Quantitative Quality Control (QC) procedures for the Australian National Reference Stations: Sensor Data

E. B. Morello, T. P. Lynch, D. Slawinski, B. Howell, D. Hughes, G. P. Timms
CSIRO CMAR
CSIRO ICT
Australia
Elisabetta.Morello@csiro.au

***Abstract*— The National Reference Station (NRS) network, part of Australia's Integrated Marine Observing System (IMOS), is designed to provide the baseline multi decadal time series required to understand how large-scale, long-term change and variability in the global ocean are affecting Australia's coastal ocean ecosystems. High temporal resolution observations of core variables are taken across the network using a standard instrument, the Wetlabs Water Quality Monitors (WQMs). These data are freely available to the scientific community and the general public and thus need to be controlled by means of a system that will ensure their consistency and fitness-for-use. This document provides insights into the Quality Control (QC) guidelines proposed for these data, towards a system that incorporates uncertainty into the standard qualitative flag system, making it quantitative.**

Keywords - Water Quality Monitors; Quality Control; temperature; salinity; Integrated Marine Observing System (IMOS); National Reference Stations (NRS)

I. INTRODUCTION

The National Reference Station (NRS) network is part of Australia's Integrated Marine Observing System (IMOS). It is designed to provide baseline multi decadal time series required to understand how large-scale, long-term change and variability in the global ocean are affecting Australia's coastal ocean ecosystems, at timescales relevant to human response.

Building on three long-term sites operating since the 1940's, the NRS network has now been expanded to include six extra sites for a total of nine, covering Australia's main coastal bioregions [1] (Fig. 1). Recent modeling work has suggested that stations are strategically positioned to observe a significant proportion of the variability of key oceanographic processes (e.g. SST, sea level and current velocity) across all of Australia's continental shelf (Oke, pers. comm.).

The NRS sampling program involves: (1) High temporal resolution data from a standard, moored sensor package, (2) Surface meteorology and delivery of sensor data in real time where surface expression is logistically feasible (four NRS), and (3) Vessel-based biogeochemical sampling and laboratory analysis.

High temporal resolution observations of core variables are taken across the network using a standard instrument, the Wetlabs Water Quality Monitors (WQMs). The WQM is a multi-sensor with on-board processing and inbuilt bio-fouling controls that measures conductivity, temperature and pressure (CTD), salinity, dissolved oxygen, fluorescence proxies for chlorophyll 'a', and turbidity.

One of the goals of the NRS data collection is to make the data available to the scientific community and the general public; this is done through the dedicated IMOS Ocean Portal (http://imos.aodn.org.au/webportal) for all stations and through the CSIRO National Reference Station sensor web site (http://www.csiro.au/tasman/nrsweb/) [2] for CSIRO-managed stations. The quality of these data thus needs to be controlled by means of a system that will ensure (1) consistency of data among and within samples, and (2) that the quality and errors of the data be documented allowing the end user to assess their suitability for purpose.

This document provides insights into the Quality Control (QC) guidelines proposed for the high temporal resolution data streams generated by the WQMs as part of the IMOS NRS network, making initial use of the delayed mode data to test our case.

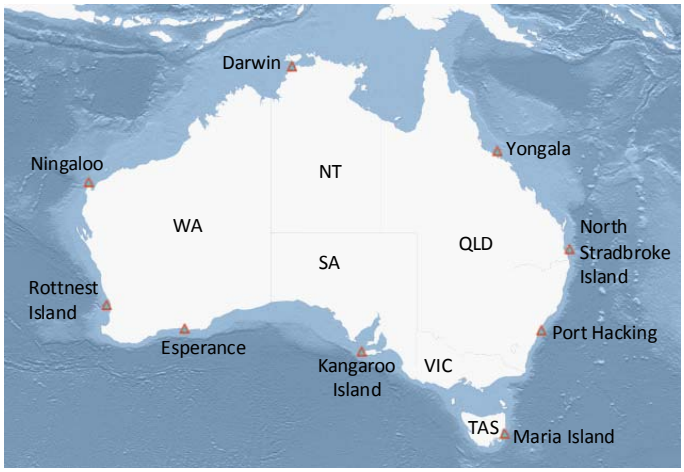

Figure 1. The nine National Reference Stations (NRS) of the Australian Integrated Observing System (IMOS).

Owing to the high frequency of sampling, large data streams are generated daily by each NRS rendering manual QC

unwieldy thus making an quantitative QC routine, that can be performed on individual or consecutive points (depending on what is being tested), more desirable. To help inform the end-user of their "fitness for purpose", the data need to be classified, with respect to selected tests, in such a manner as to indicate what procedures have been performed to ensure quality. There are two ways of doing this:

- Applying a series of qualitative "gates" or flags for the data to pass through before classifying each data point.
- Calculating a quantifiable uncertainty estimate proving the goodness of the QC carried out on the data in question.

The QC strategy proposed in this document will be advocating a hybrid approach. Many QC protocols use qualitative flags, giving the user an immediate idea of quality. The problem with this sort of system is that it can be subjective with no associated formal statistical framework based on uncertainties, upon which to judge flag assignment, and, more often, without strict definitions of what each flag means. Data are classified as either "good" or "bad" (or even "probably good") but these definitions may vary contextually based to data usage or individual preference.

One possible solution to the issue of subjectivity is to develop an agreed value, based on knowledge of the environment, which is a "gate" of uncertainty, triggering a quantitative QC flag for the data. Then, as tests are passed or failed, the data are either highly constrained by uncertainty or assumed to be accurate enough for the purpose for which it is intended. This system also does not remove any data, it merely annotates them, allowing the end user to easily sort via explicit levels of uncertainty.

II. Methods and Results

The choice was made to adopt the flag system used by the Intergovernmental Oceanographic Commission (IOC) of UNESCO [3] [4] (Table 1). This is the system of preference of many oceanographic data-collecting systems around the world from the GOOS (www.ioc-goos.org) and SeaDataNet protocols (www.seadatanet.org) to the Argo floats [5] making data quality standards comparable. The ultimate aim, though, is to incorporate uncertainty estimates of choice into the qualitative IOC flag system in an attempt to make it quantitative. This integrated quantitative QC system is foreseen to include several tests from the binning of data, impossible date, impossible location, regional range, spike and stationarity tests to a logical set-based system.

A. Test data

A common dataset generated by the Rottnest Island NRS was used to illustrate the QC methodology proposed, when necessary. The Rottnest NRS is located off Rottnest Island in WA (32.0°S, 115.4°W; Fig. 1) in water 48 m deep with WQMs at 20 m and 40 m. The dataset of choice was produced by the WQM placed at 40 m depth and relates to a four month period from 21/02/2009 (07:16:10) to 14/06/2009 (09:32:38). Data of interest include temperature (°C), salinity (PSU) and fluorescence proxies for chlorophyll 'a' ($mg.m^{-3}$).

TABLE I. Quality flag scale used by the Intergovernmental Oceanographic Commission (IOC) of UNESCO.

Flag	Meaning
0	No QC performed
1	Good data
2	Probably good data
3	Bad data that are potentially correctable
4	Bad data
5	Value changed
6	Below detection limit
7	In excess of quoted value
8	Interpolated value
9	Missing value
A	Incomplete information

B. Pre deployment tests and calibration

To maintain the accuracy of data being captured from sensors at the NRS sites, regular calibration and pre-deployment testing are required. Temperature, pressure, conductivity and dissolved oxygen calibrations are carried out at the CSIRO NATA-certified calibrations lab along with regular servicing. On-going discussions are leading towards an IMOS standard of calibration for optical instrumentation such as the WQM's fluorometer and turbidity sensor (FLNTU).

WQMs are recalibrated after 12 months of in-water time. There are currently several proposals being considered for pre-deployment checks of these sensors:

- Testing of units against a "golden standard" WQM or other CTD/DO combination which is never actually deployed.
- Testing of the two WQMs to be deployed against each other.
- Testing in a known parameter water bath, established by hydrochemical sampling.
- Testing against a solid standard.

All methods would have a preset allowable deviation. The first method requires extra instrumentation and calibration at more cost; the second is the cheapest option but does not account for sensor drift (which is likely to be in the same direction for both sensors) and has no truthing against known values; the third introduces more labour and potential delays whilst samples are analysed; and the fourth provides only a fixed point rather than a distance range from which a sensitivity curve can be determined.

C. Impossible date and location tests

The impossible date test requires that the WQM observation date and time be sensible. The first WQMs were deployed in 2008, thus the following date and time parameters are suggested.

- Year greater than 2007.
- Month in range 1 to 12.
- Day in range expected for month.
- Hour in range 0 to 23.
- Minute in range 0 to 59.

Similarly, the NRS are spread over the coastal margins of Australia at specific fixed locations (Fig. 1). The impossible location test requires that the described observation of latitude and longitude be sensible to the site. The automatic QC code flags any observations made outside of a radius of ± 0.25°S and 0.2°W around each of these points. This test has the added benefit of detecting if samples and metadata have become confused.

D. Regional range tests - climatologies

This test applies a gross filter on observed values, using reference regional climatologies and needs to accommodate all of the expected extremes encountered in the oceans.

Considering the diverse locations, and associated climates, of the IMOS NRS sites, different local/regional climatologies need to be developed for each. The first climatologies developed will be for temperature and salinity but will eventually be expanded to include dissolved oxygen, chlorophyll 'a' and turbidity. The climatologies are at least seasonal, preferably monthly, and the data used to develop them are independent of the sensor data. In cases where no independent data were to be available, appropriate proxies may be provided by the Levitus climatologies or the Australian-centric CARS (CSIRO Atlas of Regional Seas) (http://www.marine.csiro.au/~dunn/cars2009/).

In the case of our example, the Rottnest NRS, a relatively rich set of independent data was available and used, as follows:

- The Rottnest NRS bottle data for temperature and salinity, available at 10 m intervals for 0 – 50 m since 1951 approximately monthly.
- Expendable bathythermograph (XBT) data obtained from the QUOTA database (http://www.marine.csiro.au/~cow074/quota/quota.htm) for the geographic region included between 31.75°S and 32.25°S, and 115.15°W and 115.65°W.
- Additional bottle data for temperature and salinity available at 10 m intervals for 0 – 50 m for sparse months in 2009, 2010 and 2011.
- The Rottnest NRS CTD data collected monthly between 1950 and 2003 and from 2008 onwards.
- Opportunistic CTD casts and bottle data providing data for temperature, salinity and chlorophyll 'a'.
- Temperature data from the Rottnest ferry.

All these data are binned (averaging) to replicate the bottle data, as follows: 0 – 5 m for the 0 m bottle; 5 – 15 m for the 10 m; 15 – 25 m for the 20 m; 25– 35 m for the 30; 35 – 45 m for the 40 m; and 45 – bottom for the 50 m.

Once binned, the data are averaged for each month-of-year calculation: mean, minimum, maximum and standard deviation. These are then used as the bases for viable ranges as described elsewhere in this document.

The test applies the regionally expected range to measured parameters from WQMs deployed at particular sites and depths. An example of a spatially (by depth) and temporally (monthly) constrained regional ranges, or climatology, are provided for salinity around Rottnest Island (Fig. 2). Data outside the plots, will be flagged as suspect using the flag system of choice (e.g. "bad" data, Flag 4), as exemplified by the Rottnest Island salinity test data (Fig. 3). The threshold beyond which a datum is determined to be bad (Flag 4) is quantified in terms of multiples of the standard deviation, in this case 6σ: with 3σ, you assume that 99.9% of your data is good; 6σ gives a much wider margin. In this case, most of the salinity test data from Rottnest Island are within the minimum-maximum bounds of the regional monthly climatology for that depth and well within the 6σ threshold (Flag 1; Fig. 3).

E. Spike tests (rate of change)

A spike can be defined as the difference between sequential measurements, where one measurement is quite different to adjacent ones. The spike test flags for unusually large rates of change in the dataset. The algorithm used here is the same used by the Australian National Mooring Facility Ocean Gliders (ANFOG), as follows:

$$\text{Test value} = |V2-(V3+V1)/2|-|(V3-V1)/2| > \text{threshold} \quad (1)$$

where V2 is the measurement being tested as a spike, and V1 and V3 are the values above and below.

This algorithm may be used for all parameters, with different thresholds (which may be determined regionally as well). Thresholds proposed by the various protocols:

Temperature: the V2 value is flagged when

- Test value > 6.0°C for pressures less than 500 db.
- Test value > 2.0°C for pressures greater than or equal to 500 db

Salinity: the V2 value is flagged when

- Test value exceeds 0.9 for pressures less than 500 db.
- Test value exceeds 0.3 for pressures greater than or equal to 500db

Some Australian coastal areas (especially in the tropics) will have relatively broad ranges in coastal oceanographic parameters such as salinity during the 'wet' or monsoon season, and this needs to be considered.

F. Stationarity test

This test looks for all measurements being identical. The occurrence of constant values of data depends on the variable being measured, the sampling interval used and the resolution of the sensors. For temperature and salinity, the IOC [3] sets the allowable number of consecutive equal values of temperature and salinity:

$$T = 24*(60/\Delta t) \quad (2)$$

where T = allowable number of consecutive equal values; Δt = the sampling interval in minutes.

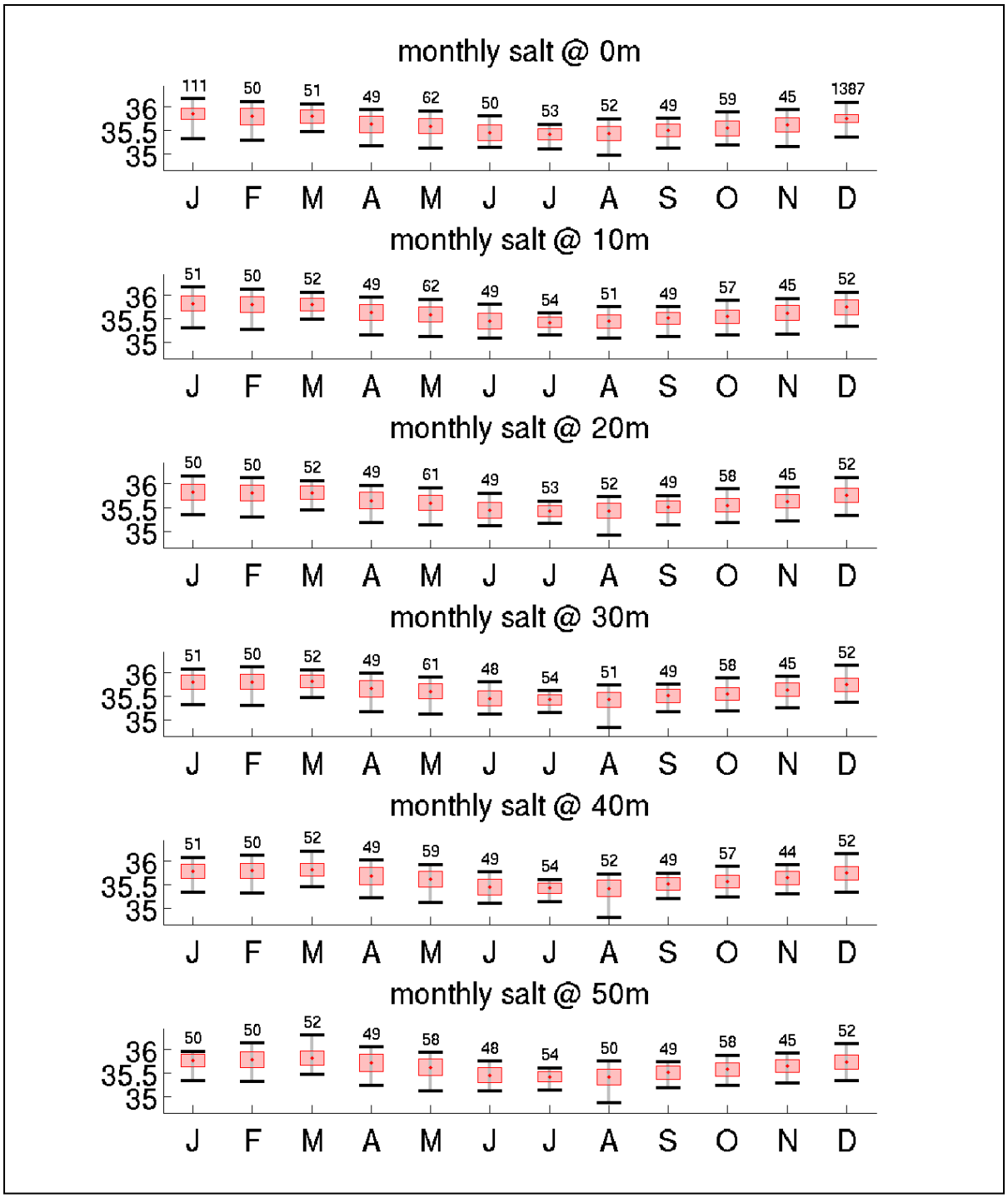

Figure 2. Monthly salinity (psu) ranges by depth at the Rottnest Island produced using a set of data independent of the National Reference Station NRS, showing mean, minimum, maximum, standard deviation and the number of observations considered for each month.

G. Logical set-based system

We are also investigating the use of a modified automatic QC system which incorporates the same thresholds as the approach outlined above but combines many of the other factors using logical set theory, also known as fuzzy logic. Logical set theory is well-suited to applications, such as quantitative data QC, where the solution is dependent on both statistical analysis and human experience. It provides a means

for combining these two types of uncertainty and estimating an overall data quality estimate.

Here we use the approach outlined in [6] to obtain both data flags and error bars for the NRS test data from Rottnest Island. In short, based on their performance with respect to a number of tests performed (in this case: range test/threshold, gradient test, cumulative gradient test and time since deployment) the data are assigned small (green), medium (yellow) or large (red) fractions of (contributions to) error for each test. The fractions emerging from each single test are added together to obtain an overall contribution to error of each data point, on the basis of which it is then categorized into the Flag system chosen. So for each data point if:

- Green fraction > 0.9; data flag = 1 (Good)
- Green fraction >0.5 and red freaction ,0.3; data flag = 2 (Probably good)
- Threshold crossed; data flag = 4 (Bad)
- Other; data flag = 3 (Bad data that are potentially correctable).

Fig. 4A presents salinity data over the test period with the data flags estimated using the logical set based system showing how most data points classify as Flag 1. Fig. 4B presents a subset of salinity data on 22 February 2009, along with estimated data flags and error bars. In this plot the fractions of error of each data point (green, yellow and red) are multiplied by the manufacturers specifications for accuracy (0.005, 0.015 and 0.1, respectively). Closely spaced data points in the latter figure result in the appearance of multiple error bars at each timestep.

III. DISCUSSION AND CONCLUSIONS

This document provides suggestions as to the best practice for performing quantitative QC of the high frequency data generated by CTD sensors of the IMOS NRS Water Quality Monitors, based, initially, on a subset of delayed mode data. The ultimate scope of this exercise will be to incorporate uncertainty into the qualitative flag system, making it quantitative.

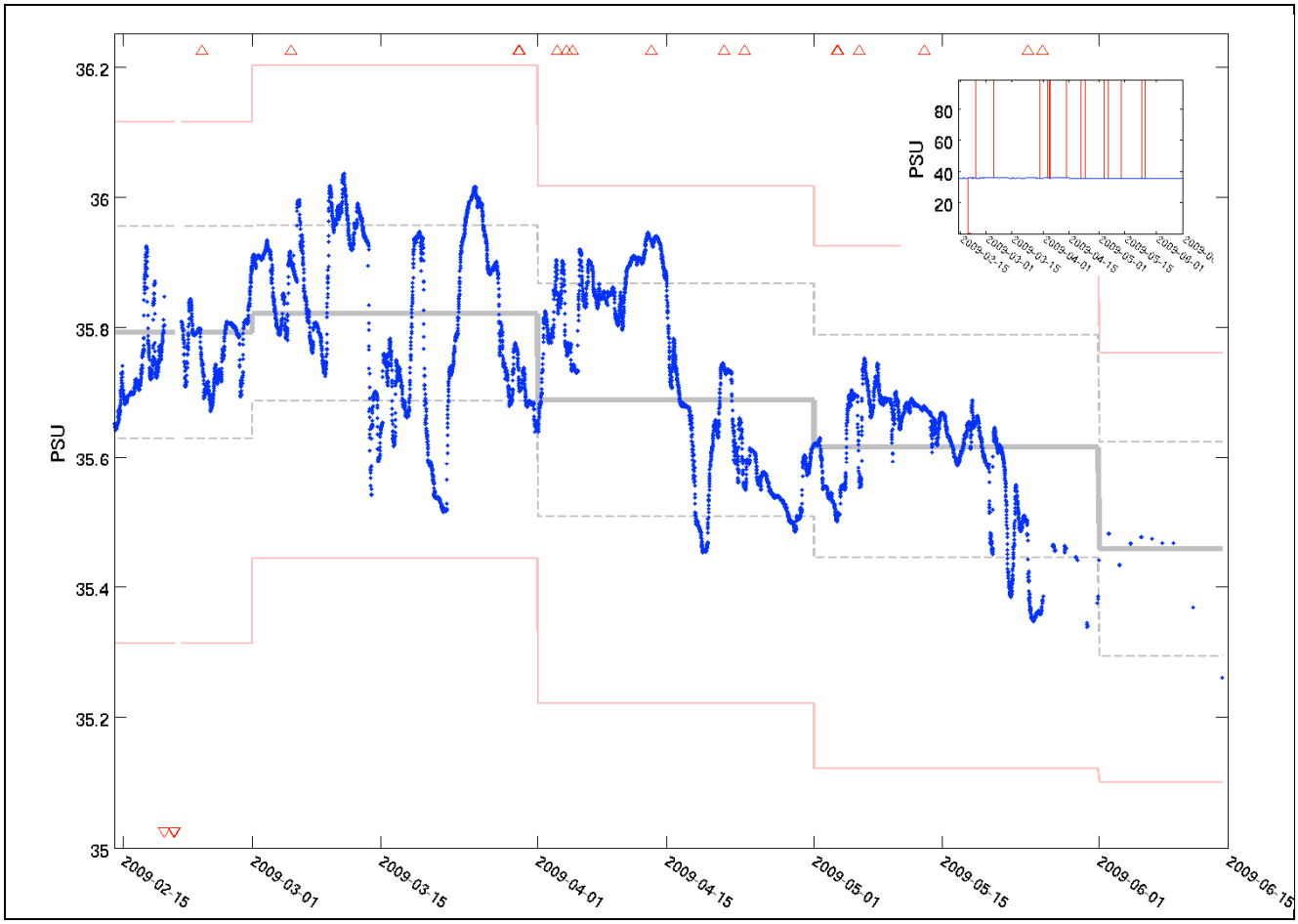

Figure 3. Variation of salinity (PSU) over time (YYYY-MM-DD) of the Rottnest Island NRS test data (blue), compared to the monthly regional climatology represented by the mean (continuous grey line), standard deviation (dashed grey line), and minimum and maximum values (continuous red lines below and above the test data, respectively). The red triangles at the top and bottom represent "bad" (Flag 4) data points which are outside the threshold of 6 standard deviations; the extent of their variation is shown by the red lines on the inset plot by comparison to "good" (Flag 1) data in blue.

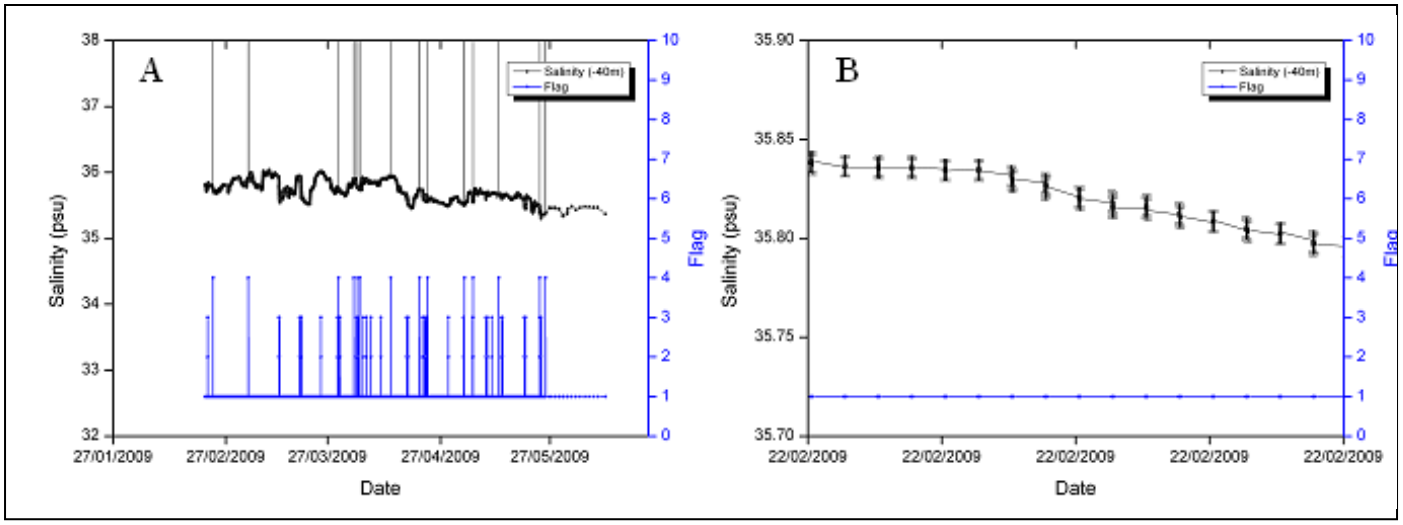

Figure 4. Variation of salinity (psu) with time (black) from the Rottnest Island NRS, using (A) all test data (February 2009 to June 2009), and (B) a subset of test data, zoomed in on 22 February 2009, along with data flags (blue) and error bars estimated using the logical set-based approach

We consider our hybrid system to be highly iterative, with continual adjustment possibly in response to research findings and changes in observable systems. However, this system does provide a strong framework for tracking both the QC decisions and any changes to how these decisions are applied. This approach is in response to the specific challenges of large long term observing programs such as IMOS. First, the program's scope, both in terms of intensity and spatial coverage, is now too broad for a single oceanographer to be able to control all aspects of QC. Second, the program is designed to operate over the long term, hence a good understanding of how data were handled in the past will be required for scientists of the future. Besides the consideration of labor cost, if accurate comparisons are to be made between datasets or over time robust and consistent rules for QC need to be developed. By developing quantitative quality control inter-observer bias is removed.

There are a number of issues in the basic QC tests being run that should be addressed in the future. The first regards the binning, or not, of delayed-mode data. Should the QC be performed on the binned (averaged) values of the 60-minute bursts collected every 15 minutes, or on the raw, unbinned, data? On one hand, if we perform QC on binned data we may be incorporating spikes (which could potentially be classified as "bad" data) and performing QC on a dataset containing bad and good data. On the other hand, by performing QC on raw data we are analyzing serially correlated bursts of data. Our proposal for the future is to perform QC on unbinned data for the physical variables (dissolved oxygen, temperature and salinity) but not for the 'biological' data (fluorescence and turbidity). The rationale behind this decision is based upon the assumption that, unlike the physical properties of temperature and salinity at microscopic spatial scales, turbidity and fluorescence sensors may be recording a series of real spikes recorded in response to a subset of the sampled population particles which are large and happen to floating past the sensor. If QC is performed before binning (i.e. removing these spikes) unknown oceanographic phenomena (potentially valuable information) may be removed from the data. Finally, statistical analysis of the distribution (e.g. a measure of skeweness) of the physical variables should be performed on them to determine the most adequate measure of central tendency to be adopted: mean or median. The reason for this being that the median is a better measure of central tendency when data are highly skewed and the mean is heavily affected by the presence of outliers.

The second issue relates to the validity of range testing and the problem of shifting baselines. Climatologies are only relevant to the present and immediate past. At some sites, such as the Maria Island NRS, the long time series has described a sliding baseline as temperature and salinity have increased over the last 60 years [7]. This means that the climatologies as regional range flags need to be assessed on an inter-annual basis.

There are a number of challenges in the development and implementation of re-locatable quantitative quality control systems for sensors. Physical location (in any or all three dimensions: latitude, longitude, depth) of sensors can make a large difference to thresholds applied to observations of particular phenomena, rate of change between observations (gradient tests), rates of biofouling of instruments (thus affecting the rate at which each sensor drifts over time) and even regional range (e.g. a sensor located in a standing wave, sink hole or eddy may observe phenomena that fall well outside the regional range).

Spike tests and gradient functions can be used to determine when one observation, differs too greatly from adjacent observations. However, neither method is sufficient for quality control assessment of step functions in time series. Whilst the gradient test flags the observation immediately after the step has occurred as poor quality data, subsequent measurements pass the gradient test. One possible partial solution to this problem is to use a sliding window from when the step function was detected (i.e. retrieval of the observation at the time-step immediately before the gradient test failed) and calculate the decay in gradient over the period in the sliding window. Although this sliding window method may be able to help us retrospectively determine QC flags and uncertainties by fitting

a curve function to a pre-determined curve or by observing a "reverse step" within the sliding window, this method does not offer a quantitative QC solution where a single step without return is observed. Finally, although the probability is low, a step function may be due to a real phenomenon (e.g. a change in current) and therefore we can never flag such data as poor with absolute certainty.

Future work on the system may include modelling sensor and sensor platform (arrays of sensors observing different phenomena) behaviour by using means such as the Bayesian quality control techniques developed by the CSIRO Tasmanian Marine Analysis Network (TasMAN) project [8] or a system using Hidden Markov Models (HMM) could, in the future, provide reliable probability distributions for sets of possible sensor readings (given the current model state). The HMM approach however will only be useful if a strong enough temporal relationship exists between different phenomena at the same location and the strength of these relationships, and, how useful they may be in Markov chain modelling, is yet to be determined.

Time warping pattern matching and function curve fitting techniques, such as those under development by the CSIRO TasMAN project [9] are another set of methods receiving attention as a means of quality control. A pattern matching system such as this may be used to find similar patterns to a slice of the current timeseries in the historical dataset in order to forecast what is likely to occur in the near future. This type of system has potential far beyond quality control such as forecasting, gap filling and event detection (either previously classified or unique).

In the initial stages, however, our emphasis is on combining the traditional qualitative flag system (IOC), with a more formal quantification of uncertainty and we have proposed some delayed mode methods in this paper including climatologies and a logical set based approach [6]. In the short- to medium-term, once delayed mode testing of these techniques has been implemented, we will then move towards executing them on near-real time telemetry. In the future, these techniques will be combined with expert analysis It must be stressed that this system does not remove data, it merely annotates and provides a explicit estimation of error. The use of expert analysis to qualitatively validate quantitative quality control procedures is a vital a 'second stage' of quality control of the NRS station data. By providing a framework of tests and climatologies, however, we hope to c greatly reduce the burden of manual quality control and at the same time improve the accuracy of the system and provide quality assurance to the assessments given.

REFERENCES

[1] T. P. Lynch, D. Mclaughlan, D. Hughes, D. Cherry, G. Critchly, S. Allen, L. Pender, P. Thompson, A.J. Richardson, F. Coman, C. Steinberg, D. Terhell, M. Roughan, L. Seuront, C. Mclean, G. Brinkman, and G. Meyers. “A National Reference Station infrastructure for Australia – using telemetry and central processing to report multi-disciplinary data streams for monitoring marine ecosystem response to climate change”. Oceans 2008 MTS IEEE Oceans, Poles and Climate: Technological Challenges. Quebec City, Canada, 15-18 September 2008.

[2] B. Howell, CSIRO National Reference Station Sensor Web. http://www.csiro.au/tasman/nrsweb/, July 2011.

[3] UNESCO, Manual of Quality Control Procedures for Validation of Oceanographic Data. Manual and Guides, 26, 437 pp., 1993

[4] UNESCO, GTSPP Real-time Quality Control Manual. IOC Manual and Guides, 22, 148 pp., 2009.

[5] ARGO, Argo quality control manual, version 2.6. Argo data management, 46 pp, November 2010.

[6] G. P. Timms, P. A. De Souza, and L. Reznik, “Automated assessment of data quality in marine sensor networks”, presented at IEEE Oceans 2010, Sydney, Australia, 24-27 May 2010.

[7] K. L. Hill, S. R. Rintoul, R. Coleman, and K. R. Ridgway, “Wind forced low frequency variability of the East Australia Current”, Geophys. Res. Lett., vol. 35, L08602, doi:10.1029/2007GL032912, 2008.

[8] D. V. Smith, G. P. Timms, P. A. de Souza, and C. D’Este, “Automated quality control for marine sensing blending statistical, causal inference and expert knowledge”, presented at IEEE Oceans 2011, Kona, Hawaii, USA, 19-22 September 2011.

[9] M. S. Shahriar, G. P. Timms, and P. A. de Souza, “Continuous anomalous pattern monitoring for marine sensor data”, presented at IEEE Oceans 2011, Kona, Hawaii, USA, 19-22 September 2011.