SCIENCE AND TECHNOLOGY ORGANIZATION
CENTRE FOR MARITIME RESEARCH AND EXPERIMENTATION

Reprint Series **CMRE-PR-2019-113**

Sensor-driven glider data processing

Daniele Cecchi, Bartolome Garau, Elena Camossi,
Alessandro Berni, Emanuel Coelho

June 2019

Originally published in:

Proceedings of the OCEANS 2015 MTS/IEEE Conference, 18-21 May 2015, Genoa, Italy, doi: 10.1109/OCEANS-Genova.2015.7271466

About CMRE

The Centre for Maritime Research and Experimentation (CMRE) is a world-class NATO scientific research and experimentation facility located in La Spezia, Italy.

The CMRE was established by the North Atlantic Council on 1 July 2012 as part of the NATO Science & Technology Organization. The CMRE and its predecessors have served NATO for over 50 years as the SACLANT Anti-Submarine Warfare Centre, SACLANT Undersea Research Centre, NATO Undersea Research Centre (NURC) and now as part of the Science & Technology Organization.

CMRE conducts state-of-the-art scientific research and experimentation ranging from concept development to prototype demonstration in an operational environment and has produced leaders in ocean science, modelling and simulation, acoustics and other disciplines, as well as producing critical results and understanding that have been built into the operational concepts of NATO and the nations.

CMRE conducts hands-on scientific and engineering research for the direct benefit of its NATO Customers. It operates two research vessels that enable science and technology solutions to be explored and exploited at sea. The largest of these vessels, the NRV Alliance, is a global class vessel that is acoustically extremely quiet.

CMRE is a leading example of enabling nations to work more effectively and efficiently together by prioritizing national needs, focusing on research and technology challenges, both in and out of the maritime environment, through the collective Power of its world-class scientists, engineers, and specialized laboratories in collaboration with the many partners in and out of the scientific domain.

NOTE: The CMRE Reprint series reprints papers and articles published by CMRE authors in the open literature as an effort to widely disseminate CMRE products. Users are encouraged to cite the original article where possible.

Sensor-driven glider data processing

Daniele Cecchi, Bartolome Garau, Elena Camossi, Alessandro Berni and Emanuel Coelho
NATO Science & Technology Organization (STO)
Centre for Maritime Research & Experimentation (CMRE)
La Spezia, Italy

***Abstract*—Gliders data are an important source of observations for oceanographers, hence harmonization and interoperability among different acquisition systems are key enablers towards efficient data sharing and effective reuse. As of today, several organizations operate gliders and produce customized results. The paper presents the effort done at CMRE, in collaboration with research partners, for the definition of a self-contained data file that eases both human readability and automated processing. By relying on community recognized standards, highly interoperable glider data sets are generated, whilst preserving readability thanks to the standardized and straightforward internal organization of variables. The paper includes also a description of the processing of the principal oceanographic sensor, i.e., CTD (Conductivity, Temperature and Depth), and an overview of the glider data management lifecycle.**

***Keywords*—Underwater gliders; interoperability; standardization; data processing; data management; data fusion**

I. Introduction

The role of underwater gliders in oceanographic data collection is continuously increasing. The data gathered from gliders are routinely used in oceanographic research, for studying mesoscale and submesoscale processes and assimilated in oceanographic forecast models [1], [3], [14]. Other applications of gliders technology are water quality monitoring, pollution detection, marine mammals monitoring [4], ocean monitoring for climate change, defense and security.

With respect to traditional profilers, underwater gliders have several advantages, including a longer endurance, the capability to be relocated, the possibility to carry multiple sensors on a single vehicle, the minimum environmental impact and the autonomy, enabling gliders to operate for long periods in large areas, accomplishing multiple missions, even without a mother ship, and allowing for an increased spatial and temporal resolution with respect to traditional oceanographic instruments.

The paper will overview the glider data lifecycle, starting from raw data processing and up to glider data management. Glider data lifecycle is presented in Section II and glider data management is introduced. Afterwards, the rest of the paper discusses in detail the aspects related to data processing and glider data standardisation.

The data processing workflow describing the different processing levels for the glider data will be presented in Section III. This processing is organized through different output levels: the raw level (L0, measurements without correction), the processed level (L1, introduction of derived variables, data as trajectories) and the gridded level (L2, data represented as vertical profiles). Differences between the Real Time and the Delayed Mode workflows will be discussed regarding the amount of data available, different processing steps, products generation. Section IV introduces the sensor-driven approach to glider data processing leveraging data fusion to improve the usability of L1 files agglomerating variables generated or derived from a specific sensor.

To illustrate the concepts previously introduced, the processing of CTD sensor data will be discussed in the V Section. The differences between gliders and traditional oceanographic data collection instruments will be highlighted. The particular case of adaptive correction of thermal lag effect [6] will be discussed, proposing a solution to store the correction coefficients together with the data. This work will also discuss the main issues regarding standardization and Automated Real Time Quality Control of glider data.

In Section VI, the data management approach implemented at CMRE for glider data is described, introducing the CMRE data management system and the demonstration experiment carried out to improve standardization of glider data. In the execution of its program of work CMRE adopts standards-based open web services, open data formats and metadata standards for achieving efficient discovery and access to scientific data sets collected during oceanographic cruise, enhancing the exploitation of glider data and facilitating the interoperability with complimentary sources of observations. CMRE's approach to glider data management is aligned with community driven initiatives that leverage open formats for data representation used in Earth Observations, in particular the Everyone's Gliding Observatory (EGO) glider NetCDF data format [11] that has been proposed within the European project GROOM (EU FP7 Glider for Research Ocean Observation and Management). This proposal addresses the compliance of glider data in NetCDF with Climate and Forecast (CF) conventions [10] used at international level.

Finally, Section VII concludes the paper outlining future research directions.

II. Glider Data Lifecycle

Fig. 1 illustrates the key phases of glider data lifecycle from a user perspective. Acquired data are processed in real time, and sent to the data centre in quasi-real time (after every glider surfacing), while delayed mode data sets are processed and stored at once. Users and models access the data available from the data centre, and use facilities for discovery such as catalogue services and web user interfaces.

Data-driven activities have to be planned since the project design stage (cf. Figure 1, in order to define how data will be preserved, permanently identified, documented, what are the

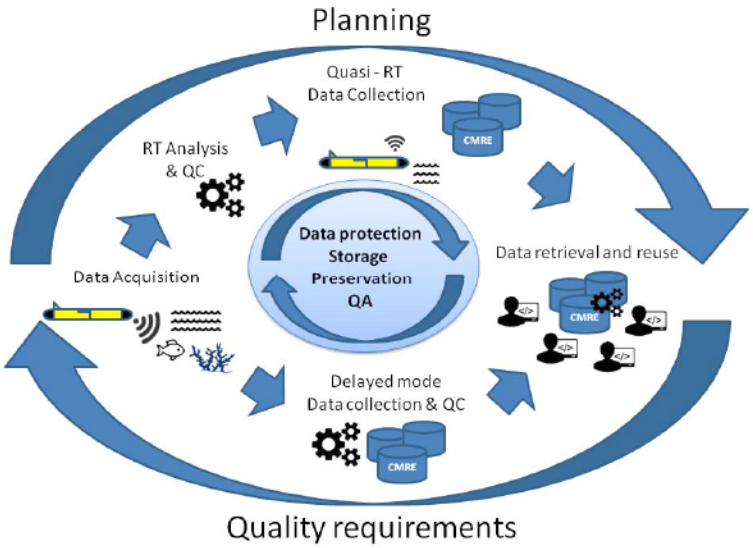

Fig. 1. Glider data lifecycle

data quality requirements and how they will be verified, how observations and derived data sets will be shared and reused along the whole life of data, and how these conditions will be periodically re-evaluated and by whom. The permanent identification of data sets is necessary to attain data provenance, to track data reuse, to obtain accreditation for shared data sets, and to support the reproducibility of experimental evaluations supporting publications.

Quality assurance is an integral part of data management. Quality checks applied to real time observations or to delayed mode data sets are performed to guarantee the data sets can be reused for assimilation and validation of ocean models. The EGO data management manual gives a reference framework for quality checks to apply to real-time and delayed mode data for vertical profiles and glider trajectories, mostly derived from the Argo Quality Control Manual [12]. Quality assurance complements data provenance and guarantees ocean modellers when reusing glider data acquired by networks of observing systems on the quality of observations, whilst enabling data providers willing to develop trusted systems to capitalise on shared high-quality information.

A fundamental aspect in data management is the definition of a consistent approach for data reuse and sharing. High quality metadata, standards and harmonised data formats play a major role to faciltate reuse, in particular in cases such as gliders', where the coverage offered by existing networks is sporadic and limited in terms of geographic area and data sharing initiatives are being put in place by oceanographic data centres in order to tackle global problems such as climate change and weather monitoring. CMRE endorses community-driven harmonization initiatives adopted at international level for glider data management to address the lack of standards [10], [11]. These initiatives, and the importance of high quality metadata for data sharing, are discussed in detail in Section VI.

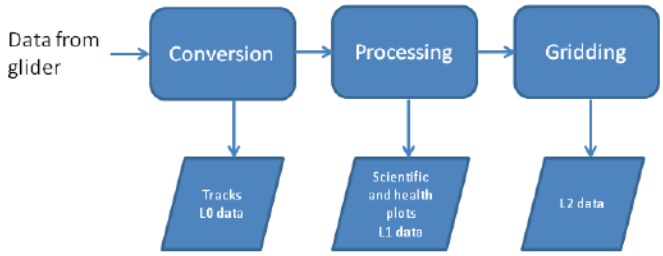

Fig. 2. Glider data processing flow.

III. Data Processing Workflow

The data processing workflow is a fundamental part of the data lifecycle illustrated in Fig. 1 and is the sequence of processing steps applied to glider raw data (in some cases available as binary data) to generate data files in a different format and with additional content. The paper focuses on the format of the generated files and on the content of the different output levels, disregarding the details of the processing steps.

The glider data processing chain implemented at CMRE produces the following outputs (cf. Fig. 2):

- Glider tracks for situational awareness
- Vehicle health status plots and scientific data automated imagery
- Glider data raw as NetCDF (Network Common Data Form) file (level 0, L0)
- Glider data processed as NetCDF file (level 1, L1)
- Glider data gridded as NetCDF file (level 2, L2)

Fig. 3 shows an example of tracks of several gliders as generated in the first block of the processing. The tracks are updated after each vehicle surfacing.

The level 0 data contain the glider raw data converted without any additional processing; all data samples are the glider

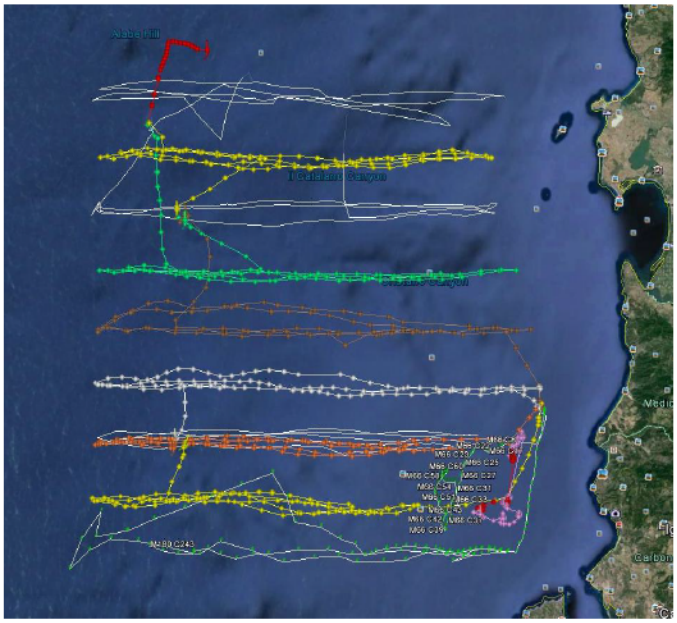

Fig. 3. Gliders tracks: different colors represent different vehicles. Points are the surfacings positions.

sensors measurements collected during the mission (errors included) stored as timeseries. This file represents the source for all the following algorithm steps and can be considered the starting point for any different processing that other users or scientists are interested to apply. Variables names are the same as in the glider; a standard name is added as variable's attribute whenever is available (cf. Section VI and Climate and Forecast conventions [10]). Units of measurement and any useful comment are stored as well as variable's attributes.

The level 1 data are the result of several processing steps starting from the level 0 data. The variables contained in level 1 files are data processed along vehicle trajectory. Each observation has its own time, latitude, longitude and depth. While time and depth are directly measured, coordinates underwater are in general estimated. Processing steps in the generation of level 1 files include some data cleaning (removal of outliers and clearly wrong measurements) and derivation of additional variables (i.e. salinity and potential density from pressure, temperature and conductivity measurements). The level 1 variables are the closest to the actual sampling (original resolution is maintained) and are different from traditional vertical profiles widely available in oceanographic historical data sets.

The level 2 data provides the data as vertical profiles interpolated at depth steps of 1 meter. An average position (latitude and longitude) and a reference time are assigned to every profile, so that they result vertical, instantaneous and regularly sampled. Derived variables are provided in level 2 files similarly to level 1. Level 2 vertical profiles are analogous to traditional CTD casts or profiling floats data and are directly usable by already existing oceanographic tools (*i.e.* Ocean Data View, ODV[1]).

The file format chosen to store glider data is NetCDF, a widely used format within the oceanographic community. For instance, it is the common format for oceanographic forecast models, because of its support to creation, access and sharing of array-oriented scientific data. A description of the NetCDF file format can be found in [13]. NetCDF files are self-describing because contains information about the data set; they are portable between different computer architectures; they are also scalable because they enable to access efficiently subsets of larger data set; they can be shared between multiple users (a single writer and several readers). The NetCDF output files contain the experiment description (as global attributes) and the variables definitions (including measurements units and standard names when applicable). The set of three levels NetCDF files are generated per glider deployment (with file size increasing with time in real time operation), after each successful glider data transmission.

A. Real Time and Delayed Mode processing

Two different workflows can be identified: the real time and the delayed mode processing. Real time data processing generates products for operational purposes, data assimilation in oceanographic or acoustic models and rapid environmental assessment. The real time chain should be automated, including if available, a Real Time Quality Control (RTQC) system. Small latency between data transmission from the asset at sea and availability of processed data is desirable. The delayed mode processing chain is in general used after vehicle recovery for more detailed analysis (using the data set with higher resolution when available), to populate historical databases and climatologies. A more accurate quality control strategy can be applied in this stage involving cross comparison between different vehicles and maybe different platforms sampling the same area.

For all processing levels, the file global attributes provide information regarding the overall content of the file and allows for data discovery. Below some attributes, like the owner of the vehicle, the specific platform, a summary of the experiment, the principal investigator, the data bounding box (including the vertical coordinate) and the temporal coverage of the data, are reported.

```
history = "2015-03-17 20:02:57: File
created"
data_type = "Glider time-series data"
featureType = "trajectory"
disclaimer = "my disclaimer"
product_description = "this file
description"
format_version = "0.99"
netcdf_version = "3.6"
source = "Glider observation"
data_mode = "D"
Conventions = "CF-1.6"
title = "myCruise"
type = "Glider data file"
references = "http://www.mysite.org"
provider = "my Institution"
summary = "myCruise glider data"
project_name = "myProject"
pi_name = "Principal Investigator"
```

[1]Ocean Data View http://odv.awi.de (Accesed march 2015)

```
cruise = "myCruise"
mission = "depl001"
processing = "post processing"
platform_code = "myGlider"
trans_system = "IRIDIUM"
positioning_system = "GPS"
platform_model = "MY GLIDER"
platform_maker = "MANUFACTURER OF MY
GLIDER"
anomaly = "none"
sensors_ctd_name = "unpumpedCTD"
sensors_ctd_manufacturer = "Seabird"
sensors_ctd_model = "SBE41CP"
sensors_ctd_serial_number = XXX
sensors_bbfl2s_name = "Fluorometer and
backscatter"
sensors_bbfl2s_manufacturer = "Wet Labs"
sensors_bbfl2s_model = "ECO Triplet"
sensors_bbfl2s_serial_number = XXX
sensors_bbfl2s_backscatter_wavelength =
"700nm"
sensors_bbfl2s_chlorophyll_ex_em =
"470/695nm"
sensors_bbfl2s_cdom_ex_em = "370/460nm"
sensors_bbfl2s_bb_cwo = "50 (clean water
offset)"
sensors_bbfl2s_chlor_cwo = "42 (clean
water offset)"
sensors_bbfl2s_cdom_cwo = "50 (clean water
offset)"
sensors_bbfl2s_bb_sf = "1.553 e-6 (scale
factor)"
sensors_bbfl2s_chlor_sf = "0.0072
ug/l/count (scale factor)"
sensors_bbfl2s_cdom_sf = "0.0907 ppb/count
(scale factor)"
launch_date = "09-Jun-2014 10:00:00"
contact = "contact@mysite.org"
distribution_statement = "Data release
is subject to xxx Memorandum Among
Participants (MAP)"
time_coverage_start =
"2014-06-09T14:29:01Z"
time_coverage_end = "2014-06-09T17:44:26Z"
geospatial_lat_min = 39.9494
geospatial_lat_max = 39.9792
geospatial_lon_min = 7.39364
geospatial_lon_max = 7.40013
geospatial_vertical_min = 0
geospatial_vertical_max = 174.63
```

Essential information regarding the scientific sensors installed onboard a specific glider are also stored as global attributes. Considering the previous example, it is possible to see that the glider is equipped with an un-pumped CTD and an optic sensor to measure fluorescence and backscatter (calibration coefficients are included in the global attributes). A very useful attribute is `anomaly` that can be used to store any event that can impact the quality of the data collected.

IV. Sensor-Driven approach

The traditional method to store glider processed data in a NetCDF file is to define a unique time dimension and to add all the other variables storing fill values at every timestamp when data are not available. Processing sensors (navigation, engineering and scientific) with different sampling rates will result in timeseries that are not aligned in time (i.e. not all the sensors have valid data at given timestamps). An example is represented by the Teledyne Webb Research Slocum glider data where navigation and engineering information are normally updated every 4 seconds and scientific sensors can be sampled at rate up to 1 hertz. The proposed approach consists in grouping variables per sensor, having the same sampling rate and reducing the required number of fill values. Data transmitted in real time are also different because a subsampling strategy is applied by the user to limit the size of files to be downloaded via the satellite link (remaining less time on surface). The aim of our work is to propose a data format easy to understand and to use for a scientist / researcher who is not a vehicle expert. As an example, the level 1 file resulting for a glider carrying two scientific sensors (a CTD and a optic one) will show a group of variables describing vehicle navigation and two groups of variables for the scientific data (one for the CTD and one for the optics data). Below some variables defining the vehicle navigation are reported.

```
double nav_time(navTime)
  long_name = "epoch time"
  standard_name = "time"
  units = "seconds since
    1970-01-01 00:00:00"
  FillValue = -1e+06
  QC_indicator = 0
double nav_latitude(navTime)
  long_name = "latitude"
  standard_name = "latitude"
  units = "degree_north"
  FillValue = -1e+06
  QC_indicator = 0
  reference = "WGS84"
  coordinate_reference_frame =
    "um:ogc:crs:EPSG:4326"
```

The attributes `standard_name` and `units` can be used to properly identify variables irrespectively of their name. The name preffix of the navigation group is `nav_`, the time dimension for the `nav_` group is `nav_time` (expressed in seconds since midnight 01/01/1970, epoch). The list of navigation group variables is:

```
double nav_time(navTime)
double nav_latitude(navTime)
double nav_longitude(navTime)
double nav_positionQC(navTime)
double nav_heading(navTime)
double nav_pitch(navTime)
double nav_roll(navTime)
double nav_depth(navTime)
double nav_distance(navTime)
double nav_profileIndex(navTime)
char nav_profileDir(navTime)
```

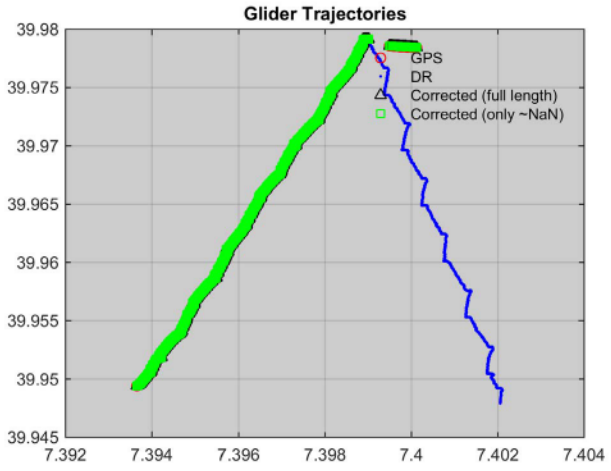

Fig. 4. Glider navigation: real time dead reckoning vs. post processing estimation.

All the variables in the list have attributes similar to the ones shown for `nav_time` and `nav_latitude`. The glider trajectory is defined by longitude, latitude and depth. Vehicle attitude is also reported because can be useful for further analysis. The depth is estimated based on pressure measurements and latitude, while longitude and latitude are known only when the glider is on surface (GPS fixes). The underwater coordinates are estimated. Depending on the available data it is possible to define two approaches: a simple one assuming linear interpolation between GPS fixes (can be adopted in real time); a more accurate method that evaluates the error between the dead reckoning navigation solution and the GPS fixes and consider the error being a linear function increasing with time when the vehicle is underwater. The navigation error is then estimated at each underwater timestamp and the dead reckoning navigation solution is corrected consequently. The vehicle attitude remains the one measured by the attitude sensor. Fig. 4 shows the dead reckoning navigation solution against the estimated navigation obtained with the more accurate approach. Quality flags are then applied to the estimated coordinates accordingly to table I (derived from the Argo quality control flags).

TABLE I. POSITION QUALITY CONTROL FLAGS

Flag	Value
0	Quality Control not performed
1	Good value (GPS fix)
4	Bad value
8	Interpolated value
9	NaN

Consider now the CTD variables group. Glider navigation data (estimated) and CTD measurements are not simultaneous. The proposed approach consists in interpolating navigation on measurements timestamps. The goal is to provide all CTD samples (and the related derived variables) with the best guess coordinates and vehicle attitude. The CTD variables group is reported below.

```
double ctd_time(ctdTime)
double ctd_cond(ctdTime)
double ctd_temp(ctdTime)
double ctd_pres(ctdTime)
double ctd_longitude(ctdTime)
double ctd_latitude(ctdTime)
double ctd_depth(ctdTime)
double ctd_pitch(ctdTime)
double ctd_profileIndex(ctdTime)
char ctd_profileDir(ctdTime)
double ctd_positionQC(ctdTime)
double ctd_distance(ctdTime)
double ctd_heading(ctdTime)
double ctd_roll(ctdTime)
double ctd_salt(ctdTime)
double ctd_pden(ctdTime)
double ctd_ptmp(ctdTime)
double ctd_svel(ctdTime)
```

All the listed variables have attributes as the variables of `nav_` group. Please note that the time dimension for the `ctd_` group is `ctdTime` and is different from `navTime`. It is possible to notice that for all samples the following data are available: time, coordinates, depth, glider attitude and derived variables like salinity, potential density and sound velocity.

The general approach to follow to process a scientific sensor is: 1) to estimate vehicle navigation and generate the navigation variables; 2) to geo-reference all sensor data and apply sensor specific processing (*i.e.* estimate derived variables). Data fusion between different scientific sensors is not currently implemented in the automated processing chain.

V. DATA PROCESSING: THE CTD CASE

One of the most common sensor that the underwater gliders are equipped with is the CTD module. In fact, it comes by default in almost all the commercial models available on the market. That means that it is one of the most extended and used sensor on this type of platform. CTD modules have been widely used throughout the years from other platforms, specially from research vessels. However, there are differences between the CTDs being operated from a ship and the CTDs mounted on gliders that impose some limitations when processing the data.

In particular, even if there are several different models of CTD on board gliders, many of them require the water parcel being measured to go through a free flushed or un-pumped pipe. The lack of pump reduces the energy consumed by the sensor, but the flow speed depends on glider surge speed which is usually unknown. This leads to uncertainty in the thermal properties of the sensor. This limitation is being mitigated, since the pumped CTD version is being recommended and installed by the manufacturers. Many gliders are still equipped with the un-pumped version, and there are a huge amount of historical data sets collected with gliders equipped with the un-pumped CTD that need to be addressed. Another limitation is that the gliders CTD sampling has a low temporal resolution in comparison to the high resolution of those operated from ships. That imposes a limitation on the features that can be

observed in the data, especially when the time scale of these features is comparable to the sampling period. Moreover, the glider CTD sampling interval is not perfectly regular, which makes more difficult the application of traditional processing techniques, because many of the filtering methods assume regularly sampled timeseries.

For all these reasons, a new processing algorithm for CTDs installed on gliders is required. The new processing chain needs to take into account all the above mentioned constraints in order to obtain better results. The approach proposed in this paper consists of several steps that are applied in sequence on the glider data. Some of these steps are processes that require a parametrization. The parameters used in the processing can come from several sources. Some of the parameters come from the sensor manufacturer and express the intrinsic response of the sensor. Other parameters are extracted from an optimization process which tries to find the best combination of values for a certain purpose. Finally, other parameters are also extracted from user's experience on the objectives of the studies to be performed with the data.

A brief description of the algorithm is following. Initially, the time series of each individual variable are low-pass filtered in order to remove noise from the signal. The time constant of this low-pass filter attends to the fastest response the sensor can provide. Any variation faster than that is assumed not to be a real variation in the signal (i.e., the sensor would not be able to reproduce it) but noise on the acquisition system instead. After low-pass filtering the input signals, the signal gradients are accentuated to revert the effect of the sensor time response. The signals also need to be shifted/aligned in time, since the same water parcel does not go through all the sensors at the same time.

Once the signals of each sensor have been processed independently, they can be combined to derive new oceanographic variables of interest. This step will require the same particular treatment to make sure that no artifacts are created. One of the most evident artifacts is the thermal lag effect [6], [7]. The conductivity cell can modify the temperature of the water parcel that is being measured in it, thus, affecting the measurement. This point has to be taken into account when deriving salinity, since the typical CTD device measures the temperature outside of the cell and the conductivity inside of it. This leads to important mismatches, especially when the probe passes through strong temperature gradients. In those situations, the temperature inside and outside of the conductivity cell can be very different and the resulting salinity derivation will be in error. A more detailed description and the procedure followed to address it can be found in [6], [7]. Fig. 5 shows a salinity vertical section obtained with the processing described above.

A. Data Quality Control

Along the various processing steps, several quality control have to be applied to make sure that the final product fulfills a certain number of constraints. A sequence of range tests, gradient tests, spike tests and stationary tests are applied on the data set to verify that the measured parameters and the derived ones are physically meaningful. The initial QC tests on the raw data are quite coarse, and they are useful to assert

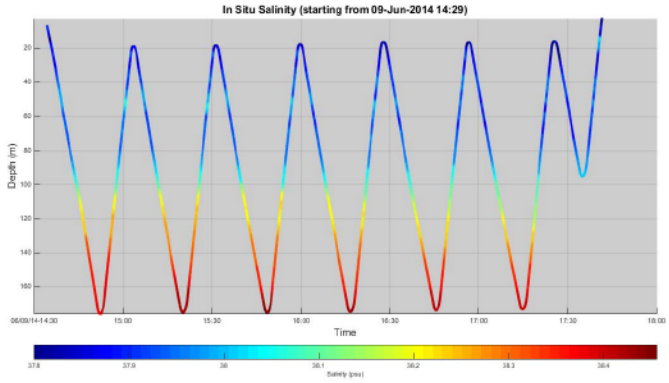

Fig. 5. Sea water salinity.

that the outliers on the data are dismissed before the processing starts. The QC applied to processed data is finer than the initial one, since the quality of data, once processed, is assumed to be higher.

These tests can be constrained by the geo-referencing of the variable. Different range tests can be applied depending on the ocean basin the data was collected in. Also seasonality has to be taken into account when constraining the quality control tests. The developed tool is highly configurable and therefore many different tests can be applied, using the above mentioned tests as basic building blocks to create more complex combinations of them.

A future improvement to the current real time QC system will be based on the generation of probability of error timeseries to be associated to the scientific variables like temperature, pressure, salinity and potential density. The availability of the probability of error can be used as a starting point to generate the quality flags, but can be also used by researchers to define their own thresholds while selecting valid data for a specific investigation or application.

Underwater gliders provide a relatively large amount of data in near real time. The data sets collected by a fleet of gliders requires the data processing chain to be automated, if operational use of this data is desired. Therefore, most of the processes mentioned before run unattended. This automatic processing allows scientists to have data sets from a fleet of gliders, processed and quality controlled in near real time, ready available for operational use.

VI. Glider Data Management

Proper documentation of data sets is fundamental to enable data reuse. Details supporting the experiment reproducibility such as acquisition systems, configuration and calibration of sensors, data processing (i.e. data quality, variable derivation, data transformation, analysis algorithms and modeling) applied, consistent description of data content, file formats, area and time of acquisition, conditions of the sea, cruise, personnel involved etc. should be included or referenced (e.g., through an e-publication or report) in the metadata.

Metadata are used by data discovery, data access and transfer services, to efficiently search for data sets matching the user requirements. Metadata standards such as the ISO 19100

family for digital geographic information have been endorsed by several Earth Observations fields including oceanography to document data and services, because they provide a reference framework for describing spatially referenced information. Moreover, these standards rely on legacy formats such as FGDC (Federal Geographic Data Committee) and Gemini, therefore most of the systems provide interfaces towards all of them guaranteeing interoperability.

Once data have been retrieved, they should be available in a format that can be easily accessed by the user. Open formats such as NetCDF and HDF (Hierarchial Data Format) have become a reference for scientific data sharing. Data represented in these format are self-describing and include additional metadata that facilitate data discovery (cf. global attribute section of NetCDF). Beside best practices, harmonisation initiatives at community level exist, such as Climate and Forecast (CF) conventions [10] used at international level for climate data that enriches NetCDF with additional metadata conventions and a standard vocabulary.

Noticeably, the glider community has endorsed such initiatives to further enhance glider data sharing. For example, the EGO glider NetCDF data format [11] is a standardisation initiative developed in the framework of the EU FP7 project GROOM by the EGO group, an open community of oceanographers around the world promoting the use of gliders for collecting ocean observations and the dissemination of glider data.

The EGO glider data management relies on previous experience of EGO partners on Argo data, and leverages NetCDF and the Unidata NetCDF Attribute Convention for data set Discovery, as well as CF specification, to define a common file format and encoding for glider observations. The EGO convention facilitates the sharing among the glider community and the aggregation of glider data collected at multiple sites, while preserving the readability from NetCDF tools, standardising names for files and observation variables, formalising the encoding of values, defining mandatory attributes. The community shares also a format file checker and scripts for converting observation to the reference format.

The EGO glider data management proposal has been endorsed at global scale by the glider community at large within the context of the Ocean Data Interoperability Platform (ODIP) framework, a global harmonisation effort that involves key Ocean Observing Systems contributing to the Global Ocean Observing System, such as the Australian Integrated Marine Observing System (IMOS), the Integrated Ocean Observing System (IOOS), and the British Oceanographic Data Centre (BODC).

A. *CMRE Glider Data Management Prototype*

The EGO NetCDF easily combines with interoperability services developed for geospatial data and standardised for Earth Observation monitoring, such as those defined by the Open Geospatial Consortium (OGC) for catalogue services (Catalog Service for the Web - OGC CSW) and sensor observations services (Sensor Web Enablement - OGC SWE), and the Open Data Access Protocol (OPeNDAP).

The CMRE data repository (cf. Fig. 6) offers discovery and access services for quasi-real-time and non-real-time data acquired during cruises, including glider observations. Discovery services are offered both for machine-to-machine processing and for human-based interactions. The catalogue web interface combines fuzzy textual search, faceted browsing and spatio-temporal filtering to help scientists find the more suitable data sets for their experiments. The metadata view is enriched with a portray service based for data set preview. Accreditated users, such as cruise partners, can directly download the data sets in different formats, including raw and processed formats (cf. Fig. 7). System access to data sets is offered through the OGC CSW, that can be queried and enables filtered access to the data sets from harvesting APIs.

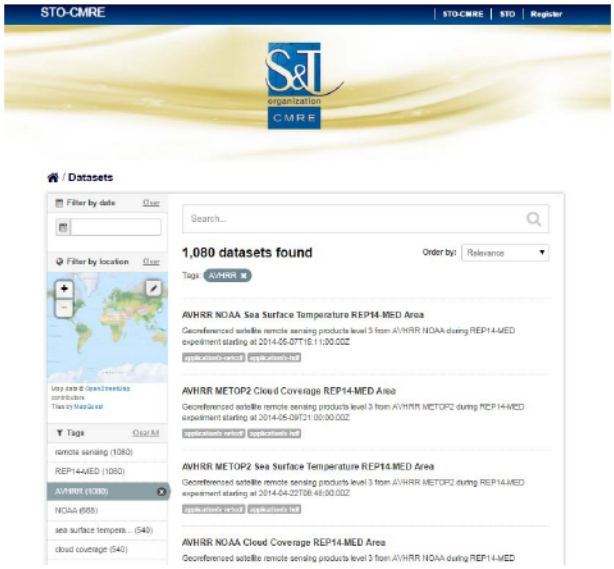

Fig. 6. CMRE data catalogue: search interface. Available at http://datacatalog.cmre.nato.int/

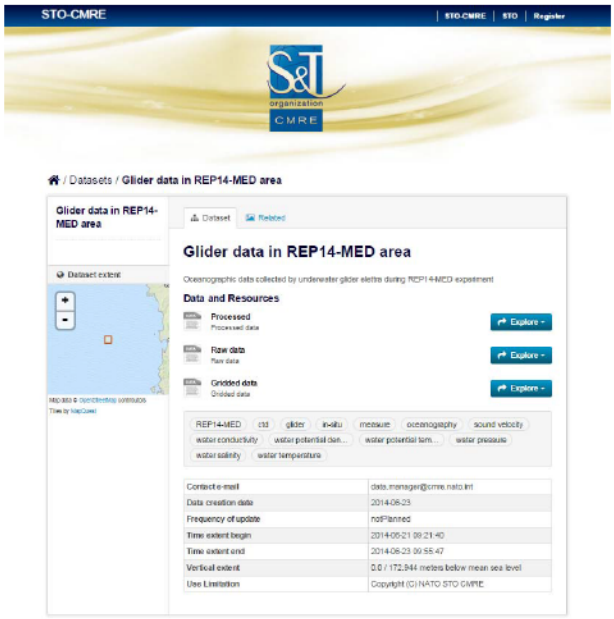

Fig. 7. Glider metadata view

For guaranteeing interoperability, metadata are automat-

ically extracted from data sets and transformed enforcing compliance to standards and application profiles used for geospatial data such as ISO 19115, ISO 19115-2 [15], [16] and the corresponding standardization initiatives adopted within NATO (AGeoP-8 [17]). This brokering approach guarantees interoperability among the integrated components of the systems, that leverage on CSW service for metadata harvesting, and facilitate federation with external services.

In a recent experiment, the data repository has been coupled with a THREDDS (Thematic Real-time Distributed Data Services) data server[2], an open source software developed by Unidata which is widely used for scientific data interoperability because it supports a variety of services for data and metadata access. THREDDS leverages NetCDF as harmonization model (namely, Common Data Model) and offers a number of brokers defined in NetCDF Markup Language (ncML) to support the ingestion of data in different formats, including output from several oceanographic models. In this experiment, simple ncML transformations have been defined and applied online to glider data sets to enforce compliance to Climate and Forecast conventions 1.6. THREDDS integrates the OPeNDAP (Open-source Project for a Network Data Access Protocol) service[3] for accessing data subsets corresponding to specific profiles, variables, spatio-temporal slices, etc. Data subsets in NetCDF can be either downloaded in locally, or accessed remotely via the Data Access Protocol (DAP) using URLs that encode the request and can be opened directly by a various clients[4], including scientific languages such as Matlab and Python avoiding local download of bigger files.

This feature facilitates data comparison, as in Figure 8, where observed and simulated glider tracks are analysed in Matlab. The data sets are accessed through OPeNDAP links.

THREDDS offers also metadata services. For instance, the ncISO server integrated in THREDDS generates ISO 19115 compliant metadata from data sets. No direct CSW service is provided by THREDDS, but the metadata automatically generated by the server can be harvested using WAF (Web Access Folder) clients and then published via CSW by other catalogue software.

In the next future, the existing prototype will be extended to support OGC Sensor Web Enablement, specifically the SOS service, to share glider observations to be used for data assimilation in model forecasting.

VII. Conclusions and Future Work

The NetCDF data file organization and metadata introduced in the paper is the result of the lessons learned at CMRE while processing glider data sets and it has proven to be sufficient for the general purpose of data description, discovery and cataloguing in the CMRE data management system. The sensor-driven approach in the organization of variables has proven to be clear enough for the final users, including CMRE scientists and scientific partners from other institutions. After these initial experiments, a larger dissemination of the proposed data format and conventions among the oceanographic community is desirable, while continuing the process of improving the format to properly handle data sets containing data collected by new sensors with different characteristics (*i.e.*, acoustic data).

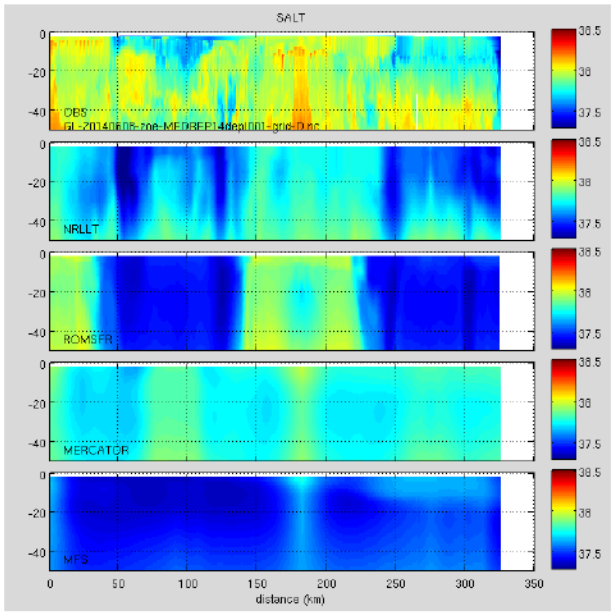

Fig. 8. Observed and simulated vertical sections

An important part of the ongoing work is the adaptive correction of thermal lag in CTD data. As previously introduced, the irregular sampling rate of the data and the presence of time gaps in the timeseries (especially in real time data sets) require further investigation of the problem. Also the sensor model needs to be reviewed and improved as new versions of the device appear. A clear and easy way to store the correction parameters inside the NetCDF files is a suggested topic for discussion in the oceanographic community.

The Automated Real Time Quality Control is probably one of the most difficult problem to address in the near future. The proposed approach in this paper would be to apply a first pass of the QC to the L0 data with the aim of identifying the clearly bad data and remove them from the following processing steps. A second pass should be then applied to the L1 data with different criteria to better identify samples with problems. Quality control tests implementation and definition of proper ranges for glider data are open issues. However, it is not easy to automate this process and make it fully unattended. An experienced person might still be required to review dubious data that cannot not be classified as definitely good or bad in the automated QC procedures above mentioned. The interaction between the automatic tools and interactive ones is a synergy to be investigated in the future.

VIII. Acknoledgment

This work was performed in the scope of project no. *ACT000507* for NATO Allied Command Transformation.

The authors kindly thank Dr R. Signell for the discussion on harmonisation and standards for glider data sets. A special thanks to S. Falchetti for her support in the review.

[2]Unidata THREDDS data server http://www.unidata.ucar.edu/software/thredds/current/tds/ (Accessed in March 2015)

[3]OPeNDAP http://www.opendap.org/ (Accessed in March 2015)

[4]OPeNDAP clients http://www.opendap.org/whatClients (Accessed in March 2015)

REFERENCES

[1] D.L. Rudnick, R.E. Davis, C.C. Eriksen, D.M. Fratantoni and M.J. Perry, Underwater Gliders for Ocean Research, Marine Technology Society Journal, Vol. 38 No. 1, 2004

[2] S. Glenn and O. Schofield, Growing a Distributed Ocean Observatory: Our View from the COOL Room, Oceanography, Vol. 22 No. 2, 2009

[3] S. Ruiz, L. Renault, B. Garau and J. Tintoré, Underwater glider observations and modeling of an abrupt mixing event in the upper ocean, Geophysical Research Letters, Vol. 39, 2012

[4] L. Suberg, R.B. Wynn, J. van der Kooij *et al.*, Assessing the potential of autonomous submarine gliders for ecosystem monitoring across multiple trophic levels (plankton to cetaceans) and pollutants in shallow shelf seas, Methods in Oceanography, Vol.10, 2014

[5] R.N. Smith, M. Schwager, S.L. Smith, B.H. Jones, D. Rus and G. Sukhatme, Persistent Ocean Monitoring with Underwater Gliders: Adaptive Spatiotemporal Sampling Resolution, Journal of Field Robotics, Vol. 28, 2011

[6] B. Garau, S. Ruiz, W.G. Zhang, A. Pascual, E. Heslop, J. Kerfoot and J. Tintoré, Thermal Lag Correction on Slocum CTD Glider Data, Journal of Atmospheric and Oceanic Technology, Vol. 28, 2011

[7] Morison, J., R. Andersen, N. Larson, E. DAsaro, and T. Boyd, 1994: The correction for thermal-lag effects in Sea-Bird CTD data. J. Atmos. Oceanic Technol., 11, 11511164

[8] A. Alvarez and B. Mourre, Optimum Sampling Design for a Glider-Mooring Observing Network, Journal of Atmospheric and Oceanic Technology, Vol. 29, 2012

[9] O. Schofield, J. Kohut, D. Aragon, L. Creed, J. Graver, C. Haldeman, J. Kerfoot, H. Roarty, C. Jones, D. Webb and S. Glenn, Slocum Gliders: Robust and Ready, Journal of Field Robotics, Vol. 24 No. 6, 2007

[10] (November 2014) Climate and Forecasts (CF) Conventions and Metadata website [Online], Available: http://cfconventions.org

[11] T. Carval (IFREMER), J. Buck (BODC), B. Garau and D. Cecchi (CMRE), EGO gliders User's manual Version 1.1, December 2013

[12] T. Carval (IFREMER), B. Keeley (MEDS), Y. Takatsuki (JAMSTEC), T. Yoshida (JMA), S. Loch (BODC), C. Schmid (AOML), R. Glodsmith (WHOI), A. Wong (UW), R. McCreadie (BODC), A. Thresher (CSIRO) and A. Tran (MEDS), Argo user's manual Version 3.1, July 2014, http://dx.doi.org/10.13155/29825

[13] (March 2015) Network Common Data Format, Unidata website [Online], Available: http://unidata.ucar.edu/software/netcdf/

[14] S. Falchetti, A. Alvarez and R. Onken, A Relocatable EnKF Ocean Data Assimilation tool for heterogeneous observational networks, IEEE/MTS Oceans 2015 Conference, 18 - 21 May 2015, Genova, Italy

[15] ISO 19115-1:2014 Geographic information – Metadata – Part 1: Fundamentals

[16] ISO 19115-2:2009 Geographic information – Metadata – Part 2: Extensions for imagery and gridded data

[17] STANAG 2586 Ed01 AgeoP-08, NATO Geospatial Metadata Profile (NGMP)

Document Data Sheet

Security Classification		Project No.
Document Serial No.	**Date of Issue**	**Total Pages**
CMRE-PR-2019-113	June 2019	9 pp.

Author(s)

Daniele Cecchi, Bartolome Garau, Elena Camossi, Alessandro Berni, Emanuel Coelho

Title

Sensor-driven glider data processing

Abstract

Gliders data are an important source of observations for oceanographers, hence harmonization and interoperability among different acquisition systems are key enablers towards efficient data sharing and effective reuse. As of today, several organizations operate gliders and produce customized results. The paper presents the effort done at CMRE, in collaboration with research partners, for the definition of a self-contained data file that eases both human readability and automated processing. By relying on community recognized standards, highly interoperable glider data sets are generated, whilst preserving readability thanks to the standardized and straightforward internal organization of variables. The paper includes also a description of the processing of the principal oceanographic sensor, i.e., CTD (Conductivity, Temperature and Depth), and an overview of the glider data management lifecycle.

Keywords

Underwater gliders; interoperability; standardization; data processing; data management; data fusion

Issuing Organization

NATO Science and Technology Organization
Centre for Maritime Research and Experimentation

Viale San Bartolomeo 400, 19126 La Spezia, Italy

[From N. America:
STO CMRE
Unit 31318, Box 19, APO AE 09613-1318]

Tel: +39 0187 527 361
Fax:+39 0187 527 700

E-mail: library@cmre.nato.int