Data Management Plan

CAREER: A Molecular Level Investigation of Halogenation as a Mechanism of Trace Gas Production and Organic Carbon Transformation at the Surface Ocean

PI: Yina Liu, Ph.D.

1. Data sharing and preservation

Data Description. Data will be generated from controlled laboratory halogenation experiments with different organic carbon (OC) substrates and field observation cruises. Each experiment and field observation will have data related to the formation of volatile halogenated organic carbon (VHOC) and halogenated organic carbon (HOC) using different mass spectrometry platforms. Other supporting data include: (*a*) experimental metadata recording experimental conditions such as experiment run time, temperature, salinity, and pH, (*b*) field metadata such as latitudes, longitude, depth, and CTD records, and (*c*) bulk chemical measurements such as excitation-emission matrices (EEMs), total dissolved organic carbon (DOC), and adsorbable Organic Halides (AOX).

Data Format. Raw mass spectrometry data is in Agilent .D format. These raw files require proprietary software such as Agilent's MassHunter. To make these data publically assessable, they will be converted into .CSV format. EEMs, DOC, AOX, and metadata will be stored in .CSV format. Protocols will be written in electronic format (Microsoft Word) and convert to .PDF format.

Metadata and Documentation. Experimental metadata recording experimental conditions, such as experiment run time, temperature, salinity, and pH, will be recorded in .CSV format. Field metadata such as latitudes, longitude, depth, and CTD records will be recorded in .CSV or ASCII text format. A readme file in .txt format will be used as data documentation, which will describe metadata associated with each experiment and field observation, as well as relevant information for the mass spectrometry, EEMs, DOC, and AOX.

Data Sharing. Cloud storage such as Microsoft OneDrive with password protection will be used to share data between project participants, e.g., the PI, Research Associate, graduate and undergraduate students, and high school teachers from Rudder High School (see Educational Objective 2 in Project Description), during the project.

Preservation. All data will be stored in electronic format in (*1*) the computer used to generate the data, such as the instrument control computer; (*2*) external hard drives, which are used to back up data from the instrument control computers on a weekly basis by project personnel; (*3*) cloud storage such as Microsoft OneDrive; and (*4*) the PI's office computer.

Designated notebooks will be used to record experimental details. These notebooks will be kept in an area accessible by the PI and participants working on the project. They will be scanned and digitized as PDF files to preserve an electronic copy of the data and metadata. All electronic files will be stored on a minimum of three hard drives.

All data will be preserved for at least three years after project completion.

Roles and Responsibilities. The PI is responsible for data quality control and assurance. Project personnel supervised by the PI, e.g., Research Associate and undergraduate and graduate students, will be responsible for data generation, data backup, archiving, and sharing.

Budget. The personnel who will be responsible for data management as described above are paid by the project. One terabyte of data storage via Microsoft OneDrive is freely available to Texas A&M researchers. The PI's lab already has data backup storage devices available for projects.

2. Data used in publications

The data, and its interpretation, will be disseminated through presentations at national meetings such as the American Geophysical Union (AGU) and the Association for the Sciences of Limnology and Oceanography (ASLO). The results and a link to the archived data will be published in peer-reviewed journals. Funds have been included in the budget to ensure that these manuscripts are freely available via the internet as open-access publications. The PI will ensure any figures and tables in publications are in machine-readable formats. Data and information gained from this project will also be disseminated through educational activities described in the Project Description.

Upon manuscript submission, quality-controlled data will be submitted to NSF's Biological and Chemical Oceanography Data management Office (BCO-DMO), where they will be accessible to the wider community. Metadata will be submitted for each experiment and field observation, including any QA/QC flags or codes used for the data. Mass spectrometry data will be converted to user-friendly formats such as .CSV files. Raw mass spectra will be freely available upon request, as they require proprietary software such as Agilent's MassHunter Software.

The data will be submitted to BCO-DMO as manuscripts reporting results from this project are submitted for publication. All manuscripts will be submitted to appropriate open-access preprint servers such as ChemRxiv and EarthArXiv under open source license when they are submitted to peer-reviewed journals, following guidelines specified by the target peer-reviewed journals. Supporting data, such as tables containing data used to generate plots and graphics presented in the manuscript, will be submitted supplemental materials, which will be freely available through the publisher upon acceptance for publication or through open-source servers such as OSF following appropriate open source license protocols. The manuscript and supplemental data information will cross-reference with the BCO-DMO data.

This project also generates data pipeline products such as algorithms in R and Matlab codes. These codes will be made available through GitHub under the Open-Source License upon manuscript submission. A compound library containing identities of HOC generated through this project will be made available through Github and public mass spectrometry databases such as MetaboLights, upon manuscript submission. Data of the natural HOC time series in the Gulf of Mexico will be accessible through BCO-DMO.

3. Data management resources

Data Volume. The data from this project include large raw mass spectrometry data. In total, data size on the order of terabytes is expected.

Data Repository. Processed, QA/QC data will be uploaded to BCO-DMO. Data pipeline codes and compound libraries will be uploaded to Github and public mass spectrometry databases such as MetaboLights.

4. Confidentiality, security, and rights

All unclassified data sharing will follow appropriate copyright and open-source protocols. All personally identifiable information, if any, will be removed prior to data submissions following NSF's policies. Any data pipeline codes such as R and Matlab codes will be shared via Github under the Open-Source License following appropriate protocols.