Chapter 8

IN-SITU OBSERVATIONS: OPERATIONAL SYSTEMS AND DATA MANAGEMENT

Sylvie Pouliquen

IFREMER, Centre de Brest, Plouzané, France

Abstract: This paper presents, through existing examples, the main characteristics of operational in-situ observing systems and the data management issues to be addressed for operational oceanography needs. It provides the main characteristics of an operational in situ observing system in comparison with a research one in term of sustainability, coverage, timeliness, implementation issues and international coordination. It highlights the main features that have to be put in place for operational system data management and differences between different architectures that are nowadays operated.

Keywords: In-situ, observing systems, data management, quality control, data formats, ARGO, GOSUD, OceanSites.

1. Introduction

Scientists, fishermen, navigators… have been observing the oceans since the middle of the 19th century for their own needs (to enhance safety, to improve transit time, to understand some phenomena, etc). But this has often been done in an unorganized way, shared only among small communities, measured over limited areas and periods of time: a lot of data have thus been lost or are too incomplete to be used by the community nowadays.

Because it has been demonstrated that ocean and atmosphere behaviour are clearly linked together, it is mandatory to observe and understand the oceans the same way it has been done for the atmosphere since the 20th century. That is why individuals, research groups, and nations have started to work together to overcome the technical and logistical challenges associated with carrying out joint routine measurements of the global-ocean.

While satellites are providing a global view of the surface of the ocean, it is important to set up in-situ systems to monitor their interior (e.g. Send this volume). Basically, the following are needed:

E. P. Chassignet and J. Verron (eds.), Ocean Weather Forecasting, 207-227.

- Autonomous instruments (moorings, drifters, profiling floats, gliders, etc) to monitor on long period of times
- Regular ship measurements to monitor long repeat sections,
- In order to have all these data available for operational models: a well-designed and robust observing system, good communication to shore to deliver data rapidly,
- Real time operational data centres,
- Suitable data protocols to distribute data to operational centres in a timely way,
- International cooperation to achieve a global coverage, set up an adequate system and maintain it in the long term.

2. Essential features of operational oceanography systems

The goal of operational oceanography is to provide routine ocean nowcasts and forecasts and analysis on timescales of days to seasons, from global to regional and coastal regions. To address the operational oceanography needs, in-situ observing systems must comply with the following requirements.

2.1 Coverage

The observing systems to be put in place are different depending on the area and the phenomena to be sampled. We usually sort observing system into 3 categories:

- Global: System designed to provide data all over the ocean (e.g ARGO for general circulation). Such a system can only be built at the international level and is complementary to observations made from space. It is built to resolve climate scale phenomena with sufficient resolution and accuracy and provides systematic upper ocean observations of a limited number of parameters (temperature, salinity, …) on a time scale from 10 days to 1 month. International collaboration is the key factor for the success of such a network because none of the countries is able to cover the globe alone, but each country has to set up elements compatible and guaranteed on the long term for their contribution to the system.

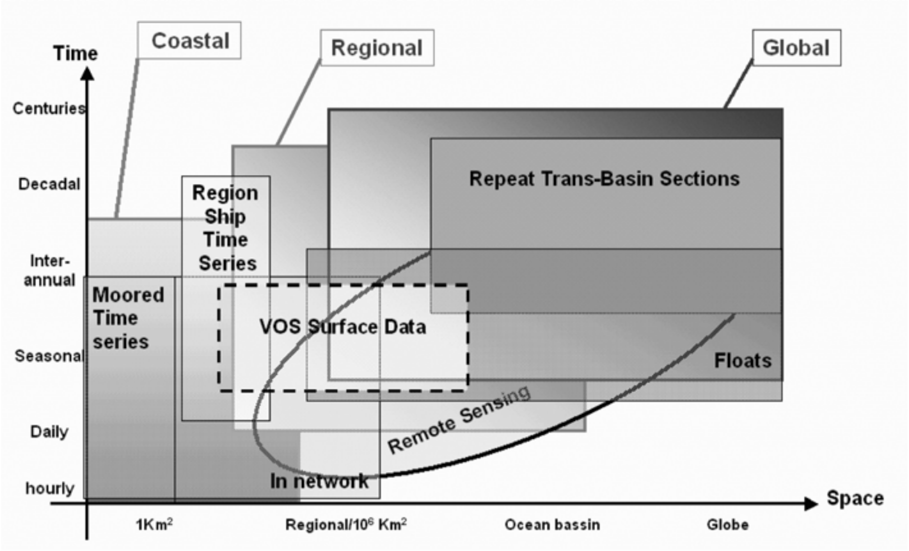

Figure 1. Platforms to use according to phenomena to sample and the kind of network to be set up (Global, Regional, Coastal).

- Regional: System designed to provide data in a specific area to monitor specific phenomena (e.g., TAO/TRITON/PIRATA Array for El Nino detection, Artic buoy network for Ice monitoring, etc). Generally it is set up in collaboration with few countries (less than 10) and the number of parameters is more important (between 10 and 20), including ocean (both physical and bio-chemical) and meteorological measurements. The time sampling is often higher rate: from hours to days.
- Coastal: These observing systems are usually set up at the national level to answer very specific questions such as coastal monitoring of the water quality or wind/wave/tide monitoring in harbour areas, etc. There is very poor collaboration among countries and these data are often used exclusively by the coastal models that have led to setting up the system. The technical issues to be solved are much more complicated such as bio-fouling (micro-organisms growing on the sensors and perturbating the measurements), interference by fishermen or other ships in area. A lot of technological work is under development in this field especially to set up cabled systems linked to shore with very high-speed networks.

2.2 Timeliness

What does real time mean for operational oceanography? The main criterion is to define the delay between measurements and assimilation beyond which the measurement adds nothing to the performance of the model. There is no unique answer: this depends on the type of models, the variables that are assimilated, the forecast product and the application for which it is produced. For instance, assimilated information of deep ocean temperature and salinity will persist within an ocean global circulation model for weeks or months and so a delay of several days in supplying data can be acceptable. On the other hand, oceans mixed layers vary on more rapid timescales in response to the diurnal heating and to storms. The impact of such data will probably not persist more than 3-5 days after assimilation, so measurements are needed within a day. As a compromise, real time for operational oceanography generally means availability within 1 or 2 days from acquisition, to allow data centres to better qualify the data even if it takes a bit more time. For climate applications larger delays are acceptable, but the length of the observation period is critical.

2.3 Agreed procedures and standards

Operational models use a wide variety of data from a diverse sources including buoys, drifters, ARGO floats, regional ships of opportunity, coastal observatories and even isolated local measurements made either by nations or scientists, as long as the data are easily available and quality-controlled in a timely way. Sea observations are very expensive and no country is able to sustain alone, the network needed by operational oceanography at the global level. Moreover it is important to design a system able to serve different communities: e.g. research, climate and operational communities. Therefore, an international coordination is needed.

In 1950, the meteorological community has set up an organisation WMO (World Meteorological Organisation) to organize this partnership for the meteorological needs. Under the auspice of Unesco, IOC(Intergovernmental Oceanographic Commission) has played an essential role in defining measurement standards and formats. The JCOMM (Joint WMO/IOC commission for operational oceanography and marine meteorology) has been set up to strengthen the role of WMO and IOC in the field of ocean and marine meteorology. It is involved in the main observing systems used nowadays by operational models:

- Surface data: DBCP (Drifters), VOS (Voluntary Observing ships),

- Sub-surface: ARGO (Profiling floats), TAO, GTSPP (Global Temperature and Salinity Pilot Project)
- Sea-Level: GLOSS (Sea Level)

Being able to collect and share the acquired data and distribute them to the user community requires significant work of normalisation/coordination on data collection and format (from metadata to profile and timeserie datasets), on quality control procedures, as well as on networking organisation to make these data circulate efficiently. Several concurrent attempts of normalization for metadata description (ISO 19115, GXML, XML-Marine, etc.) or data format and access systems are underway, both at national and international levels, but there is still no convergence towards a unique general agreement.

3. Implementation issues

When an observing system, often set up and maintained by scientific teams, moves to operational status there are some requirements that need to be fulfilled.

3.1 Sustainability

First, an operational system is sustained in different ways. This regards funding of course, as they are often expensive networks: new funding mechanisms have to be set up coming from sources other than R&D. Not all countries are organized in such a way that a transition to operational is easy: for example it is the case between NSF or NASA and NOAA in the USA, between ESA and Eumetsat for earth observations in Europe. Systems must be sustained also in terms of the operation: this goes from deployment planning, at-sea servicing (this requires ship and engineering teams to perform these activities), to data processing that has to move from R&D laboratories to operational data centres who are committed to do such tasks in the long-term. It is not always easy to find the institutions that are, in each contributing country, mandated or capable or willing to perform these tasks.

3.2 System maintenance

Second, work to maintain and operate such a network has to be coordinated at the international level with a clearly identified Project Office. This Project Office interacts with the contributing countries to update the implementation plans and secure the fundings. It coordinates the national activities with an internationally agreed framework. It interacts with other international bodies to integrate this system in a wider perspective.

3.3 Data management

Finally data processing and distribution must be designed properly to be able to deliver the data in time for operational use. First, data have to be publicly available in real-time for forecasting activities, and within a few months for re-analysis purposes. This is a revolution in a scientific community where scientists have kept data private for years until they publish and sometimes forever. This is an important data policy element to be solved by the funding agencies at national and international levels. Second is the organisation of the data flow among the different contributors in order to have an efficient data management network able to answer the operational needs listed above. For a long time, data management aspects have been neglected in projects and a too small funding was devoted to this activity both for in-situ and satellite data processing. With the arrival of operational ocean systems, the question has started to be crucial and examples like WOCE have shown that it was very energy demanding to get integrated quality-controlled data sets when it is not organized from the beginning. It is now clear that operational observing systems have to be processed by professional data centres that are sustained in the long term, that distribution has to be tailored to fulfil operational user needs

All the above has lead to the fact that attached to an observing system, there must be an effective management structure to address the implementation, coordination, data management, advertising and funding issues.

4. Prime examples of observing systems

4.1 ARGO project

To monitor and understand ocean circulation and water mass characterization on a global scale, systematic observation of temperature and salinity are essential. In the 90's, during the WOCE program, a new instrument was developed: the autonomous profiling float. Now the technology has become mature enough to start implementing an ambitious program that would deploy a large number of these instruments to cover the global ocean: the ARGO program was born. It aims to deploy and maintain an array of 3000 autonomous floats, one per 3°x3° box, measuring temperature and salinity from 2000m to the surface every 10 days for 3 to 5 years. Assimilated in models together with Sea Surface Height Anomalies from altimetry, they have become an essential network for operational models. This program started from an initiative of a group of scientists who were convinced of the importance of such a network. It was and is still partially funded on research money but a lot of work is done at the inter-governmental level to find funds to sustain such a network. A float costs about 15.000$, so the setting up of the array will cost about 50.000.000$ and about 10.000.000$ is needed each year to maintain it (700 new floats each year to replace the dead ones). These numbers do not include any additional cost to deploy these floats as the deployments are often done through free opportunities.

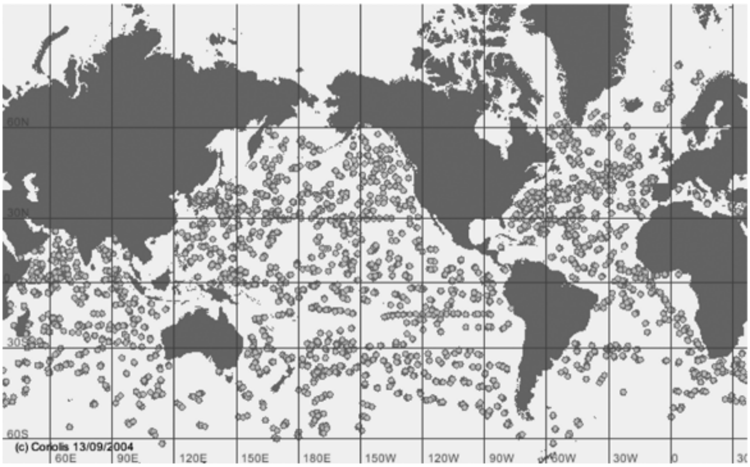

Figure 2. ARGO coverage in mid-September 2004: 1366 active platforms

In 2004 about 40 % of the network is deployed with a good coverage of the northern hemisphere and more work is to be done in the southern ocean. The keys to the success include:

- First an efficient coordination at the implementation level: deployment plans are consolidated by ocean basin to achieve uniform coverage of the float array. Good collaboration has also been set up to facilitate these deployment activities among the countries involved.
- A collaboration at the scientific and technological levels to improve the quality of the instruments, to detect deficiencies in time to avoid to deploy platforms when technological problems have been detected, and collaborative scientific work to develop delayed mode quality control methods to be used by the ARGO community.

- Finally an efficient data management system able to distribute the ARGO data in real-time within 24h from acquisition both on GTS (Global Telecommunication System used by the meteorological agencies) for meteorological community and in FTP for other operational users. This system is based on collaborative work between national data centres that process the data from the float deployed by their country, or partner countries, and a centralized data distribution through two Global Data Centres (one in CORIOLIS/France and one in FNMOC/USA). The architecture of this data management network will be presented at §5.1.1. Since 2004, the Data Management team is putting into operation the delayed mode procedures developed by the Science team.

The real challenges are now to secure funds on an operational and sustained budget to maintain this observing system. It is also to improve the technology to increase the lifetime of the platforms as well as their ability to survive in dangerous area such as partially ice covered regions.

4.2 SOO/VOS and GOSUD: Surface data

Merchant vessels are doing long ocean transects on regular basis and are good platforms to implement repetitive measurements. On the other hand, research vessels frequently traverse the oceans on routes where few other in-situ ocean observations are available. As such, they represent a natural and cost-effective mechanism to deliver routine oceanographic data along their way. These data include temperature and salinity at the surface as well as currents. Sea surface salinity (SSS) is an important parameter that is not yet measured from space.

In 2000, the GOSUD project was set up, under the IOC umbrella, as an end-to-end system for data collected from ships along their cross-ocean tracks. The goal of GOSUD is to develop and implement a data system for ocean surface data, to acquire and manage these data and to provide a mechanism to integrate with other types of data collected in the world oceans. It is complementary of the SOOP/VOS projects that under JCOMM umbrella that organize the data collection from Voluntary Observing Vessels or Ships of Opportunity. The project seeks to organize underway-surface data that are currently collected and to work with data collectors to improve data collection. These data, complementary to ARGO and OceanSites (see figure 3), will be one of the major ground truths for the calibration of the Salinity satellites SMOS and Aquarius.

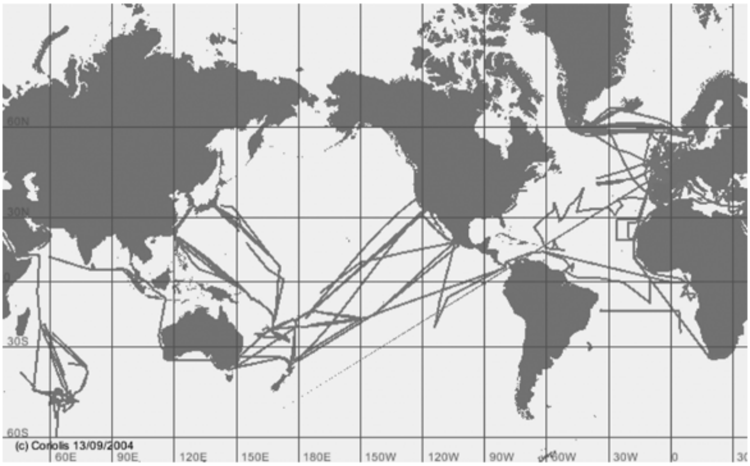

Figure 3. SSS data available for one year at CORIOLIS GDAC from 14 different vessels

In contrast to ARGO, GOSUD is not dealing with the implementation issues, which are handled by SOOP/VOS projects or national initiative (such as CORIOLIS for French research vessels). Moreover, it is building upon existing data centres that have to harmonize their quality control processing and coordinate the data distribution to ease the access to these data. The strategy used for GOSUD data distribution is similar to ARGO with distributed national data centres and two global data centres that act as "one stop shopping" points for users. 90% of the available data are distributed by Global Data Centers, as shown in figure 3.

4.3 OceanSITES

Another and complementary way to sample the space and time variability in a routine and sustained mode is to collect timeseries information at fixed locations in the ocean. The measured parameters are physical (temperature, salinity, current, etc), biochemical (oxygen, nutrients, fluorescence, carbon dioxid, etc) and atmospheric (air temperature and humidity, wind speed, etc).

In order to complement the good spatial coverage provided by ARGO and GOSUD, an international pilot project is under way for a global network of timeseries observatories, called OceanSITES. It plans to coordinate and implement a system of multidisciplinary deep-ocean sites, where sustained and publicly available data will be collected in a timeseries mode. A goal is to telemeter data in real-time (where feasible), and to interface and form synergies with the developing US OOI initiative. An important factor in turning these measurements into an operational system will be the harmonization, integration, and dissemination of the data collected. This effort is under way.

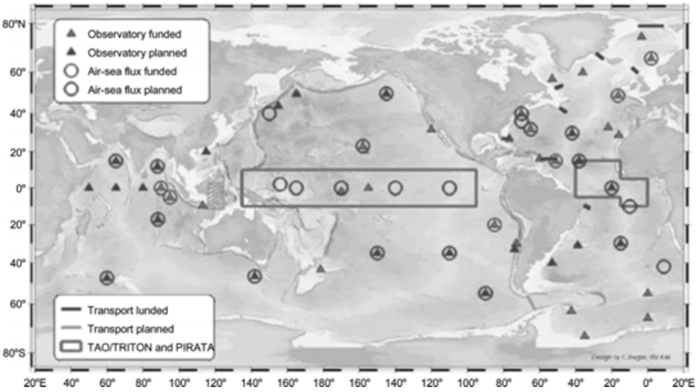

Figure 4. OceanSITES map

The current and planned state of the system is shown in figure 4. New sites have come online in the past 5 years, and an increased awareness of the importance of timeseries data has been created since OceanObs99. Two data centres are under development and a draft timeseries data format has been formulated.

At present only the surface met and TAO/TRITON/PIRATA data are used operationally, but the hope is that other data will be used for ocean forecasting once they are routinely available, once models are able to better assimilate point or integral timeseries data, and once more biogeochemical models are run operationally.

4.4 Comparison of these three systems

If we look at the criteria cited at the beginning of this paper, it is interesting to see how these three networks comply with these requirements.

	ARGO	GOSUD	OceanSITES
Sustainability	Only small part of it is sustained	Part on operational funds, part on R&D	R&D funding

Coverage	Global Network homogeneous coverage	Global to regional network. Good trans-basin coverage	Global network but very sparse coverage
Timeliness	Operational within 1 day for 85% of the data Used by operational models within Godae	From 1-2 days for vessels that have realtime transmission to month for the others. Data used for validation purposes at the moment	Operational for TAO/TRITON/PIRATA array that is used in operational models Mostly R&D for the rest of the sites because data not easily available.
International coordination	Well organised both for implementation and data management	Implementation organized at national level for implementation (France, USA, etc.). Starting to be well organized on data management level	International organisation is trying to be organized but it's hard to achieve.

5. Data management

At present, there is no consensus on data management and communication strategy for effectively integrating the wide variety of complex marine environmental measurements and observations across disciplines, institutions, and temporal and spatial scales. Data are obtained by diverse means (ships, drifters, floats, moorings, seafloor observatories, etc.), they come in very different forms, from a single variable measured in a single point to multivariate, four dimensional collections of data, that can represent from a few bytes a day to gigabytes

Even if an in-situ observing system were to make great measurements in a sustained way, if the data are not available easily to the operational users, they will not be used because they will not meet the operational modellers basic requirements: a data system for operational oceanography must provide quality controlled data, in a timely way, on a regular basis, according to procedures that are clearly documented and evolve upon common agreed decisions between user and provider.

There are three main characteristics for a data management system:

1. Its architecture
2. The quality control procedures
3. Data format and metadata attached to the data

5.1 System architectures

A data management system is designed according to the type of data handled (images/profiles/timeseries/kilobytes versus gigabytes, etc), the users access needs (individual measurements, geographical assess, integrated datasets, etc), the level of integration needed, etc.

In the past decade, with the improvement of the computer technology, the internet revolution, the increase of network speed and capacity, data management systems have been progressively moving from centralized to distributed systems. Two main architectures are nowadays commonly used:

- Distributed processing and centralized distribution: data are processed in different places and are than copied in a single place for distribution to users.
- Distributed processing and distribution: data are processed in different paces and stay where they are. To ease user access a virtual WWW portal is implemented that use networking techniques to find the data that fit the user needs.

Each system has its advantages and drawbacks, depending on the type of datasets to distribute and the contributors to the network. These different architectures will now be quickly described through examples operating at present.

5.1.1 ARGO data system: Distributed processing and centralized distribution

Within the ARGO data system, the float data processing is distributed among the contributing national data centres. They feed two global data centres (GDACs) automatically with the latest version of their float profiles, trajectories and metadata. Both GDACs are updated simultaneously to ensure consistency between the two datasets. They synchronise their holdings each night [in case a DAC (Data Assembly Centre) has updated one GDAC and not the other one].

Individual agencies (some acting on behalf of several countries) assemble the data collected from the communications system and carry out the initial processing of the data. Each file is under the responsibility of a single DAC (i.e. the data provider) who guarantees the quality and integrity of the data.

Data exchanges between DACs and GDAC are performed using a common data format. The main objective is for the users to access a unique data source (in this case, we have two servers for reliability/redundancy). A central website provides an extensive set of tools to query, retrieve, plot and compare the profiling float data dynamically. They also provide an FTP access for easy automatic data retrieval by users.

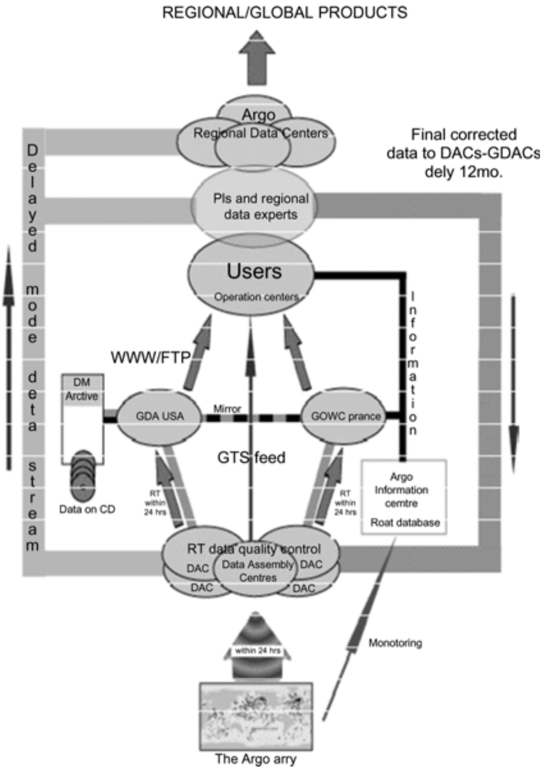

Figure 5. ARGO data management flowchart. On the right, the real time data stream; on the lef the delayed mode data stream.

In the ARGO data flow, there are two loops. One, in real time, on the right on figure 5: DAC qualify data in real time (see §5.2.1) in a semi-automated way and feed two GDACs at the same time that good measurements are put on GTS (Global Telecommunication System used by all meteorological offices). The second loop is in delayed mode, on the left on the figure 5, within one year, data are scientifically quality-controlled (see §5.2.2) and eventually corrected by scientists before being sent again to GDACs by the DACs in charge of these floats processing.

The advantages of such a system are:

- One stop shopping for the users where they get the best available data for ARGO in an unique format
- Data discovery and sub-setting tools are easy to implement as all the data are in the same place
- A robust system, as the probability that both GDACs fail is very small
- Easy to guaranty a quality of service in data delivery because GDAC have the control of all the elements in-house

The disadvantages are:

- Data are moved around the network and must rely on the "professionalism" of the DACs involved in the system to be sure that GDACs have the best profiles available.
- Additional work at DAC level to convert their data from their home format to the ARGO format. This may be hard to do for small entities.
- Data format used for data exchange cannot evolve easily as it requires coordination among all actors before implementation. Since users, especially operational ones, do not like format changes it is not such a big problem.
- If only one main server is set up than the system is fragile. Setting up a mirroring system can over pass this problem with additional synchronisation mechanisms.

5.1.2 Ocean US: Distributed data processing and distribution

The USA are developing an Integrated Ocean Observing System (IOOS) ranging from global to regional to coastal. The purpose is to integrate existing and planned observing systems that address both research and operational needs. Considering the diversity of actors and of parameters involved, this system must be a cooperative integration of independent systems that will continue their missions independently while participating in an integrated data system.

It is clear that in such a system the data processing is distributed and the data stay on physically distributed repositories, some containing huge amounts of data. The user connecting to the Ocean.US website will be able to query for data without knowing where they physically reside.

The key elements of such a system are the metadata management, the data discovery system and the data transport protocols.

- **Metadata management**: Metadata describes the data. Certain classes of metadata (variable names, units, coordinates, etc.) are mandatory to any utilization of the data, and must be tightly bound to data transport as an integral part of the delivery protocols. Other types of information, such as descriptions of measurement and analysis techniques, help to place the data in context and are essential to the overall understanding and usefulness. To be able to share data among a network it is mandatory to have a common vocabulary. Some international groups are working together to build such norms: FGDC and ISO19115 are the most common for geospatial data. As a lot of system pre-exist to Ocean US, it is mandatory to develop translation mechanism to build metadata catalogues that will be used by the Data discovery system.

- **Data discovery**: it is the way to locate data of interest for the user. This search is done by scanning the metadata catalogue. Depending on the possibilities that the system wants to offer the user, the metadata data stored in the catalogues can be more or less precise. This use of metadata is comparable to the indexing of catalogue records within a library to help users to locate books of interest. The common data discovery systems typically allow selecting the available data for a set of parameters, on a geographical area, within a period of time. In future, "data mining" techniques will offer search on semantic criteria ("I want a cloud free AVHRR image of SST over this area in March 2004 together with SST from drifters acquired in same area at same time").
- **Data transport protocols**: these are protocol between a user or a system who wants data and a data repository that stores the data. It is in this field that significant improvements have been made with Internet revolution and the increase in network speed. These protocols are mainly based on available technologies ("web services", cgi, scripts, etc) based on current transfert protocols (HTTP, FTP, etc). Each data provider needs to serve its data and metadata through a common access interface, which can be achieved with existing softwares such as OpeNDap (alias DODS) Live Access software, etc. Direct access to these interfaces may require also specific software or libraries for the user. Although the datasets are distributed through various nodes within the system, setting up a centralized query system that will redirect the user requests to the relevant node can hide this to users.

The advantages of such system are:

- Optimisation of the resources (network, CPU, Memory, etc) among the contributors,
- Data stay where they are generated preventing non compatible duplicates among the network
- Built on internationally agreed standards that guaranty its efficiency in the long term and its adaptability because it will benefit from international shared developments.

The disadvantages are:

- The system is not easy to set up because it needs a lot of international coordination, especially for metadata.
- Even more work for small contributors because it requires important computer expertise
- It can be unreliable if some data providers cannot guaranty data service on the long term. To be reliable such a system must rely on sustained data centres.

5.2 Quality control procedures

These procedures have to be adapted to the allowed delay of the delivery. In real-time, most of these QC are made automatically and only outliers are rejected. In delayed mode, more scientific expertise is applied to the data and error estimation can be provided with the data.

Data quality control is a fundamental component of any ocean data assimilation system because accepting erroneous data can cause incorrect forecast, but rejecting extreme data can also lead to erroneous forecast by missing important events or anomalous features.

The challenge of quality control is to check the input data against a pre-established "ground truth". But who really knows this truth when we know that the ocean varies in time and space, but also that no instrument gives an exact value of any parameter but only an estimation of the "truth" within some error bars.

For operational oceanography, other problems must be solved. First, the forecast requires quality-controlled data within one day. This means that only automated or semi-automated quality control procedures can be applied. Second, most of the data are processed by different actors, but used all together by operational models: this implies a clear documentation of the quality control procedures, an homogenisation of the quality flags, a reliability of different actors in applying these rules. Third, for re-analysis purpose, the models need better QC'd data for which methods employing scientific expertise are used to correct the data (drift and offset) and to provide error estimates of the corrections. ARGO quality control procedures will be discussed to highlight the different aspects.

5.2.1 Real-time quality control procedures for ARGO

Because of the requirement for delivering data to users within 24 hours of the float reaching the surface, the quality control procedures on the real-time data are limited and automatic. 16 automatic tests divided in 4 categories:

- Gross error tests: date, position, float speed at drift, temperature, Salinity
- Profile coherence: decrease of the pressure, spike detection, excess gradient between two points, density inversion, constant value or overflow for T or S
- Coherence between profiles: jump or big drift in temperature or salinity between two cycles (see figure 7)
- Grey List: For the float in this list, all profiles must be checked by an operator because their behaviour is "strange"

5.2.2 Delayed mode quality procedure for ARGO

The free-moving nature of profiling floats means that most float measurements are without accompanying *in situ* "ground truth" values for absolute calibration (such as those afforded by shipboard CTD measurements). In general pressure sensors are regarded as good even if time drift may be possible; no agreed method exist yet for ARGO but the impact of pressure drift is not negligible: 5 dbar will result in a salinity drift of 0.003psu. Temperature sensors perform pretty well and similar method could be applied to detect temperature drifts.

ARGO salinity delayed-mode procedures rely on statistical methods for detecting artificial trends in float salinity measurements. However, since the ocean has inherent spatial and temporal variability, ARGO delayed-mode quality control is accurate only to within the associated statistical uncertainties.

Using 2-stage objective mapping methods, salinity data mapped from a historical database of existing profiles can be compared to float measurements. Careful analysis of the spatial and temporal scales of the mapping gives realistic confidence levels for the mapped values. A weighted average in the vertical (giving more weight to stable water masses) results in a single salinity offset for each float profile, as compared with the mapped data. Looking at the trend of these residuals allows detection of a sensor offset or a drift and quantification within error bars.

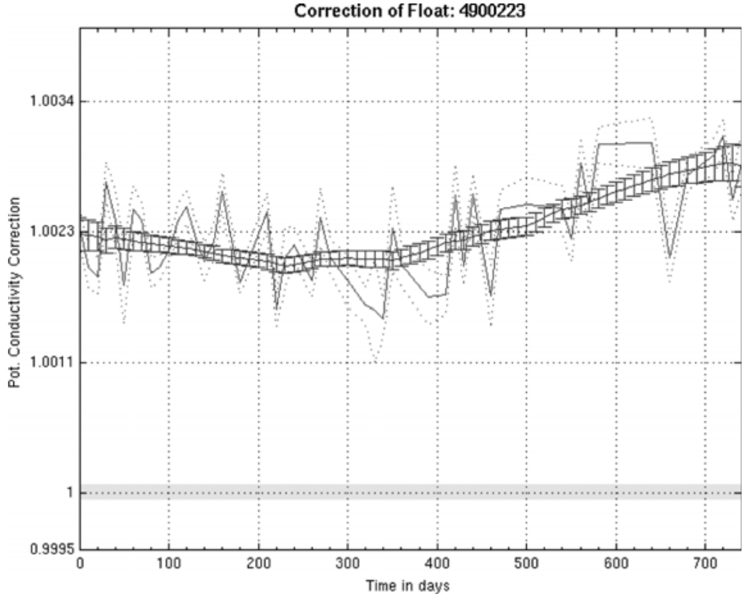

Figure 8. A float with an initial offset and which started to drift after one year at sea. The black line corresponds to individual cycle calculated corrections, The bars correspond to the proposed correction calculated by linear fit on a 6 month sliding window. When the proposed correction stays within the grey box limits (+/- 0.01PSU), no correction is applied.

Another statistical method also used to estimate sensor drift consist of calculating weekly analysis with all the available QC'd profiles coming from CTD, moorings, floats and monitoring the error residual for each float over time both in temperature and salinity by averaging these residuals on a number of levels. Such a method combines three methods: reference to climatology and history of the float as before but also collocation with other neighbouring floats. As it is unlikely that all floats in same area drift the same way, this method should help to QC float data even in areas where the climatology is poor.

5.3 Data formats

Data must be preserved in such a manner that they will still be useful in the future when the Pi that acquired the data may have moved somewhere else. They must also be distributed in a way that a user can easily merge it with other datasets relevant for his application. They must help to find the data among the network (data catalogues). That is the purpose of defining correctly distribution data format as well as the metadata (data on the data) that need to be preserved for future processing.

Data format have always been a nightmare both for users and data managers and they are both dreaming of the "Esperanto" of data format. Computer technology has improved a lot in the past decade and we are slowly moving from ASCII format (easy to use by human eyes but not for softwares), to binary format (easy for software but not shareable among platforms (Windows, Unix, etc), and self-descriptive, multiplatform formats (Netcdf, Hdf, etc) that allow more flexibility in sharing data among a network and are read by all softwares that are commonly used by scientists.

Depending on who is using the oceanographic data, the information stored in a dataset can be more or less precise. When a scientist is using data that he has acquired himself on a cruise, he has a lot of additional information (often in his head) and he is mainly interested by the measurements themselves. When he starts to share with other persons from his laboratory he has to tell them how he took the measurements, from which platform, what the sea-state was that day, what are the corrections he applied on the raw data, etc in order for his colleagues to use the data properly and understand differences with other datasets. When these data are made available to a larger community the number of necessary additional information, to be stored with the data themselves, increase, especially when climatological or long-run re-analysis are some of the targeted applications. This is why nowadays a lot of metadata are attached to any data shared among a community.

One important point for metadata is to identify a common vocabulary to record most of these information. This is pretty easy to achieve for a specific community such as ARGO, but it starts to be a bit more difficult when we want to address multidisciplinary datasets such as mooring data. To help community in this area some metadata standards are emerging for the marine community with Marine XML under ICES/IOC umbrella and ISO19115 norm.

Another important point in data format is to keep, together with the data, the history of the processing and corrections that have been applied to it. This is the purpose to the history-records that track what happened and allow going back to data centres to ask for a previous version if a user wants to perform his own processing from an earlier stage.

6. Conclusion

This paper has shown that the expectations regarding in-situ observing systems are very high and that they are not easy to set-up: in-situ observations are very expensive, diverse and made by laboratories all around the world. Some pre-operational systems, such as the TAO/TRITON/PIRATA array or ARGO float program, are managing to comply with some of the operational oceanography requirements which are sustainability in time, adequate coverage timeliness of data delivery, coordination both at implementation and data management level.

This paper has also addressed some issues related to data management such as the different data distribution architecture, the necessity of common agreed quality control procedures both in real time and delayed mode and the importance of data and metadata standardisation if we want to be able to share efficiently these data among the network.

References (including WWW sites)

Böhme L., 2003: Quality Control of Profiling Float Data in the Subpolar North Atlantic. Diploma thesis, Christian-Albrechts-Universität Kiel.

Briscoe, M., et al., 2001: Round Table 3: Access to oceanographic data, GODAE, Observing the ocean in the 21st century, p 419.

IGOS, 2004: A Strategy to Realise a Coordinated System of Integrated Carbon Cycle Observations.

Le Traon, P.Y., et al., 2001: Operational Oceanography and Prediction: a GODAE perspective, GODAE, Observing the ocean in the 21st century, p 529.

Ocean US, 2004: Data Management and Communication Plan for Research and Operational Integrated Ocean Observing Systems, http://www.dmac.oceans.us.

Robinson, I., et al, 2003: Observational requirements for inputs to European Ocean Forecasting System Models, Mersea-Strand1 EU project.

Robinson, I., et al, 2004: Measurement Challenges for European Ocean Observing System, Mersea-Strand1 EU project.

Roemmich, D., et al., 2001: The Global Array of Profiling Floats, GODAE, Observing the ocean in the 21st century, p 248.

Roemmich, D., S. Riser, R. Davis, and Y. Desaubies, 2004: Autonomous profiling floats: Workhorse for broad scale ocean observations. *Marine Technology Society Journal,* **38**, 31-39.

Send, U., et al., 2001: Oceanographic Time Series Observatories, GODAE, Observing the ocean in the 21st century, p 376.

Smith, N.R., and C.J Koblinsky, 2001: The Ocean Observing System for the 21st Century: a Concensus Statement, GODAE, Observing the ocean in the 21st century, p 1.

Wong. A.P.S., G.C. Johnson and W.B. Owens, 2003: Delayed-Mode Calibration of Autonomous CTD Profiling Float salinity Data by Theta-S Climatology. *J. Atmos. Oceanic Technol.*, **20**, 308-318.

ARGO WWW site: www.argo.net

CORIOLIS WWW site: www.coriolis.eu.org

Gosud WWW site: http://www.ifremer.fr/sismer/program/gosud/

OceanSITES www site: http://www.oceansites.org/OceanSITES/index.html

GODAE WWW site: http://www.bom.gov.au/bmrc/ocean/GODAE/

JCOMM WWW site: http://www.wmo.ch/web/aom/marprog/index.htm

JCOMMOPS WWW site: http://w4.jcommops.org/cgi-bin/WebObjects/JCOMMOPS