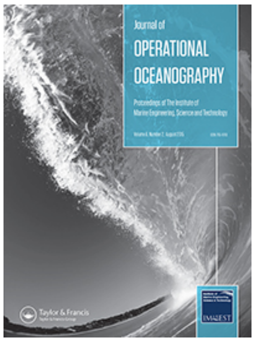


Journal of Operational Oceanography

ISSN: 1755-876X (Print) 1755-8778 (Online) Journal homepage: http://www.tandfonline.com/loi/tjoo20

# Argo real-time quality control intercomparison

R. Wedd, M. Stringer & K. Haines

**To cite this article:** R. Wedd, M. Stringer & K. Haines (2015) Argo real-time quality control intercomparison, Journal of Operational Oceanography, 8:2, 108-122, DOI: 10.1080/1755876X.2015.1087186

**To link to this article:** http://dx.doi.org/10.1080/1755876X.2015.1087186

Published online: 09 Dec 2015.

*Journal of Operational Oceanography*, 2015
Vol. 8, No. 2, 108–122, http://dx.doi.org/10.1080/1755876X.2015.1087186

Taylor & Francis
Taylor & Francis Group

# Argo real-time quality control intercomparison

R. Wedd[a*], M. Stringer[b] and K. Haines[b]

[a]*Australian Bureau of Meteorology, Melbourne, Australia;* [b]*Environmental Systems Science Centre, University of Reading, Reading, UK*

The real-time quality control (RTQC) methods applied to Argo profiling float data by the United Kingdom (UK) Met Office, the United States (US) Fleet Numerical Meteorology and Oceanography Centre, the Australian Bureau of Meteorology and the Coriolis Centre are compared and contrasted. Data are taken from the period 2007 to 2011 inclusive and RTQC performance is assessed with respect to Argo delayed-mode quality control (DMQC). An intercomparison of RTQC techniques is performed using a common data set of profiles from 2010 and 2011. The RTQC systems are found to have similar power in identifying faulty Argo profiles but to vary widely in the number of good profiles incorrectly rejected. The efficacy of individual QC tests are inferred from the results of the intercomparison. Techniques to increase QC performance are discussed.

## Introduction

The accuracy of the initialized ocean state is key to forecasting short-range ocean conditions and seasonal global climates. Accurate initialization is heavily dependent on the quality and quantity of observational data. While remote sensing provides large quantities of information about the state of the ocean surface, the ocean sub-surface is not so easily observed. Prior to 2000 the primary methods of sub-surface observation were fixed buoy systems (TAO, TOGA; Hayes et al. 1991) and ships of opportunity (SOOP; Goni et al. 2010). Both of these methods have poor spatial sampling outside of the equatorial Pacific, and SOOP additionally have irregular temporal sampling. The Argo float project, implemented in 2000, provides high-frequency observations with greater spatial uniformity (Roemmich et al. 1998).

The standard operating procedure of an Argo float is to descend to drifting depth (usually 1000 m) and follow ocean currents for a period of 8 to 10 days (Roemmich et al. 1998). The float then descends to profiling depth (usually 2000 m), and rises to the surface over a period of approximately 10 hours. Measurements of temperature, salinity and pressure are taken during the ascending phase. The float spends time at the surface (more than 10 hours for floats using the Argos communications systems, and ~30 minutes for floats using the Iridium system; Roemmich et al. 2009), during which time the recorded data is transmitted via satellite, and the next cycle begins. The measurements of a single ascent are called a ‘profile’, while each individual measurement is a ‘level’. There are generally between 200 and 2000 levels in each profile, depending on which communication system is used.

Over 3600 Argo floats are currently recording oceanic profiles with a frequency of approximately 10 days. This constitutes a large volume of important data (Current status of the Argo fleet, www.argo.uscd.edu). The highest standard of Argo quality control (QC) involves statistical analysis combined with direct scientific examination by experts. This process, known as delayed-mode QC (DMQC), is labour intensive and cannot yet be implemented in time to be utilized in real-time operational applications. Fully automated QC is required to process the data in a timely fashion. This is known as real-time QC (RTQC).

A schematic demonstrating the flow of data from Argo floats to end users is shown in Figure 1. National Argo Data Assembly Centres (DACs) collect the raw Argo profiles and implement standard RTQC tests set by the Argo Data Management (ADM) group (Wong et al. 2013). Profiles that pass the ADM RTQC are sent to the Global Telecommunications System (GTS) for download and use in operational oceanography. The DACs also send the full data, including RTQC flags, to the Global Argo Data Assembly Centres (GDACs): Coriolis (CRS) in France and the US-GODAE in the United States (US). Regional operational forecast centres download the Argo profiles from either or both of these sources and apply specialized RTQC procedures in addition to the ADM RTQC.

In 2002 the GODAE Ocean Data Quality Control Intercomparison Project was proposed (Cummings et al. 2010). RTQC data from GODAE members’ institutions was collected, with the goal of an analytical intercomparison in the future. This work compares the RTQC methods utilized by members of GODAE OceanView (Le Traon et al. 2010),

[*]Corresponding author. Email: r.wedd@bom.gov.au

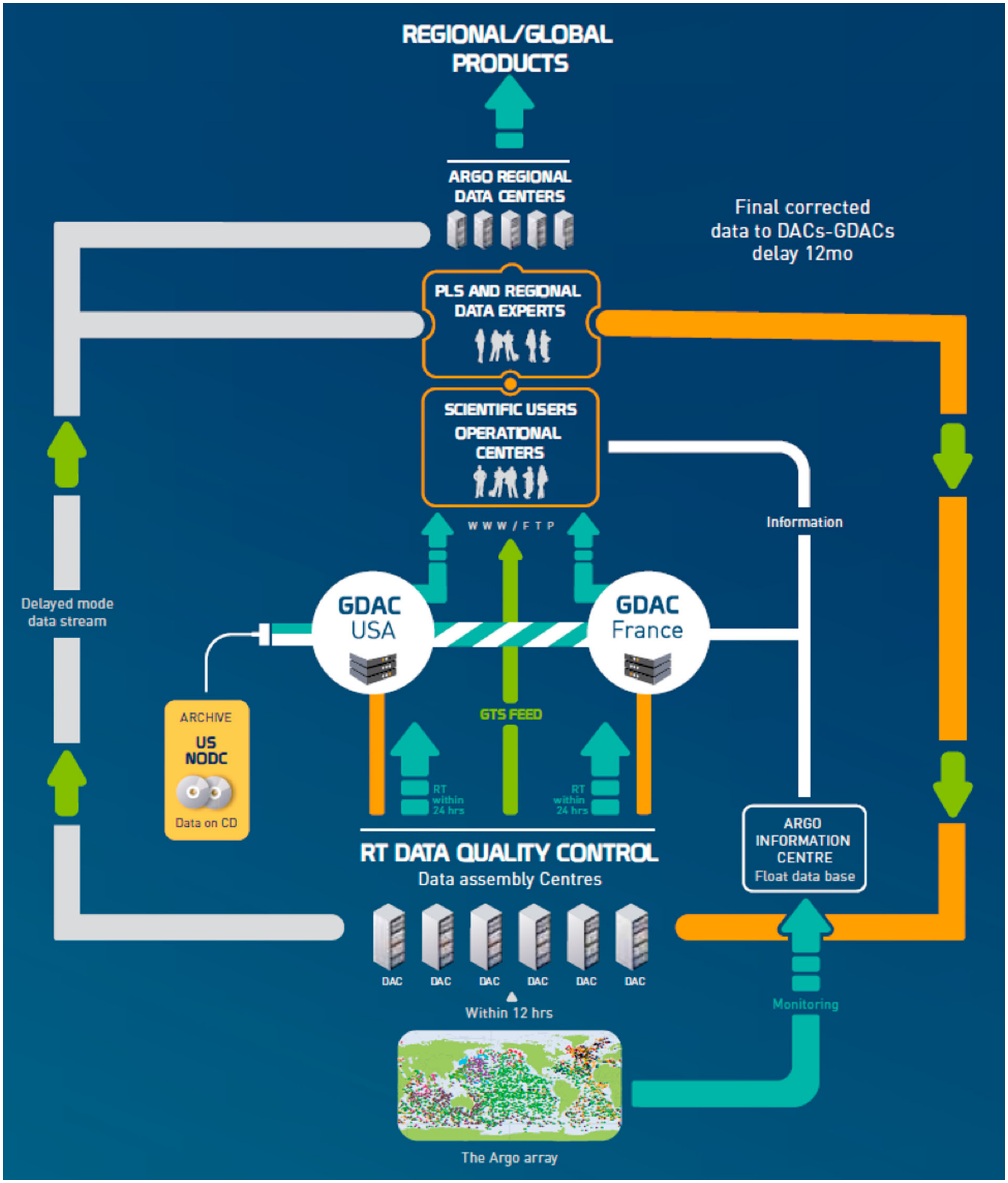


Figure 1. Schematic of data flow from the Argo float array through the DACs to operational centres via the GTS and GDACs. Source: Euro-Argo (http://www.euro-argo.eu/).

the successor of GODAE, using the collected RTQC data. Members must have available RTQC data in a readily readable format to be included. The included centres are: the United Kingdom Met Office (UKMO), the Australian Bureau of Meteorology (BoM), and the Fleet Numerical Meteorology and Oceanography Centre (FNMOC) in the US. Data has also been included from the CRS GDAC. The CRS is not an operational centre but does implement RTQC on Argo profiles for use in the French Hydrographic Service (SHOM; SHOM 2010) and the French Research Institute for the Exploitation of the Sea (IFREMER; IFREMER 2010) operational products.

The benchmark for the intercomparison is the DMQC, which becomes available at a delay of several months to several years after the profile is recorded. The results of the DMQC process are the best assessment available and are considered here to be the 'truth'. The DMQC process involves expert operators checking temperature, salinity and pressure profiles manually for proper overall shape and consistency. The profile must be internally consistent across all variables and also consistent with readings from neighbouring floats. Historical data from the float is checked to assess sensor drift. Profiles are adjusted for detected sensor drift or offset, the profile error estimates are updated and the profile and levels are assigned QC flags (Wong et al. 2013). In order to ensure sufficient DMQC data are available, only the years before 2012 have so far been considered. Data is taken from the five years from 2007 to 2011 inclusive, as before 2007 the CRS does not have data available. Profiles that have been adjusted by the DMQC are not included.

The assimilation of faulty profiles can be extremely detrimental to the quality of an analysis, especially if isolated in space and time. This is mitigated by the fact that profiles passing basic QC tests are less likely to have a catastrophic effect on an analysis: for example, a small drift in pressure will just add to systematic error, while a large temperature spike could produce spurious gravity waves that will detrimentally and persistently affect analysis accuracy. The first example could pass basic RTQC tests, while the second would be unlikely to do so. Removing catastrophically bad profiles should be the highest priority; the removal of slightly faulty profiles must be balanced against the systematic errors introduced by a lack of data. Though the Argo programme provides many profiles for analysis, the spatial and temporal coverage is still sparse in comparison to the size of the global ocean: the ~3600 floats operating currently on 10-day cycles provide an average of 360 profiles daily, or approximately one profile per M km$^2$ of ocean surface. The relative performance of the operational RTQC strategies must be assessed via the twin goals of removing the greatest number of faulty profiles while retaining as much as possible of this important observational resource.

## Data analysis

Profiles distributed via the GTS have the ADM RTQC applied before dissemination. Profiles and levels that fail are removed. The UKMO and the FNMOC take their data from this source. Profiles from the GDAC also have ADM RTQC applied but do not have any profiles or levels removed; the QC results are included in the profile meta-data. The CRS uses this data source. The BoM employs a hybrid method; profiles are downloaded from both GDACs and the GTS. Duplicate profiles are removed by selecting the version with the greatest number of QC flags regardless of test results. The GDAC version of profiles is thus preferred, as the GTS system does not provide QC flag information.

The RTQC output flags differ from centre to centre, as do the data formats. All RTQC flags and the DMQC results are converted into a binary 'good-profile/bad-profile' flag; the individual methods of achieving this are described in Appendix 1. The comparison of the binary RTQC and DMQC outputs gives a contingency table of the form shown in Table 1. Four outcomes are possible: both agree the profile is bad (CB); both agree the profile is good (CG); RTQC flags the profile as good and DMQC as bad (FB); and RTQC flags the profile as bad and DMQC as good (FG). Two complementary metrics are constructed from these outcomes: Recall (R) and Precision (P) (Van Rijsbergen 1979). Recall is a measure of success, and is defined as:

$$R = \frac{\mathrm{CB}}{(\mathrm{CB} + \mathrm{FB})}. \tag{1}$$

Precision is a measure of accuracy, and is defined as:

$$P = \frac{\mathrm{CB}}{(\mathrm{CB} + \mathrm{FG})}. \tag{2}$$

A high R indicates that the RTQC is identifying the majority of the DMQC-flagged bad profiles, and a high P indicates that it is not removing many good profiles in the process. The results for the RTQC of each centre are shown in Tables 2–4, for temperature, salinity and pressure profiles, respectively, for 2007 to 2011. The FNMOC and the UKMO do not apply QC to pressure measurements. There is a large range in the total size and the good/bad ratios of the initial data sets. Salinity profiles are the most likely to be assessed as faulty by the DMQC and pressure profiles the least likely. The centres using the GDAC system have larger initial data sets than the centres using the GTS and also have more faulty profiles.

The monthly averages of R and P for temperature, salinity and pressure for the five years studied are shown in Figure 2. The error bars represent the standard sampling

Table 1. Contingency table defining the possible outcomes of a comparison between binary DMQC and RTQC.

| | | DMQC | |
|---|---|---|---|
| | | Good | Bad |
| RTQC | Good | CG | FB |
| | Bad | FG | CB |

Note: CB = correct bad; CG = correct good; FB = false bad; FG = false good.

Table 2. RTQC results for temperature from 2007 to 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | R | P |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 182,274 | 11,949 | 170,325 | 5232 | 1185 | 175,857 | 6717 | 169,140 | 0.44 | 0.82 |
| BoM | 365,241 | 10,775 | 354,466 | 7484 | 2603 | 355,154 | 3291 | 351,863 | 0.69 | 0.74 |
| UKMO | 247,307 | 3037 | 244,270 | 338 | 3805 | 243,164 | 2699 | 240,465 | 0.11 | 0.08 |
| FNMOC | 356,797 | 5090 | 351,707 | 1285 | 7653 | 347,859 | 3805 | 344,054 | 0.25 | 0.14 |

Note: Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad and DMQC Good = the number of DMQC-flagged bad and good profiles in the initial data sets; RTQC CB and RTQC FG = the number of DMQC-flagged bad and good profiles that the RTQC rejects; Final profiles = the total post-RTQC final data set; RTQC FB and RTQC CG = the number of DMQC-flagged bad and good profiles in the final data set; R and P = the calculated values of Recall and Precision.

Table 3. RTQC results for salinity from 2007 to 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | R | P |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 182,221 | 14,426 | 167,795 | 7441 | 1222 | 173,558 | 6985 | 166,573 | 0.52 | 0.86 |
| BoM | 364,894 | 15,593 | 349,301 | 10,298 | 2872 | 351,724 | 5295 | 346,429 | 0.66 | 0.78 |
| UKMO | 246,624 | 5284 | 241,340 | 1312 | 4668 | 240,644 | 3972 | 88,352 | 0.25 | 0.22 |
| FNMOC | 356,808 | 9598 | 347,210 | 2290 | 3796 | 350,722 | 7308 | 343,414 | 0.24 | 0.38 |

Note: Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad and DMQC Good = the number of DMQC-flagged bad and good profiles in the initial data sets; RTQC CB and RTQC FG = the number of DMQC-flagged bad and good profiles that the RTQC rejects; Final profiles = the total post-RTQC final data set; RTQC FB and RTQC CG = the number of DMQC-flagged bad and good profiles in the final data set; R and P = the calculated values of Recall and Precision.

Table 4. RTQC results for pressure from 2007 to 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | *R* | *P* |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 182,232 | 10,500 | 171,732 | 9134 | 1622 | 171,476 | 1366 | 170,110 | 0.87 | 0.85 |
| BoM | 365,252 | 8653 | 356,599 | 6529 | 2410 | 356,313 | 2124 | 354,189 | 0.75 | 0.73 |

Note: Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad and DMQC Good = the number of DMQC-flagged bad and good profiles in the initial data sets; RTQC CB and RTQC FG = the number of DMQC-flagged bad and good profiles that the RTQC rejects; Final profiles = the total post-RTQC final data set; RTQC FB and RTQC CG = the number of DMQC-flagged bad and good profiles in the final data set; R and P = the calculated values of Recall and Precision.

error. For R this is defined as:

$$\frac{1}{\sqrt{T_B}}, \tag{3}$$

where $T_B$ is the total number of monthly profiles flagged as bad by the DMQC (FB + CB). For P it is defined as:

$$\sqrt{\left(\frac{1}{\sqrt{T_B}}\right)^2 + \left(\frac{1}{\sqrt{T_G}}\right)^2}, \tag{4}$$

where $T_G$ is the total number of monthly profiles flagged as good by the DMQC (FG + CG). The standard sampling error is an estimate of the uncertainty associated with measuring a generalized statistic using a random sample of size $n$ and assumes the central limit theorem (Isserlis 1918). The metrics have been smoothed with a three-month running average to reduce the noise from anomalous months and emphasize longer-term trends in the results.

The results for temperature (Figure 2(a) and 2(b)) vary widely. The BoM has a steadily increasing R with an average of approximately 0.7, and a steady P with the same average. The RTQC is identifying the majority of bad temperature profiles, while rejecting about half as many good profiles in the process. Data is available for the entire period with no breaks.

The FNMOC temperature results fluctuate between R values of approximately 0.2 to 0.4 and P values of approximately 0.1 to 0.2. The selection process is removing around a third of the bad profiles; a large number of good profiles are also being removed.

The UKMO temperature results show a distinct difference between the ASCII data of the 2007 to early 2008 period and the netCDF data after mid-2008 (see Appendix 1). The earlier results have a lower R and higher P than the later data. This pattern could be due to a broadening of the definition of a bad profile in the RTQC, rejecting greater numbers of both good and bad profiles. The gaps in the UKMO metrics indicate a lack of available data during those periods.

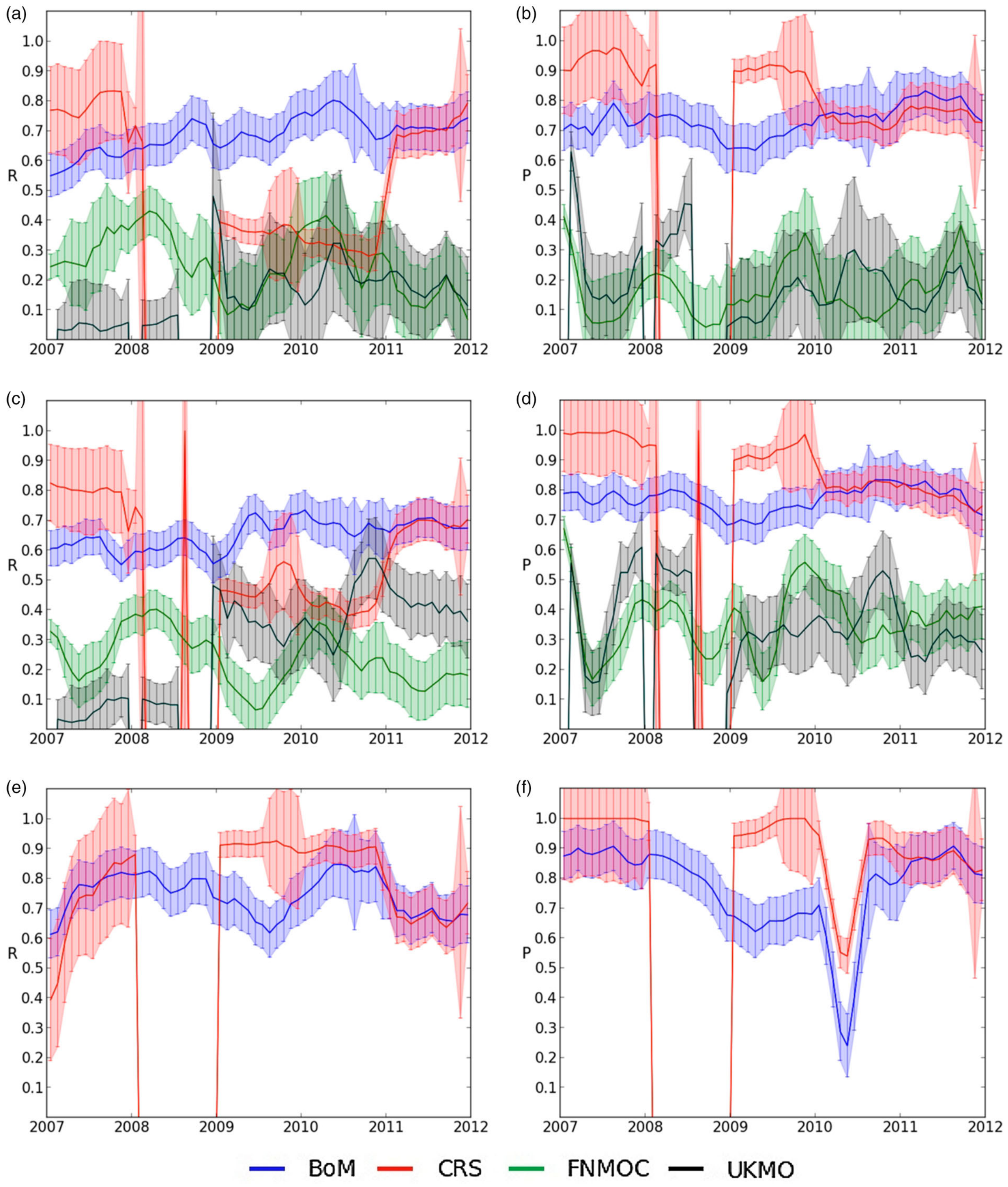


Figure 2. Recall (R) and Precision (P) for RTQC processes, shown respectively for (a) and (b) temperature, (c) and (d) salinity and (e) and (f) pressure.
Note: Error bars are the standard sampling error. A three-month running average has been applied. High scores are desirable for both R and P.

The CRS data set is the least temporally consistent of the included centres. The temperature RTQC has very good results in the first year. The data for 2007 is sparse, however, and reduces further in 2008. From March 2008 to December 2008 the metrics cannot be calculated as there are no profiles in the data that the DMQC has flagged as bad. During this time the RTQC is removing only DMQC-flagged good profiles. From 2009 P is only slightly lower than in 2007, but R drops by half. The reason for this drop is unknown. It is too large to be accounted for solely by sampling error. Data is again very sparse in the last half of 2009. In 2011 P is steady with a central value slightly below that of the BoM, while R increases rapidly at the beginning of 2011 until by March it is a similar value to that of the BoM.

The RTQC results for salinity are similar to temperature for most of the centres. The BoM identifies fewer bad profiles and discards fewer good profiles (lower R, similar P). The CRS shows similar results for salinity as for temperature in 2007/8 but removes more bad and good salinity profiles in 2009 and 2010 (higher R, similar P). The increase in R from 2010 to 2011 that was seen in the CRS temperature results is also evident in salinity. The FNMOC identifies similar numbers of bad salinity profiles as temperature profiles but removes fewer good salinity profiles (similar R, higher P). Prior to 2009 the UKMO results are slightly higher in both R and P for salinity than for temperature. From 2009, the UKMO salinity RTQC has R and P approximately twice that for temperature. The UKMO salinity R results increase significantly in mid-2010. The reason for this is unknown.

The BoM and the CRS perform better for pressure than they do for salinity or temperature. The BoM pressure RTQC results vary more than those for salinity or temperature, but average slightly higher. The R values of the CRS, though beginning low at the start of 2007, outperform those of the BoM after 2008, and the P values are extremely high. The two centres' results are very similar in 2011.

The CRS and the BoM both show a large dip in P for the pressure RTQC in the first half of 2010. This is not likely to be due to any change in the individual QC processes, as the drop is highly correlated between the centres in both phase and amplitude. Recall does not show any similar dip during this period. The monthly mean number of DMQC-identified bad profiles also remains stable for both centres. The number of profiles that are identified as bad by the RTQC increases dramatically, however, by 2 to 3 times the average of the surrounding months. The majority of these have 100% of their levels flagged as bad by the RTQC and 100% flagged as good by the DMQC. No meta-data for either centre indicate the failure of any QC pressure test; the profiles seem to have been rejected based on other information.

A comparison between the data for the months of May, the lowest point of the dip, and October, two months after P recovers, shows the difference in the number of RTQC-identified bad profiles to be made up of Autonomous Profiling Explorer (APEX) type floats that have had their pressure profiles adjusted by the DMQC. This suggests the possibility that these profiles were rejected because the floats that recorded them were on a list of those affected by the Druck pressure sensor micro-leak issue (Barker et al. 2011). These profiles could then have been cleared of any pressure drift by the DMQC. The micro-leak issue was discovered in early 2009, however, making it hard to reconcile the 2010 dip in the results.

The results shown in Figure 2 provide information regarding the temporal consistency of the RTQC processes and the performance of each centre over their individual data streams. The data sets analysed by the institutions are very different, however, as can be seen from the different data volumes in Tables 2–4. A more meaningful comparison of the performance of each RTQC process requires homogeneity in the data set.

## Intercomparison

Profiles that have undergone RTQC by each of the centres as well as the DMQC were isolated. The gaps in the CRS and UKMO early data and the small number of CRS profiles available during 2007 and the latter half of 2009 mean there are very few profiles with RTQC results from all centres before 2010. Only profiles from 2010 and 2011 were included, as the most stable, concurrent, and relevant period.

A possible bias was identified, introduced by profiles in which some levels have been removed by the Argo RTQC before dissemination on the GTS. The GTS profile in this case will have fewer levels than the GDAC or DMQC versions, which will lower the chance of GTS-based RTQC agreeing with the DMQC. To address this, all profiles in which enough of a disparity in levels exists between the DMQC and RTQC versions of the profile to possibly affect the outcome of a comparison for any centre were excluded from the intercomparison. This removes 5939 temperature profiles and 2658 salinity profiles. More detail is provided in Appendix 1.

Isolating profiles that have undergone QC by all of the centres as well as the DMQC reduces the number of available profiles significantly due to differences between the centres' data sets. The number of profiles that meet the common requirement for temperature, salinity and pressure are shown in Table 5. Also shown are the number of DMQC-flagged bad profiles in the common data, the number of these bad profiles that are identified by each centre's RTQC and the number of DMQC-flagged good profiles that each centre rejects.

The results for each RTQC process are visualized in the Roebber (2009) diagrams shown in Figure 3. This type of diagram utilizes the geometric relationship between R, P, bias and the critical success index (CSI) to display all four metrics simultaneously. R and P are defined in Equations (1) and (2). Bias is a measure of the relative frequency of selected and observed events, and is defined as:

$$\text{Bias} = \frac{\text{CB} + \text{FG}}{\text{CB} + \text{FB}}, \qquad (5)$$

Table 5. RTQC results for the common data set from 2010 and 2011.

| | | | CB | | | | FG | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Initial profiles | DMQC Bad | CRS | BoM | UKMO | FNMOC | CRS | BoM | UKMO | FNMOC |
| T | 61,641 | 568 | 99 | 107 | 95 | 100 | 182 | 94 | 368 | 845 |
| S | 62,876 | 1073 | 403 | 407 | 436 | 208 | 229 | 153 | 826 | 457 |
| P | 89,120 | 2272 | 1620 | 1655 | - | - | 1109 | 1105 | - | - |

Note: T = temperature; S = salinity; P = pressure; Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad = the number of DMQC-flagged bad profiles; CB = the number of DMQC-flagged bad profiles that each RTQC rejects; FG = the number of DMQC-flagged good profiles that each RTQC rejects.

where CB, FG and FB are defined in Table 1. Bias is indicated by radial dashed lines. CSI is a measure of accuracy when correctly identified good profiles are removed from consideration, and is defined as:

$$\mathrm{CSI} = \frac{\mathrm{CB}}{\mathrm{CB} + \mathrm{FB} + \mathrm{FG}}, \tag{6}$$

where CB, FG and FB are defined in Table 1. CSI is indicated by solid contour lines. CSI is not as relevant to this study as the other metrics due to the importance of correctly identifying good profiles. The errors in Figure 3 are the standard sampling error for R and P.

The performance of the RTQC processes show greater similarity over the common data set, especially in R. The BoM and CRS data change the most, while the UKMO and FNMOC data are close to their results in Figure 2, indicating that the GDAC-only profiles are being removed from consideration.

The R scores are not generally statistically distinguishable. All results in R for temperature profiles are within the sampling error (Figure 3(a)). For salinity the only separations above the sampling error are between the FNMOC and the other centres (Figure 3(b)). The FNMOC salinity R result is the lowest of the centres. This can also be seen in the full data set for the years 2010/11 (Figure 2 (c)). With the exception of the FNMOC, the centres are more effective over salinity profiles than temperature profiles, generally identifying twice the proportion of bad profiles. The R results for pressure over the common data set are similar to those over the total data sets (Figure 3(c)). This is expected, as both institutions use GDAC data. The BoM RTQC is narrowly more successful than that of the CRS, though the separation is not significant. All centres identify between 17% and 19% of bad temperature profiles, between 38% and 41% of bad salinity profiles (excluding the FNMOC), and between 71% and 73% of pressure profiles.

The P scores show a larger spread in performance. The UKMO and the FNMOC remove significantly more good profiles than the BoM and the CRS for both temperature and salinity, with the BoM removing the fewest in both variables (Figures 3(a) and 3(b)). All centres are more accurate for salinity profiles than temperature profiles. The CRS has lower P scores than the BoM for all three variables in the common data set, whereas in the full data the CRS shows equal or greater P scores for 2010/11 (Figures 2(b), 2(d) and 2(f)). The common data requirement is removing some profiles over which the CRS RTQC shows increased accuracy.

Similar R scores and different P scores lead to a spread in bias for both temperature and salinity. The two outliers are the BoM and the FNMOC; the BoM removes 65% fewer total temperature profiles than the DMQC (bias of 0.35), while the FNMOC removes 66% more (bias of 1.66). The lower P score of the FNMOC indicates that the extra profiles removed are flagged as good by the DMQC. For salinity, the UKMO has the highest bias (1.18), while the BoM is again the lowest (0.52).

## Discussion

The RTQC criteria employed by each of the examined centres are shown in Table 6, along with the sources of information. Details of the FNMOC RTQC test criteria were sparse in the available documentation. There is a common general strategy across the centres. Physical value tests are applied to the date and time of profile recording, and the position of the profile is required to be physical and within a defined ocean space. All institutes except the BoM set a maximum allowed drift speed. The BoM and the FNMOC require temperature and salinity values to be within a single globally-applied physical value test, while the CRS applies three geographically dependant physical value tests to profiles before 2011 and adds tests for the north-western shelves and Arctic sea regions for profiles after 2011. The UKMO does not apply a global range test but does apply a tropical waters test, rejecting any level above 1000 m that measures below 1°C.

All of the centres apply monotonicity/inversion tests to density. Salinity and temperature for levels which fail are flagged as bad. The UKMO also applies a density spike test. The FNMOC and the BoM test depth data, the FNMOC rejecting levels with duplicate depth and enforcing monotonicity, the BoM applying a global bathymetry test and rejecting levels with greater depth than that interpolated from a 2° gridded bathymetry product with enhanced resolution around Australia.

The main differences between the temperature RTQC processes lie in the sophistication of the gradient, spike and density tests and the different background and depth tests. The CRS gradient test compares the value at a level with the average value of the two adjacent levels; the maximum permitted deviation has two values, split at 500 dBar. The BoM uses a similar test, but weights the average value of the two adjacent levels according to their relative distance from the level being tested. The UKMO does not employ a gradient test; however, a more stringent spike test than that of the BoM and the CRS is implemented. The FNMOC applies a simple maximum gradient requirement, and compares with climatological gradients. The FNMOC also includes temperature spike and inversion tests; no details of these were available. All centres require increasing density between levels; the FNMOC and the UKMO have minimum density change requirements while the BoM and the CRS just require an increase. The UKMO additionally applies a density spike test.

The BoM, the UKMO and the CRS compare the profile to a background state; hybrid CSIRO Atlas of Regional Seas/World Ocean Atlas 2005 climatology (Ridgway et al. 2002; Antonov et al. 2006; Locarnini et al. 2006) for the BoM, a Bayes theorem-based (Lorenc & Hammon 1988) check against a one-day forecast for the UKMO, and an objective analysis using climatology derived from World Ocean Altas 1998 (National Oceanic and Atmospheric Administration 1998) as the background for the CRS. In 2010 the CRS updated its objective analysis process; this combined with the added physical value tests could explain the upswing in R seen in Figures 2(a) and 2(c). The FNMOC does not apply a background check.

The gradient tests seem to be reducing the number of incorrectly rejected profiles (FG) for the BoM and the CRS over the UKMO. The FNMOC gradient test is simple and might not provide the same level of discrimination as the CRS and BoM tests. The weighting of the gradient test could be a source of added accuracy for the BoM over the CRS; the difference could also be attributed to the different climatological background checks, the added depth check or the lack of speed and sensor drift tests in the BoM RTQC.

The test criteria for salinity are very similar to those for temperature. The FNMOC alone of the centres studied has lower R over salinity profiles than over temperature profiles. The poor result is difficult to interpret due to the sparse details of the FNMOC RTQC process in the available documentation. The difference in salinity criteria between the other three centres again lies in the gradient, climatology and density inversion tests. The UKMO shows the largest increase in R between salinity and

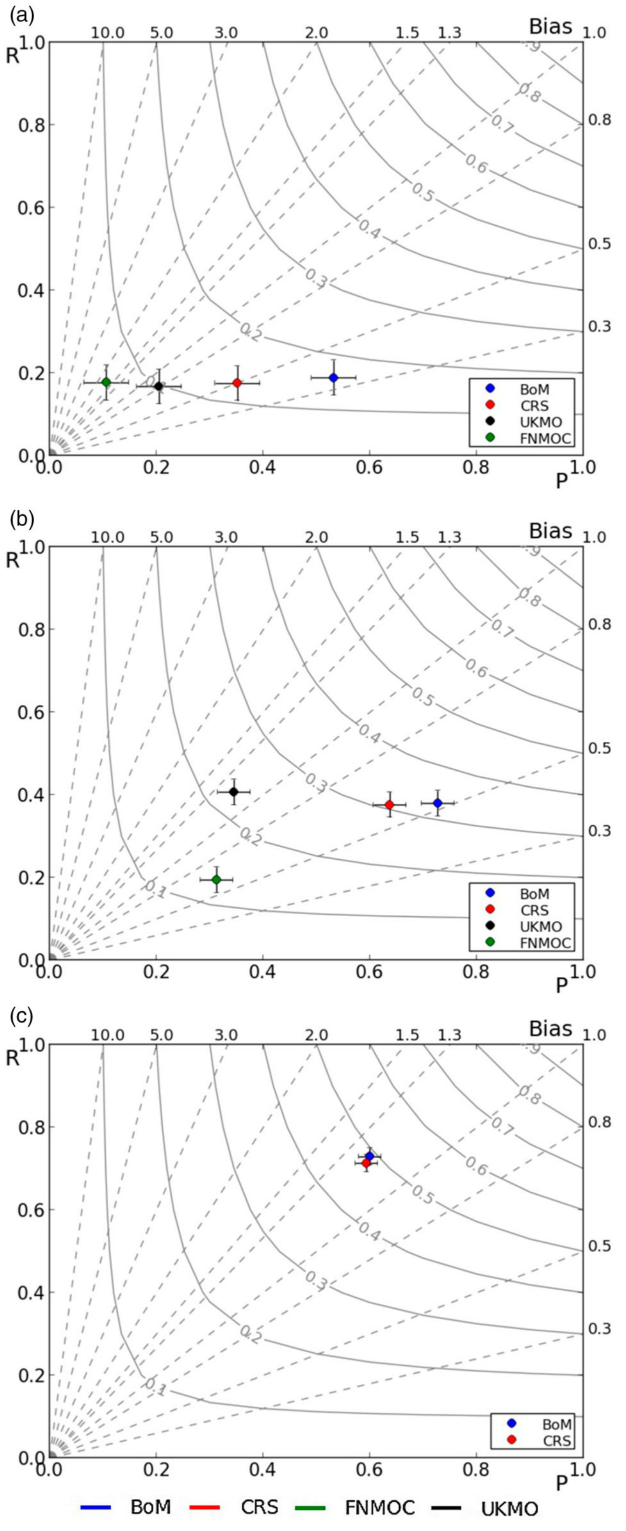


Figure 3. Roebber (2009) diagrams of RTQC results in 2010 and 2011 for (a) temperature, (b) salinity and (c) pressure.
Note: Error bars are the standard sampling error. Bias is shown in radial dashed lines. CSI is shown in solid contours. Ideal performance lies in the upper right of the diagram, where all metrics approach 1.0.

Table 6. RTQC test criteria for the CRS, the FNMOC, the BoM and the UKMO.

| | CRS | FNMOC | BoM | UKMO |
|---|---|---|---|---|
| Pressure | Physical value test:<br>Level fails if $P < -5$ dBar.<br>Monotonicity test:<br>$(k+1)$ fails if $P(k+1) \leq P(k)$.<br>Deepest pressure test:<br>Level fails if pressure is greater than 10% higher than the deepest pressure. | None. | Physical value test:<br>Level fails if it does not satisfy $0 \leq P \leq 6500$ dBar.<br>Monotonicity test:<br>$(k+1)$ fails if $P(k+1) \leq P(k)$. | None. |
| Temperature | Range test:<br>Level fails if it does not satisfy<br>$-2.5^\circ\mathrm{C} < T < 40^\circ\mathrm{C}$;<br>$21.7^\circ\mathrm{C} < T < 40^\circ\mathrm{C}$ for Red Sea;<br>$10^\circ\mathrm{C} < T < 40^\circ\mathrm{C}$ for Mediterranean;<br>For post-2010 profiles:<br>$-2^\circ\mathrm{C} < T < 24^\circ\mathrm{C}$ for North Western Shelves;<br>$-2^\circ\mathrm{C} < \mathrm{T} < 30^\circ\mathrm{C}$ for South Western Shelves;<br>$-1.92^\circ\mathrm{C} < T < 25^\circ\mathrm{C}$ for Arctic Sea.<br>Gradient test:<br>$\mathrm{X} = \lvert \mathrm{T}(k) - (\mathrm{T}(k-1) + \mathrm{T}(k+1)) / 2 \rvert$;<br>$k$ fails if $\mathrm{X} > 9^\circ\mathrm{C}$ and $P < 500$ dBar;<br>$k$ fails if $X > 3^\circ\mathrm{C}$ and $P \geq 500$ dBar.<br>Spike test:<br>$X = \lvert \mathrm{T}(k) - (\mathrm{T}(k-1) + \mathrm{T}(k+1)) / 2 \rvert - \lvert (\mathrm{T}(k-1) - \mathrm{T}(k+1)) / 2 \rvert$;<br>$k$ fails if $X > 6^\circ\mathrm{C}$ and $P < 500$ dBar;<br>$k$ fails if $X > 2^\circ\mathrm{C}$ and $\mathrm{P} \geq 500$ dBar.<br>Digit rollover test:<br>$k$ fails if $\lvert \mathrm{T}(k) - \mathrm{T}(k-1) \rvert$ or $\lvert \mathrm{T}(k) - \mathrm{T}(k+1) \rvert > 10^\circ\mathrm{C}$.<br>Stuck value test:<br>Profile fails if all profile values are identical. | Global range test:<br>Level fails if it does not satisfy<br>$-2.5^\circ\mathrm{C} \leq \mathrm{T} \leq 42^\circ\mathrm{C}$.<br>Spike test (no details).<br>Inversion test (no details).<br>Gradient test:<br>Level fails if gradient > 0.2°C/m and 4σ from climatological gradient.<br>Land-sea boundary test (no details). | Global range test:<br>Level fails if it does not satisfy<br>$-2^\circ \leq \mathrm{T} \leq 40^\circ\mathrm{C}$.<br>Gradient test:<br>$k$ fails if $\mathrm{GRAD}(k) > 10^\circ\mathrm{C}$, for $\mathrm{h} \leq 500$ m;<br>$k$ fails if $\mathrm{GRAD}(k) > 5^\circ\mathrm{C}$, for $\mathrm{h} > 500$ m, where $\mathrm{GRAD}(k) = \mathrm{T}(k) - (\mathrm{alpha1} \times \mathrm{T}(k-1) - \mathrm{alpha2} \times \mathrm{T}(k+1))$, $\mathrm{alpha1} = (h(k+1) - h(k)) / (h(k+1) - h(k-1))$, $\mathrm{alpha2} = (h(k) - h(k-1)) / (h(k+1) - h(k-1))$ and h (i) are depth values.<br>Spike tests:<br>As UKMO except:<br>$\mathrm{Ttol} = 5^\circ\mathrm{C}$ if depth $\leq 500$ m;<br>$\mathrm{Ttol} = 2.5^\circ\mathrm{C}$ if depth $> 500$ m. | Constant value test:<br>If over 90% of T levels that cover at least 100 m read identical, T profile fails.<br>Tropical waters test:<br>If depth < 1000 m and $\mathrm{T} \leq 1^\circ\mathrm{C}$ then reject level.<br>Spike tests:<br>1) If either $\lvert \mathrm{DT}(k-1) \rvert > \mathrm{Ttol}$ or $\lvert \mathrm{DT}(k) \rvert > \mathrm{Ttol}$ and $\lvert \mathrm{DT}(k-1) + \mathrm{DT}(k) \rvert < 0.5$ Ttol then $\mathrm{T}(k-1)$ is rejected as a spike.<br>2) If $\lvert \mathrm{DT}(k-1) \rvert > 0.5$ Ttol or $\lvert \mathrm{DT}(k) \rvert > 0.5$Ttol and either $\lvert \mathrm{DT}(k) \rvert$ or $\lvert \mathrm{DT}(k-1) \rvert > 0.05^\circ\mathrm{C/m}$ and $\lvert \mathrm{DT}(k-1) + \mathrm{DT}(k) \rvert < 0.25 \times \lvert \mathrm{DT}(k-1) - \mathrm{DT}(k) \rvert$ then $\mathrm{T}(k-1)$ is rejected as a spike.<br>3) If $\lvert \mathrm{DT}(k-1) \rvert > \mathrm{Ttol}$ and $\lvert \mathrm{T}(k-1) \rvert > 0.5$ and Ttol(interpolated $\mathrm{T}(k-1)$ and $\mathrm{T}(k)$) and $0 < \mathrm{DT}(k-1)$ or $\mathrm{DT}(k-1) < -3 \times \mathrm{Ttol}$ ($d <$ 250 m) then $\mathrm{T}(k-2)$ and $\mathrm{T}(k-1)$ are flagged as suspect.<br>For all tests above, $\mathrm{DT}(k) = \mathrm{T}(k) - \mathrm{T}(k-1)$ and $\mathrm{Ttol} = 5^\circ\mathrm{C}$ ( < 300 m), 2.5°C ( < 500 m), 2.0°C ( < 600 m), 1.5°C ( > 600 m); if within 20 deg of the equator then 200 m = 300 m and 300 m = 400 m; Ttol is linearly interpolated from 0 m to 300 m (400 m), step function after that. |

| | | | | |
|---|---|---|---|---|
| Salinity | Global range test:<br>Level fails if it does not satisfy<br>$2 < S < 41$ psu;<br>$2 < S < 41$ psu for Red Sea;<br>$2 < S < 40$ psu for Mediterranean;<br>For post-2010 profiles:<br>$0 < S < 37$ psu for North Western Shelves;<br>$0 < S < 38$ psu for South Western Shelves;<br>$2 < S < 40$ psu for Arctic Sea.<br>Gradient test:<br>$X = \lvert S(k) - (S(k-1) + S(k+1)) / 2 \rvert$;<br>$k$ fails if $X > 1.5$ psu and $P < 500$ dBar;<br>$k$ fails if $X > 0.5$ psu and $P \geq 500$ dBar.<br>Spike test:<br>$X = \lvert S(k) - (S(k-1) + S(k+1)) / 2 \rvert - \lvert (S(k-1) + S(k+1)) / 2 \rvert$;<br>$k$ fails if $X > 0.9$ psu and $P < 500$ dBar;<br>$k$ fails if $X > 0.3$ psu and $P \geq 500$ dBar.<br>Digit rollover test:<br>$k$ fails if $\lvert S(k) - S(k-1) \rvert$ or $\lvert S(k) - S(k+1) \rvert > 5$ psu.<br>Stuck value test:<br>Profile fails if all profile values are identical. | Physical value test:<br>Level fails if it does not satisfy<br>$0 \leq S \leq 42$ psu. | Global range test:<br>Level fails if it does not satisfy<br>$0 \leq S \leq 39$ psu.<br>Gradient test:<br>Similar to T but with limits of<br>GRAD($k$) < 1 psu, $h \leq 500$ m;<br>GRAD($k$) < 1 psu, $h > 500$ m;<br>Spike test:<br>As UKMO except:<br>Stol = 1 psu if depth ≤ 500 m<br>Stol = 0.2 psu if depth > 500 m | Temperature profile test:<br>If > 50% of the T profile is bad, the S profile is rejected.<br>Constant value test:<br>If 70% or more of S levels over at least 50 m are identical, S profile fails.<br>Spike test:<br>Similar to T but only Tests 1 and 3, and without the $0 < DS(k-1)$ or $DS(k-1) < -3 \times Ttol$ ($d < 250$ m) condition in 3;<br>Stol = 1 psu ( < 300 m), 0.2 psu( > 300 m);<br>if within 20 deg of the equator then 300 m = 400 m; Stol is linearly interpolated from 0 m to 300 m (400 m), step function after that;<br>if a T spike is detected then the corresponding S value is automatically rejected;<br>if > 4 T spikes then both T and S profiles are rejected. |
| Density | Inversion test:<br>Calc density (D) from T and S;<br>T and S level $k$ fails if $D(k) > D(k+1)$ or $D(k) < D(k-1)$. | Inversion test:<br>Level $k$ fails if $D(k) - D(k-1) < -0.025$ kg/m$^3$. | Monotonicity test:<br>If T and S tests are passed, level fails if $D(k) \leq D(k+1)$. | Monotonicity test:<br>$D\rho(k) = \rho(\Theta(k), S(k), P(k)) - \rho(\Theta(k-1), S(k-1), P(k))$;<br>if $D\rho(k) > -0.03$ kg/m$^3$ then T and S fail.<br>Density spike test:<br>If $\lvert D\rho(k-1) + D\rho(k) \rvert > 0.25 \times \lvert D\rho(k-1) - D\rho(k) \rvert$<br>then fail T and S at $k - 1$;<br>if both tests fail then T and S at $k$ and $k - 1$ fail;<br>if a profile has two or more inversions then it is discarded. |

(*Continued*)

Table 6. Continued.

| | CRS | FNMOC | BoM | UKMO |
|---|---|---|---|---|
| Bathymetry/ Depth | None. | Duplicate depth test (no details).<br>Monotonicity test (no details). | Level fails if deeper than interpolated 2 min global bathymetry formed from 2″ ETOPO2v2 (National Oceanic and Atmospheric Administration 2006) and 1 km Geosciences Australia (Webster & Petkovic 2005) products. | None. |
| Date | Profile fails if it does not satisfy:<br>Year > 1997;<br>1 ≤ Month ≤ 12;<br>Day exists in Month;<br>0 ≤ Hour ≤ 23;<br>0 ≤ Minute ≤ 59. | Profile fails if it does not satisfy:<br>1 ≤ Month ≤ 12;<br>Day exists in Month;<br>0 ≤ Hour ≤ 23;<br>0 ≤ Minute ≤ 59;<br>0 ≤ Second ≤ 59;<br>observation time must be older than the receipt time at the centre. | Identical to CRS. | Identical to CRS. |
| Position | Profile fails if it does not satisfy:<br>−90 ≤ Latitude ≤ 90;<br>−180 ≤ Longitude ≤ 180;<br>must be in ocean (ETOPO5) | Profile fails if it does not satisfy:<br>−90 ≤ Latitude ≤ 90;<br>−180 ≤ Longitude ≤ 180;<br>must be in ocean. | Identical to CRS. | Identical to CRS. |
| Speed | If drift speed > 3 m/s then flag time, position and/or float number as wrong. | Speed < 2 m/s | None. | Speed(K) = (Dist(K) – 0.5DistRes) / MAX (DTime,TimeRes), where DistRes = 20 km and TimeRes = 600 s;<br>if speed > 2 m/s or > 1.6 m/s and there is a kink in the track then a series of checks are run to determine which position correct;<br>if a buoy has > 50% of its profile positions rejected then the buoy is removed. |
| Background | Objective analysis check using climatology derived from WOA98 (National Oceanic and Atmospheric Administration 1998) as background. | None. | T and S mean within 5σ of CARS (Ridgway et al. 2002) climatology within 71 deg S–26 deg N for all longitudes and WOA (Antonov et al. 2006; Locarnini et al. 2006) in all other regions. | Bayesian background probability check; a one-day ocean forecast is used as background. |

Source: CRS (Gaillard et al. 2009; Wong et al. 2009; Pouliquen et al. 2011); FNMOC (Cummings 2006, 2010); BoM (BLUElink team); UKMO (Ingleby & Huddleston 2007).

temperature and identifies the greatest number of faulty salinity profiles, though the result is within sampling error of the BoM and the CRS. The method of the BoM again has the highest P score, although the gap between the BoM and the CRS is narrowed.

Pressure is tested less stringently than salinity or temperature by both the BoM and the CRS; the BoM applies only physical value and monotonicity/inversion tests, while the CRS also applies a deepest pressure test. The small performance difference between the BoM and CRS pressure QC can be attributed to the different ranges used in the physical value test, a negative impact from the CRS deepest pressure test, the different background checks, and/or the added bathymetry test of the BoM RTQC. The difference in performance is not statistically significant in any metric.

The performance of the RTQC methods over identical profiles is not the only consideration in determining the preferred treatment of Argo data: the choice of data source is also relevant. Tables 7–9 show the data for 2010/11 without the common QC assessment requirement but with the removal of the level-based bias for temperature, salinity and pressure, respectively. The initial data sets of the centres vary widely. Despite taking profiles from the same source, the FNMOC data set is approximately 20% larger than the UKMO. The difference is believed to be made up of US Navy Argo profiles that are not shared in real time due to security concerns. There is also a difference in size of the initial data sets of those institutions sourcing data from the GDACs. Though the BoM's hybrid data collection method is designed to collect all profiles available from all sources, the CRS has ~1000 more profiles. This is believed to be the contribution of the French Navy Argo floats, which are also not available in real time. The removal of these profiles from the common data set could explain the comparative drop in P results for the CRS between the full and common data sets.

Tables 7 and 8 show that although the BoM RTQC method gives the best result over the common data, the BoM does not have the fewest bad profiles in its final data sets. This is due to the different data streams. The GTS provides the UKMO with smaller and cleaner initial data sets than the BoM GDAC/GTS hybrid method. The UKMO final data sets also have fewer good profiles and bad profiles than those of the BoM. For temperature, the BoM's method results in 216 more bad profiles and 17,313 more good profiles. For salinity, the BoM has 423 more bad profiles and 12,878 more good profiles.

Table 7. RTQC results for temperature from 2010 and 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | R | P |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 100331 | 6065 | 94266 | 2479 | 827 | 97025 | 3586 | 93439 | 0.41 | 0.75 |
| BoM | 97677 | 3210 | 94467 | 2351 | 665 | 94661 | 859 | 93802 | 0.73 | 0.78 |
| UKMO | 81968 | 789 | 81179 | 146 | 516 | 81306 | 643 | 80663 | 0.19 | 0.22 |
| FNMOC | 100345 | 991 | 99354 | 197 | 1529 | 98619 | 794 | 97825 | 0.20 | 0.11 |

Table 8. RTQC results for salinity from 2010 and 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | R | P |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 101,904 | 7231 | 94,673 | 3527 | 884 | 97,493 | 3704 | 93,789 | 0.49 | 0.80 |
| BoM | 99,360 | 4417 | 94,943 | 3093 | 795 | 95,472 | 1324 | 94,148 | 0.70 | 0.80 |
| UKMO | 84,201 | 1567 | 82,634 | 666 | 1364 | 82,171 | 901 | 81,270 | 0.43 | 0.33 |
| FNMOC | 102,443 | 2201 | 100,242 | 419 | 779 | 101,245 | 1782 | 99,463 | 0.19 | 0.35 |

Note: Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad and DMQC Good = the number of DMQC-flagged bad and good profiles in the initial data sets; RTQC CB and RTQC FG = the number of DMQC-flagged bad and good profiles that the RTQC rejects; Final profiles = the total post-RTQC final data set; RTQC FB and RTQC CG = the number of DMQC-flagged bad and good profiles in the final data set; R and P = the calculated values of Recall and Precision.

Table 9. RTQC results for pressure from 2010 and 2011 for the full independent data from each institution.

| Centre | Initial profiles | DMQC Bad | DMQC Good | RTQC CB | RTQC FG | Final profiles | RTQC FB | RTQC CG | R | P |
|---|---|---|---|---|---|---|---|---|---|---|
| CRS | 103,133 | 5492 | 97,641 | 4618 | 1392 | 97,123 | 874 | 96,249 | 0.84 | 0.77 |
| BoM | 100,706 | 2630 | 98,076 | 1938 | 1235 | 97,533 | 692 | 96,841 | 0.74 | 0.61 |

Note: Initial profiles = the total initial data sets of profiles that have results for both RTQC and DMQC; DMQC Bad and DMQC Good = the number of DMQC-flagged bad and good profiles in the initial data sets; RTQC CB and RTQC FG = the number of DMQC-flagged bad and good profiles that the RTQC rejects; Final profiles = the total post-RTQC final data set; RTQC FB and RTQC CG = the number of DMQC-flagged bad and good profiles in the final data set; R and P = the calculated values of Recall and Precision.

The optimal balance between removing the bad profiles and retaining the good ones is system dependant. A system with an accurate ocean model will have less need for corrective data and should seek to remove as many error-inducing profiles as possible. A system with a long assimilation period will have a lower likelihood of erroneous data being isolated in space and time, and could thus include more bad profiles and rely on the mitigating effects of concurrent good data. The choice of a smaller, more accurate data set via the GTS or a larger, more contaminated one via the GDACs must be made by the individual user.

## Summary and conclusions

The performance of the real-time Argo profile QC methods used at the UKMO, the BoM, the FNMOC and the CRS were assessed and compared. The RTQC output of the centres was obtained for the years 2007 to 2011 inclusive. The data were brought into a common binary good- or bad-profile format and compared to the DMQC results. The data sets of some of the institutions were found to be temporally irregular due to changes in the RTQC criteria, recording methods and/or gaps in the uploaded data. An intercomparison was performed using those profiles recorded in 2010 and 2011 which had undergone RTQC by all institutions and the DMQC.

The RTQC techniques were found to identify similar numbers of faulty profiles; slightly more temperature and pressure profiles in the case of the BoM and slightly more salinity profiles in the case of the UKMO. The differences between the systems were not generally statistically significant given the data available. That the FNMOC identifies fewer faulty salinity profiles than any other centre can be stated confidently, but no other differences are significant. The number of good profiles rejected in the RTQC process was more system dependent. The BoM RTQC was found to remove significantly fewer good temperature and salinity profiles than the other RTQC techniques. The CRS and the BoM remove a similar number of good and bad pressure profiles.

The GTS distribution stream removes profiles that fail the Argo RTQC before dissemination, while the GDACs supply all profiles reported by Argo buoys along with the flags from the Argo RTQC. This results in very different pre-RTQC data sets for each centre, and the best performing RTQC system does not necessarily provide the cleanest final set of profiles. The FNMOC and the CRS have access to military-operated floats that are not generally disseminated in real time and their data sources cannot be compared to other centres. The pre- and post-RTQC data of the UKMO and the BoM were compared for 2010/11. The UKMO RTQC using GTS data has fewer faulty temperature and salinity profiles than the BoM system using their hybrid GTS/GDAC data source. It also results in much fewer good profiles in the final data set. Whether the removal of the extra bad profiles is worth the removal of the good profiles is dependent on the model and assimilation systems being used, and the choice must be left to the individual user.

The accuracy of operational ocean forecasting is directly related to the quality of the Argo data stream. Improving the Argo RTQC processes of operational centres will also provide material benefit to seasonal and decadal forecasts via better initialization. The current RTQC techniques remove many faulty Argo profiles, but there is scope for improvement. One possibility is to investigate the automation of DMQC tests, for example: the parametrization of expected profile shapes; drift analysis using previous profiles from the same float; or algorithmic temperature, salinity and/or pressure profile consistency checks. Statistical methods such as these can perform poorly in the presence of ocean features such as eddies, fronts and water mass boundaries, however, and care would have to be taken to avoid discarding profiles with extreme but valid attributes. Another possibility is the creation of a super-RTQC assessment based on combining the results of the individual centres' RTQCs using classical statistical methods or machine-learning techniques. The differences in the RTQC techniques could perhaps be leveraged for greater discriminatory power. This would need to be a centralized process, and the classification of profiles in real time would rely on the prompt and consistent uploading of RTQC results from operational centres.

The Argo programme continues to provide operational centres and researchers with high-quality sub-surface ocean observations. Operational RTQC systems should be subject to continual analysis and improvement as the Argo database grows. Updating the results shown here as more DMQC data becomes available will lower the statistical errors and highlight the differences between the RTQC systems, helping to inform the improvement of current RTQC systems and the design of new ones.

## Acknowledgements

The authors would like to thank Alastair Gemmel of the University of Reading and James Cummings of the US Naval Research Laboratory for their previous work on the subject of Argo quality control. The feedback and suggestions from Gary Brassington of the BoM, and Simon Good and Matthew Martin of the UKMO were instrumental in achieving the results described here.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Antonov JI, Locarnini RA, Boyer TP, Mishonov AV, Garcia HE. 2006. World ocean atlas 2005, Volume 2: salinity. In: Levitus

S, editor. NOAA Atlas NESDIS 62. Washington, DC: U.S. Government Printing Office: 182–182.

Barker PM, Dunn JR, Domingues CM, Wijffels SE. 2011. Pressure sensor drifts in Argo and their impacts. J. Atmos Oceanic Technol. 28:1036–1049.

BLUElink team. BLUElink Real-Time Quality Control File Format, Australian Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organisation, unpublished.

Cummings JA. 2006. Operational multivariate ocean data assimilation. Q J R Meteorol Soc. 131:3583–3604.

Cummings JA. 2010. Ocean data quality control. Oper Oceanogr 12st Century. Springer:91–121.

Current status of the Argo fleet [Internet]. [cited 2014 July 8]. Available from www.argo.uscd.edu

Cummings J, Brassington G, Keeley R, Martin M, Carval T. 2010. GODAE ocean data quality control intercomparison project. Proceedings of OceanObs 09: Sustained Ocean Observations and Information for Society.

Gaillard F, Autret M, Thierry V, Galaup P, Coatanoan C, Loubrieu T. 2009. Quality control of large Argo datasets. J Atmos Oceanic Technol. 26:337–351.

Goni G, Roemmich D, et al. 2010. The ship of opportunity program. Proceedings of OceanObs 9.

Hayes SP, Mangum LJ, PiCaut J, Sumi A, Takeuchi K. 1991. TOGA-TAO: a moored array for real-time measurements in the tropical pacific ocean. B Am Meteorol Soc. 72:339–347.

IFREMER. 2010. Home page. http://wwz.ifremer.fr

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

Isserlis L. 1918. On the value of a mean as calculated from a sample. J R Stat Soc. 81:75–81.

Le Traon PY, Bell M, Dombrowsky E, Schiller A, Wilmer-Becker K, et al. 2010. GODAE OceanView: from an experiment towards a long-term ocean analysis and forecasting international program. Proceedings of OceanObs 09. GODAE OceanView collaboration homepage. Available from www.godae-oceanview.org

Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE. 2006. World ocean atlas 2005, Volume 1: temperature. In: Levitus S, editor. NOAA Atlas NESDIS 61. Washington, DC: U.S. Government Printing Office: 182.

Lorenc AC, Hammon O. 1988. Objective quality control of observations using Bayesian methods. Theory, and a practical implementation. Q J R Meteorol Soc. 114:515–543.

National Oceanic and Atmospheric Administration. 1998. World ocean atlas 1998.

National Oceanic and Atmospheric Administration. 2006. 2-minute Gridded Global Relief Data (ETOPO2v2).

Pouliquen S, et al. 2011. Recommendations for in-situ data Near Real Time Quality Control. EuroGOOS.

Ridgway KR, Dunn JR, Wilkin JL. 2002. Ocean interpolation by four-dimensional least squares -Application to the waters around Australia. J Atmos Ocean Tech. 19:1357–1375.

Roebber PJ. 2009. Visualizing multiple measures of forecast quality. Weather Forecast. 24:601–608.

Roemmich D, Boebel O, Freeland H, King B, LeTraon P, Molinari R, Brechner Owens W, Riser S, Send U, Takeuchi K, Wijffels S. 1998. On the design and implementation of Argo: an initial plan for a global array of profiling floats, International CLIVAR Project Office Report 21.

Roemmich D, Johnson GC, Riser S, Davis R, Gilson J, Owens WB, Garzoli SL, Schmid C, Ignaszewski M. 2009. The Argo Program: observing the global ocean with profiling floats. Oceanography. 22:34–43.

SHOM. 2010. Home page. http://www.shom.fr

Van Rijsbergen CJ. 1979. Information retrieval. 2nd ed. London, UK: Butterworths.

Webster MA, Petkovic P. 2005. Australian bathymetry and topography grid, June 2005. Geoscience Australia Record. 12:30–30.

Wong A, Keeley R, Carval T. 2009. Argo quality control manual. Version 2.32.

Wong A, Keeley R, Carval T, et al. 2013. Argo quality control manual. Version 2.9.

## Appendix 1

The RTQC output flags differ from centre to centre. The Argo DMQC flags follow the form of Table 2a in the *Argo Data Management User Manual* (Roemmich et al. 2009; see Table A1 below), in which the flags run from A to F, based on the percentage of levels in the profile that are assessed as ‘good’ data. The BoM and the CRS also follow Table 2a. The UKMO RTQC flags follow the form of Table 2 in the same publication (see Table A2 below), and the FNMOC uses its own probability-based method. Two methods of bringing the flags into a common form for comparison were examined. The first converts all QC systems into a binary ‘good-profile/bad-profile’ flag; the individual methods of achieving this are described below. This is the form used in the majority of the analysis. The second method involves converting all RTQC flags into the form of Table 2a by calculating the proportion of levels in each profile that are rated ‘good’ according to the centre’s RTQC. No appreciable difference was found between the methods. The first method is preferred due to the possibility of some profile-based RTQC tests being excluded from the level-based RTQC results.

### *Argo Delayed-mode*

The Argo delayed-mode data was obtained from the US GODAE public server (http://www.usgodae.org/pub/outgoing/argo/geo/). The profile-based QC flags follow the form of Table A1. All profiles with greater than or equal to 50% of levels rated as ‘good’ (flag A, B or C) are considered a ‘good’ profile for the purposes of this intercomparison, and all profiles with less than 50% of levels rated as ‘good’ are considered ‘bad’. Profiles that have been adjusted by the delayed mode QC operators are excluded.

### *The UKMO*

The UKMO’s data is in the form of both ASCII and netCDF files that are uploaded to the GODAE severs daily (http://www.usgodae.org/pub/incoming/godae_qc/). The ASCII data runs from 10 April 2006 to 22 July 2008, and the netCDF data runs from 17 December 2008 to the present. Data for the gap of roughly five months were unavailable. The profiles are flagged according to Table A2. All profiles and levels rated as ‘Good’ or ‘Probably Good’ (flag 1 or 2) are considered to be ‘good’. All other flags are considered ‘bad’.

### *The BoM*

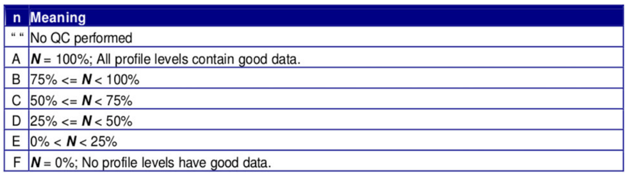
The BoM has provided daily netCDF files spanning the period 1 January 2005 to 19 June 2011 to the GODAE servers. Data for the rest of 2011 was obtained from the BoM directly. The level-based RTQC is flagged according to Table A2, and the profile-based RTQC according to Table A1. As with the DMQC data, profiles with greater than 50% of levels rated as 'good' (flag A, B or C) are considered 'good'.

### *The CRS*

The CRS uses the same format as the BoM and is treated in the same fashion. Data is available from 23 July 2009 to the present on the GODAE servers, but the data for 2009 is very sparse. Further data was obtained from the MyOcean ftp server (http://www.mycean.eu/) for the years 2007 and 2008. While files were available from the beginning of 2007 to the end of 2009, the data is very sparse at all times, except the first half of 2009.

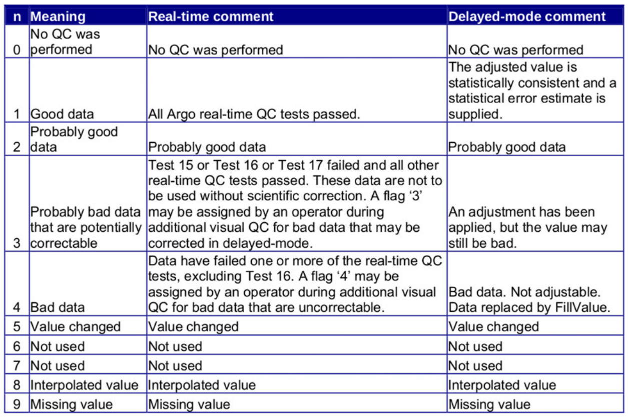
### *The FNMOC*

The raw data files from the FNMOC were unavailable; data that were processed into float-specific files were used instead. These are available on the GODAE public server (http://www.usgodae.org/pub/outgoing/godae_qc/) from 2004 to 2011. The FNMOC assigns profiles and levels separate probability values of being 'bad' from 0 to 100. Profiles with probabilities of 95 or less are considered to be 'good'. Levels with probabilities of less than 100 are considered to be 'good'.



| n | Meaning |
|---|---|
| " " | No QC performed |
| A | ***N*** = 100%; All profile levels contain good data. |
| B | 75% <= ***N*** < 100% |
| C | 50% <= ***N*** < 75% |
| D | 25% <= ***N*** < 50% |
| E | 0% < ***N*** < 25% |
| F | ***N*** = 0%; No profile levels have good data. |

Table A1. *Argo Data Management User Manual*, Table 2a.



| n | Meaning | Real-time comment | Delayed-mode comment |
|---|---|---|---|
| 0 | No QC was performed | No QC was performed | No QC was performed |
| 1 | Good data | All Argo real-time QC tests passed. | The adjusted value is statistically consistent and a statistical error estimate is supplied. |
| 2 | Probably good data | Probably good data | Probably good data |
| 3 | Probably bad data that are potentially correctable | Test 15 or Test 16 or Test 17 failed and all other real-time QC tests passed. These data are not to be used without scientific correction. A flag '3' may be assigned by an operator during additional visual QC for bad data that may be corrected in delayed-mode. | An adjustment has been applied, but the value may still be bad. |
| 4 | Bad data | Data have failed one or more of the real-time QC tests, excluding Test 16. A flag '4' may be assigned by an operator during additional visual QC for bad data that are uncorrectable. | Bad data. Not adjustable. Data replaced by FillValue. |
| 5 | Value changed | Value changed | Value changed |
| 6 | Not used | Not used | Not used |
| 7 | Not used | Not used | Not used |
| 8 | Interpolated value | Interpolated value | Interpolated value |
| 9 | Missing value | Missing value | Missing value |

Table A2. *Argo Data Management User Manual*, Table 2.

To address the bias introduced by the removal of levels from the GTS profiles, all RTQC are first translated into the form of Table A1 as described above. Profiles in which enough of a disparity in levels exists between the DMQC and RTQC versions of the profile to possibly affect the outcome of a comparison are then excluded. For example, if a profile is flagged as either 'A' or 'F' on Table A1 it will require a disparity in levels of at least 50% to alter the outcome. This is because the intercomparison takes profiles with greater than 50% good levels to be good and fewer than 50% as bad; thus, a profile with 100% good levels will require at least the same number of bad levels to change its profile rating. Conversely, a profile flagged as 'C' or 'D' requires a disparity of only a single level to call the result into question. The test is performed for each centre and profiles that fail for any centre are excluded from consideration. This process removes 5939 temperature profiles and 2658 salinity profiles from the 2010/11 intercomparison.