GTSPP Bi-Annual Report for 2009 - 2010

Prepared by Charles Sun (Charles.Sun@noaa.gov)

1.0 Introduction

The Global Temperature and Salinity Profile Programme (GTSPP) is a joint programme of the International Oceanographic Data and Information Exchange committee (IODE) and the Joint Commission on Oceanography and Marine Meteorology (JCOMM). IODE and JCOMM are technical committees of the Intergovernmental Oceanographic Commission (IOC) and the World Meteorological Organization (WMO).

Development of the GTSPP (then called the Global Temperature-Salinity Pilot Project) began in 1989. The short-term goal was to respond to the needs of the Tropical Ocean and Global Atmosphere (TOGA) Experiment and the World Ocean Circulation Experiment (WOCE) for temperature and salinity data. The longer term goal was to develop and implement an end to end data management system for temperature and salinity data and other associated types of profiles, which could serve as a model for future oceanographic data management systems.

GTSPP began operation in November 1990. The first version of the GTSPP Project Plan was published in the same year. Since that time, there have been many developments and some changes in direction including a decision by IOC and WMO to end the pilot phase and implement GTSPP as a permanent programme.

GTSPP played a key role in the WOCE Upper Ocean Thermal Data Assembly Centre and contributed to the final WOCE Data Resource DVD Version 3. GTSPP is also an accepted part of the Global Ocean Observing System (GOOS) and a participant in Climate Variability and Predictability (CLIVAR).

Tasks in the GTSPP are shared amongst the participants. Real-time data processing services are provided by the Integrated Scientific Data Management (ISDM) of Canada. The U.S. National Oceanographic Data Center (NODC) provides data processing services for delayed mode data and maintenance of the Continuously Managed Database (CMD).

Contributions to the data management portion of GTSPP are provided by Argentina, Australia, Canada, China, France, Germany, India, Italy, Japan and the USA. Scientists and data managers in these countries contribute their time and resources to ensure the continuing functioning of the programme.

2.0 Objectives

The objectives of the GTSPP are as follows.

1. To provide a timely and complete data and information base of ocean temperature and salinity profile data of known and documented quality in support of global and local research programmes, national and international operational oceanography, and of other national requirements.

2. To improve data capture, data analysis, and exchange systems for temperature and salinity profile data by encouraging more participation by member states, by locating new sources of data from existing and new instruments and implementing the systems to capture and deliver the data, by taking full advantage of new computer and communications technologies, and by developing new services and products to enhance the usefulness of the GTSPP to clients and member states of IODE.

3. To develop and implement data flow monitoring systems to improve the capture and

timeliness of GTSPP real time and delayed mode data, and to distribute information on the timeliness and completeness of GTSPP data bases so that bottlenecks in the data flow can be identified and addressed.

4. To improve the state of databases of oceanographic temperature and salinity profile data by developing and applying improved quality control systems, by implementing new data centre tests for quality control (QC) as appropriate for new instrumentation; by working with the scientific partners of GTSPP to train data centre staff and transfer scientific QC methods to the centres, and by feeding information on recurring errors to data collectors and submitters so that problems can be corrected at the source.

5. To facilitate the development and provision of a wide variety of useful data analyses, data and information products, and data sets to the GTSPP community of research, engineering, and operational clients.

3.0 GTSPP Operations

Figure 1 presents the data flows of national and international programmes within which GTSPP is placed. The boxes in the Figure represent generic centres. A given international JCOMM or IODE centre may fit within several boxes in carrying out its national and international responsibilities. The following sections discuss this figure in terms of essential elements of the GTSPP.

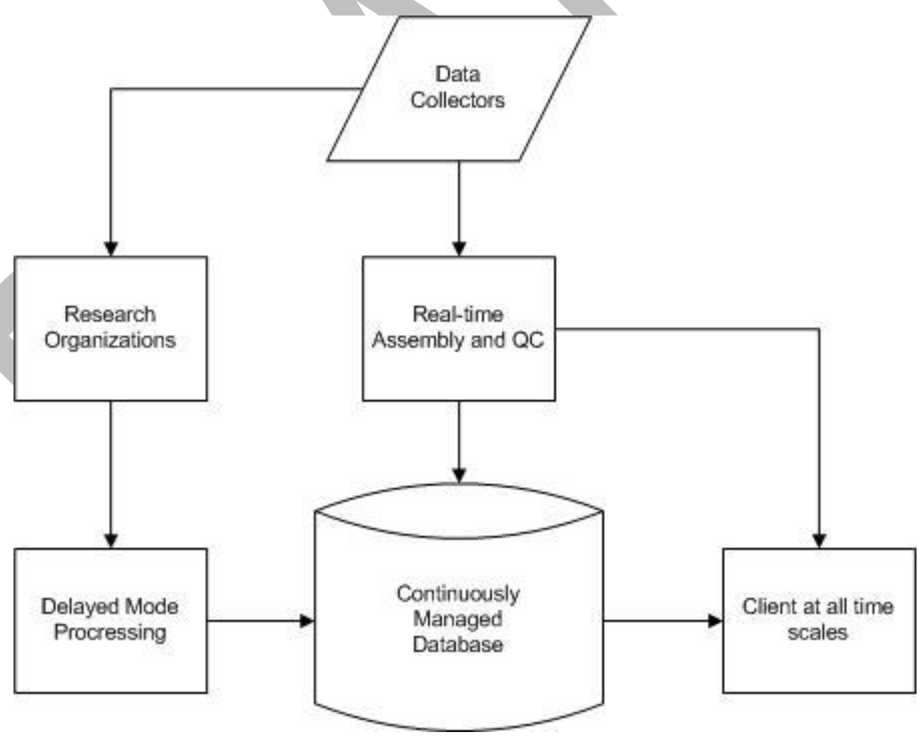

Figure 1: GTSPP data flow

3.1 Near Real Time and Operational Time Frame Data Acquisition

Near real time data acquisition within GTSPP depends on the Global Telecommunications System (GTS) of the World Weather Watch of WMO and the telecommunications arrangements for BATHY and TESAC data established by JCOMM. Copies of other real time or operational time frame data sets are acquired from any other available sources via the Internet of other computer networks. The goal is to ensure that the most complete operational time frame data set is captured.

Figure 2 is a graphic representation of the GTSPP operational time frame data flow. The "data collectors" in the top boxes follow one of two procedures. In the first case the data are provided to GTS centres that place them on the GTS within minutes to days of their collection. In the second case the data are supplied to a national organization that forwards them to the real time centre in ISDM within a few days to a month of its collection.

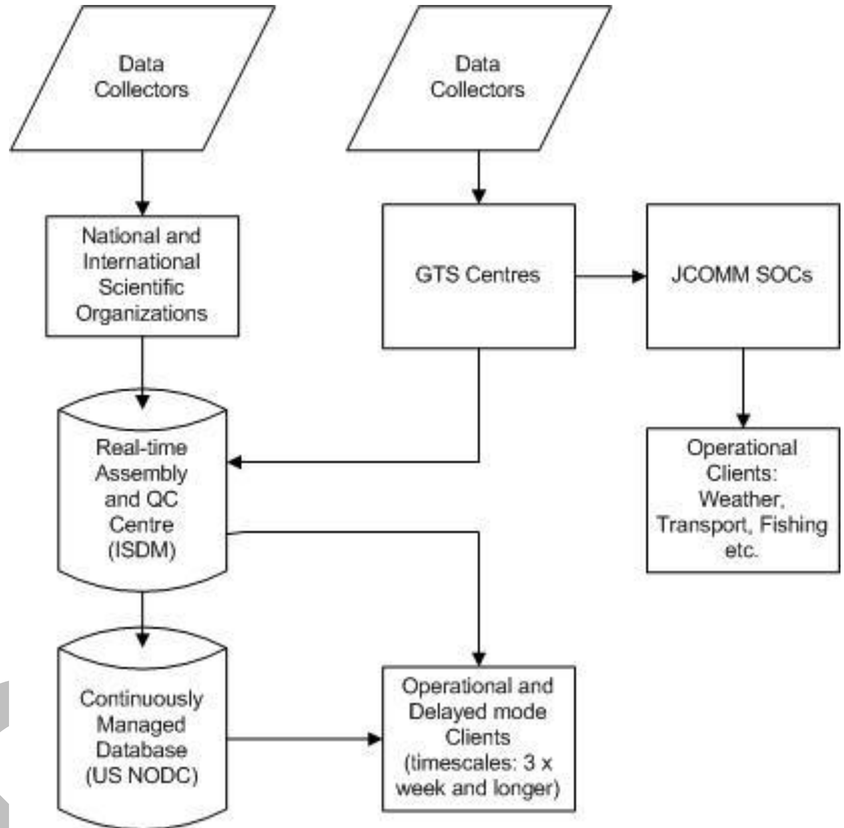

Figure 2: Real-time data flow

The real time data that are circulated on the GTS are acquired by ISDM and the Specialized Oceanographic Centres (SOCs) of JCOMM and by users of real time data who have access to the GTS. These users include meteorological and oceanographic centres that issue forecasts and warnings, centres that provide ship routing services, and centres that prepare real time products for the fishing industry.

ISDM compiles the global data set from the various sources, applies the documented GTSPP QC and duplicates removal procedures, and forwards the data to the US NODC three times per week. At NODC the data are added to the CMD on the same schedule. There are also several clients that receive copies of the data sent from ISDM three times per week. These are clients who do not need the data within hours but rather within a few days. By getting the data from the GTSPP Centre in ISDM they save having to operate computer systems to do quality control and duplicate removal.

The regular route for real-time data to the box marked "Operational Clients" in Figures 1 and 2 is not affected by GTSPP. This route provides for uninterrupted flow of data for weather and operational forecasting through the national weather services of member states. These centres need the data in hours rather than days.

3.2 Delayed Mode Data Acquisition

GTSPP utilizes, to the extent possible, the existing IODE data network and processing system to acquire and process delayed mode data. The box entitled "Delayed Mode & Historical Data" in figure 3 shows the delayed mode data flow in graphic form. The data flow into the continuously managed database is through a "Delayed Mode QC" process. This process is analogous to the QC carried out on the real-time data and conforms to the specifications of the GTSPP QC Manual. In some cases, where appropriate arrangements can be made this QC process exists and is performed in another oceanographic data centre on behalf of NODC.

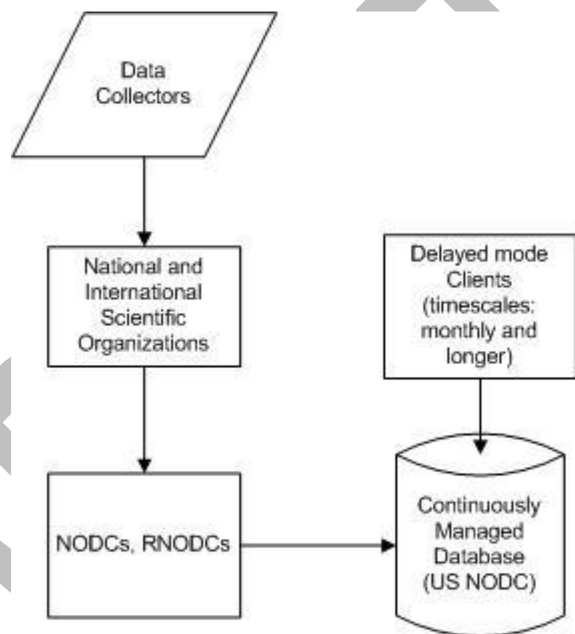

Figure 3: Delayed mode data flow

Having proceeded through the delayed mode QC process, the data then follow the same route as the real time data through the rest of the CMD process, however, on a different time schedule because of the more irregular times of arrival. During the merging of the data into the CMD, any duplicates occurring between real-time and delayed mode data sources are identified with the highest resolution copy being retained as the active CMD version.

Acquisition of delayed mode data from the Principal Investigators is a priority for the GTSPP. The goal is to get the delayed mode data into the CMD within one year or less of its collection. An excellent way for any national oceanographic data centre to support GTSPP actively is to obtain national data sets of temperature and salinity data, apply GTSPP QC procedures, and submit them to the CMD.

4.0 Progress to the end of 2010

The purpose of this section is to report on the performance of the GTSPP to the end of 2010 in meeting its objectives.

4.1 Data Volumes

The GTSPP handles all real-time and delayed mode profile data with temperature and salinity measured. Real-time data in GTSPP are acquired from the GTS in the BATHY and TESAC codes forms supported by the WMO. Delayed mode data are contributed directly by member states of IOC.

The delivery of ocean data in real-time was initiated many years ago and administered by the IOC program called the Integrated Global Ocean Services System (IGOSS). In 2001 operational oceanography programs of IOC and marine meteorological programs of the WMO were merged under JCOMM. IGOSS is now being encompassed within JCOMM. Under IGOSS, “real-time” was defined to allow data up to 30 days after collection to be included. This definition has persisted, even though the trend is to shorten considerably the delays between observation and distribution.

In JCOMM, the BATHY and TESAC code forms are the ones used most often for distribution of ocean profile data on the GTS. Figure 4 shows the progression in the use of these codes to make ocean data available. The dramatic change in mid 1999 shows the initiation of the Argo Project and the beginning of the use of TESAC to report profiles from robotic profiling floats. A review of the SOOP program in 1999 recommended a switch from broadcast sampling to line mode sampling. In principle, it was hoped that as many XBTs (exclusively reported using the BATHY code form) would be deployed along lines as formerly were deployed in broadcast mode. It is evident from the figure that the number of BATHY reports has declined since 1999 but appears to have stabilized or perhaps is slightly increasing once more.

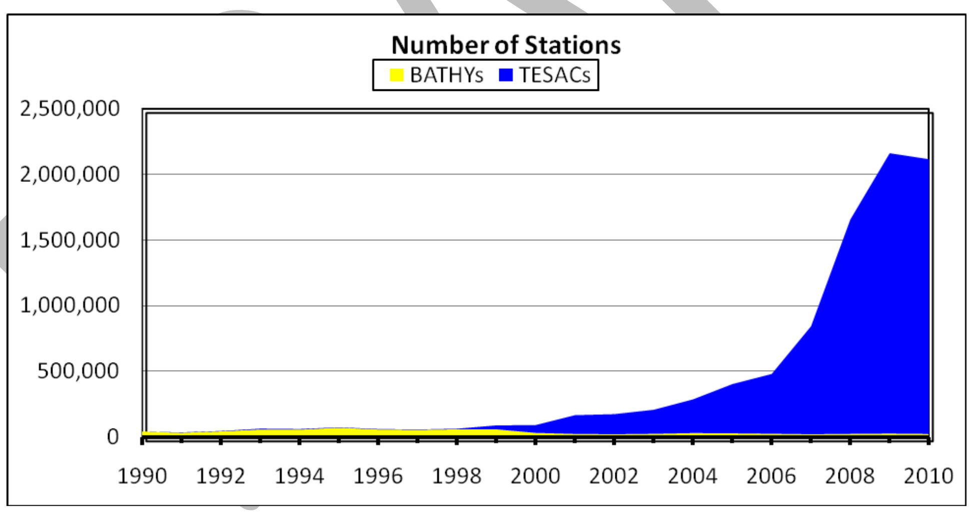

Figure 4: The number of stations reported as BATHYs and TESACs.

The next figure shows the kinds of instruments contributing data in the CMD. These have been subdivided into a few different types and presentations made of the number of stations of each type by year. Evidently, the majority of data are from profiling floats after 2000.

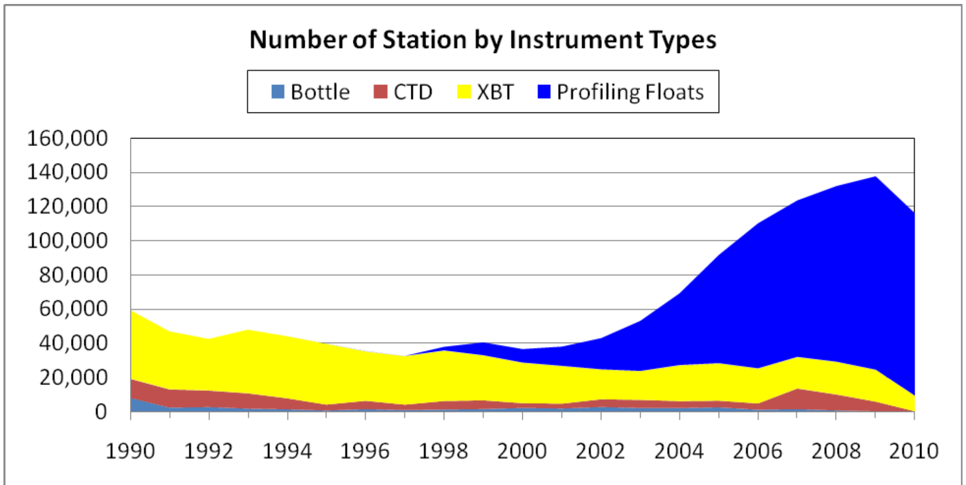

Figure 5: The number of stations by instrument type in the CMD.

The next figure shows the relative proportion of real-time to delayed mode data present in the CMD. There are a number of things to take note of in this figure. First, GTSPP deals in both real-time and delayed mode data. While it is encouraged, by no means do all of the data available in delayed mode also arrive in real-time. This means that even though there may be a significant delayed mode contribution to the CMD these may be data that were never reported in real-time and so do not replace the real-time data.

It is evident that in only a few years have the delayed mode data arrived to replace the real-time even many years after the data collection. This shows that even though it is possible to look at time lags of delayed data coming to the CMD, Figure 6 illustrates that there continue to be a significant number of high resolution stations to recover. This assumes that GTSPP is able to match the real-time data to the delayed mode profiles as they arrive. This capability is something that is touched upon as part of ongoing work reported later.

Second, as expected in the more recent years, the number of stations of delayed mode data decreases and the number of real-time increases as a proportion of the total number of stations. This, too, is typical in that it can take years for delayed mode data to reach the archives. It is precisely because of these delays that GTSPP was started and to provide the combination of real-time and delayed mode data to any user when they request the data.

Finally, the graph shows spectacular growth in the number of real-time stations from about 2000 to the present. Much of this is a direct result of the start of the Argo program. Argo profilers measure both temperature and salinity profiles usually from 2000 m to the surface. As well, there are a small number of floats now being deployed that are reporting oxygen as well. The vertical resolution varies with a typical profile having approximately 70 levels. This is all that will ever be returned from the floats and so the only difference between delayed mode and real-time profiles reported on the GTS is in increased precision of the measurements and better quality control of the data. The Argo data are also reported in real-time to the Global Data Assembly Centres of Argo, and here there is no loss of precision between real-time and delayed mode data.

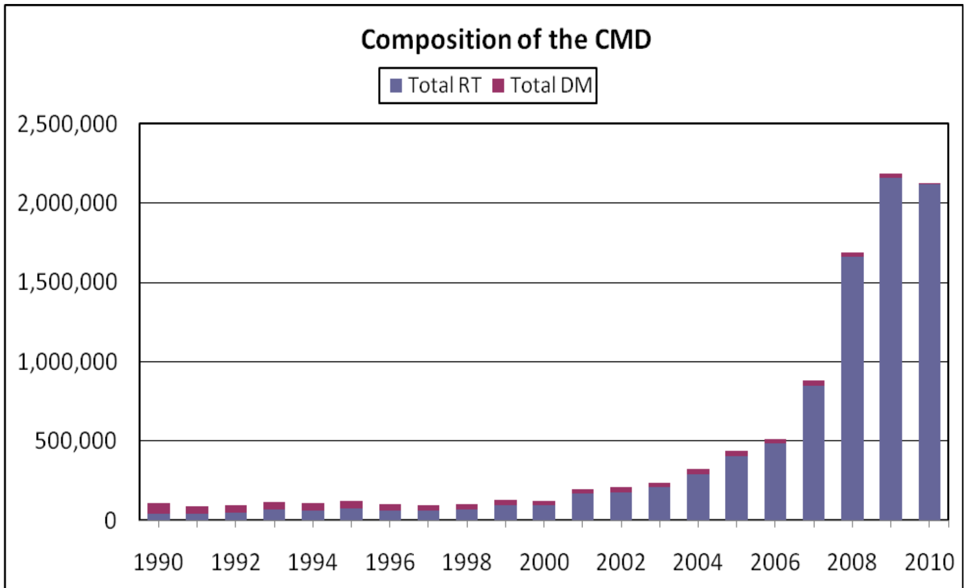

Figure 6: The volume of data in the CMD.

4.2 Timeliness of data

The management of data within the GTSPP is organized around the idea of a continuously managed database. Clients of the CMD can receive data at any time and they are of the highest quality and highest resolution available at the time of the request. Typically, the real-time data arrive first, and so become available first. As the delayed mode data arrive, they replace the real-time data or add to the total available data.

A variety of platforms report data and each of these platforms has different systems by which data get ashore and to the GTS. While it is possible to look at the timeliness of reports as a function of the variety of platform types and instruments, it is more instructive to look at platform types that to some extent represent the extremes in timeliness. To this end, data arriving from ships can be considered the least automated (and so the slowest to arrive). At the other end are those data coming from automated platforms, of which we can take Argo as an example.

It is also possible to look at the time to get data to the GTS as well as the time to make data available from the CMD. The GTSPP goal is to make data available as rapidly as possible and so it is the time to make data available that is the more important. Consequently, the difference between observation time and update time (equivalent to data being passed to GTSPP clients) is what is shown here.

Another consideration is that the real-time collection and distribution of ocean profile data continues to operate on the principle that real-time is defined as any data up to 30 days after collection date. Thus, some contributors use ships to collect data, return back to their home port and then deliver data to the GTS to still make the 30 day cutoff. Although the trend these days is to move to more rapid data dissemination, those operating under these older principles still contribute to the data flow and this will impact the timeliness statistics.

Figure 7 shows that roughly 75% of the real-time XBT data measured by ships were available within five day and almost 95% of the data available in the CMD within a month after data collection.

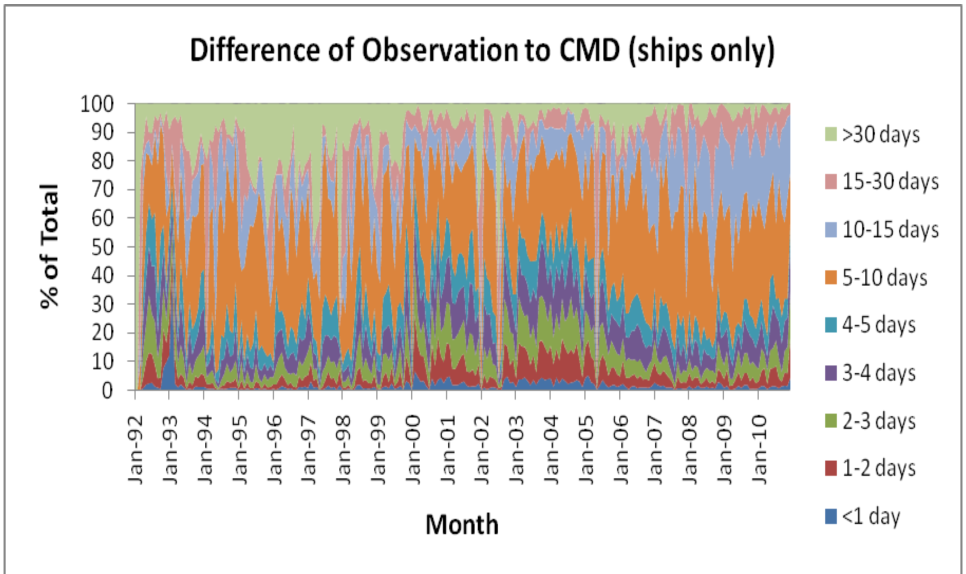

Figure 7: The time difference between observation and update to the CMD. This is generated from XBT only and only data reported from ships.

Figure 8 shows the same kind of display but now for profiling float data coming from the Argo project and reported as TESACs. For this chart, the time difference is measured between the bulletin time (the time the data were posted to the GTS) and the observation time. The use of profiling floats began earlier than 2000, but it was only at this time that a substantial number of floats began to be deployed. The Argo project has the stated goal to report all data to the GTS within 1 day of observation. At the beginning of 2007 they were hovering about the 90% mark and by the end of 2010 they were nearly 95% of the Argo data posted to the GTS. This is an improvement over last few years, and more improvements are expected.

In the case of Argo, fully automated QC procedures are carried out on the data prior to submission to the GTS. Some delays are experienced when profiles fail the automated procedures and manual intervention is required. Other delays are introduced when data are corrupted during transmission and must be recovered manually.

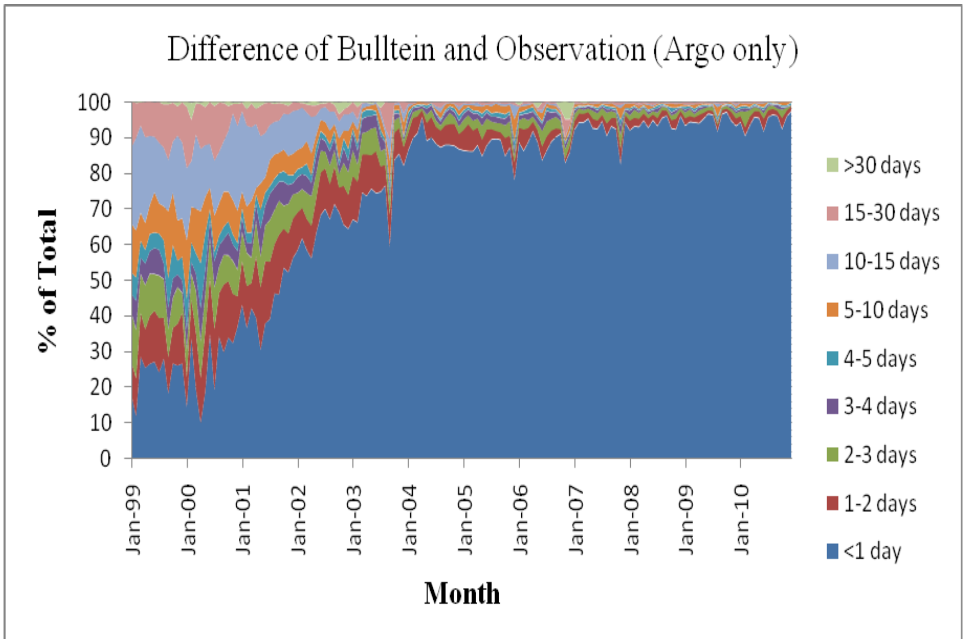

Figure 8: The time difference between observation and bulletin to the CMD. This is generated from TESACs only and only data reported from profiling floats

Of particular note in this figure is the strong dip in September of 2003. There is a similar dip in the Figure 8, but it is somewhat less noticeable. The reason for this drop was the large power failure that took place in the middle of August in 2003 in both Eastern Canada and the US. This delayed work of inserting float data onto the GTS in both the US and Canada.

Dealing with timeliness of delayed mode data is more difficult. Data can be at most 30 days old (or so) for real-time distribution. Any data older than this just does not get distributed. This makes for a clean cut off time and more importantly a clear upper limit to the volume of data expected.

For delayed mode data, the oldest date could be back to the time of the Challenger Expedition in 1873. As well, there is no known limit to the volume of data that may be received in delayed mode. Both of these make it difficult to measure success in receiving delayed mode data.

Figure 9 shows statistics derived from the delayed mode data in the GTSPP archives at NODC. The time axis shows the date of observation. The number of delayed mode data decrease from past to present consistent with what is shown in figure 9. It is also evident that in the early years of GTSPP, it was very common for data older than 5 years to be received by the project. In the mid to late 1990s, the major fraction of the data is received when they are 2 to 3 years old. In the more recent years, the delayed mode data that have arrived tend to do so within 1-2 years of their collection date.

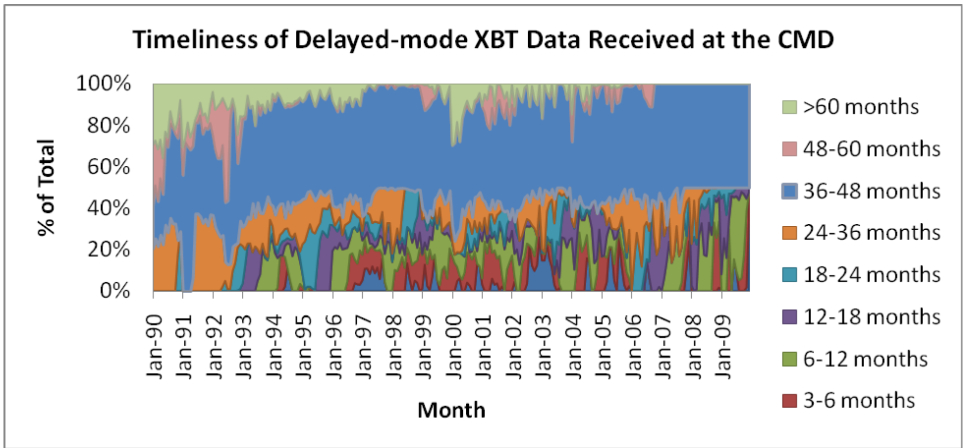

Figure 9: Timeliness of delayed mode data received at the CMD of GTSPP.

4.3 Data Quality

From the start, the GTSPP agreed to standardize the quality control procedures that were used and ensure that the quality information would be managed with the data. Within the GTSPP are both data centres and centres of oceanographic scientific expertise. Data centre QC is described in IOC Manuals and Guides #22. The first revised edition of the GTSPP Real-Time QC manual is available on-line at http://www.nodc.noaa.gov/GTSPP/document/qcmans/index.html

Scientific quality control is provided by collaborating science centres. CSIRO has produced a manual describing how to examine XBT data. It is available at the same Web page above.

The entire data resident in the CMD eventually passes through these two levels of scrutiny. Figure 10 shows the contents of the CMD where the relative volumes of data having gone through data centre QC and complete QC (science centre review) are shown.

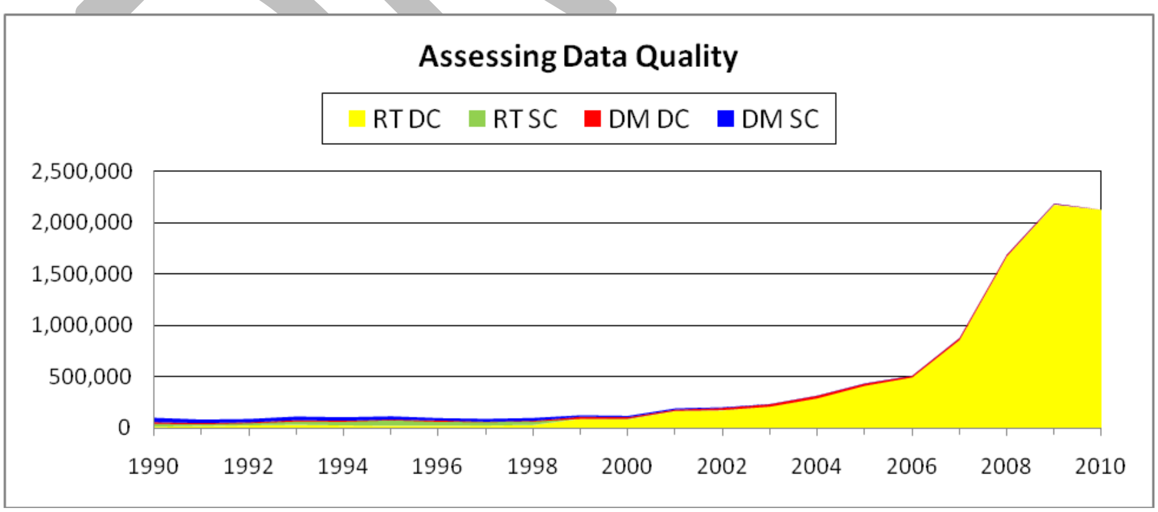

Figure 10: The numbers of real-time (RT) and delayed mode (DM) stations in the CMD having undergone quality control procedures at data centres (DC) and science centres (SC).

The review of data by science centres happens on a yearly basis, and there is always some fraction of data that escapes this process. The large jump in the numbers of stations having passed just through data centre QC in January 1999 reflects the deadlines to meet requirements for publishing the WOCE Data Resource V3. GTSPP participants continue to pursue getting the data through science centre QC.

5.0 Partnerships

5.1 Argo

GTSPP continued to handle Argo data in the same way as other data reported on the GTS and then in delayed mode. However, there is a closer association with Argo than this. The Argo data system relies on individual data assembly centers (DACs) to manage and contribute data both to the GTS and to the global data servers of Argo. Not all DACs begin operations with all capabilities in place. For some, the insertion of data to the GTS is handled by Service ARGOS while the contribution of the data to the global servers is delayed. GTSPP contributes the real-time data (having passed through GTSPP quality control procedures) to the global servers to provide at least a reduced form of the data at these servers until the originating DAC can start to send the data on their own. At the beginning of Argo, the Argo data contributed almost 18% to the data set (excluding the buoy stations data). As of December 2010, the contribution increased to about 30% at the CMD.

The quality control procedures of the GTSPP were the starting point of the automated procedures employed in the Argo program. Although the GTSPP procedures had been developed for XBT data, with suitable modifications they are reasonably effective at catching errors in float data.

The main data centers operating in GTSPP all have a significant role in Argo. The experience gained in organizing the GTSPP has been used in the design and implementation of many parts of the Argo data system.

5.2 JCOMM and GOOS

GTSPP started as a jointly sponsored program of WMO and IOC and so when JCOMM was formed it was adopted by the new commission. It reports through the Data Management Program Area but also contributes to the Ship Observation Team meetings. The experience in data management gained from GTSPP operations has been invaluable. It is an operational program that put in place a large number of elements to ensure broad support. It continues to contribute this experience in the deliberations that JCOMM are undertaking to assemble a global observation system.

In the early days of GOOS, GTSPP was recognized as an important program that was delivering on some components needed. It was for this reason that it was accepted as an Initial Observing System component.

GTSPP provides the infrastructure support in data management that is required to move the data from collectors to users in the time frames and with the level of quality and consistency that is needed. It therefore supports both JCOMM and GOOS needs.

5.3 CLIVAR

GTSPP acted as the data system in support of the WOCE Upper Ocean Thermal Data Assembly Centre. This was a natural extension to the support provided for SOOPIP. Because of this participation, GTSPP is taking part in CLIVAR. Initial contributions will be quite similar to that provided during WOCE. As the requirements for CLIVAR become clearer and different needs are expressed, operations of GTSPP will adjust.

6.0 Actions

6.1 Implementing a Unique Data Identifier

One of the most difficult problems faced by the GTSPP has been in matching real-time and delayed mode data from the same original observation. The problems stem from reduced vertical and measurement resolution reported in real-time messages and from uncertainties in positions and times as demonstrated by the levels of position and time errors shown earlier. The delayed mode data may have these errors corrected and so matching real-time to delayed mode is not simply a matter of matching ship identifier, position and time. The GTSPP developed software that considers detailed comparisons of individual station data when real-time and delayed mode positions are within 5 km distance and 15 minutes of time to each other. It assumes that errors in these quantities are not large. In a number of cases, the assumption is borne out, but not in every case. So, although a degree of success has been attained in matching real-time and delayed mode data, there is still room for improvement.

A new strategy was discussed at a GTSPP meeting in 2002. It was inspired by the Ocean Information Technology Pilot Project being undertaken by JCOMM and IODE. The solution was suggested by colleagues in Australia and hinges upon the use of a cyclic redundancy check (CRC) calculation. Since then, the GTSPP and the SEAS program in the US have been cooperating to install the necessary software to implement the solution.

The CRC has been incorporated into the US SEAS system. The CRC is a 32 bit value based on the ASCII generated BATHY message of those values following the 888 group and terminating at the equal (=) sign of the message. Development is concurrent with the development of the AOML automatic quality control software.

When an XBT is taken, SEAS shipboard software creates a binary record of the entire data stream, metadata, and computed unique SEAS ID for archive aboard ship. This is referred to as the “complete message”. The complete message is the delayed mode record sent to AOML and forwarded to NODC. SEAS shipboard software also creates a “best message” and SEAS ID for transmission to a land-based SEAS processing server.

The SEAS processing server builds two real-time messages from the best message. One is the usual BATHY record distributed on the GTS. The GTS record reaches ISDM and is incorporated into their GTSPP operation. ISDM will compute a CRC from the BATHY message using the exact algorithm used by SEAS and attaches it to the record. The other real-time message, called a real-time “archive message”, will be the same GTS record but with the SEAS ID and computed CRC of the GTS record attached. This archive record will be sent to NODC to become part of their GTSPP data management operation.

NODC receives two SEAS records from the SEAS system, the real-time archive message (SEAS ID + CRC ID) and the delayed mode complete message (SEAS ID). Comparison of the SEAS ID completes the data flow from NOAA. NODC will also receive a GTSPP record from ISDM which has the same CRC computed. Comparing the GTS CRC ID of the archive message to ISDM GTSPP record completes the GTSPP data flow.

7.0 Clients and Services

GTSPP operates World Wide Web (WWW) and Filter Transfer Protocol (FTP) sites. In addition, some clients require regular downloads of data and for these there is a subscription service.

7.1 Subscription Services

Some of GTSPP's clients require data as soon as possible after the data have been distributed. Since ISDM operates the real-time component of data assembly and as described above, carry out quality control and duplicates resolution 3 times per week. As the updated files are sent to the CMD at the US NODC, a number of clients receive global or regional updates as well. These clients include The Australian Bureau of Meteorology, The US GODAE Server, the US NAVOCEAN, the French Coriolis Project, the NEAR-GOOS Project and WDC-D. ISDM either initiates an ftp service to place files on the client site, or places files on its own ftp site for download by the client.

A similar service is offered by the US NODC with the European Centre for Mid-range Weather Forecasting and the US National Centre of Environmental Prediction being the major clients.

Both ISDM and NODC offer data downloads on a request basis. It is normal for these to be supported on an anonymous ftp site.

7.2 WWW and FTP Services

NODC maintains the Web site for GTSPP and keeps access logs for the site. An analysis of WWW and FTP logs provides the following information about our users.

On average, more than 317 and 293 GTSPP pages were accessed each day in 2009 and 2010, respectively. Of course, a number of these are by various "web-bots" harvesting information for web search engines such as Google. Beyond these, there were a number of references to the site from educational organizations both in the US and abroad. There were also referrals from international organizations such as the IOC and WMO.

It is difficult to know how much of the web traffic is composed of people clicking and moving on or those more genuinely interested in the page content. However, it is possible to track the size of downloads and the files of highest interest. So, over 2009 there were 42,820 requests (about 15% of total page requests) that downloaded between 1 – 10 Mbytes (about 16% of total bytes downloaded), 15,685 requests (about 5% of total page requests) that downloaded between 10 – 100 Mbytes (about 42% of total bytes downloaded) and 2,580 requests that downloaded between 100 Mbytes and 1 GByte (about 40% of total bytes downloaded). Over 2010, there were about about 24% of total bytes downloaded between 1 – 10 Mbytes, about 31% of total bytes downloaded between 10 – 100 Mbytes and about 40% of total bytes downloaded between 100 Mbytes and 1 GByte. The compressed, .zip and .gz, files consist of 97% and 88% of total bytes downloaded over 2009 and 2010, respectively. The usage statistics of the GTSPP FTP server show about 96% of total bytes downloaded both in 2009 and 2010.

8.0 Acknowledgements

Many agencies have played important roles in the development of the GTSPP system. Each participating agency carries out a number of functions in handling the data for the GTSPP. The most important contributors are the collectors of the original data. Without their efforts, this compilation of data and information would not have been possible. This work was partially supported by the NODC management.