Title

WHONDRS River Corridor Sediment and Water Geochemistry and In Situ Sensor Data from Machine-Learning-Informed Sites across the Contiguous United States (v5)

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 were created every month to guide sampling locations. Data from the resulting samples were 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 the 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. 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); December 2023 (v3; new and modified files); June 2024 (v4; new and modified files); and April 2024 (v5; new and modified files). See the change history section below for more details.

For details on how to navigate data packages generated by this project, see https://data.ess-dive.lbl.gov/portals/PNNLRiverCorridorSFA/About.

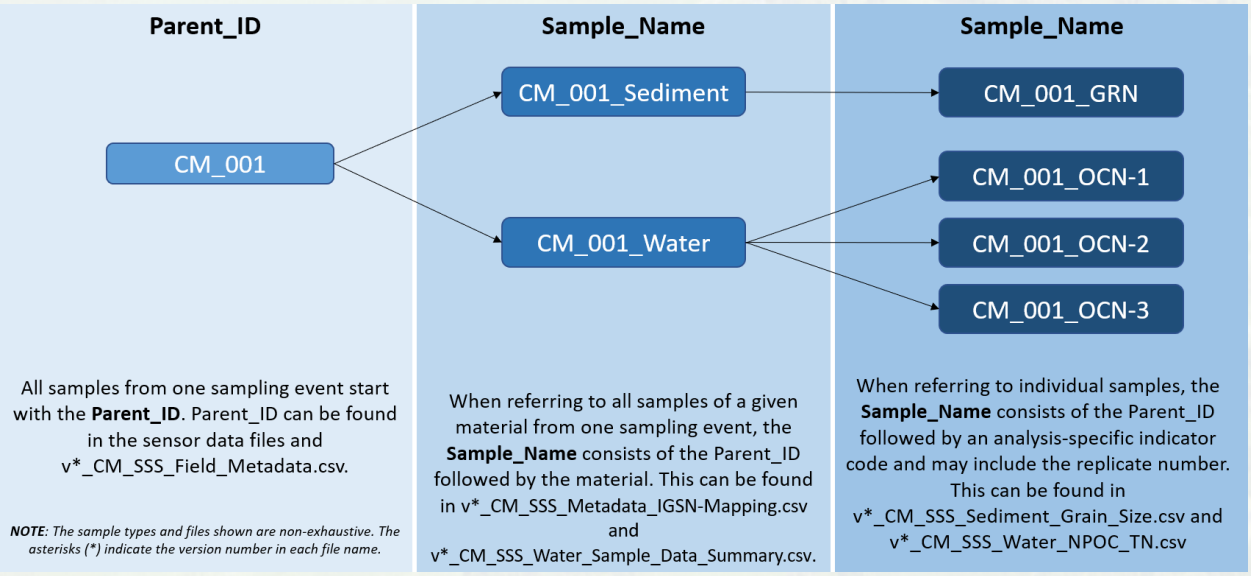
Brief Overview of Methods

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.

8. FTICR-MS data is provided in four processing levels in this data package. The unprocessed instrument files are available in a zipped folder called "CM_SSS_FTICR_Raw_Data.zip". These can be processed in newly developed public workflows (i.e., CoreMS) for data use or they can be first processed in a proprietary software (i.e., Bruker) into XML files and then processed in publicly available Formultitude (previously called Formularity; https://github.com/PNNL-Comp-Mass-Spec/Formultitude) for further data use. In previous WHONDRS data packages, we have processed the files in Bruker prior to publishing and then only published XML files and instructions for the user to process the XML files in Formultitude. In this data package, due to these new software developments, we have included raw data, pre-processed (i.e., peak picked) XML files, CoreMS output files, and fully processed data to allow for maximum data user flexibility.

9. There is a subfolder named for each individual sample within "CM_SSS_FTICR_Raw_Data.zip" folder. The sample names from the subfolder names directly map to the XML files, CoreMS output files, processed data, and to other data type same names. Within the subfolder named for each individual sample (*.d subfolder), there is a .d file. This file is named with a shortened sample name used for data generation (the shortened name should be ignored) but is the same sample as the *.d subfolder name.

10. We have provided a static version of our CoreMS processing instructions and scripts that were used for this dataset. These instructions and scripts may change over time as this new pipeline evolves. The most up to date instructions and scripts can be found at https://github.com/danczakre/RCSFA_CoreMS/tree/main/. A release was created on GitHub to document the version pulled for this data package.

11. We do not recommend combining processed data, whether it was run through Formultitude or CoreMS. If you wish to combine data from across datasets, you need to go back to at least the .xml files and reprocess the data using your pipeline of choice. For convenience, we have included the CoreMS outputs which **can** be combined with other CoreMS outputs released in our data packages, though this is not ideal. We are happy to work through these considerations with you if you have questions.

12. During solid phase extraction (SPE), FTICR-MS samples were diluted to achieve a standard NPOC concentration of 1.5 milligrams carbon per liter. For the water samples (labeled with "ICR" in the sample name), the water NPOC values were used for this calculation (labeled with "OCN" in the sample name). For example, CM_001_OCN-1 NPOC values were used to determine the dilution for CM_001_ICR-1. For the sediment samples (labeled with "SED" in the sample name), the sediment NPOC values were used

for this calculation (labeled with "SED" in the sample name). For example, CM_001_SED-1 NPOC values were used to determine the dilution for CM_001_SED-1

Data Package Structure

This dataset is comprised of two folders of field photos and videos, one folder of raw Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) data 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) surface water and sediment dissolved organic carbon (DOC, measured as non-purgeable organic carbon, NPOC) data and averages; (2) surface water and sediment total nitrogen data and averages; (3) surface water major cations and anions and averages; (4) sediment grain size data; (5) sediment iron (II) data and averages; (6) wet sediment mass, dry sediment mass, water mass, and wet sediment volume in incubation and sediment ICR vials; (7) sediment incubation respiration rate data and averages; (8) normalized respiration rate data and averages; (9) methods codes; (10) sediment specific surface area; (11) sediment percent carbon and nitrogen; (12) sediment gravimetric moisture and averages; (15) sediment X-ray diffraction (XRD) data; (16) sediment adenosine triphosphate (ATP) and averages; (17) a subfolder with sediment incubation respiration data, scripts, and plots; (18) surface water and sediment FTICR methods; and (19) a subfolder of 9 Tesla (9T) FTICR-MS data. This folder contains five subfolders, one containing the sediment .xml data files, one containing the water .xml files, one containing the sediment CoreMS output files, one containing the water CoreMS output files, and the other containing instructions and scripts for processing the files in CoreMS (https://github.com/EMSL-Computing/CoreMS).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, .xml, .d, .html, .Rmd, .py, .cal, .json, .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).

XRD data were generated at the Environmental Molecular Sciences Laboratory, a DOE BER User Facility (EMSL; https://ror.org/04rc0xn13), under the EMSL User Proposal 61025 and 60624.

FTICR-MS data were generated at University of Arizona by Malak Tfaily's lab (https://malaktfaily.com).

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.

We thank the United States Forest Service, Washington Department of Fish and Wildlife, Washington Department of Natural Resources, Cowiche Canyon Conservatory, Washington State Parks and

Recreation Commission (Scientific Research Permit #210901), and the Confederated Tribes and Bands of the Yakama Nation for access to field locations where the SSS samples were collected. We also thank 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.

WHONDRS consortium members were asked to provide any acknowledgments for the CM 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 *February 2023*	Original data package publication
Version 2 *June 2023*	**NEW FILES** • CM_*_DO_Temp.csv • CM_SSS_Sediment_Fe.csv • CM_SSS_Sediment_Fe_Respiration_Summary.csv • CM_SSS_Sediment_Normalized_Respiration_Rates.csv • CM_SSS_Sediment_Normalized_Respiration_Rates_Summary.csv • CM_SSS_Sediment_Water_Mass_Volume.csv **BULK CHANGES TO FILES** [1] indicates (meta)data were added to the corresponding file for samples collected November 14, 2022 – April 24, 2023. **MODIFIED FILES** **CM_SSS_Metadata_IGSN-Mapping.csv** (v2) • Corrected coordinates and/or site IDs for parent IDs SSS001, SSS004, SSS006, and SSS022. **CM_SSS_Field_Metadata.csv** (v2) [1] • Corrected coordinates and/or site IDs for parent IDs SSS001, SSS004, SSS006, and SSS022. **CM_SSS_Sediment_Incubations_Respiration_Rates.csv** (v2) [1] • Corrected the data with updated QAQC approach. **CM_SSS_Water_NPOC_TN.csv** (v2) [1] • Corrected NPOC limit of detection (LOD) and data for error in LOD calculation. **CM_SSS_Water_NPOC_TN_Summary.csv** (v2) [1] • Corrected NPOC limit of detection (LOD) and data for error in LOD calculation.

	CM_SSS_dd.csv (v2) • Added rows for new column headers from new data types. • Updated grain size column headers to remove "00000" **CM_SSS_Methods_Codes.csv** (v2) • Added rows for methods codes of new data types. • Updated column headers to remove "00000". **readme_BSLE.pdf** (v2) • Updated version number in data package title • Updated data package structure to include new data types. • Added new versioning information. **CM_SSS_flmd.csv** (v2) • Added rows for new files. • Added version number to updated files. **CM_SSS_Sediment_Incubations_DO_vs_Incubation_Time_Plots.pdf** (v2) [1] **CM_SSS_miniDOT_Plots.pdf** (v2) [1] **CM_SSS_miniDOT_DO_Temp_Summary.csv** (v2) [1] **CM_SSS_Sediment_Grain_Size.csv** (v2) [1] **CM_DO_INC.csv** (v2) [1] **CM_FieldPhotos.zip** (v2) [1]
Version 3 *December 2023*	**NEW FILES** • CM_SSS_Sediment_Specific_Surface_Area.csv • CM_*_DO_Temp.csv **BULK CHANGES TO FILES** [2] indicates (meta)data were added to the corresponding file for samples collected April 24, 2023 – October 10, 2023 [3] indicates a column was added with international generic sample numbers (IGSN) **MODIFIED FILES** **Folder Structure** • 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. **SSS_FieldPhotos.zip** (v2) • Corrected site IDs in file names. **SSS_FieldVideos.zip** (v2) • Corrected site IDs in file names. **CM_045_DO_Temp.csv** (v2) • Corrected data as it did not previously include necessary preprocessing. **CM_SSS_Field_Metadata.csv** (v3) [2] • Corrected site IDs for 26 sites. **CM_SSS_Metadata_IGSN-Mapping.csv** (v3) [2] • Corrected site IDs and resulting parent IGSNs for 26 sites. **CM_SSS_Sediment_Sample_Data_Summary.csv** (v2) [2, 3] • 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.

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

	CM_SSS_Sediment_Incubations_Respiration_Rates.csv (v3) [2, 3] **CM_SSS_Sediment_Fe.csv** (v2) [2, 3]
Version 4 *June 2024*	**NEW FILES** • CM_SSS_Sediment_ATP.csv • CM_SSS_Sediment_CN.csv • CM_SSS_Sediment_Gravimetric_Moisture.csv • CM_SSS_Sediment_XRD.csv • CM_SSS_Water_Ions.csv **BULK CHANGES TO FILES** [4] indicates that the dissolved oxygen (DO) data was corrected. An offset was applied to account for sensor drift but was miscalculated previously. The corrected offset has now been applied to the DO data. [5] indicates DO saturation data was removed. The data was miscalculated previously. **MODIFIED FILES** **CM_SSS_Sediment_Water_Mass_Volume.csv** (v3) • Updated column name from "Wet_Sediment_mL" to "Wet_Sediment_Volume_mL" **CM_SSS_Sediment_Sample_Data_Summary.csv** (v3) • Added columns for new data. **CM_SSS_Water_Sample_Data_Summary.csv** (v4) • Added columns for new data. **readme_CM_SSS.pdf** (v4) • Added new versioning information. • Updated version number in data package title. • Updated data package structure to include new data types and updated folder structure. **CM_SSS_flmd.csv** (v4) • Added rows for new files. • Updated version number to corrected, added, and/or updated files. **CM_SSS_dd.csv** (v4) • Added rows for new column headers from new data types. **CM_SSS_Methods_Codes.csv** (v4) • Added rows for methods codes of new data types. ***_DO_Temp.csv** (v2 and v3) [4, 5] **CM_SSS_miniDOT_Plots.pdf** (v4) [4, 5] **CM_SSS_Water_DO_Temp_Summary.csv** (v4) [4, 5]
Version 5 *April 2025*	**NEW FILES** • *.xml • *.d • CM_SSS_Sediment_FTICR_Methods.csv • CM_SSS_Sediment_NPOC_TN.csv • CM_SSS_Water_FTICR_Methods.csv • CoreMS_MergeProcess.Rmd • CoreMS_Runner.py • Format_CoreMS_Outputs.R

	• getLambda.R • Instructions_for_Processing_FTICR_Data_using_CoreMS.pdf • *.corems.csv • *.corems.cal • *.corems.json • CM_SSS_CoreMS_Processed_ICR_Mol.csv • CM_SSS_Sediment_CoreMS_Processed_ICR_Calibration.csv • CM_SSS_Sediment_CoreMS_Processed_ICR_Data.csv • CM_SSS_Water_CoreMS_Processed_ICR_Calibration.csv • CM_SSS_Water_CoreMS_Processed_ICR_Data.csv **MODIFIED FILES** **CM_SSS_Sediment_Fe.csv** (v3) • Updated data for a sample that was rerun. • Updated data to report summarized analytical replicate data as biological replicate data to match other data types. **CM_SSS_Sediment_Sample_Data_Summary.csv** (v4) • Added columns for new data. • Updated summary values for reprocessed data **CM_SSS_Sediment_Water_Mass_Volume.csv** (v4) • Added data for additional samples **readme_CM_SSS.pdf** (v5) • Added critical details for new data types • Updated new versioning information. • Updated version number in data package title. • Updated data package structure to include new data types. **CM_SSS_flmd.csv** (v5) • Added rows for new files. • Updated version number to corrected, added, and/or updated files. **CM_SSS_dd.csv** (v5) • Added rows for new column headers from new data types. **CM_SSS_Methods_Codes.csv** (v5) • Added rows for methods codes of new data types.

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