**Data Management Plan**

**1. Overview.** This data management plan (DMP) is part of the proposal "Collaborative Research: Individual based approaches to understanding krill distribution and aggregation".

**2. Expected Data**

**Data**. The following data will be generated: 1) Krill swimming trajectories and behavioral responses to flow; 2) krill aggregation structure; 3) simulation models and code; processed simulation outputs of aggregation structure; 4) publications (e.g., journals, conference papers, and presentations).

**Data Formats.** All data will be stored on PCs and laptops during experiments and archived permanently on DVDs after the experiments. This includes all electronic files of individual krill and aggregation visualizations (.mov, .tif, .jpg formats). Processed image data on swimming kinematics and aggregation structure will be archived as .xls and .txt files, and will be sent to the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http:bco-dmo.org/) for general access. This data also will include raw image files used in publications or products.

***Lab notebooks*** containing the raw data generated by this project, detailed descriptions of procedures and methodological approaches, deviations from protocols, specific equipment, and chemical reagents utilized for this project will be cataloged by all project participants. Notebooks and digital copies thereof will serve as permanent records of the project and will be available upon request from NSF program managers.

**Data Generation and Acquisition.** Data generated through computer simulations is expected to comprise ~500 fully validated production runs that will produce archival quality data.

**Software**. The software developed in this project includes the simulation code and post processing visualization and analysis codes. Software will be primarily developed in MATLAB and NetLogo. Code, demos, and documentation will be developed as open-source and shared in a GitHub repository.

**Meta Data.** Meta-data describing the data will be collated to facilitate storage, discovery and archiving. This meta-data will be obtained by several means: (a) The simulation software will automatically generate, along with the simulation results, the metadata for the outputs. The metadata includes the file name for the results, location, owner, create time, model version, runtime, inputs, hardware information, etc. Each simulation will be launched using a configuration file that specifies model configuration and parameterization; this information will be stored in the metadata. The output filename will appear as ModelName-VerNo-Date(YYYYMMDD)-Time(HHMMSS); (b) descriptive metadata based at a minimum on the Dublin Core schema; and (c) preservation meta data, versioning, and provenance.

**3. Data Storage and Preservation**

Data that is expected to be retained by the project for sharing and later archiving will include all publication data, all input and output data for all simulations, all computer codes, raw image files from of individual krill and aggregation visualizations, processed data from these visualizations.

**Storage and Backup during Project.** The PIs will be responsible for operational data storage and backup during project execution. On individual PI machines the spreadsheets are backed up in near real time using TimeMachine or a similar program. The system administrators at the institutions who manage the clusters where these simulations are run (see Facilities and Resources Document) in this project will provide secondary data backups. The backups are

expected to take place daily/weekly/monthly and will be stored locally.

**Data Capacity and Volume.** While we expect to initiate many simulations each year as part of this project, many of these simulations will not be successful (due to accuracy, convergence or stability issues). Output files will vary considerably in size based on the model run, from < 1 GB to ~ 50 GB. We expect to complete ~500 simulations with adequate accuracy, convergence and quality to have archival value, requiring 5-10 TB of storage space. Raw digital videos of individual and krill aggregations will require 2-3 TB of storage. Processed data will be roughly 100 MB.

**Security.** All data will be stored, during and post project execution, within the Bigelow and GT campus network infrastructure, which employs firewalls and secure authentication and authorization methods for login and access.

**Long-Term Archiving and Preservation.** All raw image data will be stored on publically available storage sites at the PI institutions (e.g. https://smartech.gatech.edu/), which will include a reference to the paper that contains the analysis of this data. Within two years of data collection, data will be transferred to BCO-DMO for public access and long-term storage. After the project has been completed, arrangements shall be made to transfer data at Bigelow Laboratory from short-term storage to a long-term archival system.

**Length of archival.** Data, tagged as described in Section 2 will be stored for five years, or until it has been successfully uploaded to, a nationally or internationally funded database

**Roles and Responsibilities.** The PIs will have decision making authority over all data management. The PIs will draft the overall data management policy during the first three months of the project. And will be reviewed twice a year.

**4. Data Retention**

**Operational Data.** Following the conclusion of the proposed project, data for simulations and publications will be retained for a minimum of five years by the PIs, as described in section 3.

**Archival Data.** The archival lifecycle and retention policy for archived data will be managed by the PIs as described in section 3. The retention period for this will be determined at the conclusion of the project.

**5. Data Sharing and Dissemination** The project will be highlighted on the websites of the PIs at Bigelow and GT, where we will also provide links to digital data repositories'. The PIs will collaborate closely on NSF Annual Reports to provide updates to NSF program managers, and to the general public. The PIs will collaborate on data management in their respective laboratories. Sharing files between the laboratories we will use in-house storage facilities which are password protected. Distribution of data to BCO-DMO will be done within two years of their collection via password-secure FTP. All data transferred will be made publicly available 24 months after the completion of the project as documented above. Prior to this, the data will be shared on a secure FTP website with any requestors. The software tools developed under NSF support will be released under an open-source license (to be determined) and available in a GitHub repository. As always, in our presentations, we encourage interested scientists to consider using the apparatuses or instruments that we have developed for mimicking and quantify oceanographic features in the laboratory for their own research.

**6) Curriculum Materials:** Digital presentations and large-format research posters will be made available on the websites of each PI for educational use following the guidelines from NSF on data release and fair use. The Bigelow PIs expect to incorporate results from the proposed research into their undergraduate course lecture materials.