A Component of the U.S. Global Change Research Program

U.S. GLOBEC Data Policy

U.S. Global Ocean Ecosystems Dynamics

Report Number 10

February 1994

U.S. GLOBEC

Global Ocean Ecosystems Dynamics

A Component of the U.S. Global Change Research Program

U.S. GLOBEC Data Policy

Report Number 10

February 1994

This document describes the U.S. GLOBEC data policy. It developed from meetings and discussions among the Data Management Subcommittee of the U.S. GLOBEC program and from input from funded U.S. GLOBEC scientists. Chair of the subcommittee is Leonard Walstad. Other members of the committee are Ronald Fauquet, Glenn Flierl, Don Olson, Peter Ortner, Sharon Smith and Peter Wiebe.

Produced by

U.S. GLOBEC
Scientific Steering Committee Coordinating Office
Division of Environmental Studies
University of California
Davis, CA 95616-8576
Phone: 916-752-4163
FAX: 916-752-3350
Email: T.POWELL (Omnet)
hpbatchelder@ucdavis.edu (Internet)

Additional copies of this report may be obtained from the above address

Table of Contents

U.S. GLOBEC Data Policy

Philosophy and Motivation

The fundamental objectives of U.S. GLOBEC are dependent upon the cooperation of scientists from several disciplines. Physicists, biologists, and chemists must make use of data collected during U.S. GLOBEC field programs to further our understanding of the interplay of physics, biology, and chemistry. Our objectives require quantitative analysis of interdisciplinary data sets and therefore data must be exchanged between researchers. To extract the full scientific value, data must be made available to the scientific community on a timely basis.

Precedent and perception have resulted in a disparity of data collection, storage, and archival methods. This makes the exchange of data difficult and may suppress dissemination of data. The U.S. GLOBEC Scientific Steering Committee seeks to enhance the value of data collected within the U.S. GLOBEC program by providing a set of guidelines for the collection, storage, and archival of these data sets.

The policy detailed below applies to all U.S. GLOBEC investigators. Field data, retrospective data sets, and numerical experiments must all be included in the U.S. GLOBEC database.

National Requirements

As a component of the U.S. Global Change Research Program (USGCRP), U.S. GLOBEC must subscribe to the data management requirements of the U.S. Global Change program. These requirements are provided at the end of this document for reference. The document available from the U.S. Global Change Office also includes an annex, which should be consulted**.**

The second and seventh USGCRP policy statements address the need for exchange of data between researchers. A period of exclusive use is permitted, though the data should be made available when they become widely useful. The annex of the USGCRP expands upon this policy with the statement

> " In the past, some Principal Investigators have retained data for indefinite periods, and this has inhibited their widespread use. This practice should be eliminated through active consideration of the tradeoffs between widespread distribution of data sets and the need to assure data quality and validity. The guiding principle is

> that as soon as data might be useful to other researchers the data should be released, along with documentation which can be used by the other researchers to judge data quality and potential usefulness."

This clearly limits periods of restricted access to the time during which data is not generally useful. There is no provision for granting a period of exclusive use to provide the Principal Investigator with an opportunity to delay exchange until papers describing the data have been published.

Statements 4, 5, and 6 address the need to provide easy access. Statement 3 identifies the need to designate an archive for all relevant data. The policy must prevent the loss of important data sets. The annex notes that many data sets, especially biological, have no archive.

Objectives of the U.S. GLOBEC Data Policy

The U.S. GLOBEC data policy consists of twelve concise statements addressing the collection, sharing, and archival of data within U.S. GLOBEC programs. Preceding these statements is text which seeks to provide some details and the motivation behind the specific policy statements. In setting forth these statements and establishing a data management office, the Steering Committee intends to increase the value of data collected in support of our mutual scientific objectives. The Steering Committee will not attempt to force an investigator to comply with these policy statements, but does wish to encourage and organize full and accurate communication. Plans for data collection must be communicated prior to execution of a field experiment to insure that all necessary data are collected. Data collected during field programs or in laboratory experiments, organized for retrospective studies, or produced from a model must be shared with the scientific community to maximize the scientific value of the data.

This document primarily addresses data collected during ship based field experiments. Extrapolation to other types of studies is expected. The policy statements anticipate that data will be organized about cruises; when measurements are not made aboard ship, the data should be organized about a period of time during which a sensible unit of data is collected. For instance, telemetered time-series data might be organized by month and near-shore benthic data might be naturally organized on a seasonal schedule. Investigators tend to organize their data into periods which are natural for their purposes. If a particular time period is appropriate for your study, then use this period for organizing and submitting data to the U.S. GLOBEC Data Management Office (DMO) while keeping in mind the need for timely submission. Investigators conducting retrospective studies should also recognize that the data organized for their purposes should in

general be submitted to the DMO. Data culled from generally unavailable or difficult to access archives are of great value to the community. Model results which would be useful to the interpretation of field data or comparison with later model studies should also be included. Potential candidates for submission include annual or seasonal fields of flow, temperature, and salinity, Reynolds stresses, particle trajectories, initial/boundary conditions, and surface fluxes.

Policy Statements

Quality and Methodology

Data quality is of fundamental importance, yet the demands of individual disciplines vary. Recognizing this variation and the cost and effort that would be required if standards were imposed uniformly, a set of data quality requirements has been defined. Rather than specify requirements for data collection, the U.S. GLOBEC Steering Committee has chosen to assign responsibility for selection of methods, equipment, and calibration procedures to the principal investigators funded to make measurements as a component of U.S. GLOBEC.

All principal investigators are required to submit plans for the collection of data prior to execution of their sampling program. In general, these plans are expected to be similar to the information provided in proposals submitted prior to funding. The purpose of this requirement is to provide a common resource for the participating scientists to evaluate the suitability of the expected data set for achieving their scientific objectives. A single description of the expected data sets, a "data plan", will be derived by the Data Management Office from the submitted plans of individual investigators or groups of cooperating investigators. Where a group of investigators is cooperating in managing and collecting data, a single responsible scientist should be identified for each measurement type. The Steering Committee may also review the data plan to evaluate the applicability of the data set to U.S. GLOBEC goals beyond the specific experiment.

To provide the opportunity for comparison with historical data, measurement techniques should be consistent with techniques used to collect the existing data unless there is significant scientific justification for change. When new techniques are adopted, methods for relating the new data to existing data should be developed. This requirement extends to regional comparisons as well. For example, measurements made in eastern boundary currents should be designed in consideration of the existing large database for the California Current.

1. Investigators must select methods and equipment which are adequate to insure that data quality is sufficient for the objectives of the U.S. Global Change Research Program and U.S. GLOBEC. At least three months prior to execution of a sampling program, the principal investigators will document the procedures that will be used to collect and process samples and data. This documentation will be submitted to the U.S. GLOBEC Data Management Office which will derive a single data plan for each cruise or field season. The derived plan will be included in the U.S. GLOBEC database for access by participating scientists. Of particular interest are the following considerations, and each must be specifically addressed by the principal investigator in describing collection and analysis methodology:

   i. Measurements to be made and the anticipated precision and accuracy of each measurement.

   ii. A description of the sampling equipment sufficient to permit an assessment of the anticipated raw-data quality. Typical descriptions will include where appropriate: navigation, timekeeping, sensor make and model, net opening and mesh size, rate of retrieval, mooring configuration, and similar information appropriate to the types of samples to be collected. Note that where the data collection equipment is well known or documented in generally available technical reports or the published literature, the need for documentation will be substantially reduced and may be satisfied by identifying the system or referring to the appropriate documentation.

   iii. A description of the analysis methodology sufficient to permit an assessment of the anticipated analyzed-data quality. Typical descriptions will include where appropriate: filter size and type, sample preservation technique, counting method, numerical algorithm, incubation procedure and similar details as appropriate to the measurements planned.

   iv. A discussion of the means by which the measurements to be taken could be compared with historical observations or with regions which are thought to have similar ecosystems. When the sampling method is critical to the interpretation and utilization of a data type, a description of sampling methods used in the region or in similar regions during past experiments must be included. Where the planned sampling method differs from the previously used measurement technique, the principal investigator must either demonstrate that a quantitative comparison will be valid or provide justification for the change in technique. The Steering Committee supports collection of calibration data

where the requirement for comparison with historical data is in conflict with the modern scientific objectives.

v. Ancillary measurements needed to achieve the investigator's scientific objectives.

2. Documentation of the measurement and analysis techniques used to produce the data set must be submitted with the data to the U.S. GLOBEC Data Management Office. This documentation will be similar in form to the documentation submitted prior to sampling.

3. The investigator is responsible for estimating the accuracy and precision of each measurement and recording this information in the database.

4. The overall objectives of the USGCRP and U.S. GLOBEC demand knowledge of the physical setting of the ecosystem. To this end, physical data must be acquired with biological measurements. In general, the following measurements must be made and included in any biological data set:

   i. location and time (to within 200 m in the horizontal, 2 db in vertical, and 1 minute in time)

   ii. temperature and salinity (to within .02°C, and .02 ppt.)

Of course, remote sensing (e.g., acoustic sampling) makes it impossible to determine the physical environment at the location of the measurement. Where possible, temperature and salinity sensors should be combined with biological sensors or profiles should be taken between tows.

5. The investigator is responsible for insuring that the quality of the data available to the community is of as high a standard as possible. Specifically, corrections or improvements made subsequent to submission of the data to the U.S. GLOBEC DMO must be submitted to the DMO. The DMO will endeavor to inform users of the data of any corrections or improvements.

Data Exchange and Archival - Methods and Schedule

A data system must facilitate the exchange of data and insure the long-term existence of the data set. National requirements for submission of data to the National Oceanographic Data Center

(NODC) must be satisfied either by the investigator or by a data management office. Because some of the data types will be new and because there is no existing data center which will permit both submission and retrieval of interdisciplinary data sets, a data management office is needed to facilitate exchange and cooperate with NODC on establishing a national capacity for the exchange of interdisciplinary data sets. While this office will accept responsibility for submitting data to NODC, the primary objective of this office is to provide a mechanism for the exchange of interdisciplinary data sets.

The reader is reminded that it is not ethical to publish data without proper attribution or co-authorship. **Beyond this, the U.S. GLOBEC Scientific Steering Committee believes that the intellectual investment and time committed to the collection of a data set entitles the investigator to the fundamental benefits of the data set. Therefore, publication of descriptive or interpretive results derived immediately and directly from the data is the privilege and responsibility of the investigators who collect the data. The purpose of a data archive is to facilitate collaboration between scientists, the combination of multiple data sets for interdisciplinary and comparative studies, and the development and testing of new theories. Any scientist making substantial use of a data set should communicate with the investigators who acquired the data prior to publication and anticipate that the data collectors will be co-authors of published results.** This extends to model results and to data organized for retrospective studies. As possible, the U.S. GLOBEC Data Management Office will encourage and facilitate the ethical and courteous use of data within the archive. In particular, the U.S. GLOBEC DMO will maintain a list of all data access and will notify those who access the data of our commitment to the principle that data is the intellectual property of the collecting scientists.

Data collected for U.S. GLOBEC field programs will be diverse and there is a substantial emphasis on the application of emerging technology. Therefore, the schedule for submission of data products must differentiate between types of data and provide a mechanism for flexibility where application of the data submission requirements is impractical. While these requirements must be followed, the spirit of the USGCRP Data Policy is that the data be made available whenever it is of general use. In some cases, this may require multiple submissions of the data set. This will be necessary when a portion of the data is not available promptly or if calibrations need to be changed after the original submission of the data.

Data sets consist of both the actual measurements and also descriptive data, sometimes referred to as metadata. Metadata consists of location, time, units, accuracy, precision, method of measurement or sampling, investigator, reference to publications describing the data set, a description of the processing of the data, etc. This information is often crucial for correct interpretation of the measurements. Therefore, U.S. GLOBEC databases must include all relevant metadata in a form which can be used efficiently by analyzers of the data. As the primary user of the data, the principal investigator is uniquely qualified to determine the relevant information needed to make use of the data.

U.S. GLOBEC field programs will frequently involve the coordination of several investigators making independent measurements in a cooperative sampling plan. Some information will be common to all investigators; time and location are needed for each measurement. Users of the data will need to know the full suite of measurements and the sequence in which the measurements were taken. Also, consistency of the data set is of paramount importance; measurements taken at the same time and location should have identical time and spatial coordinates recorded in the database. Of particular concern is the use of time to determine location from the navigation log. Careful maintenance of consistent timekeeping is critical and investigators are required to document the procedures which will be used to insure that temporal and spatial errors are controlled. The U.S. GLOBEC Steering Committee strongly recommends the use of a logging system which will record the underway data (navigation, and where available meteorology, near-surface temperature and salinity, and any other data collected automatically). These data should be integrated with data records made by other sampling instrumentation. This will greatly simplify the task of inventorying the data set and insuring the most accurate navigation possible. Whether or not an electronic logging system is used, responsibility for maintaining and reporting a log of all measurements lies with the chief scientist of the experiment.

6. Within three (3) months after collection, a detailed inventory of measurements made during the cruise or field season must be submitted to the U.S. GLOBEC DMO by the chief scientist of the experiment in cooperation with the participating principal investigators. This inventory will include the time and location of each measurement and a schedule for submission of full or partial data sets. Of special concern is the inventory of biological samples; all information necessary to retrieve a specific sample must be recorded in the database. Also, any anticipated problems with the data should be reported at this time.

7. Measurements which do not involve manual analysis and which would be useful to the science community must be submitted by the principal investigator within six (6) months

after collection. Metadata should include any procedures that were followed to correct errors, remove noise, or otherwise modify the collected data.

Plankton samples inherently present special problems with respect to data policy. The data submission in the case of readily producible statistics, such as displacement volume, and easily producible data, such as silhouette photographs, may be available within the time frame above. The longer time frame associated with sorting plankton requires a more flexible policy which is tied to the completion of a significant portion of the sample suite from a cruise. That is, when an investigator completes the analysis of a set of samples to the degree they form a useful measure of conditions observed during a cruise, the data should be submitted. A data set becomes useful to the community at the same time that the investigator begins to use the data for ocean science. Investigators must plan resources and technician time to accomplish these primary data reduction tasks within one year from the end of the cruise during which the samples were collected.

8. All other measurements and any standard analyses of these measurements must be available to the community within one year after collection. Standard analyses include the displacement volume, species counts, and silhouette photographs of net tows, displacement volume and grain size distribution of sediment trap samples, and any other similarly producible derived data. This is not a requirement that these standard analyses be conducted. Principal investigators are responsible for selection of the types of analyses appropriate for the scientific objectives of the experiment. We expect that these analyses will be specified in the proposal and in the planning document described in policy statement 1. Any analysis similar to those listed above and produced from U.S. GLOBEC samples by the investigator or by any other scientist must be submitted to the U.S. GLOBEC DMO. Metadata must include any procedures that were followed to analyze the samples, correct errors, remove noise, or otherwise modify the collected data.

The primary responsibilities of the DMO will be to accept data from U.S. GLOBEC investigators, to verify the data has been properly transmitted, to report on the status of data submissions to the Program Manager and the Steering Committee, and most importantly to facilitate the interdisciplinary exchange of data. Also, the DMO will provide standards for the creation of the database particularly concerning the types of operations supported by data objects. These standards will be designed to conform with the data policy and to insure that the structure and appearance of the database is relatively consistent between separate contributions.

We believe that a useful database must support extension to new data types and be distributed. The wide range of data types expected within U.S. GLOBEC and the emphasis on technology application which will lead to new data types suggests that current database technology is insufficient. Object oriented methodology, which is currently emerging in programming languages and database implementations, appears to satisfy our need for multiple and easily extensible data types. Distributed databases have the advantage that the data collector is directly involved with creation of the database. If the database system is well designed, then the data collector may use the same system to access the data that is being used by the research community. This would be a significant improvement over the present situation.

To satisfy our objectives for a database which is distributed and which can handle arbitrary data types, U.S. GLOBEC is cooperating with JGOFS and the community efforts under the auspices of The Oceanography Society. These are evolving systems and important issues have not been fully resolved, however, the initial U.S. GLOBEC data system will be the JGOFS system. Important issues include cost and accessibility which will be assessed during the first year of the DMO. Each principal investigator and chief scientist should consider using the JGOFS system. Transferring the data to the DMO will be greatly simplified since all that is necessary with a distributed data system is the name and location of the database. The DMO will take responsibility for obtaining the data when investigators use the JGOFS type database.

Archival will be accomplished on two levels. The DMO will serve as the initial archive and for the length of U.S. GLOBEC, data will be available on-line from the DMO. In addition, the DMO will be cooperating with NODC to insure that the data is transferred to a permanent archive. NODC is committed to providing an accessible archive for all ocean data. When measurements are taken in foreign waters, the DMO will be responsible for communicating data reports to the State Department as required.

9. Investigators will either submit data to the Data Management Office or place it on-line as a U.S. GLOBEC distributed database. Standards for submission formats and development of the database will be specified by the DMO in support of the objectives of the data policy. The DMO will verify that the data is properly represented in the database and report on the status of data submission to the Program Manager and the Steering Committee at each Steering Committee Meeting.

10. The DMO will serve as an intermediate archival location and data source, will transfer data to the NODC, and will prepare the necessary documentation for data collected in foreign

waters. The DMO will communicate the data policy to all producers and users of U.S. GLOBEC data. In particular, the rights of the data collectors, organizers, and producers of the data will be communicated to those who access the database.

Sample Preservation

Investigators are responsible for maintaining biological samples for at least twenty (20) years. A representative subset of the sample must be preserved in reagent grade alcohol for later genetic analysis. Preservation techniques are to follow currently accepted practice for the particular type of sample. For example, preservation of macro-zooplankton in a buffered formaldehyde solution at a temperature between 10°C and 25°C is generally adequate[1]. In addition, investigators should anticipate that other scientists will need physical access to these samples and may need to sub-sample the original samples. In general, requests for access to the samples should be approved if the objectives of the study are compatible with those of the USGCRP and U.S. GLOBEC. Investigators are responsible for the maintenance of the archive and therefore may reject requests which would damage the samples. Disputes will be settled by the parties submitting letters stating their positions to the Program Manager and to the U.S. GLOBEC Steering Committee. As the Smithsonian is the official repository of biological samples in the United States, samples from U.S. GLOBEC funded field programs should be offered to the Smithsonian before disposal.

11. Biological samples will be preserved following currently accepted practice for the particular contents. Sub-samples of a representative subset of the samples must be preserved in reagent grade alcohol for later genetic analysis. These samples will be retained for a period of 20 years and shared with the community as requested. Institutional representatives should be made aware that these samples must be stored for this extended period at a controlled temperature. Prior to disposal, the samples must be offered to the Smithsonian.

Modification of Policy

Recognizing that ours is an evolving field, there may be a need for modification of policy in the future. Specification of an archive will be done in the near future. Other issues may arise. The Steering Committee reserves the right to change the Data Policy. Any changes will be made with respect for the resource needs of investigators with regards to the processing and distribution of

1 Zooplankton Fixation and Preservation, H. F. Steedman, Editor, UNESCO Press, 1976, p. 145.

information. When changes in data policy would require substantial increases in equipment, supplies, or personnel, current investigations will not be expected to comply with the changes.

Some investigators may wish to be exempted from all or part of the data policy requirements. The only reason for exemption is a lack of general usefulness of the data collected and there may be data sets which are not of general usefulness within the time allotted. In these cases, the investigators should submit a request for exemption to the Program Manager and Steering Committee for review and a decision.

12. Requests for exemption from the data policy should be submitted to the Program Manager and the U.S. GLOBEC Steering Committee.

Data Management for Global Change Research Policy Statements[2]

The overall purpose of these policy statements is to facilitate full and open access to quality data for global change research. They were prepared in consonance with the goal of the U.S. Global Change Research Program and represent the U.S. Government's position on the access to global change research data.

1. The Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets.

2. Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.

3. Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearing-house process should be established to prevent the purging and loss of important data sets.

4. Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.

5. National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.

6. Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.

7. For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case, the funding agency should explicitly define the duration of any exclusive use period.

2 These are the US Global Change Research Program data management policy statements. A booklet detailing the policy is available from The U.S. Global Change Research Program, National Science Foundation, Washington, DC 20550, (202) 357-9715.