Methods in Oceanography 9 (2014) 17–33

Contents lists available at ScienceDirect

Methods in Oceanography

journal homepage: www.elsevier.com/locate/mio

Full length article

Quality Control (QC) procedures for Australia's National Reference Station's sensor data—Comparing semi-autonomous systems to an expert oceanographer

Elisabetta B. Morello[a,*], Guillaume Galibert[b], Daniel Smith[c], Ken R. Ridgway[d], Ben Howell[c], Dirk Slawinski[e], Greg P. Timms[c], Karen Evans[d], Timothy P. Lynch[d]

[a] *CSIRO Oceans & Atmosphere, 41 Boggo Road, Brisbane, 4001 QLD, Australia*
[b] *IMOS eMarine Information Infrastructure, University of Tasmania, private bag 21, Hobart TAS 7001 Australia*
[c] *CSIRO Digital Productivity & Services, Castray Esplanade, Hobart, 7001 TAS, Australia*
[d] *CSIRO Oceans & Atmosphere, Castray Esplanade, Hobart, 7001 TAS, Australia*
[e] *CSIRO Oceans & Atmosphere, Underwood avenue, Floreat, 6014 WA, Australia*

HIGHLIGHTS

- We describe a hybrid quality QC combining QC flags and a fuzzy logic approach.
- We apply the system to high frequency data from IMOS National Reference Stations.
- We compare the results to those produced by an independent manual QC (expert).
- The hybrid system flags 'bad' data well but did not accurately match expert QC.
- The system is a robust low-pass filter requiring further expert review of 'bad' data.

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
22 August 2014
Accepted 8 September 2014

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 oceanographic variables are taken

* Corresponding author. Tel.: +61 7 3833 5947.
E-mail address: elisabetta.morello@csiro.au (E.B. Morello).

http://dx.doi.org/10.1016/j.mio.2014.09.001

Keywords:
Sustained observing
Coastal oceanography
Climatology
Quality control
Fuzzy logic
IMOS

continuously across the network's nine moored stations using a Water Quality Monitor (WQM) multi-sensor. The data collected are made freely available and thus need to be assessed to ensure their consistency and fitness-for-use prior to release. Here, we describe a hybrid quality control system comprising a series of tests to provide QC flags for these data and an experimental 'fuzzy logic' approach to assessing data. This approach extends the qualitative pass/fail approach of the QC flags to a quantitative system that provides estimates of uncertainty around each data point. We compared the results obtained from running these two assessment schemes on a common dataset to those produced by an independent manual QC undertaken by an expert oceanographer. The qualitative flag and quantitative fuzzy logic QC assessments were shown to be highly correlated and capable of flagging samples that were clearly erroneous. In general, however, the quality assessments of the two QC schemes did not accurately match those of the oceanographer, with the semi-automated QC schemes being far more conservative in flagging samples as 'bad'. The conservative nature of the semi-automated systems does, however, provide a solution for QC with a known risk. Our software systems should thus be seen as robust low-pass filters of the data with subsequent expert review of data flagged as 'bad' to be recommended.

1. Introduction

Over the last decade, programs such as the Array for Real-time Geostrophic Oceanography (ARGO) (Busalacchi, 2010) and the Integrated Marine Observing System (IMOS) (Hill et al., 2010) have delivered substantially more data for oceanographic research than has ever been available before. The result is that the amount of data that needs to be quality controlled has increased substantially. For example, the National Reference Station (NRS) network, which is part of the Australian National Mooring Network (ANMN) facility of IMOS, grew between 2007 and 2010 from three monthly water quality monitoring sites, which were established in the 1940–1950s (Thompson et al., 2009), to nine highly instrumented sites (Fig. 1) (Lynch et al., 2001).

In addition to an expanded water sampling program, sensors at the NRS now collect a range of physical, chemical and biological data, including conductivity, temperature and pressure (CTD) and derived salinity, dissolved oxygen, fluorescence proxies for Chlorophyll *a*, turbidity and current velocities. Besides an increase in types of data collected, this instrumentation of sites also means that the temporal frequency of sampling has increased by up to five orders of magnitude for individual parameters. For example, temperature and salinity, which were historically measured a maximum of three times per season at each site, are now measured up to 480,000 times per season (Lynch et al., 2001). The establishment of multiple sites with sensors at multiple depths is progressively resulting in the generation of Big Data (i.e. datasets whose sizes make it impossible for commonly used software tools to manage, and process the data within a operative time frames) (Mayer-Schönberger and Cukier, 2013).

For the end-user to be able to assess the suitability of data collected by others, knowledge of any procedures performed on the data to assess quality is required. There are two ways in which the quality of data can be assessed:

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

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

Most oceanographic and meteorological data collection systems subject the data they collect to multi-stage QC procedures (Cummings, 2011). The simplest foresee a two-step process where an automated QC is performed soon after the data are collected (real-time), followed by a semi-automated or a manual analysis involving an expert (Doong et al., 2007; Gronell and Wijffels, 2008; Fiebrich et al., 2010; ARGO, 2012). Often quality is determined without an 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"), and across observing systems definitions vary contextually (NOAA, 2004, 2005, 2006; EuroGOOS, 2010). The NRS network is designed to provide the baselines of multi-decadal time series against which more spatially replicated short-term studies can be referenced and remote sensing products validated to improve understanding of how large-scale, long-term change and variability in the global ocean are affecting Australia's coastal ocean ecosystems (Lynch et al., 2001). One of the goals of IMOS is also to make data collected by the NRS freely available through a dedicated IMOS Ocean Portal (http://imos.aodn.org.au/webportal, under the heading "IMOS – Australian National Mooring Network (ANMN) Facility – WQM and CTD burst averaged data products"). Good quality data are fundamental to the NRS network's mission success. In order to obtain this, a quality control (QC) system is required that will ensure (i) consistency of data among and within samples, and (ii) quality and errors associated with data collected are documented allowing the end user to assess their suitability for purpose.

This exposes us to the quandary of Big Data: while review of data by expert professional oceanographers with strong local knowledge of a sampled area is ideal, the high frequency of data generated by the NRS on a continuous basis makes manual QC labour intensive and expensive (Ingleby and Huddleston, 2007; Gronell and Wijffels, 2008). Moreover, servicing and upload of data from individual NRS stations is distributed across four Australian marine institutes, making consistent application of QC flags a further challenge. Standard data handling and QC processes that can be semi-automated and performed consistently on either individual or consecutive points (depending on what is being tested) are thus necessary to uniformly deal with the NRS sensor data streams.

We describe a hybrid QC strategy that envisages two different automated approaches: a traditional, first-stage, "qualitative" flag-based method that feeds into an experimental "quantitative" system able to calculate uncertainty estimates. Our qualitative data QC uses eleven consecutive tests to classify data according to the flag system of the Intergovernmental Oceanographic Commission (IOC) of UNESCO (UNESCO, 1993, 2009) (Table 1). Our quantitative system employs fuzzy logic (Zadeh, 1965) to develop agreed values which are "gates" of uncertainty. Fuzzy logic is particularly suited to modelling the uncertainty of QC assessments because it allows data quality inferences of partial truth to be made. Further, it can be used to mimic the "fuzzy" intuition of human experts so is well-suited

Table 1
Data flag scheme 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

to oceanographic data where the solution is dependent on both statistical analysis and human experience making complex decisions with different sources of evidence. The expectation is that this system be good enough to allow a significant reduction of the manual second-pass QC by integrating expert knowledge into the fuzzy system.

Our aim is to assess both our standard qualitative QC and the experimental quantitative fuzzy logic QC for the high frequency NRS data. Though the methods described are applicable to most other sensors across the NRS and the ANMN, as a test case we used three years of NRS temperature, salinity and depth data collected from moored multi-sensor Water Quality Monitors (WQM; Wetlabs, Philomath, USA). To evaluate the goodness-for-purpose of each automated approach, the results were compared to those emerging from a manual QC of the same dataset undertaken by one of the authors (KRR), an oceanographer with extensive subject matter expertise and knowledge of the physical oceanography that the NRS observe (Ridgway and Condie, 2004; Ridgway, 2007).

2. Methods and results

2.1. Qualitative approach to data QC

Our qualitative QC system assesses data according to eleven consecutive tests, described in order of execution in Sections 2.1.1–2.1.11, and classifies them according to the IOC flag system (Table 1). Data are not removed, modified or corrected, just flagged. The flags used are 1 (good data), 3 (bad data that are potentially correctable) and 4 (bad data). The final flag assigned to a data point will be the highest obtained as a result of all the tests it was subjected to. Owing to the sequential nature of tests, the discriminations made between flags 3 and 4 are very important. If one test flags a data point as bad (flag 4) that datum will not be subjected to subsequent tests and will carry that flag with it. Whereas if a datum is given a flag 3, it will be subjected to subsequent tests and there will be the possibility of changing the flag to 4, should a subsequent test require it. This is useful for instances when failure of a test is not necessarily critical (e.g. a typographical error in the metadata resulting in an impossible date). The tests that are assigned a flag of 4 in case of failure are tests that imply that data would be meaningless or impossible if outside the determined thresholds (e.g. a datum collected when the instrument is out of the water has no meaning, is flagged 4 and precluded from the remaining tests). On the contrary, the tests whose failure generates a flag 3, are generally less robust and more prone to flagging real outliers, which could be of some interest from a scientific (rather than operational) point of view. A flag 3 serves to attract the user's attention on a critical portion of a dataset.

Our qualitative QC procedures are implemented across the entire NRS and ANMN facility through a centrally version-controlled toolbox (“IMOS toolbox”) developed using the software Matlab (R 2010b, Mathworks Inc., Natick, MA, 2010) and Java (Java Runtime Environment 1.6, Oracle Corporation, Santa Clara, USA, 2009); more information and download links for the toolbox can be found at http://code.google.com/p/imos-toolbox/.

Following is a description of the eleven qualitative tests, in the order in which they are applied in the toolbox.

2.1.1. Impossible date

The impossible date test allows checking of observation dates to ensure that they are sensible and match with the time period of instrument deployment. The test is based on the value of the year only. The first WQMs were deployed in 2008, thus all records with associated dates earlier than 01/01/2008 or beyond the collection date of the data (i.e. into the future) fail the test and are assigned Flag 3.

2.1.2. Impossible location test

The impossible location test requires that the described observation of latitude and longitude of the data collected matches the site. Any observations with locations outside a radius of 2.5 km from the nominated location of the station from which the data were collected fail the test and are assigned a Flag 3. This test also allows for the detection of any discrepancies between data and metadata information.

2.1.3. In/out-water test

Instruments are often turned on prior to deployment and, depending on conditions, are not turned off for many hours after recovery. The in-water/out-of-water test identifies those data collected prior to the deployment of the instrument and after the time of recovery. Between these two times, data are deliberately left non-QC'd.

This test is based on metadata associated with deployments and these can vary depending on different practices (IMOS moorings are deployed by 4 different institutions in 5 geographic nodes all over Australia). As a result in-water and out-of-water times are determined in order of preference from:

- Time of the first/last good datum (expert-derived information, assuming no failure has occurred during the deployment time period),
- Time the sensor is first/last in position (metadata collected on board),
- Time the sensor is first deployed in the water/retrieved and on deck,
- Time the sensor is switched on/off.

Data that fail this test are assigned a Flag 4.

2.1.4. Global range test

The global range test flags data which are beyond the possible bounds of parameter ranges. The parameter ranges used are derived from the ARGO Program QC manual (ARGO, 2012) and for temperature, salinity and depth collected by the WQMs they are:

Temperature: −2.5–40.0 °C
Salinity: 2–41.0 psu
Pressure (depth): −5–12000 dbar (m).

Data that fail this test are assigned a Flag 4.

2.1.5. Regional range

The regional range test is similar to the global range test, but incorporates annual upper and lower boundaries which are defined for each NRS location (Table 2) This was done using frequency distributions of each parameter (e.g. temperature and salinity) calculated from all available and QC'd *in-situ* sensor data (Fig. 2). Within the toolbox a default increase in the boundaries obtained from Fig. 2 is set to capture plausible outliers (Table 2). For salinity these are set to ±3 psu and for temperature ±5 °C.

Data that fall outside the boundaries set for the NRS location are assigned a Flag 4.

2.1.6. Impossible depth

Each NRS comprises a moored sensor package that is suspended in the water column at a nominally fixed depth. Due to oceanographic conditions, however, actual depths of sensors vary as the mooring is 'pulled down' or 'knocked over' by currents. For example, at the Maria Island NRS mooring, which is

Table 2
Ranges for temperature (°C) and salinity (psu) applied in the regional range test.

NRS	Temperature (°C)		Salinity (psu)	
	Min.	Max.	Min.	Max.
Maria Island	5.0	25.0	27.5	38.9
North Stradbroke Island	9.0	33.0	31.3	38.7
Kangaroo Island	6.0	25.1	31.7	39.9
Esperance	10.2	26.5	32.2	39.0
Rottnest Island	9.5	38.0	31.8	39.3
Ningaloo	16.4	35.0	29.9	38.3
Darwin	20.0	38.0	27.75	38.7
Yongala	15.8	44.7	22.0	38.6
Port Hacking	6.8	30.5	29.8	39.2

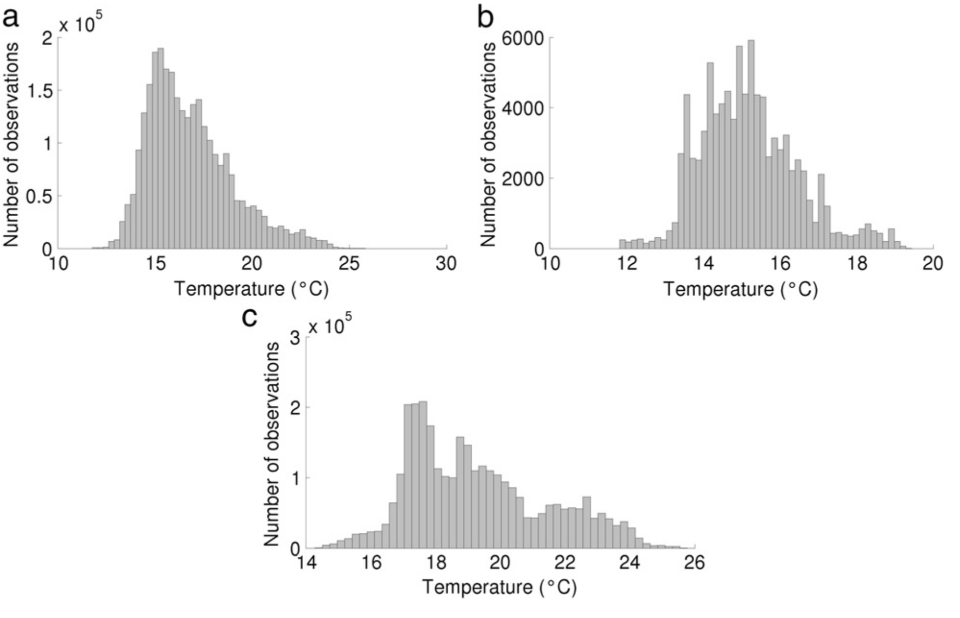

Fig. 2. Cumulative histograms used to derive the regional ranges for temperature at Port hacking using (a) temperature loggers 0–112 m depth (28/10/2009–06/03/2012), (b) ADCPs 108–112 m depth (29/03/2011–02/03/2012) and (c) WQMs 0–110 m depth (03/05/2010–02/03/2012).

considered a taut mooring with minimal knock down, instruments normally moored at a depth of 20 m, have been observed to drop by 20 m at this 100 m site during a strong current (e.g. 2 m s^{-1}) event (D. Hughes pers com). Our impossible depth test compares the depth recorded by the instrument to a designed depth range that the sensors may move across. Depth values recorded that are deeper than the instrument depth range fail the test (Flag 3). Values above the water surface are flagged by the global range test (Section 2.1.4) and the regional range test (Section 2.1.5).

The depth range is determined by a formula which relates the nominal depth of the sensor to the depth of the site via a coefficient (1). The coefficients are the offset bounds in metres for the instrument at its nominated depth. Based on the relatively shallow depths of the NRS moorings the default values for coefficients for the upper and lower bounds of the depth range are 3 m and 5 m, respectively, allowing for more variability on the deeper range as moorings are more likely to be knocked down rather than up by strong current events.

The coefficient, however, can be user-defined, allowing tuning of the flag based on the engineered tautness of the mooring, local conditions and the depth of the site.

$$\text{Range} = \text{Instrument nominal depth} \pm \textit{coeff} * \left(\frac{\text{Site depth}}{\text{Instrument nominal depth}}\right). \tag{1}$$

In addition, because instruments cannot be deployed exactly at a specific depth, any possible mismatch between the nominal depth and the actual depth needs to be accounted for. To do so, we compare the previously obtained range with an error range for the instrument nominal depth (2) and we consider the union of both ranges.

$$\text{Instrument nominal depth range} = \text{Instrument nominal depth} \pm \text{constant margin}. \tag{2}$$

The resulting range gives the minimum (depthMin − 3) and maximum (depthMax − 4) allowed depths.

$$\textit{depthMin} = \min(\text{Range, Instrument nominal depth range}) \tag{3}$$

$$\textit{depthMax} = \max(\text{Range, Instrument nominal depth range}). \tag{4}$$

Finally, these allowed depths must remain within the water column so minimum depth cannot be deeper than the Pressure (Depth) global range (Section 2.1.4) minimum value and maximum depth cannot be greater than the site nominal depth +10%. For example, if given depthMax in (4) is 120 m and the site nominal depth is 100 m then in the end we will have depthMax of 110 m.

2.1.7. Spike test

Unusually large rates of change where one measurement is quite different to adjacent ones in a sequence are defined as 'spikes'. For physical measures, these spikes are usually not actual data but are artefacts or failures from the sensor. This test assesses temporally adjacent readings from the same sensor and employs an algorithm (5) used by the Australian National Facility for Ocean Gliders (ANFOG, 2010) and the European Global Ocean Observation System (EuroGOOS, 2010) which can be described as follows:

$$\text{Test value} = |V_n - (V_{n+1} + V_{n-1})/2| - |(V_{n+1} - V_{n-1})/2| > \text{threshold}. \tag{5}$$

Where V_n is the data point being tested, and V_{n+1} and V_{n-1} are the values preceding and following it. Thresholds, determining pass/fail, used in the algorithm vary depending on the parameter and can be user-defined and modified to take into account, for example, regional differences. For temperature and salinity, standard ARGO/IOC (ARGO, 2010) threshold values are used: temperature > 6.0 °C and salinity >0.9 psu. Those data whose values are beyond these thresholds, fail this test are assigned Flag 3.

This test is used for physical data and is continuous across sampling periods conducted by the WQM (60 s bursts every 15 min). However, neighbouring data points are not compared beyond a 60 min period between them. The maximum time span of 60 min is introduced to avoid comparing data points that are not part of the same temporal string.

2.1.8. Rate of change test

The rate of change test is similar to the spike test but considers both the data point before and after the datum of interest as well as expected rates of change taken from a previous "good" baseline dataset. If the sum of absolute differences between the data point of interest and respectively the data point before and the data point after is four times the standard deviation of the baseline data, the datum is flagged (Flag 3). The algorithm used is:

$$\text{Test value} = |V_n - V_{n-1}| + |V_n - V_{n+1}| \leq 2(2\sigma V). \tag{6}$$

Where V_n is the value of the data point being tested, V_{n-1} and V_{n+1} are the values preceding and following it, and σ is the standard deviation of the baseline dataset. The baseline dataset is set as a default to the first month of good data "so far", i.e. data that has not been assigned a flag greater than 2 (this can be user-defined).

In some cases, there may be no previous or successive point in the time series, but there may still be an anomalous difference between two neighbouring points, e.g. at the beginning of a time series. This is a particular case where their absolute difference is compared to two standard deviations ($2\sigma V$) only:

$$\text{Test value} = |V_n - V_{n+1}| \leq 2\sigma V. \tag{7}$$

Similarly to the Spike test, a 60 min time frame beyond which neighbouring data points will not be compared has been set.

2.1.9. Stationarity test

Given the high precision of the sensors and natural environmental variability, continuous identical readings are unlikely to occur and are probably sensor or transmission failures. This test allows for these sequentially identical measurements to be identified.

The test uses an algorithm set by the IOC (UNESCO, 1993), which determines the maximum number of identical measurements allowed to occur within a data stream based on the sampling interval used.

$$T = 24 * (60/\Delta t)\,. \tag{8}$$

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

Data that fail this test are assigned Flag 4.

2.1.10. Climatology test

The climatology test is similar to the global and regional range tests (Sections 2.1.4 and 2.1.5) but applies finer seasonal scales on range boundaries. Observations are compared to a set of regional reference climatologies (Sections 2.1.10.1–2.1.10.3) to identify erroneous data-points. Observations found to be more than six standard deviations from the monthly mean are determined to be outliers and are assigned Flag 3.

Considering the diverse locations, and associated climates, of the NRS sites, climatologies used in this test are required to be described for each location. Where possible, climatologies at each site are based on independent historical data collected either at that site or at adjacent sites over long enough time periods to generate robust means. Where such observations are not available mean fields are determined from data collected from each NRS site itself. In the following sections we describe three approaches to determining the reference climatology.

2.1.10.1. Climatologies calculated from long-term station data. At three of the NRS sites (Maria Island, Port Hacking, Rottnest Island), historical water sample data have been collected at 0, 10, 20, 30, 40 and 50 m depths over time periods of 60–70 years (http://imos.aodn.org.au/webportal/). These were used to build reference climatologies for those sites. Ocean parameters for each calendar month were interpolated onto a 10 m vertical grid over the full depth of the site. Data were then binned temporally at each depth level at an approximate fortnightly spacing (bin size = 365/26 ∼= 14.04 days). The mean, median, standard deviation, minimum, maximum and number of samples per bin (i.e. per 10 m depth interval per fortnight) were then calculated. The climatology built for temperature data at Maria Island at 10 m depth as an example (Fig. 3).

2.1.10.2. Climatologies calculated from regional data. Where historical data are not available for a NRS, we used data collected from adjacent areas to build reference climatologies for these sites. Temperature and salinity climatologies are thus built using the least-squares technique (loess) developed for the CSIRO Atlas of Regional Seas (CARS) climatology (Ridgway et al., 2002). Custom climatologies were required to be calculated because the CARS resolution of gridded field at 0.5° was too coarse for our coastal NRS sites. Mean seasonal cycles were obtained after allowing for the location in space and season of the regional data; the data covariances were also adjusted as a function of horizontal and vertical data density, bathymetry and land barriers. The latter factor reduces leakage of structure between deep and shallow regions and produces more realistic coastal gradients. The validity of these regionally-derived climatologies was tested using the reference climatologies generated from long-term historical data (Section 2.1.10.1), exemplified for Maria Island (Fig. 4).

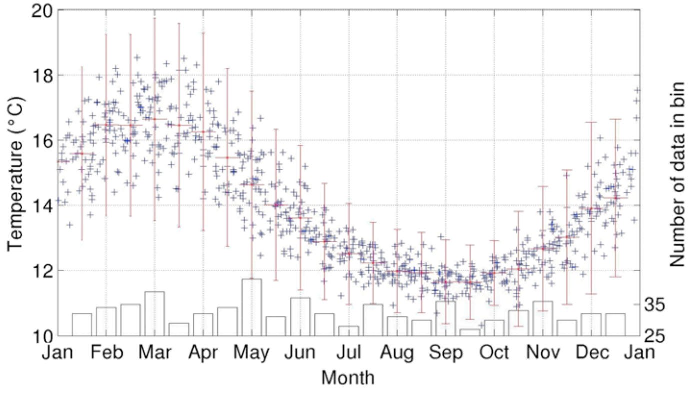

Fig. 3. Example climatology built with the 10 m temperature data (blue crosses) from Maria Island NRS. The number of observations contained in each fortnightly bin is in green, the mean and ±6 standard deviation are in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The regional climatologies were found to capture the seasonal changes in both temperature and salinity and closely match the long-term climatology results (Fig. 4(a)–(d)); the two climatologies being highly correlated (Fig. 4(g)–(h)). The root mean square difference between the two climatologies (for both properties) was in the order of 10%–20% of the monthly standard deviation.

2.1.10.3. Climatologies calculated from mooring data. At some sites the independent regional data distribution is limited and insufficient to construct a climatology. In these cases, climatologies were calculated using observations from the WQM sensors deployed at the site. To do this, sensor data collected across a day (24 h) are averaged and then the same methodology outlined in Section 2.1.10.1 used to calculate the climatology (Fig. 5).

2.1.11. Salinity from pressure and temperature flags test

As salinity is derived from temperature and pressure (depth) it is important to ensure that flags on derived data are correctly associated with flags of instrument-derived observations. This test takes any flags from temperature and pressure, and assigns a flag at least equal to the maximum of the temperature and pressure flags to the salinity observation.

2.2. Maria island fuzzy sets

2.2.1. Fuzzy theory

In contrast to classic logic, which exploits a binary set of truth (1) or falsity (0), fuzzy logic utilises a set of functions that map binary sets onto an interval (0.0–1.0) representing partial truth (or confidence). This is achieved with fuzzy sets (x_i, f_i), which are comprised of an input variable x_i that is mapped onto an interval (0.0–1.0) using the membership function $f_i(x_i)$. A larger value of $f_i(x_i)$, indicates an increased grade of membership. Logic rules can be used to formulate the relationship between I_i and Q_i using their associated membership functions m_i and k_i, respectively. An example of a logic rule is as follows:

$$\text{If } I_A \text{ is } m_A \quad \text{or} \quad I_B \text{ is } m_B, \quad \text{then } Q_C \text{ is } k_C. \tag{9}$$

The character operators of logic rules determine the mathematical operation that will be performed between the values of the membership functions. The term '*is*' represents the mapping of the inputs onto the interval (0, 1) via each of their respective membership functions i.e. $m_A(I_A)$, $m_B(I_B)$. The '*or*' operator of the Mamdani system (Mamdani and Assilian, 1975) equates to finding the maximum value between $m_A(I_A)$ and $m_B(I_B)$ in (9). The resulting membership value is then mapped onto the output membership function k_C to produce the fuzzy output Q_C. The output can then be 'defuzzified' by computing a geometrical attribute of Q_C, such as the bisector that is described in Section 2.2.5.

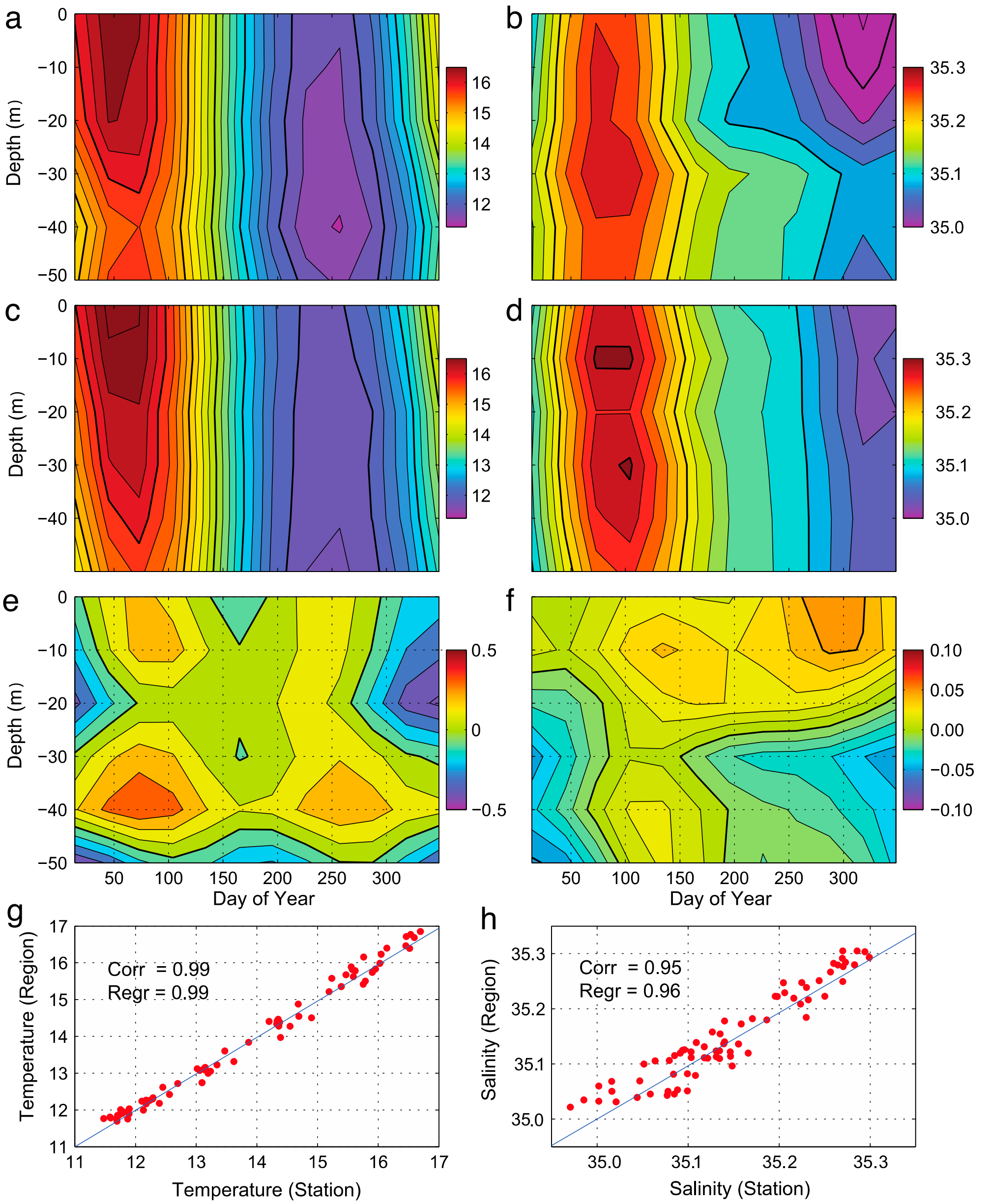

Fig. 4. Temperature (°C) and salinity (psu) climatologies from the Maria Island NRS at each depth for each day of the year from the long-term monthly sampling (a and b, respectively) and from the regional climatology (c and d, respectively). The difference plots are shown in (e) and (f). The station climatologies are plotted against the regional climatologies in (g) temperature and (h) salinity, the lines are the linear best-fit to the data. The correlation and regression coefficients are also shown.

Fuzzy logic has been widely used in the control systems of a number of consumer devices including vacuum cleaners, cameras and washing machines (Reznik, 1997). For instance, fuzzy controllers are used in washing machines to optimise water volume and washing time based upon sensor information provided about the weight of clothes and the impurity of water. In this case, fuzzy rules map the relationship between the sensory inputs of water impurity and clothes quantity to output functions

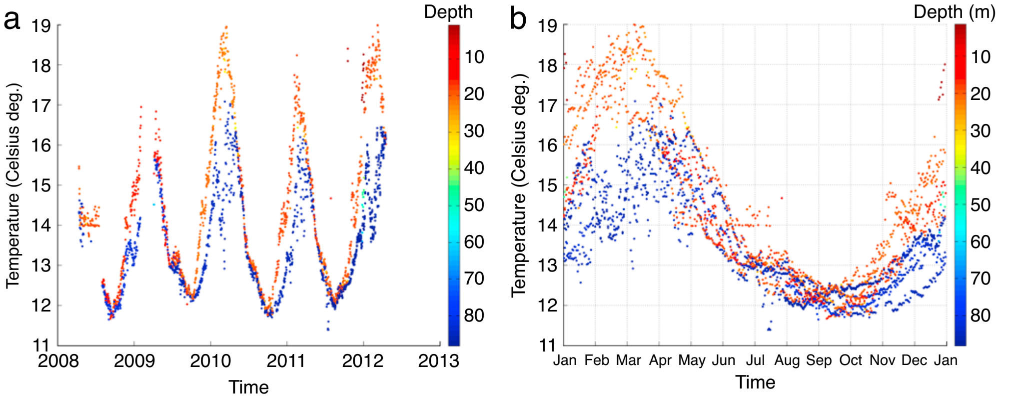

Fig. 5. (a) WQM raw temperature data for the Maria Island site and (b) associated calculated climatologies at 20 m (red) and 90 m (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that control the water volume of water and washing time. The usefulness of fuzzy logic has been attributed to its ability to represent information uncertainty (or imprecision), and incorporate human experience into the real devices. These two characteristics of fuzzy logic are desirable properties for a quality control scheme for sensor data.

2.2.2. Data

The parameters of the quantitative assessment scheme were computed from the statistical distributions of the quality tests calculated across 4 years of historical temperature and salinity observations acquired from WQMs at the Maria Island NRS at 20 and 90 m (available from the CSIRO Atlas of Regional Seas, CARS, http://www.cmar.csiro.au/cars). The dataset that was then used to evaluate and compare the qualitative and quantitative schemes and the manual expert QC, comprised approximately 23 months of observations from the same WQMs between the 24/11/2009 and 24/10/2011 (found at http://imos.aodn.org.au/webportal). The temperature and salinity observations were sampled at 1 Hz.

2.2.3. Fuzzy inputs

Fuzzy sets were formulated for three QC tests: spike (x_s), rate of change (x_{rc}) and climatology (x_c) (Sections 2.1.7, 2.1.8 and 2.1.10, respectively). The input variables of temperature and salinity were mapped to three overlapping membership functions of trapezoidal shape, of low (l), medium (m) and high (h) test uncertainty (Fig. 6). The bases (parallel sides) of the trapezoid provided complete membership to the function, whilst the legs provided partial membership overlapping with the legs of adjacent membership functions. The spike and rate of change input variables were calculated according to the tests specified in (5) and (6) respectively, whilst the input variables for the climatology test were calculated as the absolute difference between the test sample and the mean of the corresponding date aligned climatology bin b (i.e. $|V_n - \mu_b|$). Fuzzy sets were parameterised from the defined thresholds of the qualitative tests and analysis of historical data-sets. Specifically, the lower vertex of the "High" membership base was set to the threshold value of its corresponding test (i.e. indicating there was a very high confidence of the sensor reading being erroneous). A cumulative histogram of each test variable x_s, x_{rc} and x_c was then computed over four years of historical observations. The observations of the test variable that exceed the test threshold were omitted from the cumulative histogram. The remaining membership functions of each test variable were then parameterised by using the associated histogram (Table 3). The upper vertex of the "Low" membership base was computed at the test value of the 98th percentile of the histogram. The lower vertex of the "Medium" membership base was computed at the test value of the 99th percentile of the histogram. The upper vertex

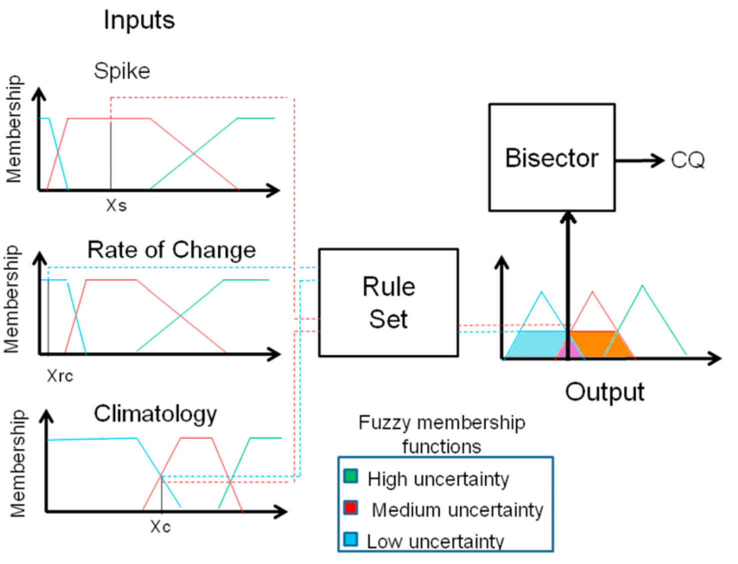

Fig. 6. Schematic of the fuzzy quality assessment system for temperature observations at Maria Island NRS. The space of each test input (Spike, Rate of Change and Climatology) is represented by three membership functions of "Low", "Medium" and "High" uncertainty. The membership of each function lies on the interval (0–1). Inputs are mapped onto these functions to compute their contributions. Fuzzy logic rules (see Section 2.2.4) are then used to combine the contributions of equivalent membership functions across the three tests. These combined memberships are then mapped to the output membership functions of "Low", "Medium" and "High" uncertainty. The output membership functions are "defuzzified" using the bisector operator that finds the x-axis value that breaks the area under the output functions into two equal parts. This value becomes the quality value CQ.

Table 3
The fuzzy membership intervals for the quantitative tests for temperature and salinity observations at the Maria Island NRS.

Variable	IMOS test	Low	Low → Medium	Medium	Medium → High	High
Temperature	Vertical spike (s)	$x_s < 0.07$	$0.07 \le x_s < 0.2$	$0.2 \le x_s \le 2$	$2 < x_s \le 6$	$x_s > 6$
	Rate of change (rc)	$x_{rc} < 0.5\sigma_{rc}$	$0.5\sigma_{rc} \le x_{rc} < 1.5\sigma_{rc}$	$1.5\sigma_{rc} \le x_{rc} < 3\sigma_{rc}$	$3\sigma_{rc} \le x_{rc} < 4\sigma_{rc}$	$x_{rc} > 4\sigma_{rc}$
	Climatology (c)	$x_c < 3\sigma_c$	$3\sigma_c \le x_c < 4\sigma_c$	$4\sigma_c \le x_c < 5\sigma_c$	$5\sigma_c \le x_c < 6\sigma_c$	$x_c > 6\sigma_c$
Salinity	Vertical spike (s)	$x_s < 0.05$	$0.05 \le x_s < 0.15$	$0.15 \le x_s \le 0.5$	$0.5 < x_s \le 0.9$	$x_s > 0.9$
	Rate of change (rc)	$x_{rc} < \sigma_{rc}$	$\sigma_{rc} \le x_{rc} < 2\sigma_{rc}$	$2\sigma_r \le x_{rc} < 3\sigma_{rc}$	$3\sigma_r \le x_{rc} < 4\sigma_{rc}$	$x_{rc} > 4\sigma_{rc}$
	Climatology (c)	$x_c < 3\sigma_c$	$3\sigma_c \le x_c < 4\sigma_c$	$4\sigma_c \le x_c < 5\sigma_c$	$5\sigma_c \le x_c < 6\sigma_c$	$x_c > 6\sigma_c$

of the "Medium" membership base was computed at the test value of the 99.5th percentile of the histogram.

Apart from the 'High' uncertainty membership, the fuzzy parameters of this model were derived from the probability distributions of historical observations. When historical records are available over sufficiently long periods at a site, statistical tests for observations can be utilised. The advantage of fuzzy logic, however, is its flexibility to encode other information such as expert opinion, or in this case, test thresholds from established programs in conjunction with statistics. Furthermore, when historical observations are unavailable at a site, it could be used to provide an initial inference of data quality based upon expert opinion alone. The system can capture the uncertainty associated with such opinion through the fuzzy sets.

Table 4
Fuzzy rules used to map the quality test input functions to the output membership functions.

Fuzzy logic rule	Mathematical operations
If x_s is low and x_{rc} is low and x_c is low then Q is low uncertainty	$Q_l = K_l(u_l)$ $u_l = mean(S_l(x_s)RC_l\,(x_{rc})\,C_l(x_c))$
If x_s is medium and x_{rc} is medium and x_c is medium then Q is medium uncertainty	$Q_m = K_m(u_m)$ $u_m = mean(S_m\,(x_s)\,RC_m\,(x_{rc})\,C_m(x_c))$
If x_s is high or x_{rc} is high or x_c is high then Q is high uncertainty	$Q_h = K_h(u_h)$ $u_h = max(S_h(x_s)RC_h\,(x_{rc})\,C_h(x_c))$

2.2.4. Fuzzy rules

Three logic rules were formulated to express the connection between the input memberships and output memberships (K_i) of each test (Table 4). Average membership of the three input variables was calculated using 'and' operators, whilst the maximum membership value of the three inputs was calculated using 'or' operators. High uncertainty membership of tests were combined with the 'or' operator, on the basis that the presence of high uncertainty for a single test provides evidence to suggest the sample is of dubious quality, independent of the uncertainty levels of the other tests. In the case of the low and medium uncertainties, however, the mean of the corresponding test membership was computed to provide a combined result at each uncertainty level.

2.2.5. Fuzzy outputs

The output fuzzy set consists of three overlapping triangular membership functions of low, medium and high uncertainty (Fig. 6). The crisp quality assessment outputs span between 0 and 1 for all uncertainty extremes so when all the tests were completely associated with a 'low uncertainty' the output was 0 and when all the tests were completely associated with a 'high uncertainty' the output was 1.

The fuzzy input memberships were applied to the logic rules in Table 4 in order to map to the specified output membership functions. The point at which the output area of the combined memberships is divided into two equal intervals is known as the bisector (Fig. 6). A continuous quality assessment (CQ) is computed as the bisector of the area under the output of the combined memberships such that:

$$\sum_{i=l,m,h}\sum_{l=0}^{CQ} Q_i x_l = \sum_{i=l,m,h}\sum_{l=CQ}^{end} Q_i x_l. \tag{10}$$

This is then used to aggregate test membership into a CQ and provide a quality assessment.

2.2.6. Comparison of the expert QC with qualitative and fuzzy logic schemes

The Maria Island NRS temperature (Fig. 7(a)) and salinity datasets were also quality assessed by one author (KRR) who, as a physical oceanographer with extensive experience in data collected at the Maria Island NRS, provided an 'expert' assessment against which the qualitative (derived from the tests described Section 2.1) and quantitative QC schemes could be assessed (Fig. 7), with data flagged as '1' when considered 'bad' data and as '0' when considered 'good' data.

The expert (Fig. 7(b)), qualitative flag (Fig. 7(c)) and fuzzy logic QC (Fig. 7(d)) assessments were shown to be highly correlated across salinity and temperature series, consistently identifying extreme outlier samples with significant uncertainty. These outliers were flagged as Flag 4 by the qualitative test scheme (Fig. 7(c)) while the fuzzy logic test scheme consistently produced CQ values above 0.85 (Fig. 7(d)). This indicates that both semi-automated QC schemes were capable of flagging samples that were clearly erroneous. In general, however, the quality assessments of the two semi-automated schemes did not completely match the expert quality assessments generated by the oceanographer (Table 5). Table 5 quantifies the overall number (and percentage) of expert assessed 'good' and 'bad' samples that were in agreement with the quality assessments produced by each of the semi-automated schemes. The two QC schemes were far more conservative in flagging samples as

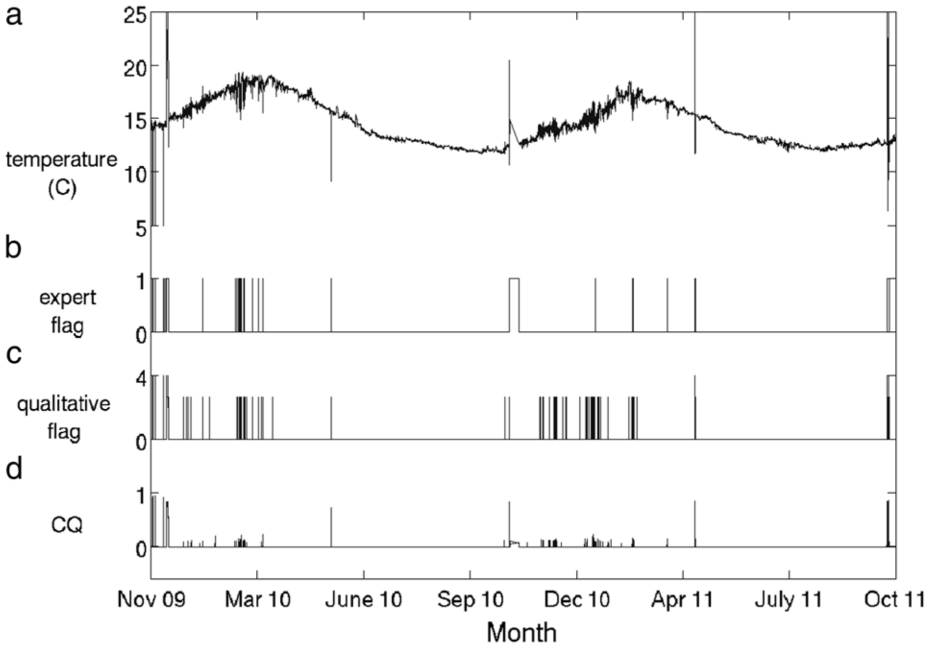

Fig. 7. Comparison of the quality assessment schemes against the expert assessment. (a) Raw temperature observations acquired at Maria Island NRS at 20 m from 24/11/2009 to 24/10/2011; (b) expert assessment; (c) qualitative assessment scheme (QC flag derived from the application of the tests described in Section 2.1); (d) quantitative assessment scheme (fuzzy logic CQ, derived as described in Section 2.2).

Table 5
A comparison of the quality assessments produced by two semi-automated QC schemes (qualitative and quantitative) with the benchmark quality assessments produced by an oceanographer expert for temperature and salinity observations acquired at Maria Island NRS (at 20 and 90 m) between the 24/11/2009 and 24/10/2011. The number (%) of correctly assessed observations refers to the samples where the semi-automated quality assessments were in agreement with the corresponding expert based quality assessments of 'good' or 'bad'.

	Qualitative tests (QC flags)		Quantitative tests (fuzzy logic)		Expert assessment	
	Number (%) of observations correctly assessed as 'bad'	Number (%) of observations correctly assessed as 'good'	Number (%) of observations correctly assessed as 'bad'	Number (%) of observations correctly assessed as 'good'	Number of observations assessed as 'bad'	Number of observations assessed as 'good'
Temperature (20 m)	3,962 (23.4%)	3,939,927 (99.9%)	4,334 (25.6%)	393,955 (99.9%)	16,932	3,943,900
Temperature (90 m)	115,810 (62.4%)	2,495,604 (66.3%)	103,370 (55.7%)	2,796,780 (74.3%)	185,590	3,764,100
Salinity (20 m)	9,227 (34.1%)	3,930,148 (100.0%)	9,579 (35.4%)	3,930,148 (100.0%)	27,059	3,930,148
Salinity (90 m)	16,631 (59.0%)	2,667,776 (68.0%)	15,137 (53.7%)	2,899,210 (73.9%)	28,188	3,923,200

'bad' when compared to the oceanographer. The percentage of 'bad' samples that were correctly classified by the QC (23.40%–62.43%) was consistently lower than the percentage of 'good' samples that were correctly classified in each of the data-sets (66.30%–100.00%). Neither of the QC approaches performed particularly better than each other when compared to the expert dataset. The major difference between the two semi-automated schemes was that for the fuzzy scheme, the level of uncertainty varied between samples, so that samples that were not clearly erroneous were flagged with a lower level of quality uncertainty.

3. Discussion

IMOS's network of National Reference Stations is aimed at providing freely accessible, long-term, high-frequency oceanographic data to the wider community of marine and atmospheric scientists as well as to the general public. To do this, data are collected by a series of sensors moored at nine coastal sites around Australia. There are numerous challenges in the development and implementation of re-locatable quality control systems for sensors. For instance the 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, rates of change that can occur between observations, and rates of biofouling of instruments. The continental scale of the network thus affects both the rate at which each sensor drifts over time and regional ranges observed by sensors. At smaller scales, sensors located in a standing wave, sink hole or eddy may also observe phenomena that fall well outside a nominated or identified regional range. The quality of each data point recorded by these instruments thus has to be controlled and annotated, before the data are disseminated.

The quality control approaches described here were developed in response to the specific challenges of large, long-term observing programs such as IMOS. Ideally, to address all of these issues and others, quality control checks should be performed by a local expert, but both the quantity and scale (temporal and spatial) of data generated by this system makes this type of consistent manual control difficult. Semi-automated systems integrating manual QC with automated procedures have been shown to provide significant added value to the QC process (Gronell and Wijffels, 2008; Fiebrich et al., 2010), but they still require the time of an expert. In this paper we describe and compare two different methods of performing quality control, which do not entail the direct involvement of an expert, with the ultimate intention of providing the scientific community with a platform of known performance for exploration of the data. We view our QC scheme as providing a method of data annotation, rather than removal, allowing the end user to easily sort data-streams via an explicit level of uncertainty. Our overall intention is thus to flag data as suspect, not to remove data that are flagged as bad. This system also provides a robust framework for tracking both the QC decisions and any changes to how these decisions are applied.

Our initial expectation in undertaking the development of the two assessment schemes was that by providing a framework of tests and climatologies we could (i) greatly reduce the burden of manual quality control; (ii) improve the accuracy of the system and; (iii) provide quality assurance to the assessments given. Comparison of the results from the different systems, however, identified that while they were able to detect gross errors, our procedures were much more conservative compared to the subtle understanding of an expert undertaking manual QC. A better representation of the capabilities of human expertise is required by both schemes.

The differences observed between the qualitative (QC flag) and quantitative (fuzzy logic) schemes and expert assessment suggest that the former two are, to some degree, limited in replicating the capabilities of a human expert. One important limitation is their inability to assess samples across broader temporal and/or spatial contexts. A human expert will generally not assess samples independently, but as a temporally connected group of samples, or spatially, with respect to other sensors. Hence, the quality of each sample is heavily influenced by other observational sources or the uncertainty associated with past and future samples. In contrast, the proposed semi-automated schemes assess each sample with a high degree of independence.

For this reason, unlike the expert analysis, our qualitative and fuzzy logic tests are not subjective. In many cases, the qualitative tests are binary sets based upon a quality metric of the data stream i.e. in-water *vs.* out of water, stationary data over a long period in a naturally variable environment or globally erroneous data such as water temperatures greater than 100 °C. Our approach is a hybrid that uses fuzzy logic to combine quality tests and produce a continuous quality assessment (between 0 and 1) to represent more complex situations where the quality of the samples is unclear. The conservative nature of the semi-automated systems provides a solution for QC with a known risk. Our software systems should thus be seen as robust low-pass filters of the data with subsequent expert review of data flagged as 'bad' to be recommended.

The continuous quality variable generated by the current approach acts as an "index". While this provides some utility (e.g. the user may, over time, develop an understanding of which index value acts as an appropriate threshold to select data of suitable quality for their application), it is not the most desirable approach. Ideally the continuous quality variable would be used as a starting point to generate error bars on the data (Timms et al., 2011), allowing the user to select data points which were fit for purpose. The real advantage of a fuzzy-logic based approach (or an approach utilising Bayesian statistics) is that it provides a mechanism for combining statistically-based measures of uncertainty with insight from a human expert.

Quality assessment may depend on the use of the dataset. From an operational point of view, varying degrees of strictness in the quality control could be implemented via the use of different thresholds applied to both the qualitative and quantitative components of the system. Thus if in need of a conservative approach, high thresholds should be set on the CQ index whilst if in need of more data and less concerned by how doubtful they are the threshold applied to the CQ index could be lowered. When putting together a quality control system such as the one we describe, the choice of the most appropriate threshold may be an iterative process.

Finally, our climatology flags are somewhat unusual compared to our other tests as they assume a stable climatology, which we know to be false. Currently, most regions in global ocean are undergoing some form of long-term change, particularly apparent for our data site (Hill et al., 2008), and also exhibit variability on shorter time scales with phenomena such as the El-Niño Southern Oscillation (Ridgway, 2007). Unless short-medium term variability and longer-term change are accounted for in calculated climatologies they have limited capacity to provide reliable bounds for QC. These tests thus need to be either (a) regularly re-calibrated to take into account drift in the climate signal or (b) seen as an indicator for climate change. At the moment we use six standard deviations from the mean as a threshold for determining if data are 'good' or 'bad' with respect to the climatology test. This makes the test very weak, as 99.9% of a normally distributed variable should be within three standard deviations. Although a threshold of six standard deviations from the mean will only flag completely erroneous data-points (e.g. related to instrument malfunction), it ensures that real anomalies will not be discarded.

In conclusion, as it is structured now, the hybrid strategy we present here is a sophisticated first level QC strategy that can be applied to both real-time and delayed-mode data but that still requires a more in-depth analysis in a second instance which will still require an expert. Future work should include a more structured involvement of experts in the automated test inputs, i.e. in the setting of flag thresholds and in training the fuzzy system to incorporate local knowledge. Further, modelling the sensor and sensor platform (arrays of sensors observing different phenomena) behaviour by using methods that provide probability distributions for sets of possible sensor readings (given the current model state) will contribute significantly in advancing the system. Bayesian quality control techniques which incorporate observed system behaviour and expert belief of physical phenomena (Smith et al., 2012) and time warping pattern matching and function curve fitting techniques (Shahriar et al., 2011) are potential approaches which can be explored. Pattern matching systems can be used to find similar patterns to a slice of the current time series in the historical dataset in order to forecast what is likely to occur in the near future (Shahriar et al., 2011).

Acknowledgements

The authors would like to thank two anonymous reviewers for their valuable contribution towards significantly improving the paper; and Daniel Hugo for comments on an earlier draft. Thanks to Jeff Dunn for providing excellent advice on development of climatologies; to Ming Feng for assisting in the identification of bounds for QC tests; to Dave Hughes for providing advice on the behaviour of moorings in real conditions. Funding for the development of the NRS was provided by the Australian Government through the National Collaborative Research Infrastructure Strategy (2007–2011), the Education Investment Fund (2009–2013) and the Collaborative Research Infrastructure Scheme

(2013–2014). Co-investment has been provided by the Queensland State Government, NSW State Government, the Northern Territory Government, CSIRO, AIMS, SIMS and SARDI.

References

ANFOG, 2010. Australian National facility for Ocean Gliders Data Management User's manual. Version 3. The University of Western Australia, p. 22.

ARGO, 2012. Argo quality control manual, version 2.7. Argo data management, 3 January 2012, p. 47.

Busalacchi, A., 2010. Celebrating a decade of progress and preparing for the future: ocean information for research and application, in: Proceedings of the OceanObs'09 Conference (Vol 1): 45–52. http://dx.doi.org/10.5270/OceanObs09, 10pp.

Cummings, J.A., 2011. Ocean data quality control. In: Schiller, A., Brassington, G.B. (Eds.), Operational Oceanography in the 21st Century. Springer, Dordrecht, p. 736.

Doong, D.J., Chen, S.H., Kao, C.C., Lee, B.C., Yeh, S.P., 2007. Data quality check procedures of an operational coastal ocean monitoring network. Ocean Eng. 34, 234–246.

EuroGOOS, 2010. Recommendations for in-situ data near real time Quality Control. EG10.19, 23 pp. http://www.coriolis.eu.org/content/download/4920/36075/file/Recommendations%20for%20RTQC%20procedures_V1_2.pdf.

Fiebrich, C.A., Morgan, C.R., McCombs, A.G., Hall, P.K., McPherson, R.A., 2010. Quality assurance procedures for mesoscale meteorological data. J. Atmos. Ocean. Technol. 27, 1565–1582.

Gronell, A., Wijffels, S.E., 2008. A semiautomated approach for quality controlling large historical ocean temperature archives. J. Atmos. Ocean. Technol. 25, 990–1003.

Hill, K., Moltmann, T., Proctor, R., Allen, S., 2010. The Australian integrated marine observing system: delivering data streams to address national and international research priorities. Mar. Technol. Soc. J. 44 (6), 65–72.

Hill, K.L., Rintoul, S.R., Coleman, R., Ridgway, K.R., 2008. Wind forced low frequency variability of the east Australia current. Geophys. Res. Lett. 35, L08602. http://dx.doi.org/10.1029/2007GL032912.

Ingleby, B., Huddleston, M., 2007. Quality control of ocean temperature and salinity profiles — historical and real-time data. J. Mar. Syst. 65, 158–175.

Lynch, T.P., Morello, E., Middleton, J.F., Thompson, P., Feng, M., Richardson, A., Doherty, P., Roughan, M., Morris, B., Tilbrook, B., Steinberg, C., James, C., Allen, S., Hill, K., Moltmann, T., 2001. Scientific Rationale and implementation plan for the Australian National Reference Station (NRS) Network. A report to the IMOS Office. http://imos.org.au/anmndocuments.html.

Mamdani, E.H., Assilian, E.M., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. 7 (1), 1–13.

Mayer-Schönberger,, Cukier, K., 2013. Big Data, A Revolution That Will Transform How We Live, Work and Think. John Murray, ISBN: 1848547900, p. 256.

NOAA, 2004. First Workshop Report on the Quality Assurance of Real-Time Ocean Data, December 2005, 89pp, Stennis Space Centre, MS, USA.

NOAA, 2005. Second Workshop Report on the Quality Assurance of Real-Time Ocean Data, July 2005, 48pp, Norfolk, VA. CCPO Technical Report Series No. 05-01.

NOAA, 2006. Fourth Workshop Report on the Quality Assurance of Real-Time Ocean Data, 124pp, Woods Hole, MA, USA.

Reznik, L., 1997. Fuzzy Controller Handbook: How to Design Them, How They Work. Newnes, Oxford, UK, p. 287.

Ridgway, K.R., 2007. Long-term trend and decadal variability of the southward penetration of the east Australian current. Geophys. Res. Lett. 34, L13613. http://dx.doi.org/10.1029//2007GL030393.

Ridgway, K.R., Condie, S.A., 2004. The 5500-km-long boundary flow off western and southern Australia. J. Geophys. Res. 109, C04017. http://dx.doi.org/10.1029/2003JC001921.

Ridgway, K.R., Dunn, J.R., Wilkin, J.L., 2002. Ocean interpolation by four-dimensional least squares — application to the waters around Australia. J. Atmos. Ocean. Technol. 19, 1357–1375.

Shahriar, M.S., Timms, G.P., de Souza, P.A., 2011. Continuous anomalous pattern monitoring for marine sensor data, presented at IEEE Oceans 2011, Kona, Hawaii, USA, 19–22 September 2011.

Smith, D.V., Timms, G.P., de Souza, P.A., D'Este, C., 2012. A Bayesian framework for the automated online assessment of sensor data quality. Sensors 12 (7), 9476–9501.

Thompson, P., Baird, M.E., Ingleton, T., Doblin, M.A., 2009. Long-term changes in temperate Australian coastal waters: implications for phytoplankton. Mar. Ecol. Prog. Series 394, 1–19. http://dx.doi.org/10.3354/meps08297.

Timms, G.P., de Souza Jr., P.A., Reznik, L., Smith, D.V., 2011. Automated data quality assessment of marine sensors. Sensors 11, 9589–9602.

UNESCO, 1993. Manual of Quality Control Procedures for Validation of Oceanographic Data. In: Manual and Guides, vol. 26. p. 437.

UNESCO, 2009. GTSPP Real-time Quality Control Manual. In: IOC Manual and Guides, vol. 22. p. 148.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8, 338–353.