formats follow ESS-DIVE Reporting Formats.

## Critical Details

1 – Naming structure: 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.




2 – Model-guided sampling: At the end of each month of sampling, data collected during the month (dissolved oxygen, temperature, and pH) were used to produce new machine learning models to predict sediment respiration. The models, in turn, provided 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") was created and sent to collaborators. 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 – Model-guided maps: The iterative maps generated each month and sent to collaborators can be found at the following link: https://tinyurl.com/CM-Published-Maps.

4 – Time series sites: 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 were each sampled bimonthly for one year (six times total) 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 was weighed and added to a 40 milliliter glass vial in the laboratory. Unfiltered water from the same sampling site was added until there was no headspace. Partial pressure measurements of oxygen were taken for up to two hours, after which two milliliters of the sediment/water slurry were taken from the vial for iron (II) analysis. The remaining sediment and water in the vial was 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 was 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.

7 – Multiple sampling times: Most sites were sampled once for this study (with the exception of NEON sites, see number 4 above), however one site was sampled twice (MP-102338). Sediment was not collected during the first sampling event (CM_026), so some data types are missing. The site was revisited (CM_039) and the full suite of samples and data types were collected during the second visit. The data user may decide to eliminate CM_026 from analyses or keep it.

## 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) readme; (5) international generic sample number (IGSN) mapping file; (6) field protocols; (7) a subfolder with sample data; and (8) a subfolder with sensor data. The sample data subfolder contains (1) dissolved organic carbon (DOC, measured as non-purgeable organic carbon, NPOC) data and averages; (2) total nitrogen data and averages; (3) sediment grain size data; (4) sediment iron (II) data and averages; (5) wet sediment mass, dry sediment mass, water mass, and wet sediment volume in incubation vial; (6) sediment incubation respiration rate data and averages; (7) normalized respiration rate data and averages; (8) methods codes; (9) sediment specific surface area; and (10) a subfolder with sediment incubation respiration data, scripts, and plots. The sensor data subfolder contains (1) a subfolder with miniDOT dissolved oxygen and temperature data and plots; (2) miniDOT dissolved oxygen and temperature summary data; and (3) miniDOT installation methods. All files are .csv, .pdf, .R, .jpg, .jpeg, .png, .mov, or .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 with their corresponding Site IDs:

- MART: Research activities were conducted in part on the Wind River Experimental Forest within the Gifford Pinchot National Forest.
- MP- 100379: Philadelphia is part of Lenapehoking, the ancestral homelands of the Lenape peoples.
- MP-102398: Land surveyed is the ancestral homelands of the Nookhose'iinenno (Arapaho), Tsis tsis'tas (Cheyenne), and Nuuchu (Ute).
- MP-100749 and MP- 100747: Georgia Coastal Ecosystem LTER, OCE-1832178.
- SP-70 and SP-72: Eastern Shoshone, Shoshone-Bannock.
- MP- 102944: Funded by Oregon Watershed Enhancement Board. On the traditional lands of the Confederated Tribes of the Siletz, Confederated Tribes of the Grand Rhonde, and the Clatsop-Nehalem Confederated Tribe.
- MP- 100607: Holiday Creek is located on the traditional territory of the Monacan Indian Nation.
- SP-45: Lafayette Blue Springs State Park.
- MP-102420: NSF DEB-2016749.
- MP-100019: New Hampshire Agriculture Experiment Station.
- SP-35: Rayonier (land owner; https://www.rayonier.com/).
- MP- 101276: US Department of Energy, Office of Science, Biological and Environmental Research, Subsurface Biogeochemical Research, Watershed Dynamics and Evolution SFA at ORNL.
- MP- 103224: Watershed Dynamics and Evolution SFA at ORNL.
- MP- 101584: Traditional lands of the Oceti Sakowin (Dakota, Lakota, Nakoda) and Anishinaabe Peoples.

## Contact

James Stegen, james.stegen@pnnl.gov

## Change History

Approach to change history and versioning:

Updates to **data package** version: When any file within a data package is updated, the data package version number is updated. The data package version number is indicated in the title of the data package, the data package folder name, and in the change history table below. You can access previous versions of the data package by sending a request to ESS-DIVE.

Updates to **individual file** versions: As files are changed, the file version number is also updated. The file version number is indicated in the file name, file level metadata (flmd) file, and the change history table below. The version number on an individual file may not match the version number of the data package. For example, v3 of a data package may include v2 of an individual file.

The change history below describes each file revised during versioning. If you are interested in seeing the exact cells within a file that have changed, you can utilize the daff package in R (https://github.com/edwindj/daff) to compare a previously downloaded file to a newly downloaded file.

In the change history table below, the sub-headers and bullets indicate the type of change in each file:

- New files: Describes new files added that were not present in previous data package versions

- Bulk changes to files: Describes a change to many files within the data package. The indicated superscript with be added to each file name that the change applies to.
- Modified files:
    - Corrected: Describes existing information modified or removed to prevent sharing of incorrect information
    - Added: Describes new information inserted into an existing file (e.g., appending new columns/rows)
    - Updated: Describes modifying existing information to maintain accuracy though version changes. (e.g., changing version number to new version number)

Change history:

| Data Package Version | Changes |
|---|---|
| **Version 1**<br>*February 2023* | Original data package publication |
| **Version 2**<br>*June 2023* | **NEW FILES**<br>• CM_*_DO_Temp.csv<br>• CM_SSS_Sediment_Fe.csv<br>• CM_SSS_Sediment_Fe_Respiration_Summary.csv<br>• CM_SSS_Sediment_Normalized_Respiration_Rates.csv<br>• CM_SSS_Sediment_Normalized_Respiration_Rates_Summary.csv<br>• CM_SSS_Sediment_Water_Mass_Volume.csv<br>**BULK CHANGES TO FILES**<br>[1] indicates (meta)data were added to the corresponding file for samples collected November 14, 2022 – April 24, 2023.<br>**MODIFIED FILES**<br>**CM_SSS_Metadata_IGSN-Mapping.csv** (v2)<br>• Corrected coordinates and/or site IDs for parent IDs SSS001, SSS004, SSS006, and SSS022.<br>**CM_SSS_Field_Metadata.csv** (v2) [1]<br>• Corrected coordinates and/or site IDs for parent IDs SSS001, SSS004, SSS006, and SSS022.<br>**CM_SSS_Sediment_Incubations_Respiration_Rates.csv** (v2) [1]<br>• Corrected the data with updated QAQC approach.<br>**CM_SSS_Water_NPOC_TN.csv** (v2) [1]<br>• Corrected NPOC limit of detection (LOD) and data for error in LOD calculation.<br>**CM_SSS_Water_NPOC_TN_Summary.csv** (v2) [1]<br>• Corrected NPOC limit of detection (LOD) and data for error in LOD calculation.<br>**CM_SSS_dd.csv** (v2)<br>• Added rows for new column headers from new data types.<br>• Updated grain size column headers to remove “00000”<br>**CM_SSS_Methods_Codes.csv** (v2)<br>• Added rows for methods codes of new data types.<br>• Updated column headers to remove “00000”.<br>**readme_BSLE.pdf** (v2) |

| | |
|---|---|
| | • Updated version number in data package title<br>• Updated data package structure to include new data types.<br>• Added new versioning information.<br>**CM_SSS_flmd.csv** (v2)<br>• Added rows for new files.<br>• Added version number to updated files.<br>**CM_SSS_Sediment_Incubations_DO_vs_Incubation_Time_Plots.pdf** (v2) [1]<br>**CM_SSS_miniDOT_Plots.pdf** (v2) [1]<br>**CM_SSS_miniDOT_DO_Temp_Summary.csv** (v2) [1]<br>**CM_SSS_Sediment_Grain_Size.csv** (v2) [1]<br>**CM_DO_INC.csv** (v2) [1]<br>**CM_FieldPhotos.zip** (v2) [1] |
| **Version 3**<br>*December 2023* | **NEW FILES**<br>• CM_SSS_Sediment_Specific_Surface_Area.csv<br>• CM_*_DO_Temp.csv<br><br>**BULK CHANGES TO FILES**<br>[2] indicates (meta)data were added to the corresponding file for samples collected April 24, 2023 – October 10, 2023<br>[3] indicates a column was added with international generic sample numbers (IGSN)<br><br>**MODIFIED FILES**<br>**Folder Structure**<br>• All sample data and related methods information were put into a "Sample_Data" subfolder. All sensor data and related methods information were put into a "Sensor_Data" subfolder.<br>**SSS_FieldPhotos.zip** (v2)<br>• Corrected site IDs in file names.<br>**SSS_FieldVideos.zip** (v2)<br>• Corrected site IDs in file names.<br>**CM_045_DO_Temp.csv** (v2)<br>• Corrected data as it did not previously include necessary preprocessing.<br>**CM_SSS_Field_Metadata.csv** (v3) [2]<br>• Corrected site IDs for 26 sites.<br>**CM_SSS_Metadata_IGSN-Mapping.csv** (v3) [2]<br>• Corrected site IDs and resulting parent IGSNs for 26 sites.<br>**CM_SSS_Sediment_Sample_Data_Summary.csv** (v2) [2, 3]<br>• Corrected means for both iron columns. The mean previously included data from 2.5 milliliters and 10 milliliters incubations. It now only includes data from the 10 milliliters incubations.<br>• Added columns with summary of new data types.<br>• Updated file name. Previously named "CM_SSS_Sediment_Normalized_Respiration_Rates_Summary.csv".<br>**CM_SSS_Sediment_Water_Mass_Volume.csv** (v2) [2]<br>• Corrected average mass of 50 milliliters tubes used to calculate the dry sediment and water masses. The value used now reflects the average of the whole data set, rather than the partial data set previously published. This affects all samples with the INC_QA_000 methods deviation. |

| | |
|---|---|
| | **CM_SSS_Water_NPOC_TN.csv** (v3) [2, 3]<br>• Corrected values for samples that were rerun. Original values were accidently published previously.<br>• Corrected values that were recalculated with an adjusted calibration curve.<br>• Updated unit basis for clarity.<br>• Updated structure of text string when data value is below LOD or above/below standard curve and replaced deviation code describing the text string.<br>**CM_SSS_Water_Sample_Data_Summary.csv** (v3) [2, 3]<br>• Corrected means by recalculating with the corrected values in CM_SSS_Water_NPOC_TN.csv.<br>• Updated file name. Previously called "CM_SSS_Water_NPOC_TN_Summary.csv".<br>**CM_SSS_flmd.csv** (v3)<br>• Added rows for new files.<br>• Updated version number to corrected, added, and/or updated files.<br>**CM_SSS_Water_DO_Temp_Summary.csv** (v3) [2, 3]<br>• Updated column name from "Sample_Name" to "Parent_ID".<br>• Updated formatting to include metadata header rows.<br>• Updated file name. Previously named "CM_SSS_miniDOT_DO_Temp_Summary.csv".<br>**readme_CM_SSS.pdf** (v3)<br>• Added new versioning information.<br>• Added methods section.<br>• Added information to the critical details section.<br>• Added new acknowledgements.<br>• Updated version number in data package title.<br>• Updated data package structure to include new data types and updated folder structure.<br>**CM_SSS_dd.csv** (v3)<br>• Added rows for new column headers from new data types.<br>**CM_SSS_Methods_Codes.csv** (v3)<br>• Added rows for methods codes of new data types.<br>**CM_DO_INC.csv** (v3) [2]<br>**CM_FieldPhotos.zip** (v3) [2]<br>**CM_SSS_miniDOT_Plots.pdf** (v3) [2]<br>**CM_SSS_Sediment_Incubations_DO_vs_Incubation_Time_Plots.pdf** (v3) [2]<br>**CM_SSS_Sediment_Grain_Size.csv** (v3) [2, 3]<br>**CM_SSS_Sediment_Normalized_Respiration_Rates.csv** (v2) [2, 3]<br>**CM_SSS_Sediment_Incubations_Respiration_Rates.csv** (v3) [2, 3]<br>**CM_SSS_Sediment_Fe.csv** (v2) [2, 3] |

**Note**: An asterisks (*) in a file name indicates multiple files with the same file and file naming structure. See the flmd more details about specific files and the meaning of the asterisks.