

# ESSENTIAL ELEMENTS OF THE DATA AND INFORMATION MANAGEMENT SYSTEM (DIMS) FOR THE OCEAN OBSERVING SYSTEM FOR CLIMATE

**David M. Legler**

*Florida State University, Tallahassee, FL, USA (legler@coaps.fsu.edu)*

**Paul Freitag**

*NOAA/PMEL, Seattle WA, USA*

**Penny Holliday**

*WOCE/CLIVAR IPO, Southampton Oceanography Centre, Southampton, UK*

**Bob Keeley**

*Marine Environmental Data Service, Ontario, CANADA*

**Syd Levitus**

*NOAA, NODC/WDC-A, Silver Spring, MD, USA*

**Ron Wilson**

*Data Information Unit, University of Delaware, Lewes, DE, USA*

**ABSTRACT** - Existing data management philosophies, structure, and approaches have evolved in conjunction with the ocean observation system. During the past two decades management of ocean data and data products has focused on developing research-driven data systems in support of specific programs. Some of these systems, such as TAO system, have now become operational and are well suited for ocean observing systems for climate. Others such as WOCE have demonstrated that distributed managed data systems can produce unprecedented amounts of quality-controlled data as well as value-added products. However, many components of the existing ocean data system are insufficiently mature or complete to satisfy the very different and demanding needs of an observing system for climate. An ocean observing system for climate necessitates the development of a carefully managed data and data product system that considers many diverse in-situ streams and remotely sensed data as well as numerical model and synthesis-based products. All of these sources have common essential elements: data (product) and metadata assembly/collection, quality-review, and data distribution/access. Currently these elements for each source are in various stages of development towards contributing to the ocean observing system for climate. This paper describes these essential elements of the ocean data management system in terms of their historical context, their potential contribution to the ocean observing system for climate, as well as the advancements and resources necessary for them to meet the requirements for the future ocean observing system. Additionally, archeology efforts and overall data management oversight and coordination are discussed as additional efforts required to transition the current transient data management efforts to a coordinated permanent system capable of providing high quality data and value-added data products.

## 1. MOTIVATION FOR DATA AND INFORMATION MANAGEMENT SYSTEM (DIMS)

The present scenario of ocean data and information management is based primarily on transient structures for finite period research projects (e.g. TOGA, TOGA-COARE, WOCE) and

existing activities of long-standing intergovernmental organizations such as GLOSS, IGOSS and IODE. The developmental nature of these systems has resulted in a data system not optimally designed for climate or (with the exception of ENSO observing systems) for real-time reporting. These research projects utilized operational organizations and developed extended systems to manage the wider range of data and products required for their individual objectives. This has resulted in limited interaction between these systems and real-time observing programs and activities that consequently restricted universal and open access to all available data. A future ocean DIMS for climate must oversee and coordinate a broad range of activities establishing real-time observation reporting, productive quality-review mechanisms, and facilitating access to data sources for all users to ensure maximum return from the investment in ocean observations.

### **1.1 Historical Perspective**

When considering the way forward in ocean data and information we draw on the experiences of existing systems. This section provides an overview of philosophy and implementation of present systems, and Section 1.3 highlights their successes and weaknesses.

The Joint IOC-WMO IGOSS is the world-wide system for the collection, rapid exchange and analysis of oceanographic (and some meteorological) data, and the timely preparation and dissemination of ocean products and services to various marine users. Research ships, fishing vessels, merchant ships, fixed platforms and drifting buoys, collect surface and subsurface temperature, salinity and current data for this system.

The IOC IODE system is a collection of national data centers which archive and distribute data and information on many ocean variables from the disciplines of physical, chemical, geological and biological oceanography. It concerns itself almost exclusively with scientific data that often arrive years after data collection.

Other sources of in-situ data include those from the TAO system that has proven to be invaluable for monitoring ENSO variability and climate forecasting. This system is now operational and expansions in other oceans (e.g. PIRATA) are underway.

The most recent long-term experiment, WOCE, implemented a data management system based on a series of distributed DACs directed by scientists with expertise and familiarity in their respective data stream of responsibility. The philosophy was for each DAC to provide quality-controlled data in order to achieve the highest standard of data possible and the maximum added-value through generation of products. Central to the WOCE data system is the DIU which tracks the status of the observations, data and products. Ultimately, and in conjunction with established national data centers, the WOCE data system will unite data from the different DACs to form the WOCE Data Resource.

In the US, a network of DAAC's are sources of research and operational remote sensor data. Other functionally equivalent centers are established internationally; for example the AVISO project in France. These centers are sponsored by national space agencies such as NASA and ESA to provide data in real-time and delayed mode, and generally have developed a wide variety of products such as improved retrieval algorithms and value-added data sets.

In the atmospheric community, PCMDI has provided information and data about numerous atmospheric general circulation models (AMIP) and coupled atmosphere-ocean models (CMIP). There are no centers formally acting as information and/or data distribution centers for ocean modeling/syntheses products (there is at least one in the planning stages and will be discussed later). The volume of data/information to assemble and serve makes the creation of such a server a non-trivial and expensive task.

### **1.2 Existing DIMS**

WOCE was a major stimulus to data and information management. The investment in the WOCE DIMS has returned many benefits. It was responsible for changing the outlook of physical oceanographers from individuals or laboratories who analyze, then privately file their data, to a community which thinks globally and delivers data to the public domain and the global archive. But there are other programs spurred by WOCE (e.g. JGOFS) that have changed data management

practices for other ocean disciplines. This progress has been especially remarkable in several areas: better data assembly; improved quality-control methodologies; faster data delivery; easier and broader data distribution; better integration of data from a variety of disciplines; and the development of an effective DIMS.

WOCE will have collected about 9000 deep (one-time) hydrographic stations, three times the number available previously (Chapman, 1998). This increase is also mirrored by other observational systems. The data assembly process was very effective. By the end of WOCE (2002) as much as 90% of in-situ data collected for WOCE will be available to users, arguably one of the greatest achievements of the program.

Quality-review methodologies were also developed for newly managed streams (e.g. Smith *et. al.*, 1996) and improved for other streams. For example, the GTSP in association with WOCE has developed quality control procedures for operational and delayed mode XBT data (Bailey *et. al.*, 1994; Daneshzadeh *et. al.*, 1994, IOC, 1990). IGOSS also identified errors in the fall-rate equation of XBTs and developed correct equations.

Faster delivery of some data streams was achieved in conjunction with IGOSS and GLOSS for sea level data for satellite validation and for drifter and upper ocean thermal data for use in numerical models. Other streams, e.g. floats, are now starting to report in real-time.

Taking advantage of modern electronic communication technology, dissemination of data has improved dramatically. In-situ and remotely-sensed data from WOCE are available on CD-ROM's. Additionally, as a response to user needs, IGOSS created an electronic product bulletin board on the Internet.

The JGOFS data management system has made substantial progress towards managing biological, physical, chemical, and geological data in one comprehensive system. Much has been learned from their work that can form the basis for future improvements.

WOCE DIMS activities are coordinated by a Data Products Committee which meets regularly to review program and users' requirements, review the progress of all DACs in processing their data and product generations, and take action where necessary to insure WOCE data flowed to users. This committee (and similar committees formed in other programs) was a critical ingredient to the success of the WOCE DIMS.

A significant weakness in the current ocean observation system is the delay between observation and availability of the data to the community of users. Some parts (e.g. sea level) of the current DIMS can report data in real time (e.g. satellite altimetry, SST, and winds). However, for all streams there are significant delays in providing access to the high quality, high-resolution data. The causes of delays are manifold. Some data are not sent by existing operational systems because of technical deficiencies and some because of insufficient commitment. Data are sent after analysis on shore, either to ensure calibration or simple chemical analyses. Quality control procedures take time and cause delays. Proprietary-use periods for data (ubiquitous in some areas of the ocean research community) means data are not available until long after collection. Some data are never processed due to lack of funds, or lost instrumentation. Data are lost or delayed by resource constraints throughout the system.

Management for some data streams is relatively non-existent. Mechanisms exist within IGOSS for the transmission of salinity data both from the ocean surface and as profiles. The amount of such data is dismal. Last year (1998) approximately 6000 salinity profiles from the entire world were exchanged by IGOSS. There were only about 30,000 surface salinity observations available. Despite this being an important activity, lack of resources have resulted in salinity having the lowest percentage of data submitted in the WOCE data resource.

Data integration remains an elusive goal. Ideally all of the data collected from a particular spot of the ocean, whether from remote, or in-situ instruments or from models should be comparable. We should be able to assemble all data from meteorological, biological, physical, etc. disciplines and compare and contrast them. The WOCE CDs are approaching this in a phased way, first by assembling the data, then encouraging a common data structure for all DACs.

### 1.3 Motivation for DIMS

Building an ocean observing system for climate requires a corresponding DIMS. The foundation of this management structure has been established, but existing systems will need to be modified to suit future requirements. Data systems must foster and further develop the perspective of the community of data providers (including those associated with a limited scope such as process experiments) and users to ensure all data reach the public domain rapidly (eliminating the requirement for future data archaeology), and in a form necessary to address climate issues. A sustained data product system must monitor and coordinate a wide range of activities within many diverse in-situ and remotely sensed data streams as well as numerical model and synthesis-based products. **These diverse streams all have common essential elements: data (product) and metadata assembly/collection, quality-review, and data distribution/access.** These elements along with an overall coordination and oversight body comprise the DIMS. Currently these elements for each source are in various stages of development towards contributing to the ocean observing system for climate. Here we discuss their potential contribution to the ocean observing system for climate, as well as the advancements and resources necessary for them to meet the requirements for the future ocean observing system. Overall data management oversight and coordination are central to all efforts transitioning the current transient data management efforts to a coordinated permanent system capable of providing high quality data and value-added data products.

## 2. DATA AND METADATA ASSEMBLY

### 2.1 Models of ocean data collection, transmission and assembly

The first essential element of an effective data system is the secure and managed assembly of observational data. This assembly element is comprised of the steps necessary to get the observational data to a responsible organization or redistribution system. In the past, this process often involved seeking out data collectors, making personal contacts, building trust to secure data submission and release, and even harassment when necessary. Future systems may need this type of activity if they are going to access some types (e.g. hydrography) of data in large quantities. However, the delays in assembly of data through this mode of operation are unacceptable. We focus our discussion on the more likely prospect that most of the ocean data will be more openly available in near-real-time. The communication process that forms the backbone of the assembly process may have a significant impact on the amount of data delivered as well as the timeliness of the delivery. Here we explore the TAO observation system to highlight desirable attributes of the data/metadata assembly element.

The TAO Array (McPhaden, *et al.*, 1998), consisting of nearly 70 ATLAS moored buoys spanning the equatorial Pacific, measures oceanographic and surface meteorological variables to provide climate researchers, weather prediction centers, and scientists around the world with real-time data. The ATLAS moorings measure various surface meteorological variables as well as subsurface temperatures down to a depth of 500 meters (Milburn *et al.*, 1996). In addition to daily mean observations the buoys also transmit the most recent hourly surface meteorological observations.

Data from the array are telemetered in near-real-time via Service Argos utilizing the NOAA polar-orbiting satellites. The TAO Project at PMEL processes the buoy data nightly, where calibration coefficients and quality controls are applied, and the data are made available to the international scientific community. In a parallel-processing path, Service Argos submits ocean temperature and surface wind, relative humidity and air temperature onto the GTS. This data path requires that PMEL share calibration and operational data with Service Argos. PMEL also identifies and informs Service Argos of data that should not be submitted to the GTS.

There are four primary factors that limit the delivery of real-time full-resolution data products: transmission costs; transmission logistics; quality-review; and loss via distribution.

Improvement in each of these areas will need to be addressed in all future DIMS to insure timely delivery of high-quality data.

The cost of transmitting data can be significant in terms of power requirements and/or financial resources required to telemeter the data to a receiving station. In order to reduce battery requirements and Service Argos charges for the TAO array, ATLAS moorings transmit for 8 hours per day. Because the buoys all report at the same local time, this biases the observations to these particular time periods. The complete high temporal resolution data (hourly or 10-minute) are available in delayed mode after recovery of the moorings. Within the next few years technological advancements in telemetry will make it possible for more data to be transmitted and a greater proportion included in operational models. For the Argos system (Ortega and Woodward, 1999), 2<sup>nd</sup> and 3<sup>rd</sup> generation satellites provide: increased sensitivity of satellite receivers and transmitters which transmit only when a satellite is in view, which could lower the transmitter power required on the buoy and allow an increase in the hours of transmission without increased battery size; increased data rates, so that more data could be sent in a shorter time. The GOES satellite system may be a more attractive option for moorings if improvements in power requirements and positioning can be made. Other new telemetry options which offer increased data rates include several low earth-orbit satellite systems (*e.g.*, OrbComm) but the cost to user of these systems is still unknown.

Due to a combination of the ATLAS sampling and transmission schedule, telemetry paths of the NOAA satellite system, and Service Argos processing time, there is a finite time lag between observations and availability of data on the GTS. This is likely to be a factor in all data streams. For telemetry via Service Argos, increased numbers and more uniform global coverage of satellite down-link stations should decrease this time lag.

The TAO data are passed through quality-review procedures each night and posted for open distribution. This automatization is key to realizing real-time data delivery. As the delayed mode TAO data are recovered from the moorings, they are also reviewed and the data archive updated. The TAO data archive remain dynamic, undergoing review as necessary when new problems are identified.

Lastly the means of data distribution limit data availability. For example, not all GTS users receive identical amounts and types of ocean data. Multiple monitoring taps at various points on the GTS will help to insure all possible data are delivered (Molinari, 1999). One possible alternative to this distribution model is to simply retrieve needed data from appropriate data centers (the data at these assembly centers are likely to be the best quality and most complete), thereby avoiding potential data losses through the GTS.

The assembly process must pay critical attention to metadata. Particularly for climate studies, the methodologies, instruments, and other measurement-environment conditions are critical components of the observational record. It becomes increasingly more difficult to gather metadata as time elapses after observations are recorded. Thus it is very helpful to prompt the collection of this information by providing templates of required information to the observers. The data and metadata information must be linked to reduce the chances that they become de-coupled.

## **2.2 Satellite and Model Data Assembly**

Ocean observations from satellites are usually assembled through dedicated data communications systems and delivered to appropriate data centers (*e.g.* DAAC's) which process the measurements into a variety of products and act as data distribution sites. Some real-time satellite measurements are delivered (via sometimes independent routes) to NWP centers and agencies which often do their own processing and produce products independently. The efficiency of these assembly processes is very good and well supported through development of extensive documentation, read-software, and a support system for inquiries and assistance.

Access to information and digital results of ocean models and synthesis efforts has never been fully realized. PCMDI has developed advanced methods and tools for the diagnosis, validation and intercomparison of GCM's and has recently begun comparing global coupled models (CMIP). The DIU documented the availability of model output streams; however, model-to-model and model-to-observation comparisons could be better facilitated through a more comprehensive ocean modelling resource.

### 3 QUALITY-REVIEW

#### 3.1 Why is data quality assessment necessary?

The purpose of data quality assessment (i.e. quality control or quality review) is to determine if measured values are free of errors. It documents the instruments used, reporting errors discovered, and the evaluation process itself. The assembly of the reviewed data becomes an added value product. The quality assessment process also provides information about the precision of the measurements as well as the limits of use of the data. A user, in deciding if the data can be used for a particular purpose, can weigh the data quality. If quality control is carried out well, users of the data can distinguish real signals untainted by errors. Additionally, when it is done reliably, quality control can also reduce duplication of effort.

Effective quality control requires 1) sensible suite of test procedures accepted by users of the data; 2) clear and well-documented descriptions of the test procedures, and 3) results be clearly associated with the data values. If any one of these is missing, the value of the quality control is diminished.

Quality control applies not only to observed measurements but also equally to the independent variables associated with the measurements. It is as important to have reliable positions of a temperature profile, for example, as to have the profile itself.

Quality control is not a one step process. At its most basic, it tries to identify all contamination by instrument error, by measurement technique, by transmission faults and by other mistakes. Errors can be subtle or gross. The gross errors are the easiest to identify and can often be picked up by automated techniques. As errors become subtler, more experience and knowledge are required to detect the errors. Quality control is also of greater value to users if both indicators of the quality of the data (i.e. data are considered good or bad) and reasons for the setting of the indicators are associated with the measurements.

Quality control should not be confined to the scientists who make the measurements. Autonomous instruments will make a significant number of measurements. These measurements will pass through communications systems, through assembly centers (nodes) and on to users with diverse interests. Some data will be needed within hours or days of collection. The scientific scrutiny, that is the process that can identify subtle errors, may not operate within this time constraint. Thus first-order (automated) checks must be developed to remove gross errors and the more careful scrutiny from scientists working in non-operational time frames will add scientific quality control to the data. The results of their work must be brought back into the data collection so that other users can gain from their knowledge and work.

Members of the international data exchange systems all carry out some form of QC on the data they receive. Their efforts are varied, as is the documentation of what they do. Scientists also carry out QC, but with occasional exceptions, their work is targeted for their own purposes and is not used by others. There has been a substantial level of redundancy in the ocean data management system, far more than is appropriate.

WOCE was the first truly large-scale oceanographic programme. The range of variables measured and the global extent required a new cooperation in collecting and managing the data. Data assembly centers were established based on type of data (i.e. data streams). This concentrated the quality control process to one or a few places where expertise existed and which were trusted with the task. The global QC system took a significant step forward and has initiated a convergence of QC techniques at least within data streams.

Within WOCE, the relative costs of data management to data collection have been roughly estimated to average 1% (Lindstrom, pers. comm.). In retrospect, the relative investment in data management was recognized as being too small; however, the data resulting from WOCE are the most comprehensive, highest quality collection ever assembled; Much valuable experience has been acquired in organizing international programs and how cooperation can repay data management investment.

### 3.2 Value of Quality Review

Part of the role of quality control is to identify systematic problems in the data and to notify collectors so steps can be taken to correct them. This is especially critical in a sustained observing system where constant vigilance will be needed to ensure consistent high quality observations. A solid QC system exploits the knowledge of instrument characteristics to scrutinize the data. A good example is the reporting of earth-relative surface winds from automated systems. Newly-introduced automated systems on several research vessels were not recording the proper variables to calculate earth-relative winds and were mistakenly reporting ship-relative (or a number of other variants). Feedback to ship technicians has improved the situation [Smith *et. al.*, 1999].

The scientific value of any quality control program can be measured in a number of ways. Because the data system is responsible for detecting suspect data it reduces duplication of effort. Centers such as the current meter DAC have been able to distribute data sets free from erroneous data points that had gone unnoticed by originators. Scientific analysis of contemporary sea level data led to the discovery of instrumental drift in the TOPEX/POSEIDON altimeter measurements that could have been mistaken for temporal changes in sea level [Mitchum, 1998].

Improving the quality of measurements is achieved through end-to-end data management: preparation, equipment purchase, and training before the measurements are made and evaluation of the data immediately after collection. Examples of successes include the relatively higher accuracy of WOCE hydrographic measurements—more accurate than previous standards. As a result, WOCE deep ocean salinities have revealed subtle temporal changes that were not possible to detect previously [Bryden *et. al.*, 1996]. Important also was the setting of standards, such as done by the Hydrographic Programme, to be achieved [WHP, 1994]. Close scrutiny by data quality experts, for example, provided an unquantifiable incentive for investigators, even the most experienced, to be careful in their collection and reporting processes. Where data scrutiny was absent, such as for WOCE bathymetry or surface salinity, data have not been carefully collected or neglected and no improvements have been made. Additionally, some sensors such as for surface salinity require careful and continual care; no amount of post-cruise management will rectify data from poorly maintained instruments. Overall, the quality control process identifies those measurements that have problems. The improvements are gained by the effective consultation between assembly centers and data collectors.

There are two key lessons here. First, that collection of high quality data is achieved through training, preparation and continuous evaluation in comparisons to existing data. Post collection quality-control can improve data quality only through consultation with collectors. Second, having a data evaluation system in place not only identifies suspect data but provides an incentive to maintain high quality measurements and can help laboratories improve their techniques.

### 3.3 Evaluation: Current and future efforts.

Quality review methodologies and infrastructure are now developed for most data streams (bathymetry and sea surface salinity notably excepted). Scientific quality control comes at more than just a financial cost. It takes time and (for WOCE especially) can result in the delay of availability of data. The scientific QC of WOCE XBT data took longer than anticipated, partly due to delays in data submission, partly due to developments of the system. Future QC efforts must considerably reduce the delay between reception of data and delivery of evaluated data. This will be achieved by the automation of review procedures.

Many operational centers and modelers (including those assimilating ocean data) have the capability to label anomalous data. Some of these will be due to previously undetected errors while others will be genuine extremes or data outside an envelope of acceptable (to the model) values. These results have not been widely distributed to the community, yet have large potential value. The successful incorporation of this information into the standard QC products requires the cooperation of the user and the assembly center (or other entity with responsibility for providing scientifically reviewed data). This cooperation is more than just exchanging results. The centers must understand

their respective roles and develop an plan where they will agree on respective responsibilities they have in the quality control process in order to maximize the use of all quality reviews.

A lesson from previous incarnations of the ocean data management system is that a distributed data system can sustain relatively diverse levels of data management activities and desirable partitioning of tasks and costs of data management.

Quality control systems do not operate individually. They are usually a single node of a larger system. Better coordination necessitates the development of partnerships with data collectors, data centers, scientists and users. Additional cooperation must be sought with data transmission specialists and agencies as well as archival centers for ocean data. Finally, there must be the commitment of each of these groups to the success of the effort. This requires a substantial effort and therefore for each data stream one agency or group should lead the efforts assisted by any number of contributing groups. There must be a clear reason for the system and a clear advantage for agencies to accept these responsibilities. They must engage the providers and users of data in order to succeed.

#### **4. DATA AND PRODUCT ACCESS/DISTRIBUTION**

The requirements and needs of a (potentially) wide range of users are the most important considerations for ocean data distribution systems. Data characteristics that can be tailored to fit these needs include format, temporal and spatial resolution, data gridding (yes or no), and data type; e.g. digital values or products (e.g. contour plots or profiles). Climate researchers may prefer daily or monthly mean data, while those doing satellite validation would like higher temporal resolution data. Distribution systems can provide both, but at the cost of database volume, processing time and maintenance. Adding gridded products means choosing a gridding method (sometimes a highly subjective decision- a method appropriate for one investigation may not be appropriate for others).

Users need to know whether the data they seek are available before downloading and processing large volumes of data. Therefore, some general overview of data availability such as data coverage maps, time charts, or browse products is helpful.

Because metadata is so critical for climate studies, data distribution systems should emphasize it and make it as available as the data. Characteristics that should be included are: the measurement methods, sensor type, calibration frequency, estimated accuracy, location, frequency and time period of the measurements, and a contact for further information or feedback.

Environmental data sets have become increasingly large and more complex. This increase in the size of ocean data and products makes it impractical to download entire data sets (to insure completeness). Even when technology advances provide the means for distribution of complete data sets, it is becoming increasingly important to facilitate searching and/or quick identification of required data.

Data archives are not static. Data are continually being added, and replaced as higher quality, higher resolution data replace less desirable forms. It is crucial that there be a method of notifying users of the changes to the archives. The goal is to have a single, authoritative source for the data. While older versions of an archive (or part of an archive) may exist in the scientific community, a user should be able to readily identify if the version of the data that he has is the most recent and best available, and what changes have been made since last accessing the archive.

These user requirements must then be matched to the capabilities of the data distribution center. For significant amounts of data (e.g. hydrographic), historically there have been relatively long delays before distribution can occur due to lack of appropriate permissions and data use issues (this will be addressed by another paper). The hardware resources necessary to serve large data sets are relatively modest, but infrastructure costs and server development and maintenance (developing and updating web and FTP sites, adding new data entries, maintaining server) are much larger. Additionally, adding sophisticated search capabilities usually requires additional funds and talent to design and implement efficient query and server capabilities. Costs for all these efforts must be factored into future data management requirements.

Emerging technologies in fast networks, distributed data access and data visualization are examples of methods that may enhance analysis of environmental data sets and model outputs. These include 3D visualization, interactive and immersive visualization, interactive steering of a simulation and analysis of model outputs while the model is running. The development of these methods requires sustained investments in fast networks, as well as hardware and software development. These technological advances will have the greatest impact in shared, collaborative environments.

Taking advantage of advances in electronic (e.g. FTP, WWW) and removable media (e.g. CD-ROM), the distribution of ocean in-situ and remotely-sensed data has taken ocean data distribution to new levels of availability. The approaches, formats, and data (and product) delivery modes are extremely diversified and range from the simplistic FTP'ing of flat ASCII files to the complex data sets tailored through interactive forms. These all focus around the "model" of transferring data from a designated server as an independent task (for the purposes of discussion, the server can be a CD-ROM with a browser interface).

Other approaches such as specialized software tools at ECMWF and the Bureau of Meteorology (Australia) query and pull the latest data from specified databases. The Distributed Ocean Data System (DODS) which links a data-handling application with disparate datasets in remote locations [see [www.unidata.ucar.edu/packages/dods/](http://www.unidata.ucar.edu/packages/dods/)]. DODS uses the client-server model, with a client sending a data request across the Internet to a server, which responds by sending the requested data. The client is an application that speaks through DODS constructs to obtain data. The DODS server retrieves data from specified datasets. DODS has the flexibility to extract portions of on-line data sets from within a number of programming interfaces (e.g. FORTRAN, Matlab) and thus does not require the individual task of downloading data files.

An effective data access system makes transfer of information simpler, more efficient, and in a form more readily usable by investigators. For example, standardization of formats (e.g., netCDF, IEEE Binary, GRiB) makes integration and use of multiple data files easier for many users. Likewise, subsetting (selecting specific time periods, spatial areas, or measured quantities) increases efficiency by reducing access time and processing efforts required by researchers interested in only a portion of the domain of a given database. New network technologies such as DODS makes it possible for users of common data analysis and display programs (e.g., MATLAB, Ferret) to access remote databases as if they were local, removing the need for transfer and storage of data on the user's computer system.

A new paradigm of ocean data availability is under development at FNMOC-Monterey, CA. In order to meet the real-time needs of the GODAE community, a dedicated data server is being installed to facilitate access to real-time ocean observations contained in the GTS streams (as well as from delayed mode archives of ocean observations and products), and select operational products from a variety of atmospheric and oceanic modeling centers. It will also distribute GODAE products. This unique server may act as a catalyst for embracing real-time distribution of quality-reviewed data and is a great opportunity for implementing many of the changes for ocean data distribution discussed in this section.

## **5 DATA ARCHAEOLOGY AND RESCUE**

Unfortunately, not all ocean observations were considered by organized DIMS as described previously. To support studies of the role of the ocean in climate change, scientists need access to the most complete oceanographic databases possible. However, the number of observations currently available is woefully small for many variables and oceans. Therefore substantial (and relatively costly) efforts are underway to recover historical data that are at risk of being lost if no rescue efforts are initiated.

The international oceanographic community has had a long history of exchanging oceanographic data that begins with the founding of ICES. It was ICES policy to gather and publish oceanographic profile data in its *Bulletin Hydrographique* and publish plankton data in *Bulletin Planktonique* beginning in 1907-1908. Until the past twenty years, oceanographic data were not exchanged in real-time via the GTS, as has been the case for some meteorological data. Hence, a

great deal of historical oceanographic data are available only in manuscript form. Many of these data are at risk of being lost due to decay or neglect.

Most of the data made available in real-time are XBT's which provide only temperature data. As noted by Cooper (1988) and Acero-Schertzer *et al.* (1997), salinity data are necessary in order to correctly model tropical ocean currents.

The GODAR project was initiated in 1993 as a project of the Intergovernmental Oceanographic Commission (Levitus *et al.*, 1994). The goal of this project is to locate (archaeology) and digitize and/or transfer to modern electronic media (rescue), i.e. Fig. 1 and Fig 2. One of the principles of this project is that all data "rescued" are made available internationally without restriction. Since its inception, the GODAR project has added approximately two million profiles, 120,000 chlorophyll profiles, 600,000 plankton observations as well as other data have been made available internationally without restriction. The GODAR temperature profiles account for approximately 40% of the total number (5.2 million) profiles in the NODC/WDC-A archives.

Specifically, the data were first made available as part of the *World Ocean Atlas 1994 CD-ROM set*. The most recent data are available as part of the *World Ocean Database 1998 CD-ROM set*. Data distribution plots and statistics describing the existing databases are available as part of the *World Ocean Database 1998 Atlas Series* (Boyer *et al.*, 1998a,b,c; Conkright *et al.*, 1998a,b; Levitus *et al.*, 1998a,b; O'Brien *et al.*, 1998).

There are an estimated one million Nansen casts and more than 600,000 profiles that are known to exist and need to be acquired. Experience suggests that there may be at least as many profiles to recover as have been so far. Thus archeology and rescue activities need to continue to continue to build this foundation for climate study.

## 6 COORDINATION OF ACTIVITIES

The global observing systems, GCOS, GOOS, and GTOS, are in the process of specifying observational requirements. Networks are to be implemented to collect data that will permit the detection, understanding, and prediction of climate and other change, including both natural and anthropogenic elements. These observing systems will collect physical, chemical, and biological variables that have been identified by science panels, with the appropriate expertise, as necessary for the observing systems to meet their goals. The data requirements include both in situ and satellite observations and will call for long term observations that have been collected using "best practices" that will ensure comparability.

In the future GOOS and GCOS will be a major source of high quality ocean data for a large variety of users. In fact many of the physical oceanographic data collection activities that were developed in WOCE and TOGA have been identified as part of the initial operating system for GOOS and GCOS, and will provide measurements in support of CLIVAR as well. Thus in practice the observing systems for climate requirements for the various ocean programs will be closely linked by sharing existing observing systems, data management systems, and expertise.

These programs will need an observation-to-user (i.e. end-to-end) data management system. Only this will provide the needed mechanism for coordinating the activities and increasing the participation of member states. GOOS, GCOS, and the WWO, IOC and WMO have formed a new Joint Commission (JCOMM) which is expected to provide this mechanism.

In general terms the JCOMM has responsibilities to:

- Further develop the observing networks of IOC and WMO to meet the needs of the IOC and WMO programmes and in particular of GOOS, GCOS and WWO.
- Implement end to end data management systems to meet the needs of the present operational systems and the global observing systems.
- Encourage national and international analysis centres to prepare and deliver the data products and services needed by the national and international users.
- Work with existing archival bodies to ensure the long term archival of data sets for future users.

JCOMM does not have resources to carry out data management functions. JCOMM works by soliciting support from IOC Member States and WMO members to take on these activities.

JCOMM and its sub groups will have significant responsibilities for implementation of the long term observing networks that will serve the needs of IOC and WMO programs and in particular GCOS and GOOS. The science panels of GOOS and GCOS will provide scientific expertise for network design and specifications of parameters to be measured. Coordination and links with CLIVAR, IGBP, etc will have to be established, primarily through interaction of the science panels of GOOS and GCOS with the equivalent bodies in CLIVAR and IGBP for example.

The other important linkage will be with the JDIMP and the WOCE/CLIVAR CDTT. JDIMP is responsible for oversight of data management in GCOS, GOOS, and GTOS. The task team is responsible for planning and coordinating WOCE/CLIVAR data management.

The role of the science teams is critical. They respond very quickly to new ideas, and consequently their efforts working closely with the observations contribute significantly to improvements of the DIMS.

In summary, The science panels will develop the data collection and management requirements. The other groups develop the mechanisms and then jointly recruit the centers to do the work.

## 7 DIMS MODEL FOR FUTURE OCEAN OBSERVING SYSTEM

The DIMS for ocean observations for climate will grow in complexity and breadth to encompass new as well as previously-unmanaged data streams, data obtained through archeological rescue, and finally model and synthesis products sets that will serve as the foundation for climate investigations. All of these sources share essential elements: data (product) and metadata assembly/collection, quality-review, and data distribution/access. Along with a coordination and oversight organization, these will compose the DIMS. Each of these elements is in various stages of development towards contributing to the ocean observing system for climate.

A DIMS patterned on a model of distributed data centers has numerous advantages such as distribution of costs and range of expertise. These will take responsibility for specific data streams and areas of expertise. Different models of the DIMS may exist for different data streams. **The DIMS will likely evolve around two-modes to differentiate between the requirements of the operational forecast modelers, who need as many data as possible in near-real-time, and the climate modelers, who need delayed-mode data of the highest quality but not necessarily instantly.** Building this dual mode system is going to be an evolutionary process. As either new methods of disseminating data or faster methods of QC become available, these can be implemented as appropriate. New QC methods will not be developed at the same rate for different data streams. There will likely always be some delay before final hydrographic data are available, for instance. A dynamic archive of the latest and best data offers researchers the knowledge of what data are available and the opportunity to access the best available dataset.

There are several areas of needed change (improvements) in the nature of the DIMS elements. In the area of data (product) and metadata assembly/collection there is a need to make ocean observations available in near-real-time to allow their inclusion in climate prediction and nowcasting systems. This necessitates a shift towards automating data management tasks and tailoring them for real-time-user requirements. The key to developing this capability is identifying groups with necessary resources, access to operational data streams, and long-term stable funding to take advantage of the future paradigm of unrestricted ocean data circulation. A few streams (e.g. sea level, upper-ocean-thermal) are meeting some of these requirements, but need to refocus their methodologies to fully embrace the notion of near-real-time processing. Compiling metadata will become even more critical to the needs of a climate observation system for the ocean.

The quality review element must make two fundamental changes. Quality-review must adapt to new user requirements for real-time reporting. This necessitates automating quality-control algorithms and methodologies. For nearly all data streams, this will force quality-review to evolve into parallel efforts; real-time and delayed-modes. The second fundamental change for the quality-

review element is increasing the information utilized to assess the data stream and flag erroneous data. Many modeling and synthesis activities produce potentially valuable quality assessments of input data. In close cooperation with modeling centers, this information may become a significant source of knowledge to determine error characteristics of the data streams. Additionally, data from multiple (i.e. instrument specific) streams for the same variable (e.g. sea surface salinity from thermosalinograph, floats, and profilers) should all be considered in the review process.

As envisioned, the ocean observing system will grow, the data will become more qualified, and subsequently more and larger products will be developed. Consequently, data distribution will become a much more important DIMS element than previously realized. A data information center has been demonstrated to be effective in providing a centralized information depot. There are several paths that could be chosen to facilitate wider distribution and easier access to ocean data. The concept of a data warehouse, i.e. tailoring custom datasets to match user and specifications (and provider limitations), has been offered as one approach, but there are many others.

Observation systems where the DIMS elements are still in their infancy (e.g. surface and upper ocean salinity-ARGO) will need special attention to encourage dynamic and well-supported efforts. The experiences from archeological rescue efforts can contribute to developing mechanisms and motivation for data curation. DIMS for satellite observations are a rich source of ideas, resources, and expertise that can be infused into the ocean DIMS to encourage improvements as well as stimulate product and synthesis development (some of this has taken place in WOCE, but on a very limited scale). Numerical model and synthesis-based products are a vital component of climate studies, yet there is no comparable management structure to foster the utilization of these resources. Finding suitable support and talent to develop this program represents a significant challenge to the DIMS.

Finally, effective DIMS oversight, coordination, and outreach are the keys to building a permanent and cost-efficient ocean observation system. In the coming era of ocean data management, responsibilities for all DIMS participants will increase. The overall coordination of data processing must be more organized and more carefully monitored to insure qualified data are available. **Working in cooperation of data providers and data users, the DIMS must demonstrate the value of its activities.** The value of this DIMS lies in the synergy of using many sources of data easily for individual climate studies, reducing redundant QC, and insuring maximum return on investments made in the observation system.

The costs for the future DIMS are unspecified. In-situ ocean DIMS costs have historically been typically on order 1% of the respective observation program. The volume and breadth of the ocean observation network are expanding as are the responsibilities outlined in this paper. The activities necessary to transition ongoing activities towards this future DIMS will not begin until more support provides the necessary impetus. Sustaining this funding at modest levels (i.e. 10% of observational program costs) should be an integral part of every observational program and not left as an afterthought.

Current DIMS must turn towards transitioning the current (mostly) transient data management efforts to a coordinated permanent system capable of providing high quality data and value-added products necessary for climate analysis. The OCEANOBS99 conference and subsequent activities will identify contributing ocean observation systems for climate. Development of a single (joint with JCOMM, etc.) data management implementation plan must then be completed mapping out the plan for the system. Identifying necessary resources for this implementation plan will then follow. Through these steps and with adequate support a coordinated and cost-efficient permanent DIMS capable of providing high quality data and value-added data products will be realized.

#### **Acknowledgments**

Many members of the WOCE/CLIVAR Data Products Committee contributed to this report. We especially thank Piers Chapman and Jim Crease for their insightful comments. We appreciate the DPC's continuing efforts to refine the existing ocean DIMS system through hard work, dedication, and an undeniable *esprit de corps*.

## 8 REFERENCES

- Acero-Schertzer, C.E., D.V. Hansen, and M. Swenson, 1997: Evaluation and diagnosis of surface currents in the National Centers for Environmental Prediction's ocean analyses. *J. Geophys. Res.-Oceans*, **102**, 21037-21048.
- Bailey, R., A. Gronell, H. Phillips, G. Meyers, and E. Tanner, 1994: CSIRO Cookbook for Quality Control of Expendable Bathythermograph (XBT) Data. Report 221, CSIRO Marine Laboratories, Hobart, Tasmania, 75 pp.
- Boyer, T.P., J. Antonov, J., S. Levitus, M.E. Conkright, T. O'Brien, C. Stephens, D. Johnson, R. Gelfeld, 1998a: *World Ocean Database 1998, Volume 3: Temporal Distribution of Expendable Bathythermograph Profiles*. NOAA Atlas NESDIS 20, U.S. Government Printing Office, Wash., D.C., 170 pp.
- Boyer, T.P., M.E. Conkright, S. Levitus, C. Stephens, T. O'Brien, D. Johnson, R. Gelfeld, 1998b: *World Ocean Database 1998, Volume 4: Temporal Distribution of Conductivity-Temperature-Depth Profiles*. NOAA Atlas NESDIS 21, U.S. Government Printing Office, Wash., D.C., 163 pp.
- Boyer, T.P., M.E. Conkright, S. Levitus, D. Johnson, J. Antonov, T. O'Brien, C. Stephens, R. Gelfeld, 1998c: *World Ocean Database 1998, Volume 5: Temporal Distribution of Ocean Station Data (Bottle) Temperature-Salinity Profiles*. NOAA Atlas NESDIS 22, U.S. Government Printing Office, Wash., D.C., 108 pp.
- Bryden, H.L., M. J. Griffiths, A. M. Lavin, R. C. Millard, G. Parrilla, W. M. Smethie, 1996, Decadal changes in water mass characteristics at 24 degrees N in the subtropical North Atlantic ocean. *J. Climate*, **9**, 3162-3186.
- Chapman, P., 1998: The WOCE Data Resource. *Bull. Amer. Meteor. Soc.*, **79**, 1037-1042.
- Conkright M.E., S. Levitus, T. O'Brien, T.P. Boyer, C. Stephens, D. Johnson, L. Stathoplos, 1998a: *World Ocean Database 1998, Volume 6: Temporal Distribution of Ocean Station Data (Bottle) Nutrient profiles*. NOAA Atlas NESDIS 23, U.S. Government Printing Office, Wash., D.C., 296 pp.
- Conkright M.E., L. Stathoplos, T. O'Brien, T.P. Boyer, C. Stephens, D. Johnson, S. Levitus, 1998b: *World Ocean Database 1998, Volume 8: Temporal Distribution of Ocean Station Data (Bottle) Chlorophyll Profiles and Plankton Data*. NOAA Atlas NESDIS 25, U.S. Government Printing Office, Wash., D.C.,
- Cooper, N.S., 1988: The effect of salinity on tropical ocean models. *J. Phys. Oceanogr.*, **18**, 697-707.
- Daneshzadeh, Y. C., J. F. Festa, and S. M. Minton, 1994: Procedures Used at AOML to Quality Control Real-Time XBT Data Collected in the Atlantic Ocean. NOAA Technical Memorandum ERL AOML-78, NOAA/AOML, Miami, FL, 44 pp.
- IOC, 1990: GTSP Real-Time Quality Control Manual. IOC Manuals and Guides No. 22, UNESCO.
- Levitus, S., R. Gelfeld, T. Boyer, and D. Johnson, 1994: *Results of the NODC and IOC Data Archaeology and Rescue projects. Key to Oceanographic Records Documentation No. 19*, National Oceanographic Data Center, Wash., D.C., 67 pp.
- Levitus, S., M.E. Conkright, T.P. Boyer, T. O'Brien, J. Antonov C. Stephens, L. Stathoplos, D. Johnson, R. Gelfeld, 1998: *World Ocean Database 1998a, Volume 1: Introduction*. NOAA Atlas NESDIS 18, U.S. Government Printing Office, Wash., D.C., 346 pp.
- Levitus, S., T.P. Boyer, M.E. Conkright, D. Johnson, T. O'Brien J. Antonov, C. Stephens, R. Gelfeld, 1998b: *World Ocean Database 1998, Volume 2: Temporal Distribution of Mechanical Bathythermograph Profiles*. NOAA Atlas NESDIS 19, U.S. Government Printing Office, Wash., D.C., 286 pp.
- McPhaden, M.J., A.J. Busalacchi, R. Cheney, J.R. Donguy, K.S. Gage, D. Halpern, M. Ji, P. Julian, G. Meyers, G.T. Mitchum, P.P. Niiler, J. Picaut, R.W. Reynolds, N. Smith, K. Takeuchi,

- 1998: The Tropical Ocean-Global Atmosphere (TOGA) observing system: A decade of progress. *J. Geophys. Res.*, **103**, 14,169-14,240.
- Milburn, H.B., P.D. McLain, and C. Meinig, 1996: ATLAS buoy-Reengineered for the next decade. In: *Proceedings of IEEE/MTS Ocean'96*, Fort Lauderdale, FL, September 23-26, 1996, 698-702.
- Mitchum, G.T., 1998. Monitoring the stability of satellite altimeters with tide gauges. . *J. Atm. and Oceanic Tech*, **15**, 721-730
- Molinari, R. L., 1999: Lessons learned from operating global ocean observing networks. *Bull. Amer. Meteor. Soc.*, **80**, 1413-1419.
- O'Brien, T., M.E. Conkright, S. Levitus, T.P. Boyer, C. Stephens, D. Johnson, L. Stathoplos, O. Baranova, 1998: *World Ocean Database 1998, Volume 7: Temporal Distribution of Station Data (Bottle) Oxygen Profiles*. NOAA Atlas NESDIS 24, U.S. Government Printing Office, Wash., D.C., 235 pp.
- Ortega, C. and W. Woodward, 1999: New Argos Capabilities for Global Ocean Monitoring. *Sea Technology*, **40**, No. 5, 59-66.
- Smith, S. R., C. Harvey, and D. M. Legler, 1996: Handbook of Quality Control Procedures and Methods for Surface Meteorology Data. WOCE Report 141/96, COAPS Report 96-1, WOCE Data Assembly Center, COAPS, Florida State University, Tallahassee, FL 32306-2840, 56 pp.
- Smith, S. R., M. A. Bourassa, and R. J. Sharp, 1999: Establishing more truth in true winds. *J. Atm. and Oceanic Tech.*, **16**, 939-952.
- WHP, 1994: WHP Operations and Methods Manual, Part 3.1.3. WOCE Hydrographic Programme Office, WHOI, Woods Hole, MA 02543, WOCE Report 68/91, 144 pp.

## ACRONYMS

AMIP	Atmospheric Model Intercomparison Project
ARGO	Broad-scale global array of temperature/salinity profiling floats
ATLAS	Automated Temperature Line Acquisition System
CMIP	Coupled Model Intercomparison Project
DAAC	Distributed Assembly and Archive Center
DAC	Data Assembly Center
DIMS	Data and Information Management System
DIU	Data Information Unit
DODS	Distributed Ocean Data System
DPC	WOCE/CLIVAR Data Products Committee
ENSO	El Niño Southern Oscillation
ESA	European Space Agency
FNMOC	Fleet Numerical Meteorology and Oceanography Center
GCM	General Circulation Models
GCOS	Global Climate Observing System
GLOSS	Global Sea Level Observing System
GODAE	Global Ocean Data Assimilation Experiment
GODAR	Global Oceanographic Data Archaeology and Rescue
GOOS	Global Ocean Observing System
GTOS	Global Terrestrial Observing System
GTS	Global Telecommunications System
GTSP	Global Temperature and Salinity Profile Programme
ICES	International Council for Exploration of the Sea
IGBP	International Geosphere-Biosphere Programme
IGOSS	Integrated Global Ocean Services System

IOC	Intergovernmental Oceanographic Commission
IODE	IOC Intergovernmental Data and Information Exchange
JCOMM	Joint Commission on Oceanography and Marine Meteorology
JDIMP	Joint Data and Information Management Panel
JGOFS	Joint Global Ocean Flux Study
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PIRATA	Pilot Research Moored Array in the Tropical Atlantic
PMEL	Pacific Marine Environmental Laboratory
QC	Quality Control
TAO	Tropical Ocean Atmosphere
TOGA	Tropical Ocean Global Atmosphere
TOGA COARE	TOGA Coupled Ocean-Atmosphere Response Experiment
WHP	WOCE Hydrographic Programme
WMO	World Meteorological Organization
WOCE	World Ocean Circulation Experiment
WWW	World Weather Watch
XBT	Expendable Bathythermograph

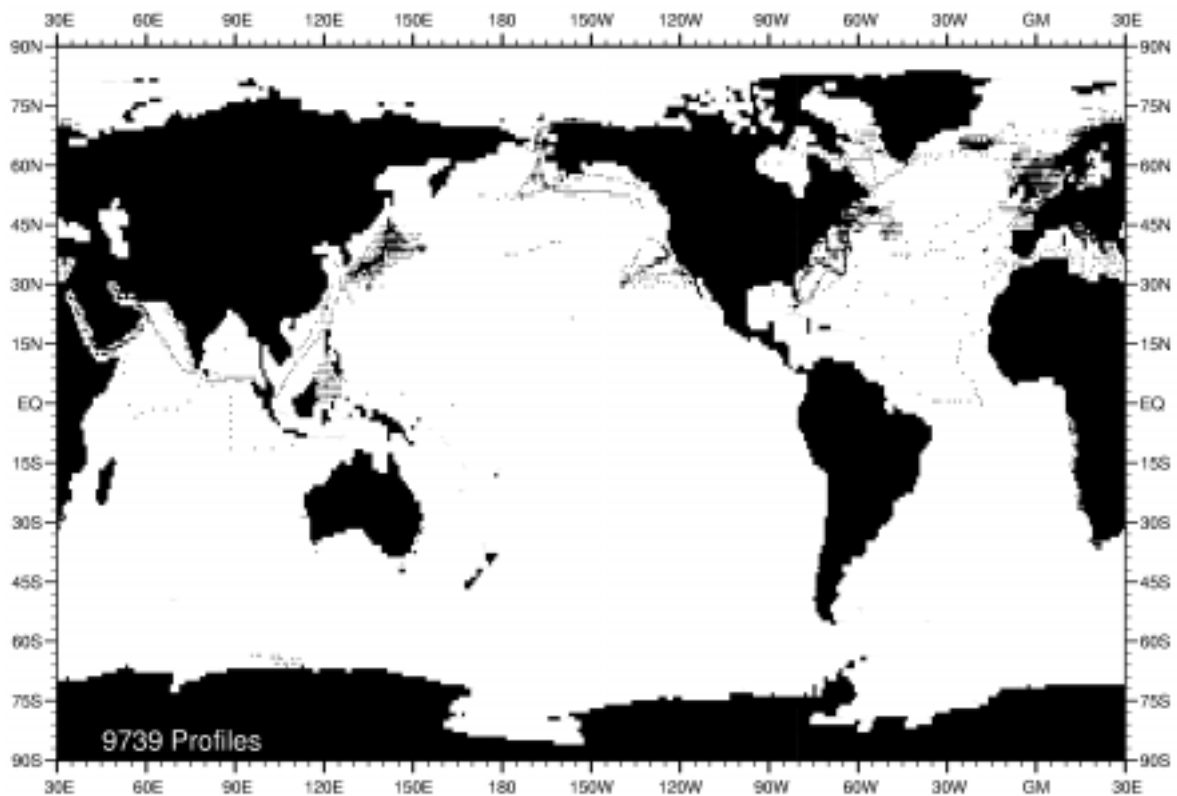


Figure 1. NODC temperature profile data for 1948 before data archeology and rescue project.

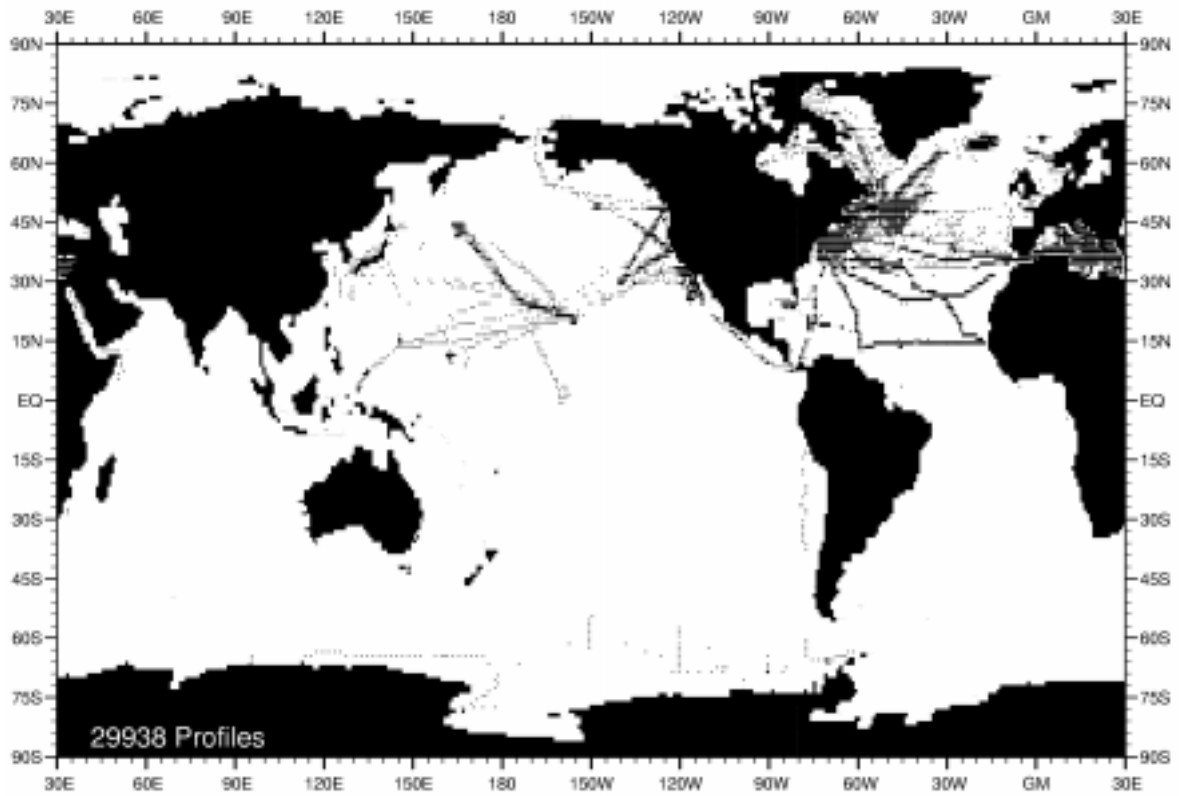


Figure 2. Additional NODC temperature profile data for 1948 acquired and made available as a result of the data archaeology and rescue project.