DATA MANAGEMENT PLAN: Collaborative Research: Biogeochemical and Physical Conditioning of Subantarctic Mode water in the Southern Ocean

Data Management Plan: W. Balch, BLOS ; **N. Bates**, BIOS; **M. Long**, OS, CGD/NCAR; **D. McGillicuddy**, WHOI; **P. Morton**, Florida State University

Our data management plan is based on guidelines established by the National Science Board and the NSF and covers dissemination and sharing of materials and data that are expected to be collected as part of the research described in the above named proposal. This project entails both the collection of observational data and generation of model output, the data management plan is divided into two sections.

OBSERVATIONAL DATA

(1) Types of data, samples, physical collections. We will collect data associated with two cruises to the Southern Ocean (Durban, S. Africa to Crozet Island region to Durban, S. Africa in January 2018 and 2019). Field data will be collected at a total of 152 stations per cruise. Water sampling will be done at 76 stations. Protocols will include CTD profiles typically to 500m+ with about one cast per day to 1000 m. Depth-profile data will be binned to 1m intervals). Half the stations will involve no water collection (i.e., Niskin bottles not tripped) and the other half of the stations will include water sampling using a 12-place rosette with 30 L Niskin bottles for the seawater carbonate system (TA, DIC [Bates]), chlorophyll *a*, particulate organic carbon (POC), particulate inorganic carbon (PIC), coccolithophore microscopy/enumeration, FlowCAM imaging cytometry, SEM, and ^{14}C photosynthesis/ calcification (only taken at daily pre-dawn water cast) [Balch]. Water samples will also be drawn for nutrients (nitrate, nitrite, phosphate, silicate and ammonium), dissolved O_2 and salts. Underway DIC/TA samples and autonomous pCO_2/pH data will be collected [Bates]. Ship's ADCP and Video Plankton Recorder (microplankton, optical and hydrographic data) will be collected [McGillicuddy]. Dissolved and particulate trace metals will be sampled using Niskin-X trace-metal-clean sample bottles [Morton].

In targeted eddies, samples from 20m depth will be taken from the eddy center with a 30L GO-FLO bottle and water immediately sampled for PIC, POC, nutrients, dFe, carbonate chemistry, BSi, chlorophyll, coccolithophore counts and FlowCAM abundance of phytoplankton functional groups [e.g. pennate and centric diatoms, dinoflagellates, nanoeukaryotes along with other rarer groups), single cell F_v/F_m and metagenomics analysis. Deckboard incubators will be maintained at the ambient temperature with recirculating surface seawater supplemented by a heater/chiller, as needed. Incubators will be exposed to 50% incident irradiance. Following time-zero sampling, the remaining water will be divided into triplicate polycarbonate bottles followed by 6 treatments: no addition (control), +nitrate (20μM), +silicate (20μM), +iron (+2nm), a combined iron+silicate to examine nutrient co-limitation and 0.2μm-filtered sub-euphotic SAMW to examine the whole community response. The grow-out experiments will be incubated at ambient water temperature. After 4 d, the bottles will be sampled for nutrients, dissolved and particulate trace metals, POC, PIC, Chl, BSi, microscopy, and FlowCAM, TA and DIC, single-cell F_v/F_m and metagenomics analysis.

(2) Standards to be used for data and metadata formatting and content. Tabular data and metadata from CTD casts will be in Sea Bird formatted files. Other variables sampled from water casts will be kept primarily as tabular data in Excel spreadsheets; SEM and light microscope images will be stored as TIF files. The data will be made available online via the BCO-DMO data system (http://bco-dmo.org/data/) and BCO-DMO will submit all data to the NODC for long-term archive.

(3) Mechanisms for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. For sharing files between the Balch, Morton, McGillicuddy and Bates laboratories we will use password-protected Dropbox software. Distribution of data to BCO-DMO will be done within two years of their collection via password-secure FTP. All data generated as a result of research performed at the respective laboratories, or with each respective laboratory's funds, shall be the intellectual property of the respective laboratory, as well as the investigator or co-investigator overseeing the research. As such, each laboratory will retain an implied copyright for these data.

(4) Policies and provisions for re-use, re-distribution, and the production of derivatives. Data from this work will be submitted to BCO-DMO where they will be publically available two years after collection. Peer-reviewed papers will be written about these data, along with details of their collection and data interpretation.

(5) Plans for archiving data, samples, and other research products, and for preservation of access - Each investigator is responsible for maintaining data associated with their own research group's activities, and in accordance with any institutional requirements of the PI's home institution. Basic practices for key areas are outlined. **Lab notebooks.** All information connected with initial data collection, analysis, and results shall be kept in a lab notebook and in cases digital notebooks may be used. Regardless of media, these notebooks shall be stored securely. Hard-copy notebooks shall be situated to enable institutional retrieval. Digital notebooks will follow short-term and long-term policies of each laboratory outlined below. All research notebooks of each investigator are property of their home institution. **Short-term storage and data management. Data volumes.** Data storage shall be appropriately and redundantly stored using computer hardware and software that is available to the each laboratory, and may include in-house and/or off-site resources. The PI and Co-Is are responsible for due diligence with respect to short-term storage of data. Additionally, all data shall be retrievable from primary media or back-ups, as well as reasonably protected from accidental loss due to corruption, power loss, or failure of computer hardware. **Data security.** Data shall be stored on either off-network mobile devices (external hard drives) or off-site cloud resources. Password protection will be utilized. This data must be made available to senior institution officials in the case that any institutional liability issues should arise. **Data backup.** Stored data shall be backed-up weekly. Balch will use the Bigelow Laboratory high-performance elastic computing system backed by 200TB of high performance storage. This system facilitates the Laboratory's long-term data management strategy as well as short-term hourly backups. The McGillicuddy and Bates laboratories use external drive backups. **Deposit and long-term preservation. Long-term strategy.** 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 each home institution from short-term storage to a long-term archival system. **Length of archival.** Data will be kept in long-term storage for at least five years, or until it has been successfully uploaded to and made publically available through a nationally or internationally funded database specific to that data.

MODELING DATA

Data Types. The modeling data created by this project will be output from the Community Earth System Model (CESM). As such, we will abide by the CESM Data Management Plan, available here: http://www.cesm.ucar.edu/management/docs/data.mgt.plan.2011.pdf. Output will be produced from (1) Development Simulations and (2) Production Simulations. Development simulations entail runs conducted to evaluate model behavior during the development cycle, including for the purposes of tuning and testing performance. Data from these runs does not typically have much use beyond completion of developments and will thus not be archived or made publically available. Data from Production Simulations is what will be used to conduct the proposed research; these data will be made publicly available and archived. **Data Format**: The model output data will be in netCDF format, and the file size will range from 1 GB to 150 GB. The data files will be named and structured using the following convention: CaseName.ComponentName.OutputType.Date.nc. See here for additional information: http://www.cesm.ucar.edu/models/cesm1.0/filename_conventions_cesm.html. **Metadata**: File level metadata for each data file will be recorded automatically through the header section of the netCDF file, which is also compliant to the Climate and Forecast (CF) metadata convention. **Access to Data and Data Sharing Practices and Policies**: During the project lifetime, all data generated during the project will be stored and available for project use and for other interested parties on a request basis that recognizes proprietary access to the project members. The final results of the project will be moved into the public domain and be made publicly available through the Earth System Grid (https://www.earthsystemgrid.org). While public access is free of charge, a registration process with the Earth System Grid will be required. This is to help in tracking the data's distribution and usage. **Policies for Re-Use, Re-Distribution**: Those who are interested in using the project's final data can obtain access via Earth System Grid as described in the above section. Users are expected to cite the project and the corresponding data according to the citations that will be established by the project team.