Quality Control of Data

TVS Udaya Bhaskar
Head, Ocean Data Management (ODM) Division,
INCOIS, Ministry of Earth Sciences (MoES)

About Quality

- Manage Quality: Data-quality management is a process where protocols and methods are employed to ensure that data are properly collected, handled, processed, used, and maintained at all stages of the scientific data lifecycle.
- Quality Control: Quality control (QC) of data refers to the application of methods or processes that determine whether data meet overall quality goals and defined quality criteria for individual values.

Why is "Quality Control" important?

- It is critically important that good quality data flow from the observing sites to data users.
- Good quality data are also important for verification and research purposes.
- All forecasts and warnings start with data, and data must be of the best possible quality.

Data Errors and need for QC

- Every measurement includes the actual parameter value plus any errors resulting from or affecting the measurement method.
- Therefore, an "error" is the difference between the "observed" or "measured" value of some parameter and the "actual" value of the parameter.
- Data errors, are in one of two categories:
  - Random errors: Distributed more or less symmetrically about zero, do not depend upon the measured value
  - Systematic errors: Distributed asymmetrically about zero error, tend to bias the measured value either above or below the actual value

Type of Data sets & QC methods

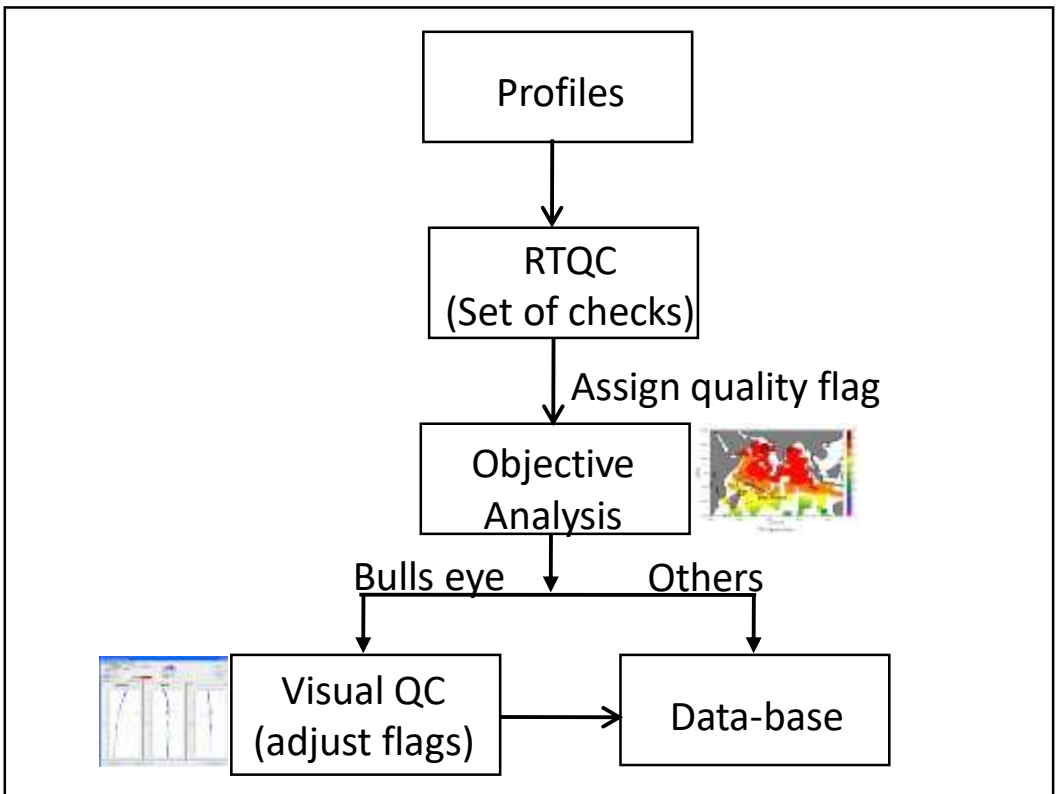
- Argo data from profiling floats.
- CTD data from Ocean Research Vessel (ORV).
- XBT/XCTD data from Research Vessels.
- Specific Project related data viz., ARMEX, BOBMEX, ICRP etc.
- Time series data.
  - Moorings, Drifters, Current Meter Arrays, ADCPs etc.
  - Surface temperature, salinity, currents, winds etc.

Types of Quality Control

- QC are generally of the form
  - Real Time Quality Control (RTQC) : To be done immediately after obtained.
  - Delayed Model Quality Control (DMQC): To be done by expert in the field by checking the data.
  - Visual Quality assessment (by expert)
  - Patterns based Quality Assessment (Machine Learning based)
  - Clusters based Quality Assessment (Machine Learning based)
  - Polygon and Alpha convex hulls based.

IOC Manuals for treating data

QC Flow Chart



Real Time Quality Control

Metadata Quality Check

1. **Platform identification:** Each platform should have an unique valid identifier provided by World Meteorological Organization (WMO).
2. **Impossible date/time test:** Month in the range of 1 to 12; date must in the expected range for the month; hours in range 0 – 23; minutes in the range 0 – 59.
3. **Impossible location test:** The latitude (longitude) must be in the limits -90 to 90 (0 to 360).
4. **Position on land test:** The position must be located in the ocean. ETOPO2 bottom topography can be used for this test.
5. **Impossible speed test:** Surface and subsurface drift speeds must not exceed 3 m s^{-1}.

Quality Check on data

6. **Global range test:** Temperatures must be in the range of -2.5° to 40.0° C and salinity must be from 2 to 41 psu.

7. **Regional range test:** Temperatures from floats in the Red Sea (Mediterranean Sea) must range from 21.7° to 40.0° C (10.0° - 40.0° C) and salinity ranges must be from 2.0 to 41.0 (2.0 to 41.0 psu).

8. **Pressure increasing test:** The pressure must increase monotonically.

9. **Spike test:** $|V_2-(V_3+V_1)/2| - |(V_3-V_1)/2|$ for a value V_2, where V_1 and V_3 are the values above and below V_2, which may not exceed prescribed limits. Above 500 dbar, the limit for temperature (salinity) is 6°C (0.9) and below 500 dbar the limits are 2°C (0.3).

10. **Top and bottom spike test:** This test is used to check the top and bottom values.

Quality Check on data

11. Gradient test: The test value $|V_2 - (V_3+V_1)/2|$ for a value V_2 may not exceed prescribed limits. Above 500 dbar, the limit for temperature (salinity) is 9.0°C (1.5) and below 500 dbar the limits are 6.0°C (0.5).

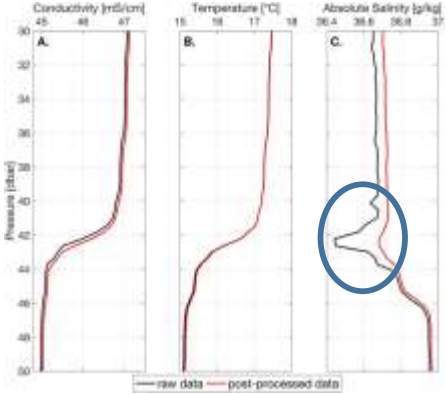
12. Stuck value test: This test checks for constant temperature or salinity values throughout the profile

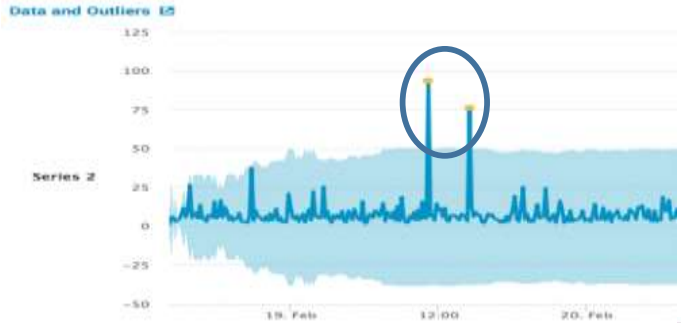
13. Density inversion : This test computes the density at all pressure levels from the observed temperature and salinity values and tests for hydrostatic stability.

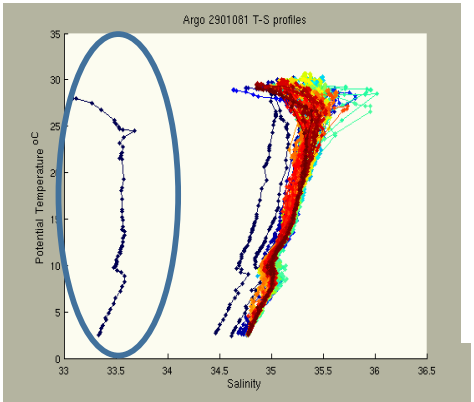
14. Gross salinity or temperature sensor drift: If the average temperature (salinity) from the last 100 dbar of two adjacent profiles exceeds 1°C (0.5), then the profile is considered to be bad.

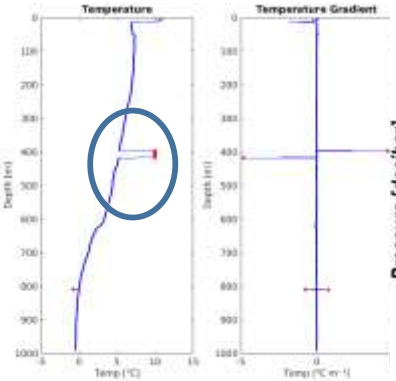
15. Frozen profile test: If five consecutive profiles with very small differences throughout the entire water column (i.e., of the order of 0.001 for salinity and of the order of 0.01°C for temperature) they are candidates for the gray list.

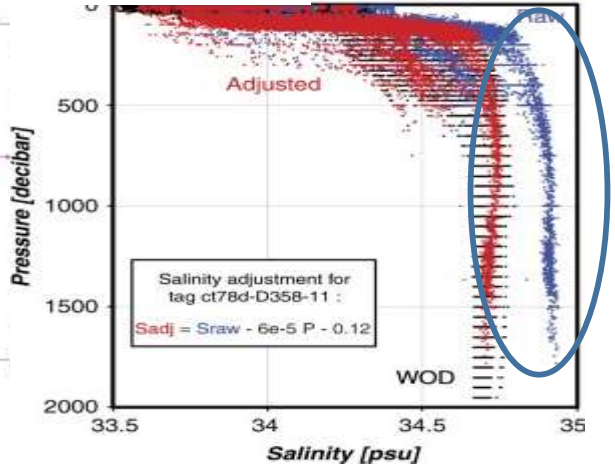
Typical examples



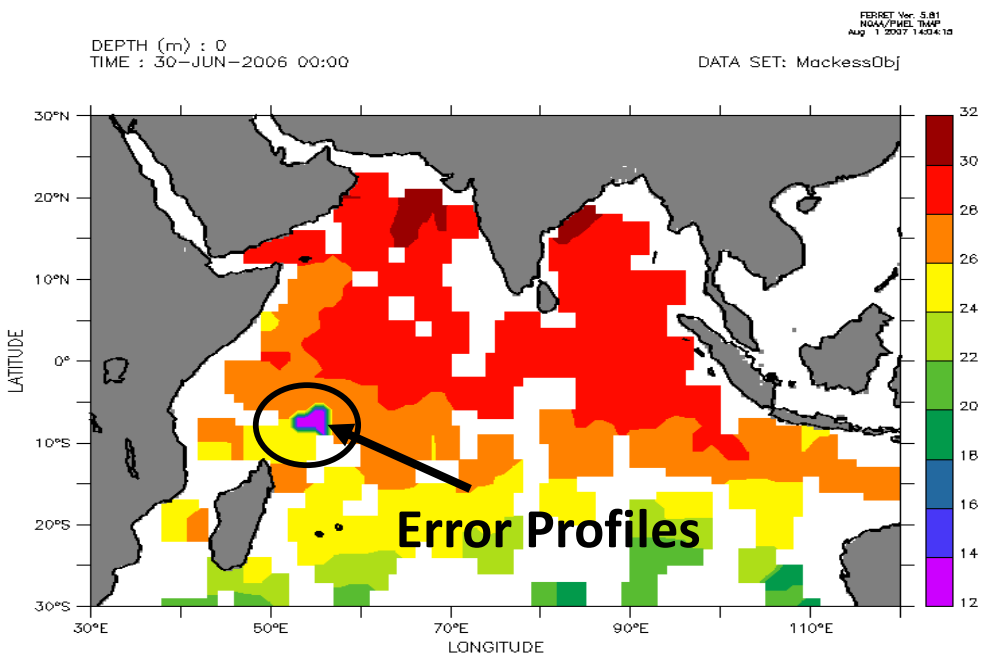








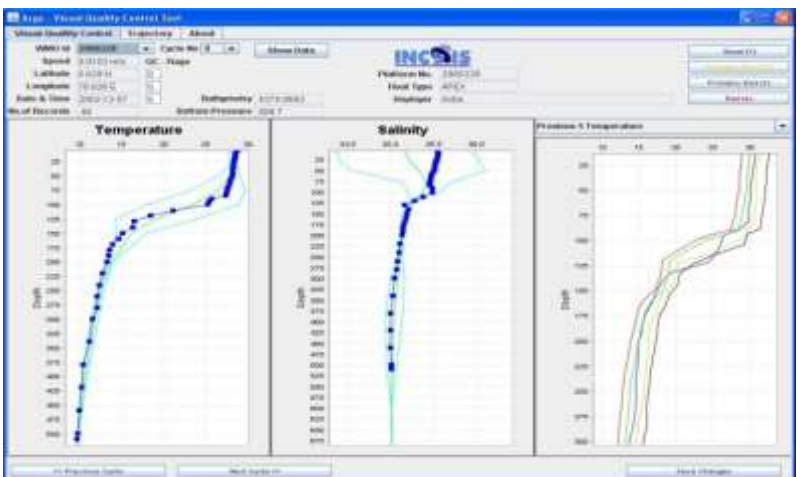
Objective Analysis



Temperature

Visual Quality Control of Profile data

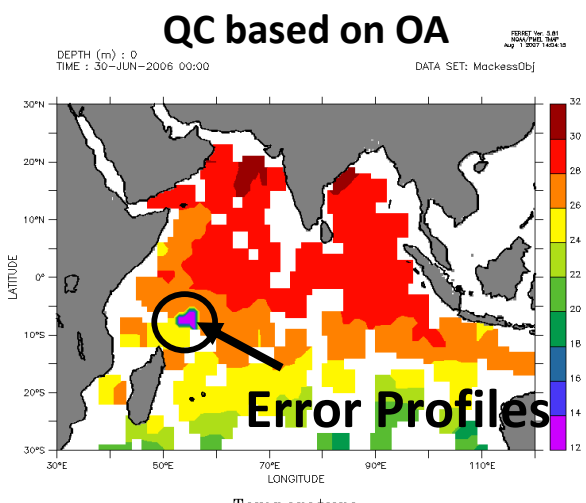
- Visual Quality Control Tools is used for QC of profile data. WOA18 is used in the background for the QC process.
- Profiles falling out of 2.5*sigma are set as bad.

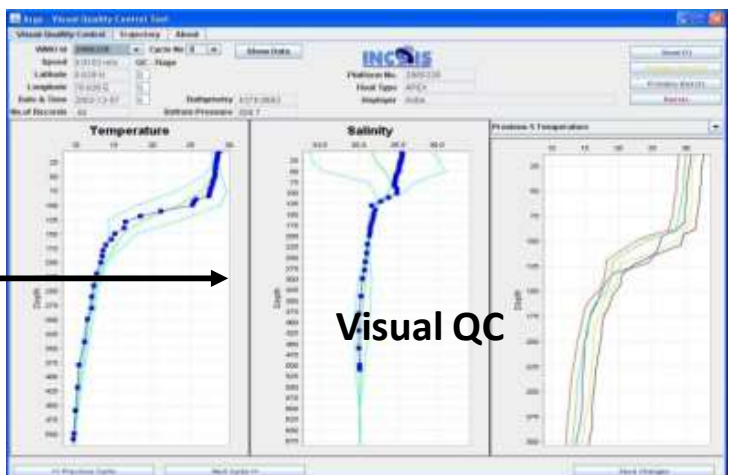


T&S Profiles (Argo Floats, XBT, XCTD, CTD)

Three way QC of ocean profiles

Automatic QC



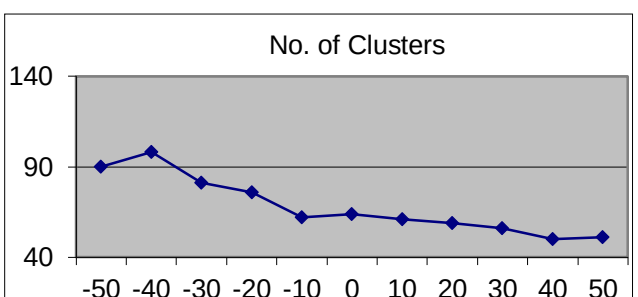


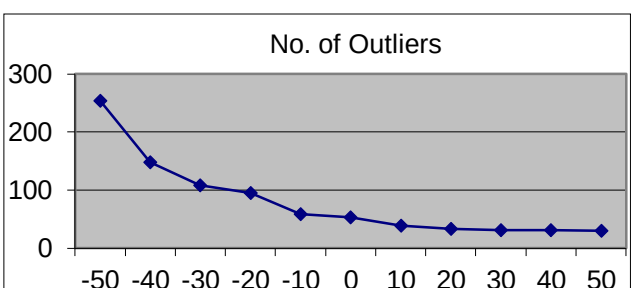
Udaya bhaskar et al., 2013 IJMS,
Udaya bhaskar et al., 2012, IJEE

Cluster based Quality Control

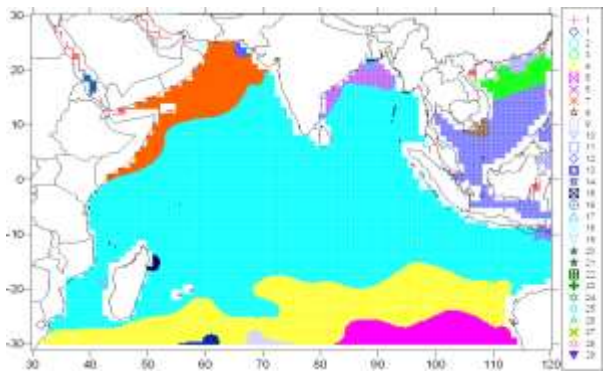
Quality control of oceanographic data sets based on spatio-temporal clustering

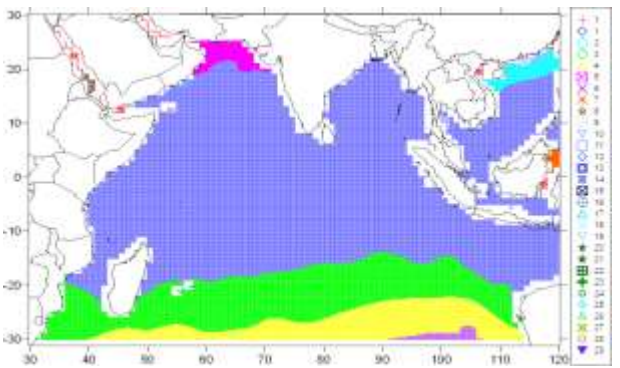
- Clustering method with WOA18 climatology.
- Tested for all standard levels
- Sensitivity studies were done.

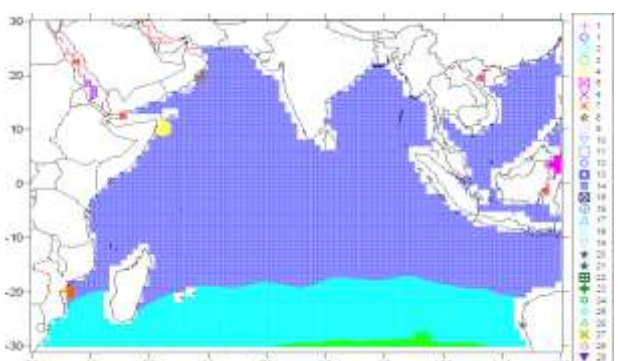


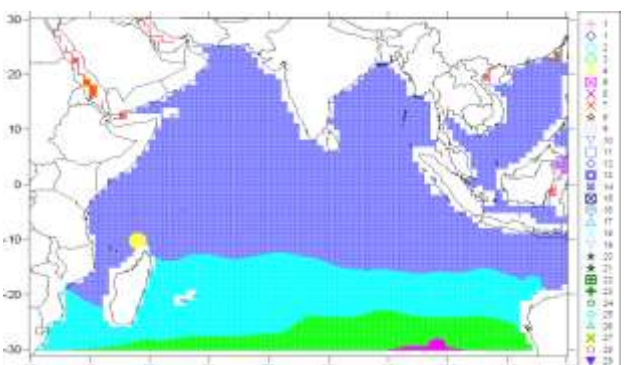


Clusters



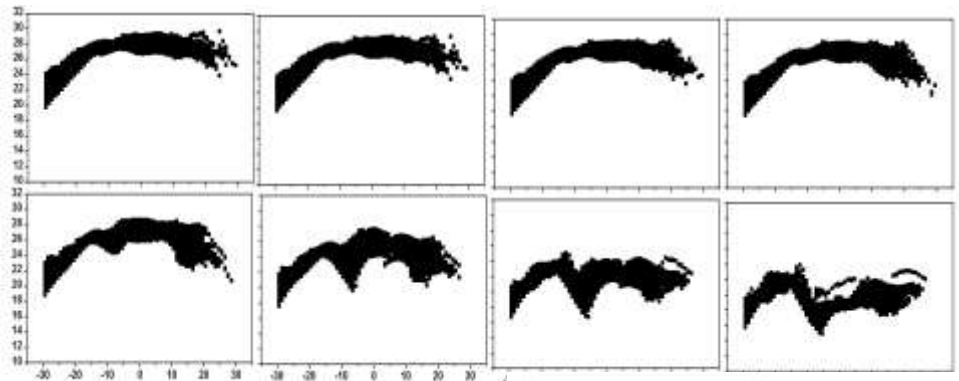




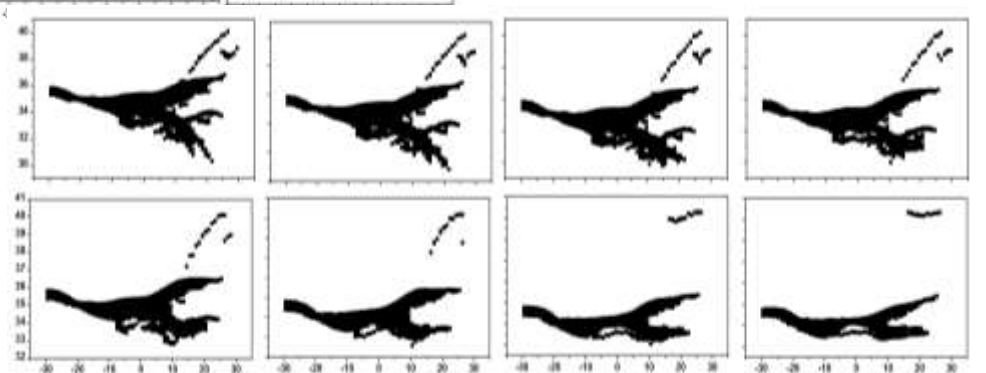


Pattern based Quality Control

Patterns of Temperature and salinity at different depths (Source: WOA18)

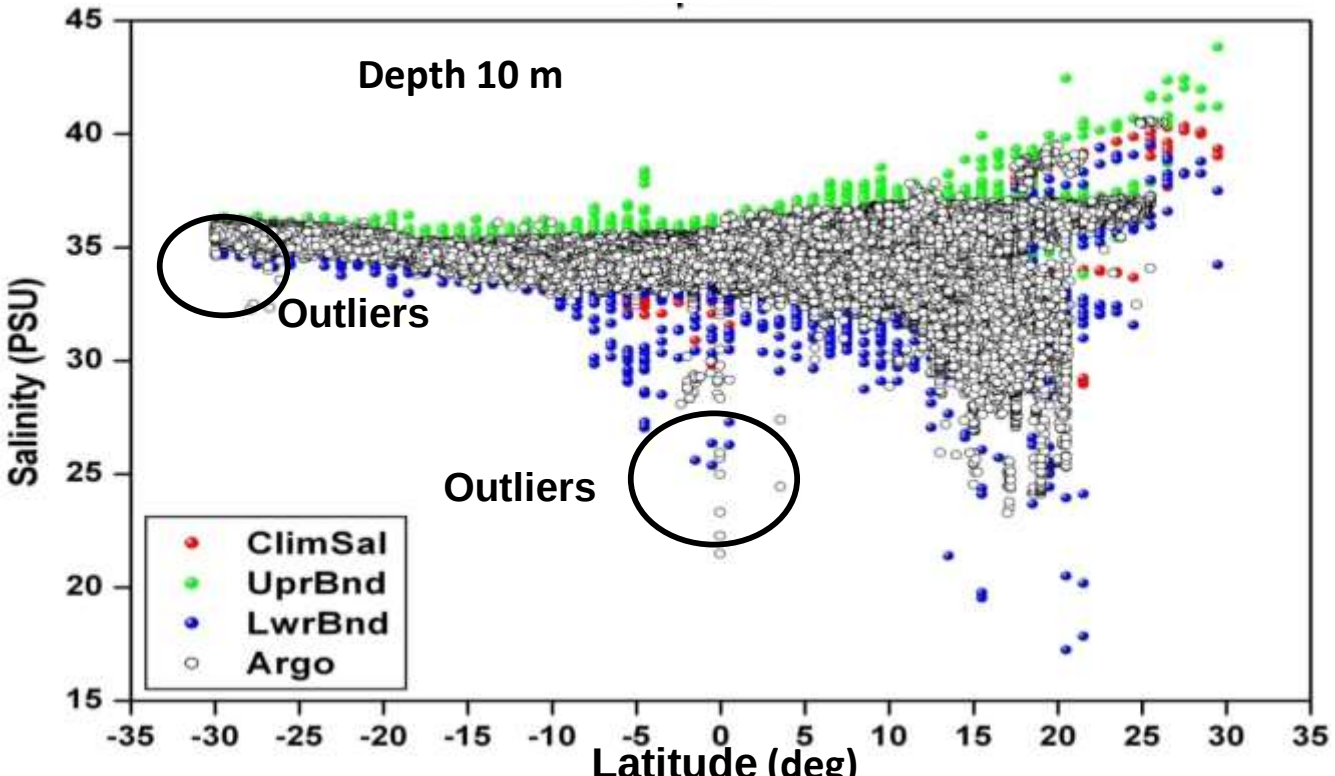


Latitude (Deg N)

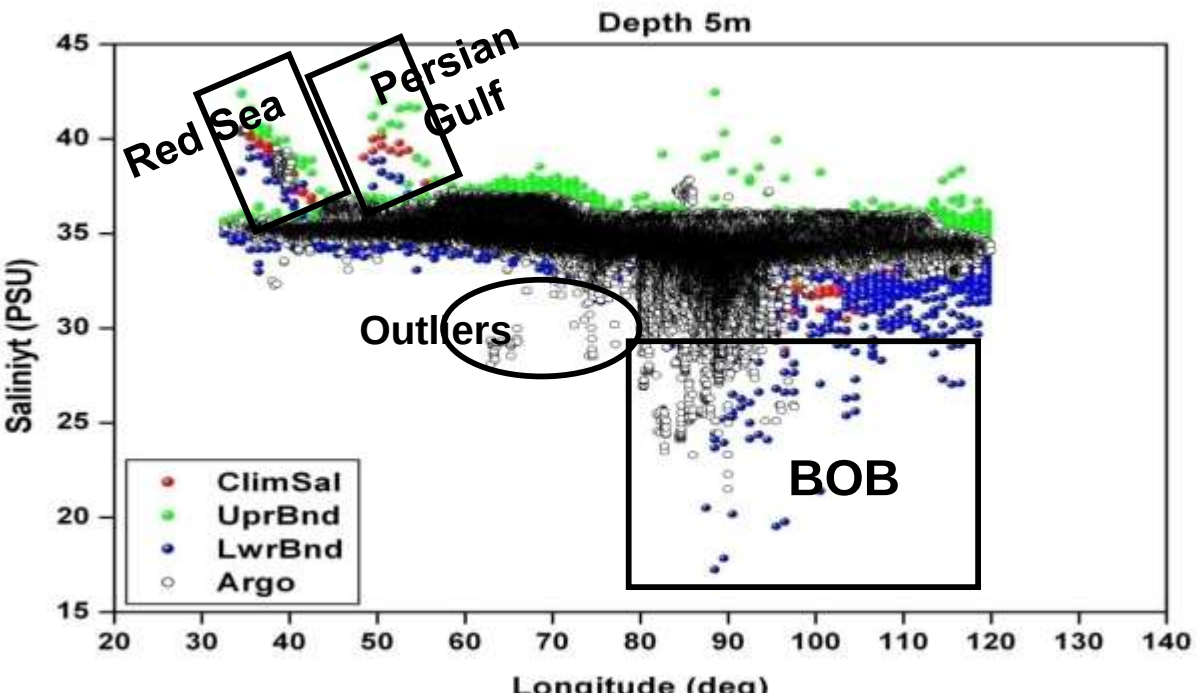


Latitude (Deg N)

Salinity Pattern

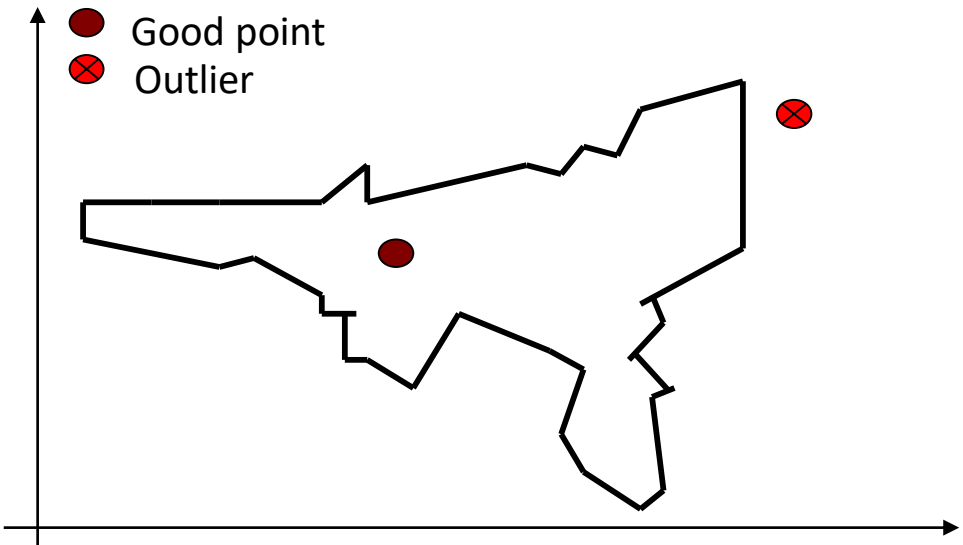


Salinity Pattern

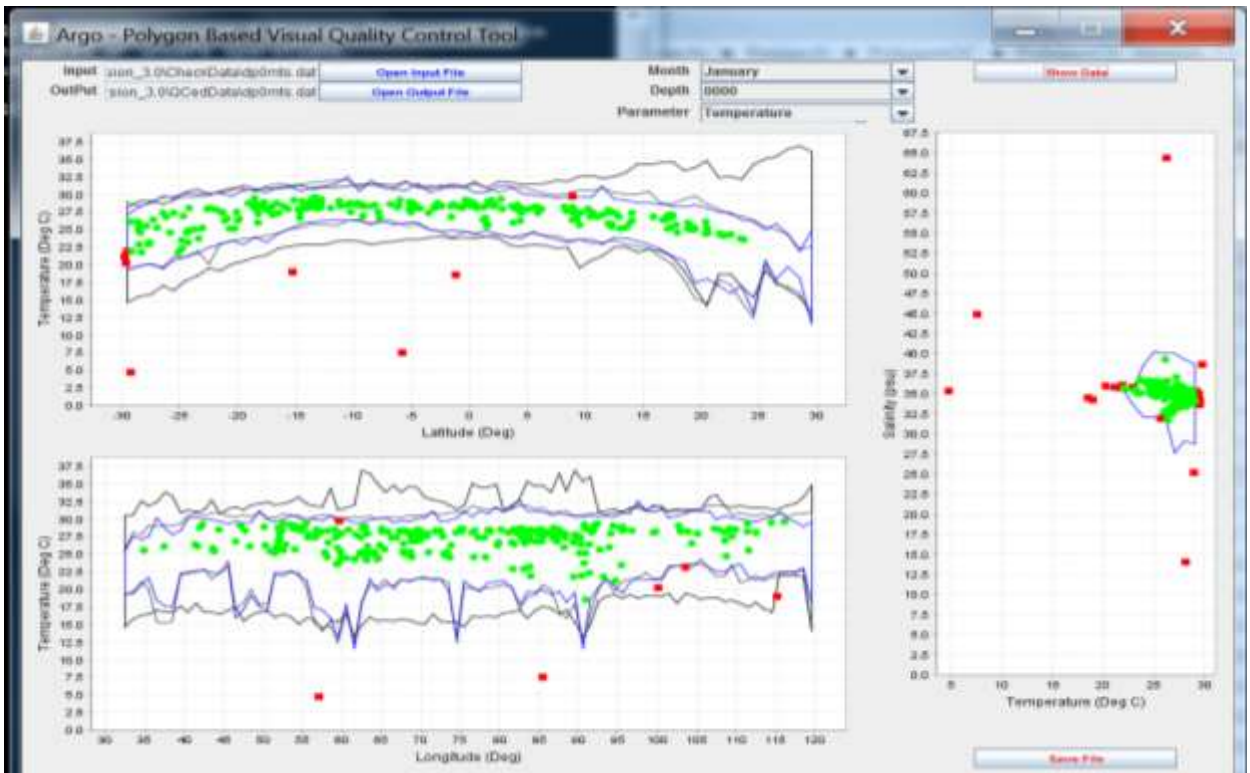


QC based on Lat/Lon Clusters

- Climatological mean and standard deviation are used to build n-sided (dimensional) polygon.
- Outliers are detected based by judging whether the points are falling with in the polygon or out side.
- These are re-examined for any special events and then injected into the database.



The n-sided polygon is used for checking outliers. Now the data points whose quality is to checked are introduced and checked whether the points fall with in the n-sided polygon (good points) or outside the polygon (bad points).


Argo - Polygon Based Visual Quality Control Tool
Input
OutPut
Open Input File
Open Output File
Month
January
Depth
0000
Parameter
Temperature
Show Data
Temperature (Deg C)
Latitude (Deg)
Temperature (Deg C)
Longitude (Deg)
Salinity (psu)
Temperature (Deg C)
Save File

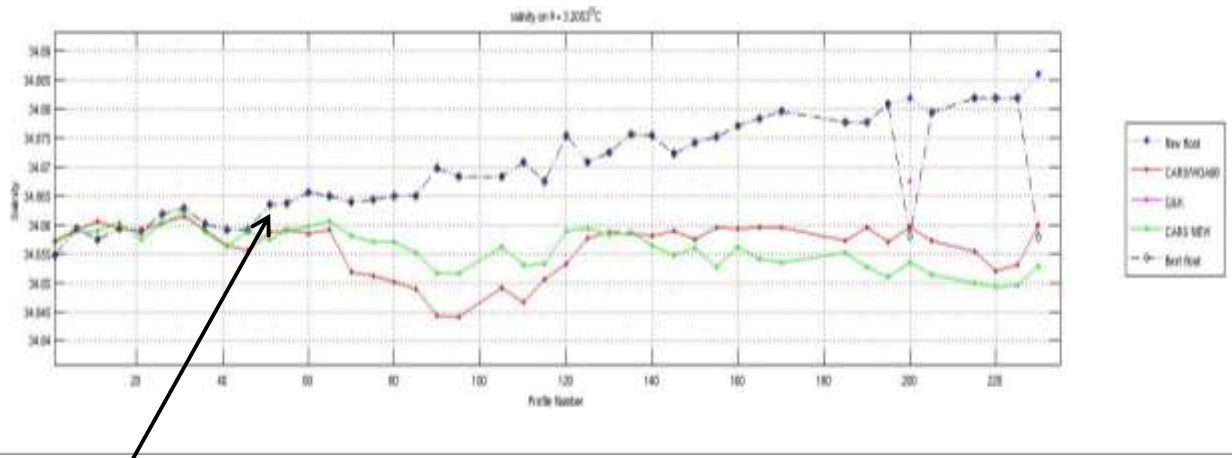
Delayed Mode Quality Control (DMQC)

- Why DMQC?
  - Once an Argo float is deployed in the ocean, it is very difficult to calibrate its sensors or to monitor its condition under operation.
  - Argo target accuracies for measurement are 5 dbar for pressure, 0.005°C for temperature, and 0.01 for salinity (Argo Science Team, 2000).
  - The former two objectives could be achieved over a four year float life using technology available
  - How ever Salinity measurements were expected to be liable to experience some drift and offset, probably due to bio-fouling.
- Hence DMQC is required…

What we do in DMQC

- Compare with available ship based CTD and also different climatologies.
- In DMQC we check for:
  - Salinity drifts.
  - Tri-Butyl Tin Oxide (TBTO) problems which causes freshening on salinity in initial profiles.
  - Salinity hooks.
  - Surface pressure Offsets problems.
  - Thermal lags problems.
  - Truncated Negative Drifting Pressure (TNDP).

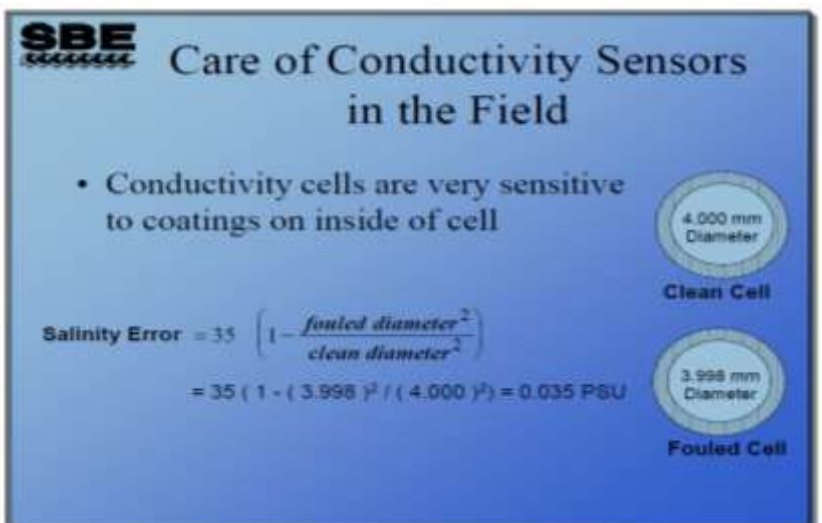
Salinity drifts



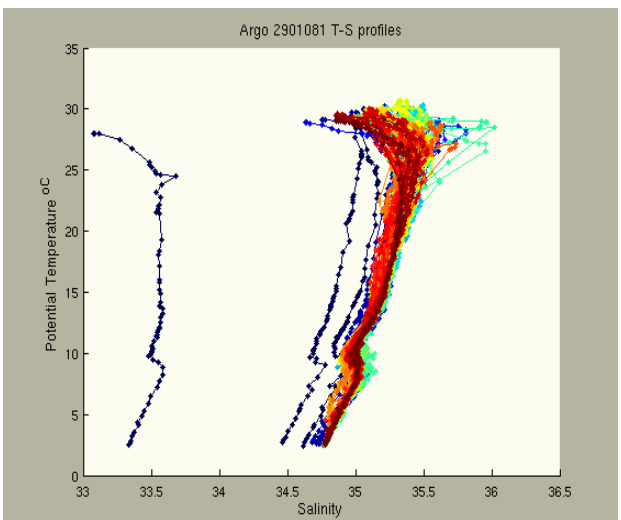
This float is observed to posses drift in salinity starting from cycle 44 onwards

TBTO issues

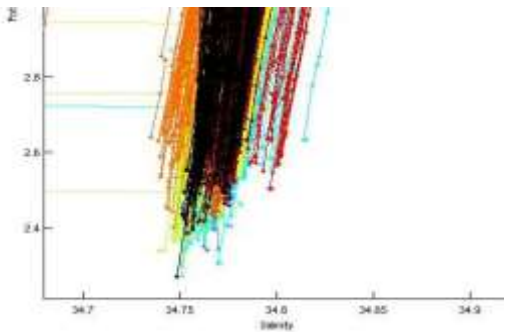
- Anti-fouling agents causing the cell dimensions to change there by causing errors in the salinity measurements.



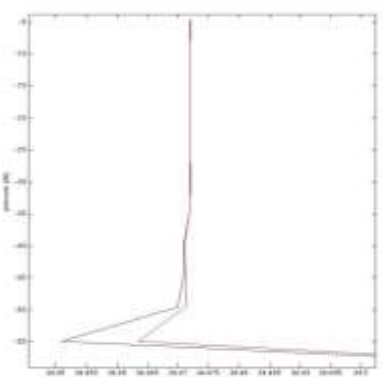
(Courtesy: SBE training manual)



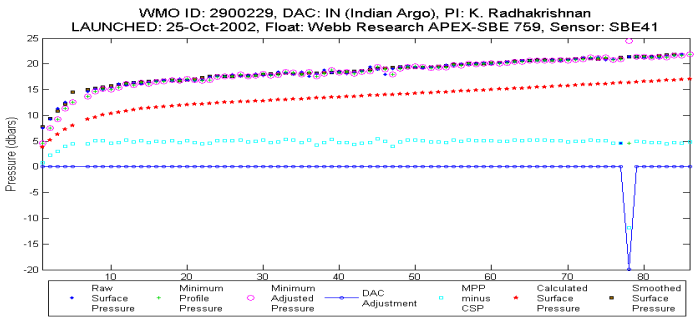
- This happens some times due to trapping and non-flushing of sea water.



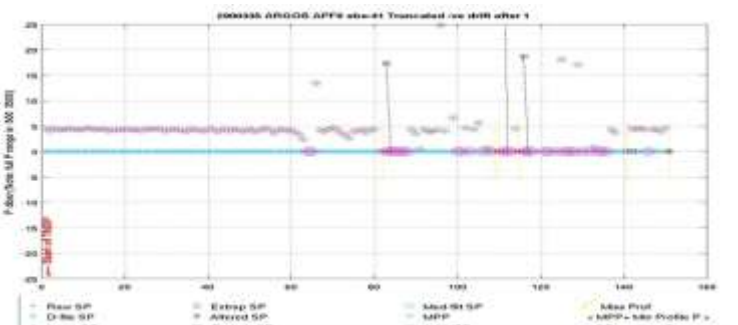
- Thermal lag problems: salinity spiking at the Mixed layer.



- Pressure sensor offset:



- TNDP issues :



Outcome of QC

- Data with QC flags classifying the data:
  - 0 – No QC done
  - 1 – good data
  - 2 – Probably good data
  - 3 – Potentially correctable bad data
  - 4 – Bad data
  - 5 – Interpolated data
- No data nor the records are eliminated.
- The end user has the choice to use the flags or perform his own methods before using the data.

Reference Publications

- Quality control of oceanographic in situ data from Argo floats using climatological convex hulls, T.V.S.Udaya Bhaskar, R.Venkat Shesu, Timothy P. Boyer, E. Pattabhi Rama Rao, MethodsX 4(2017)469–479.
- R. Venkat Shesu, T.V.S. Udaya Bhaskar, E. Pattabhi Rama Rao, M. Ravichandran, B. Venkateswara Rao, An improved method for quality control of in situ data from Argo floats using α convex hulls, MethodsX, Volume 8, 2021, 101337, ISSN 2215-0161, https://doi.org/10.1016/j.mex.2021.101337.
- A note on three way quality control of Argo temperature and salinity profiles - A semi-automated approach at INCOIS, TVS Udaya Bhaskar, R Venkat Shesu, E Pattabhi Rama Rao, R Devender, International Journal of Earth Sciences and Engineering, Vol 5(6), pp 1510 - 1514, 2012.
- GUI based interactive system for Visual Quality Control of Argo data, TVS Udaya Bhaskar, R Venkat Seshu, E Pattabhi Rama Rao, R Devender, Indian Journal of Geo-Marine Sciences, Vol 42(5), pp 580 - 586, 2013.
- Detecting and Correcting the Degradations of Sensors on Argo Floats Using Artificial Neural Networks, TS Raju, T.V.S. Udaya Bhaskar, JP Kumar, KS Deepthi, Computer Communication, Network and Internet Security, 299 - 308, Springer Lecture Series.
- Use of Convex Hulls for detection of outliers in oceanographic data pertaining to Indian Ocean, Ch Murali Krishna, TVS Udaya Bhaskar, M Kranthi Kiran, International Journal of Advances in Electronics and Computer Science, Vol 3 (8), 2016.

Summary

- Quality Control: refers to methods or processes about overall quality goals.
- Good quality data are also important for verification and research purposes.
- Data errors, are Random and Systematic.
- Many conventional and bulk treatment methods are in place.
- No data is rejected but flag is assigned which can be used by individual users.


Thank you