DATA MANAGEMENT SYSTEM FOR SURFACE DRIFTERS

Robert Keeley[(1)], Mayra Pazos[(2)], Bruce Bradshaw[(3)]

(1) Integrated Science Data Management, Department of Fisheries and Oceans, 200 Kent St., Ottawa, Canada, K1A 0E6, Email: Robert.Keeley@dfo-mpo.gc.ca
(2) National Oceanic and Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratory, 4301 Rickenbacker Causeway, Miami, FL 33149, Email: Mayra.Pazos@noaa.gov
(3) Integrated Science Data Management, Department of Fisheries and Oceans, 200 Kent St., Ottawa, Canada, K1A 0E6, Email: Bruce.Bradshaw@dfo-mpo.gc.ca

SUMMARY

The large scale deployment of surface drifters has its origins in the FGGE programme in the late 1970s. Over the course of the next 25 years, surface drifters have become routinely used in all oceans so that now there are more than 1250 operating. The data management systems to handle the data returned from these platforms started with FGGE and have evolved. The main difference between the early years and now from a data system perspective is the volume and timeliness of reporting the data, though more variables are also being measured. Presently there are more than 1 million records reported in real-time and such high volumes are straining processing systems that were originally set up for fewer records and manual scrutiny of the data. The future system will need more sophisticated algorithms for automatic detection of possible errors and these must be developed in cooperation with the scientific and instrument communities. Data must move more quickly through the processing streams and become available to users. Despite the large volumes, there is high demand for on-line access to the entire archive. Strategies must be developed to meet this demand without large data requests swamping data servers. Part of this will involve providing data browse and selection tools that permit a user to refine their request to only the data of interest. Finally, there is an increasing need to use data from different sources to examine a variety of scientific and societal problems. Standard vocabularies for naming variables, instruments, methods, etc., must become widely used to allow easier interoperability of data.

1. INTRODUCTION

The surface drifter program began during the First GARP (Global Atmospheric Research Programme) Global Experiment (FGGE) and in earnest in the late 1970s. The data from these instruments are used to support short-term (seasonal to inter-annual) climate predictions, climate research, ocean monitoring, calibration of satellite SST, barometric observations for improved weather forecasting, observations of ocean-atmosphere interaction during hurricane passage and for educational and outreach efforts. Prior to the start of this programme there was no data management system for the data from drifters. Shortly after FGGE, the drifting Buoy Technical Coordinator was created and Canada became involved handling the data. In mid 1985, Canada (formerly MEDS but recently renamed to Integrated Science Data Management, ISDM) applied to the International Oceanographic Data and information Exchange (IODE) committee of the Intergovernmental Oceanographic Commission (IOC) and was accepted to be the Responsible National Oceanographic Data Centre (RNODC) for surface drifter data. As an RNODC, ISDM agrees to manage the international archive of all surface drifter data. This includes any data reported in real-time over the Global Telecommunications System (and therefore drifters that have World Meteorological Organization (WMO) identifiers) and whatever data are submitted to them in delayed mode. During the World Ocean Circulation Experiment (WOCE) a Surface Velocity Programme Data Assembly Center (SVP-DAC) was identified as needed and Canada combined operational data management activities with the Atlantic Oceanographic and Meteorological Laboratory (AOML) in Miami, U.S.A. AOML formed the Global Drifter Program's Drifter Data Assembly Center (DDAC) to apply delayed mode quality control (QC) to the surface drifter data. Canada continued to receive and archive the real-time data and acts as the long term archive for the delayed mode data. The Drifting Buoy Cooperation Panel (DBCP) was created and later became a part of the Joint Commission on Oceanography and Marine Meteorology (JCOMM) when JCOMM was formed in 2001. With the end of WOCE, cooperation between Canada and AOML has continued although updates do not arrive as often as occurred during WOCE, due in most part to the increase in data volume.

2. ORGANIZATION OF ACTIVITIES

The management of the surface drifter data stream is relatively simple. The largest fraction of surface drifter data circulates on the Global Telecommunications System (GTS), the system operated by the World Meteorological Organization (WMO) for broadcasting meteorological and oceanographic data to meteorological centres around the world. This distribution system has been promoted by the DBCP as a way to ensure widespread distribution of the

observations. In recent years, some countries have provided internet access, in addition to distribution on the GTS, to the data collected by the buoys they operate. The ISDM role is to capture all of the data from the GTS to ensure a complete (global) and long term archive of the data and provide dissemination to users.

The Global Climate Observing System (GCOS) objectives laid out in 1992 and recently updated placed a target of 1250 buoys, uniformly distributed in the world's ocean. The target number of buoys was reached in 2005, but there are still challenges to achieve the global coverage.

Most drifters reporting data on the GTS use the Argos system, although Iridium satellites have gained popularity in recent years. There is also the facility of Local User Terminals (LUTs) to download data from the satellite transponders used by Argos. These LUTs sometimes also report data onto the GTS but their reported buoy positions tend to be less accurate than those from the Argos system because they usually lack the most recent satellite ephemeris data. However, they sometimes report data more quickly.

3. DATA FORMATS AND VOLUMES

The GTS has strict rules for the format of data. Originally data were reported in character code forms. These have evolved over the years to what is currently the BUOY code form. A number of years ago, a binary form for data reporting was designed for surface drifter data. This uses the generalized BUFR data structure with a template used to simplify encoding and decoding. Now all surface drifter data are reported in both BUFR and BUOY code forms. Reporting in BUOY will cease sometime in the 2011 time frame.

Consistent formatting ensures that material is presented well and makes the issue of software maintenance relatively simple. The main complication has been in the changing content with the changes in the character and now binary forms used to report the data.

Surface drifter data, especially sea surface temperature and increasingly surface air pressure data are used by meteorological models for numerical weather prediction. These models have strict time windows to accept data and so the requirement is to place drifter measurements on the GTS as soon as received. Because of the characteristics of the Argos system, buoys can report partial or complete messages multiple times on a satellite pass. As well, because the surface location of the buoy is computed from the Doppler signal of the transmission, the quality of the position depends on the relative positions of satellite to surface drifter and signal strength. This results sometimes in small differences in reported positions from one message to another. What appears on the GTS is that in the space of a few minutes, there can be multiple messages from the same buoy, with content differing in the separate messages. That is, the positions may be different, the observed values may be different, and the suite of reported variables may be different or the same between messages.

It is an easy matter to capture all of these multiple messages. As an ocean archive centre, Canada seeks to assemble the data, remove redundancies, apply consistent quality control procedures and distribute the best version possible of the data. Removing the redundancies is particularly important since users in delayed mode do not usually want to see all of the duplications. Real-time users simply consolidate all of the observations and allow models to use the assimilation scheme to sort out the differences.

The number of messages being received has been steadily increasing even after the target number of buoys has been reached. The volume of data now received is well over one million messages each month. ISDM processing collapses the number of reports where locations and times are the same, but different messages have different content. The experience is that this reduces the number of messages by nearly a third. Fig. 1 shows the growth in the size of the archive both in the number of archived messages and the number of bytes of information.

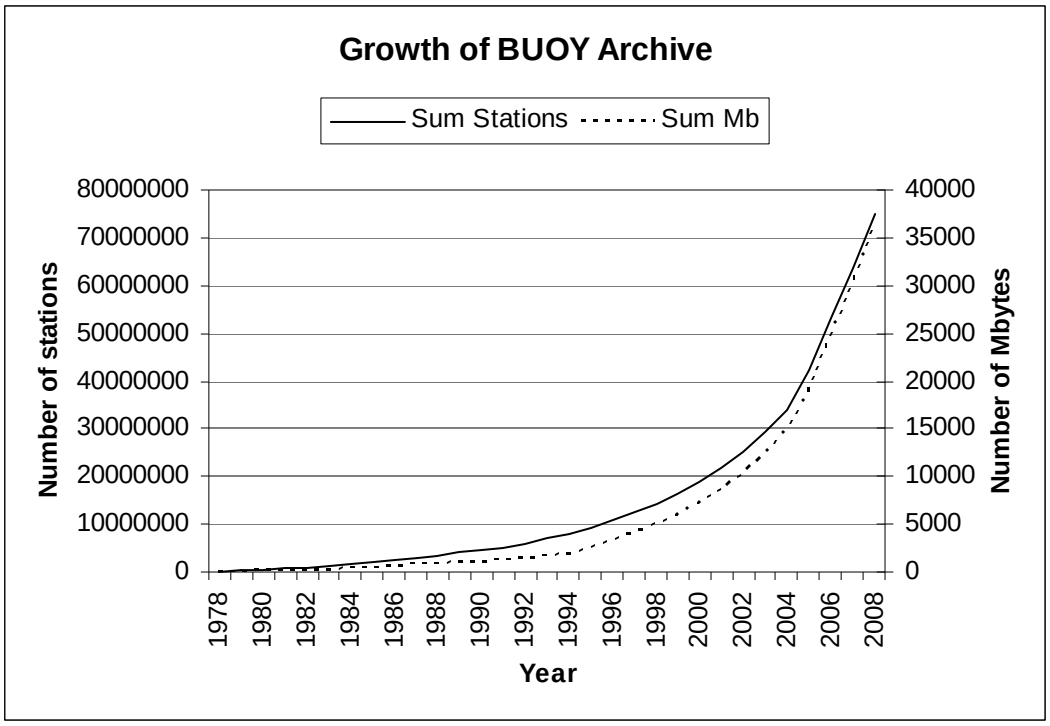

Figure 1: Growth of the real-time data archives.

4. OPERATIONS AT ISDM

ISDM currently uses the data reported in BUOY code as the basis of its archives. We have software for managing the data coming in BUFR form, but this is not yet in production. The BUOY code form has evolved over the years and this has caused significant work to reprocess the archives to change the way data are stored. The major change took place when the message content was changed to allow for transmission of the time of the position determination separately from the time of observations. Because of choices made in how to archive this information, and concerns expressed by DBCP, ISDM was required to reprocess a significant fraction of the archive. In the end this meant reprocessing many millions of records.

In 1985 ISDM became a designated RNODC for surface drifters. This was based on experience gained during FGGE and the establishment of capabilities to manage data coming from this platform. One of its first actions was to ensure that it had data in its archives from the end of the FGGE program until 1985. This required the recovery of many thousands of records from the real-time data stream from the start of 1980 until the end of 1985 and was done with the assistance of the US National Climate Data Center and the managers of the International Comprehensive Ocean-Atmosphere Data Set (ICOADS). All of these data were processed through the existing system.

The ISDM system has evolved but still relies on the GTS as the primary data source. ISDM is not directly connected to the GTS because it does not operate within the national meteorological services of Canada. Instead an arrangement with the appropriate government agency copies the BUOY messages from the GTS data stream and sends by ftp a file of retrieved data every 15 minutes to an ISDM server. These files are accumulated and once a day the content is processed through software that decodes the BUOY code form and transforms the data into the ISDM internal format. Messages that cannot be decoded are placed in a separate file in the same format as received. Each working day, the contents of this file are reviewed to recover whatever can be used. Sometimes this means removing parts of a message that failed the decoding process.

The daily files of decoded data accumulate over the course of a month before further processing. This procedure was adopted because there was no evident need for a more rapid turn around in data. Times have changed and ISDM has received requests for more rapid data dissemination. These and other factors (increasing volumes of data and needing to change system software) will likely cause a speed up of processing.

The initial step in monthly processing is to split off a calendar month of data from the accumulating data file. All subsequent processing is on this calendar month. The split actually occurs a few days after the end of a

month. Data proceeding the target month are removed from the file; the number of records lost this way is very low. Data from the most recent month are returned to the accumulating file.

The monthly file is processed to remove the duplications and near duplications described earlier. Messages from the same buoy are sorted by time of observation. Groups of messages received during the same satellite pass are identified by the small time differences between them. Messages that have identical times and locations are then examined to identify identical observed values or where different messages have different observed values. Where possible, multiple messages are then collapsed into a single message at the given observed time and location.

Data received from LUTs are also considered. Messages received from LUTs are associated with the same messages from Argos system. This is done by matching buoy identifier and observation time. Position is not used since positions provided by LUTs are often less reliable than those from Argos. Messages received from LUTs that are identical in content and time are deleted from further processing.

The monthly file after processing is then passed to a quality control procedure. This is a mix of automated and manual scrutiny using a graphical application. Buoy tracks are constructed from the times of position and locations. Time series of air pressure and surface temperature are displayed. Other variables are also displayed if present in the GTS message. Drift speeds are calculated from consecutive buoy positions and times. Where these speeds exceed 5 m/s, the position is marked with a quality flag that indicates it is suspicious. Positions are marked rather than times because the position is calculated from a Doppler measurement whereas time is taken from a clock. The suspect positions are removed from the display, but retained in the data file.

Observed data pass through a simple range check of the value. The technician views the graphical displays and looks for spikes in the observed values. Where these occur he can set a quality flag to express his degree of doubt. He proceeds to examine these monthly time series until he is satisfied that the data portrayed look reasonable.

On completion of the quality control, the data are added to the ISDM archives. All observed values that pass through initial duplicates processing are retained in the archive regardless of the quality flag. All observed values as well as times and positions are assigned quality flags. In addition, a history of the processing steps through which the records passed is retained as well as any changes that might have been made in the data. If such changes occur, the changed value is marked by a special quality flag, but the original value is retained in the processing history record.

5. DATA SERVING

The ISDM archives contain many millions of records (a position and time plus observed values constitute one record) as indicated in Fig. 1. On-line displays of the last 30 days of data are shown on the ISDM web site (http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/drib-bder/gts-smt/index-eng.asp). Because of the data volume and server capabilities, the data are not placed on-line for user self serving. It is our experience that some data requests are very large (e.g. all data from the Pacific Ocean) and if the data were placed on-line we would either need to subdivide the request into many files or the server would be overtaxed. Instead users wanting data can request them through an on-line form which generates an email for ISDM staff to respond to. Nevertheless, it is clear that improvements are needed for more automated access to our archives.

6. SVP AND AOML OPERATIONS

AOML operates the Drifter Data Assembly Center as part of the Global Drifter Program. Raw data are received daily and monthly from Argos for AOML programs, and also for programs belonging to other national and international partners. These data are quality controlled and interpolated in delayed mode at about 3 month intervals. Details of the quality control and interpolation procedures carried out at AOML can be found in Hansen and Poulain (1996) (http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/drib-bder/svp-vcs/res/qc-cq-eng.htm). AOML makes the interpolated, quality controlled data available through a web interface, but there is no web access to the raw or quality controlled non interpolated data. . These data are offered to researchers, operators and users in general when requested via e-mail; AOML has not been funded to provide data distribution. ISDM operates in partnership with AOML in maintaining the SVP archive and its continuation after WOCE. Data are received from AOML usually about once a year in a number of separate files containing different levels of processing. A Directory file contains one line per buoy with metadata information. A Raw file containing each buoy identifier with the raw data. A "P" file contains a list of buoy identifier, position times and location. An "S" file contains a list of buoy identifier, sensor time and data from different sensors. A “K” file indicates buoy identifier, and contains uniformly interpolated data to 6 hour intervals, including velocity, speed and error estimates from the measured values.

"P" and "S" files have been quality controlled with bad positions and SSTs discarded. "S" files have more records than their counterpart "P" files because there are more sensor data transmitted per satellite pass than there are positions. ISDM recombines the P and S files into a single file using position times to link to the closest sensor times from the S files.

Other processing is also required to reconcile differences in how AOML and ISDM archive data. The data then are placed in an archive separate from the data received from the GTS. The size of this archive is shown in Fig. 2.

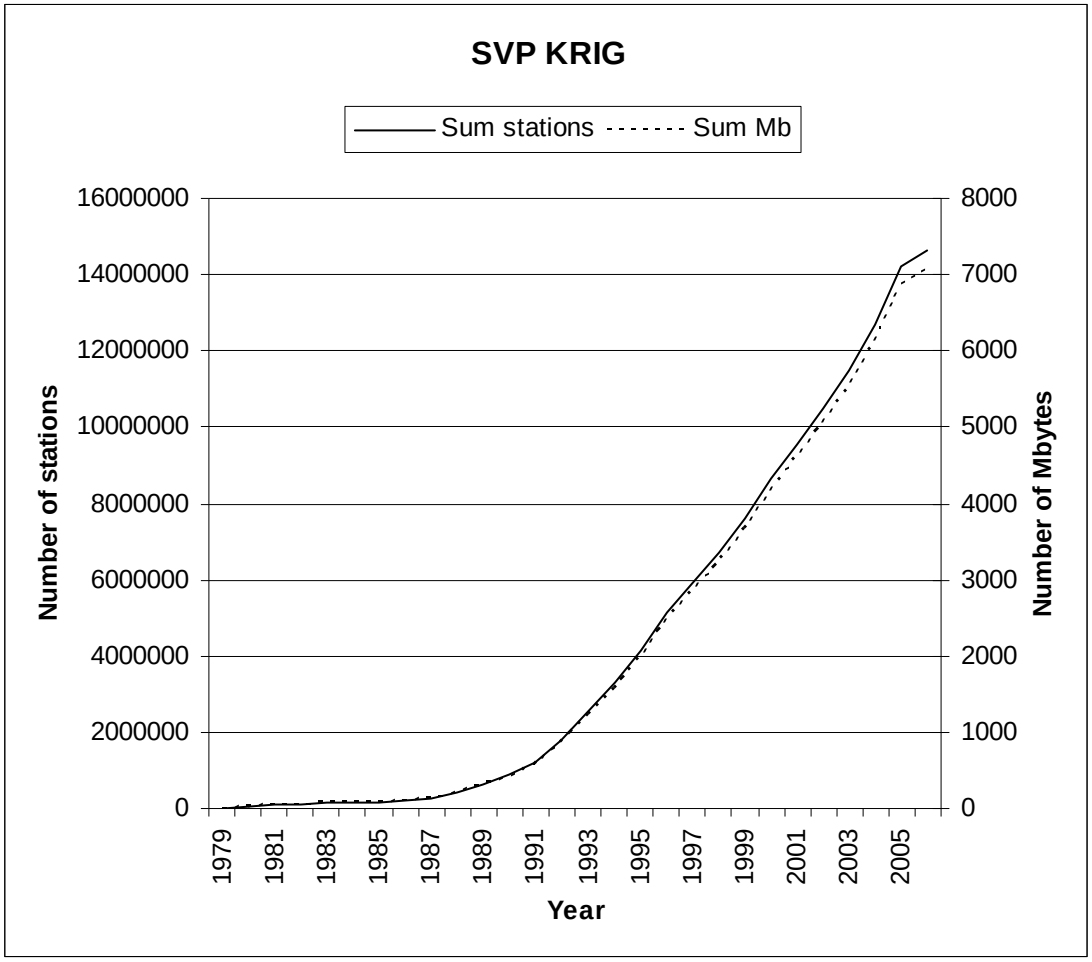

Figure 2: Growth and size of the SVP Archive

At the same time, files (in ASCII form) from all different processing levels are built by year and ocean basin from the data just received and these are placed on the ISDM web site (http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/drib-bder/svp-vcs/index-eng.asp . These may be downloaded as needed.

7. DATA DISSEMINATION AND USERS

While the main objective at the start of the data system for surface drifters was to protect the data from loss, access to the data has also been supported. At present, users can access web pages at both ISDM (http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/drib-bder/index-eng.htm) and AOML (http://www.aoml.noaa.gov/phod/dac/gdp.html) where products and data are available. At ISDM about 8 GB of data were downloaded by 8,000 visitors in the past 6 months, with 60% from outside of Canada. In addition many of the visualization pages, even those that are static monthly maps, are routinely accessed as well. Each month all of the data captured by ISDM are routinely uploaded to two international organizations.

Users are scientific and operational although most operational users get the data directly from the GTS if they require the data quickly. Scientific users range from those wanting data from a small area or time frame, to others who want data from large areas and long time frames (significant fractions of the entire archive). Typically, ISDM handles through manual means about 2 requests a month for surface drifter data.

8. LESSONS LEARNED AND CHALLENGES

A number of observations can be made about the data handling of surface drifter data and some comments made about the future.

- Recording the processing history, any original values, and all data but with quality flags has been very helpful. The processing history allowed us to identify more precisely where problems have occurred even many years after the data were processed. By preserving all data but with quality flags, ISDM has been able to carry out reprocessing of the archives as changes in code forms and mistakes were found. If data assessed as bad or suspicious had been removed, it would have been necessary to go back to original records as received which would have made the reprocessing a much more time consuming process.
- ISDM chose a data structure that used indirect referencing for reported data and information. In simple terms, a code is associated with a value. The code indicates what is being stored and the value provides its content. We have used this to store both observations, using for example TEMP to indicate water temperature followed by the observed value, and information, such as IST$ to indicate the instrument used followed by its name. By doing so, we are able to adjust to the changing content reported in GTS messages by simply adding new codes for new information. Without this flexibility, we would have needed to change archive structures and even processing software many times over the years.
- Having achieved the GCOS target of 1250 buoys likely means that the growth in data volumes will ease somewhat. However, the goal of uniform buoy density in the oceans has not been achieved and will require more drifters than the 1250 target figure. Such data volumes are straining existing processing procedures and these procedures will require changes.
- Changes in processing are also required to provide a faster delivery of fully processed, quality controlled data. This will mean more complete processing more rapidly than on a monthly basis and so changes in how the QC is accomplished. Presently it takes nearly the same amount of time to process one full month of data from a buoy as would be required if we did one week or one day This is because we view data buoy by buoy and the number of buoys is independent of the number of days of data. It is likely that we will need to carry out a more automated version of QC for quick data delivery.
- Accepting and processing data encoded in BUFR will change the content we must process and may allow other tests to be conducted on the data. This will mean software changes as well.
- Working with AOML in the context of the WOCE and afterwards has been a very good experience for ISDM. It provides contact with scientists working closely with the data and this is always a good thing for a data centre to do. In retrospect, it would have been better to coordinate our two processing systems. As it stands, the differences between the two cause a significant amount of processing at ISDM.
- Close contact with the observing programme, through the DBCP are also very important to a data centre. These contacts provide links to the programmes placing buoys in the water, to Service Argos and the processing that occurs there and helps us to understand why data are reported the way they are and so allows us to adjust our processing to exploit or compensate for what happens. Again, close contacts between the data archive centre and the deployment program have been very beneficial.

Looking forward, the following will be objectives for improving management of the data.

- Despite the data volumes, on-line access to raw, edited and fully quality controlled data needs to be made. Such access needs to provide data selection tools to allow users to fully refine the quantity of data desired.
- Closer cooperation with the scientific community will be needed to develop and prove more robust QC systems as a necessary condition to dealing with increasing data volumes.
- To support fully accessible archives there will be a need to visualize at least the times and locations of the available or selected data. Visualization of the content of the potential data set (i.e. what variables are present and to what degree on each buoy) is also highly desirable.
- More than one data access and visualization system will be needed. Just as the clientele of an archive are varied, so are their needs for such access. Use of web services will assist to meet some of these demands.
- Because surface drifters will continue to be used for sensors more and different from the current suite, any data system will need to be agile in adjusting to the new variables. Such ability currently exists, but as systems migrate to more modern software, this ability needs to be retained.
- It is important for the surface drifter data systems to collaborate and conform to internationally accepted vocabularies for variables, instruments and other information. This will permit easy and reliable comparisons of the same variable from different

platforms. The data system will need to incorporate these vocabularies as they are developed.

A number of the challenges and changes noted above are being addresses already. We have adapted some quality control procedures to become more automated. However, collaboration with a scientific organization to explore other strategies will be helpful. Software has been developed already and is in final testing for dealing with data served on the GTS in the BUFR format. We are in the process of verifying that all information transmitted in traditional character codes is also coming in BUFR before we make the switch over to dealing in BUFR only. ISDM has and will continue to support a staff member to attend DBCP meetings. Software to improve visualization of the archives currently exists. The main hurdle is to allow such visualization for a large archive without significantly degrading server response times.