On the use of climatological convex hulls for detecting outliers and quality control of oceanographic data sets

TVS Udaya Bhaskar
DMG, INCOIS
Hyderabad – 90

4th IQuOD meeting

3 – 5th October 2016 Tokyo, Japan

Outline of talk

- Guidelines for QC and some of the flaws.
- Building convex hull from climatology
  - Mean field + standard deviation fields
- Use of Jordan Curve Theorem for building polygon.
- Use of Jarvish March algorithm + PIP algorithm to detect outlier.
- Application of the proposed model to identify outlier in oceanographic data sets.
  - Demonstration of sample cases.

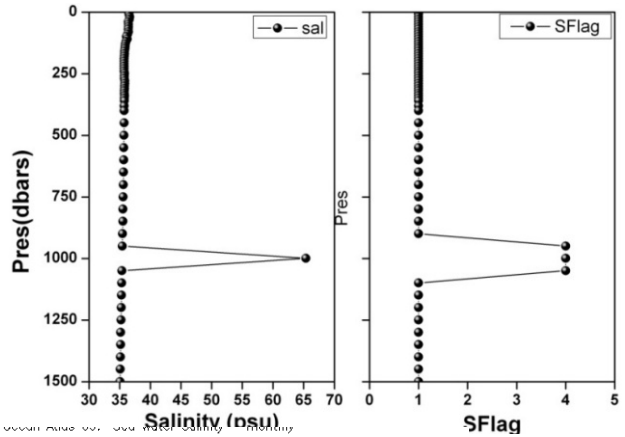
Guide lines for QC of data

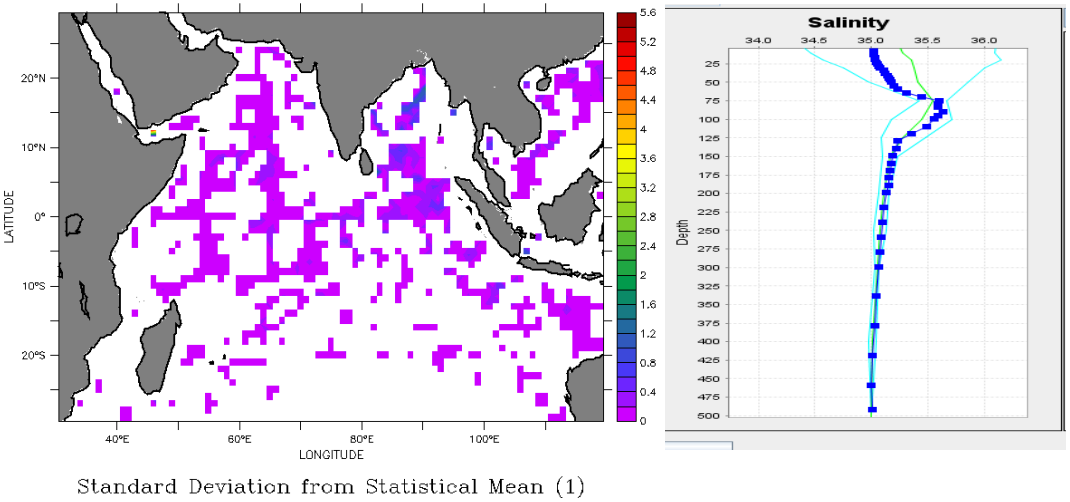
- In practice the oceanographic data is Qced based on preset QC methods
  - Eg: IOC Manual #22 gives guide lines for QC of oceanographic data.
  - This checks for density inversions, spikes, stuck values, offsets etc.
- Visual QC based on climatological means and standard deviation fields is also employed for performing QC of oceanographic data sets.

  (Eg: NODC, USA uses monthly, seasonal and annual means and stdev)

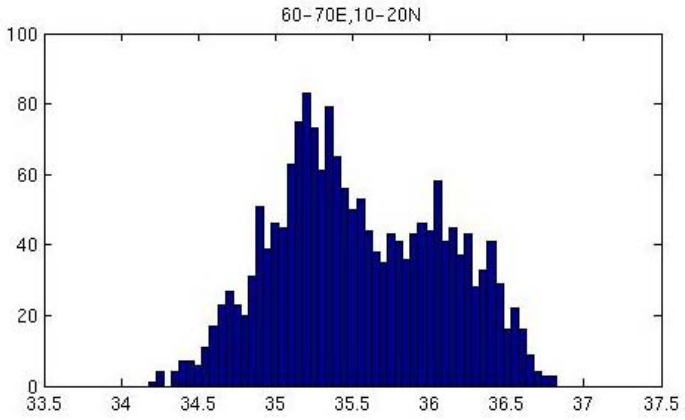
Flaws in these methods

- Udaya bhaskar et al, (2012) describes some of the flaws in the standard checks automated checks.
  - Good records above and below the spiky data is wrongly flagged as bad even though they are good.
- VQC cannot be effectively performed if the standard deviation field is missing.
  - Slight variation from mean too cannot be justified.
- Problems when data is bio-modally distributed contrary to the assumption that data is normally distributed while performing the VQC with mean and standard deviations. One of the mode may be set as bad.





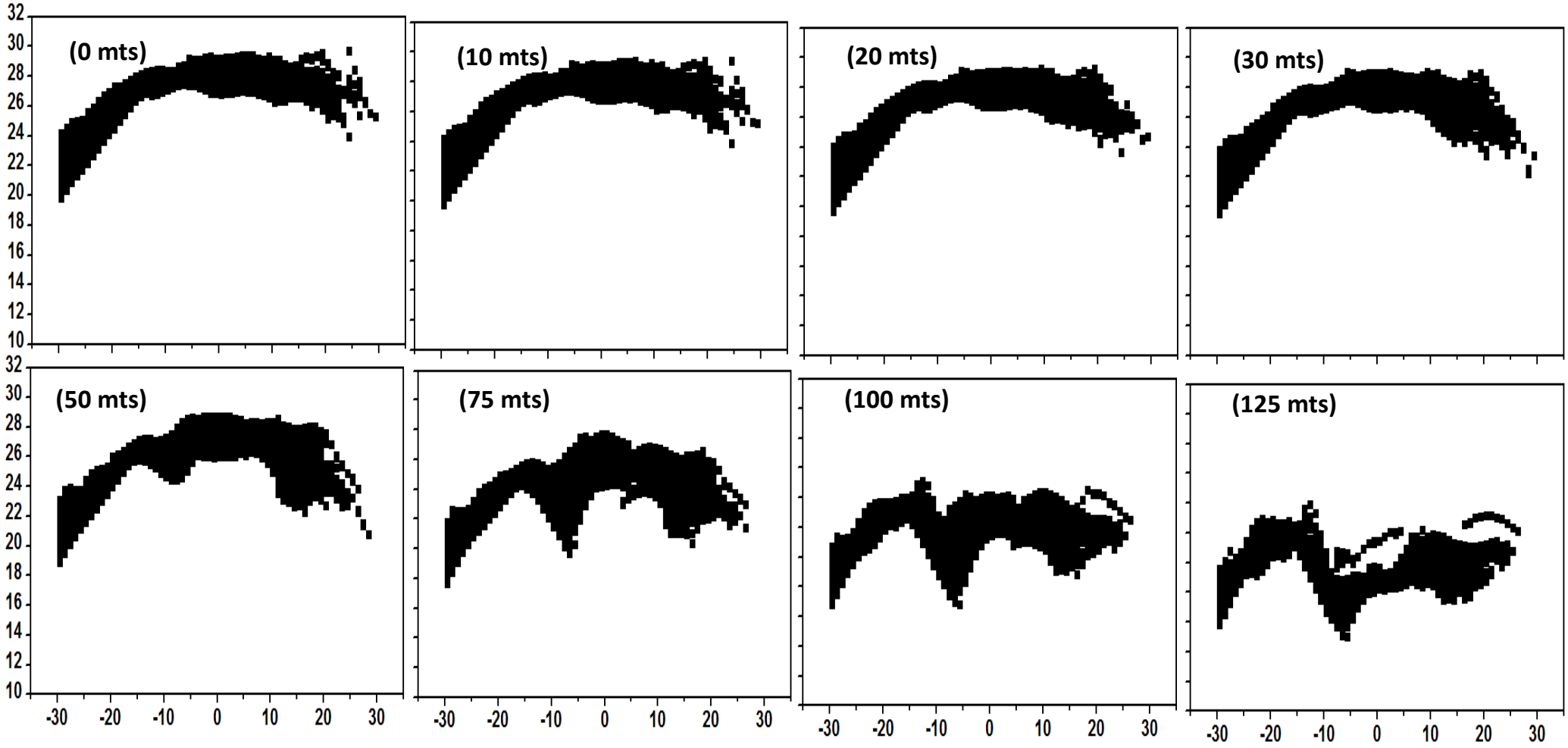
Standard Deviation from Statistical Mean (1)



The proposed method

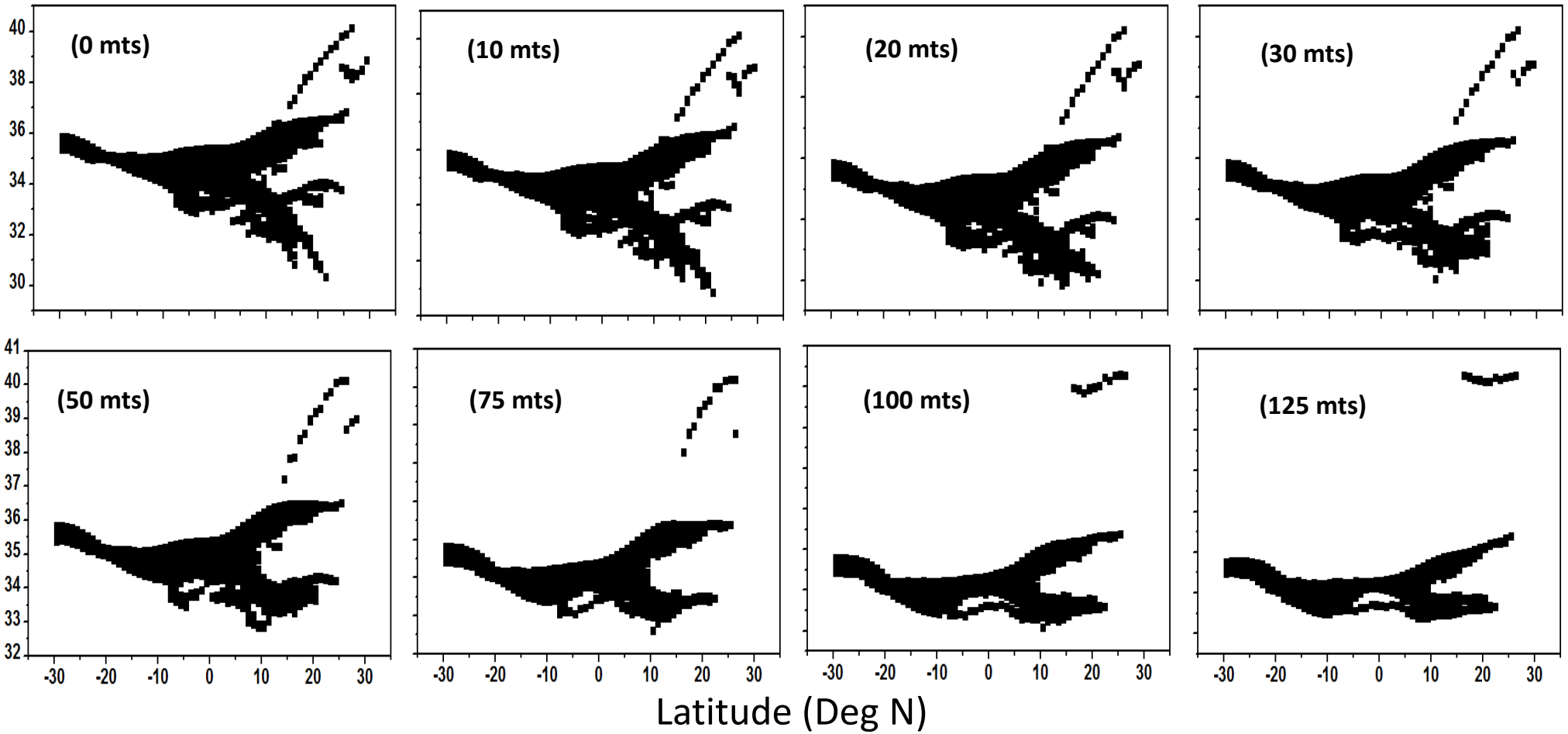
- Observe patterns from the climatological data sets while plotting parameter against latitude, longitude.
- Build polygons using these `n` topples of (lon,param) and (lat,param).
- Based on Jordan Curve theorem separate the space into two parts.
- Use point in polygon (PIP) algorithm to separate the good and bad points and perform QC.

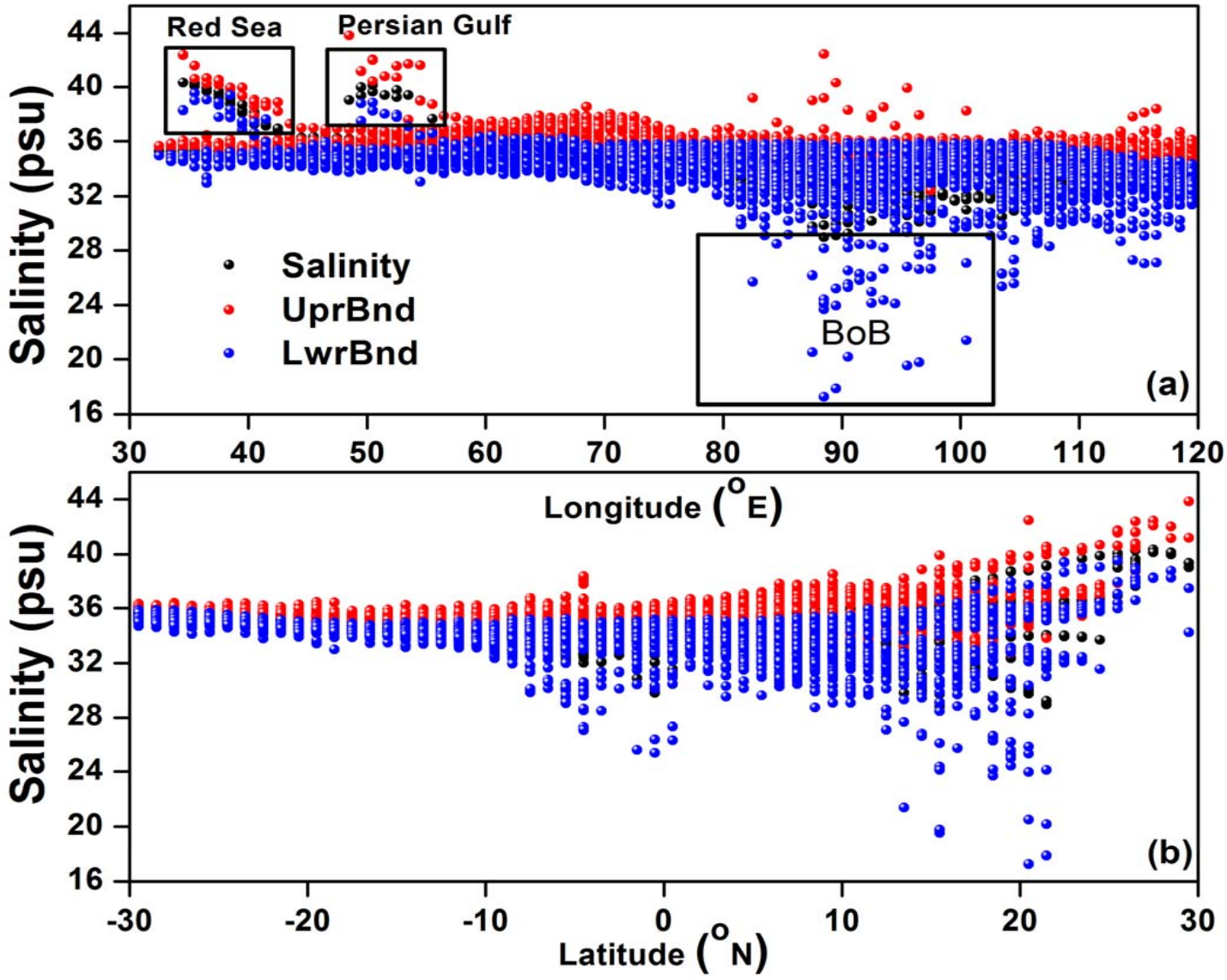
Patterns of Temperature at different depths (Source: WOA13)



Latitude (Deg N)

Patterns of Salinity at different depths (Source: WOA13)

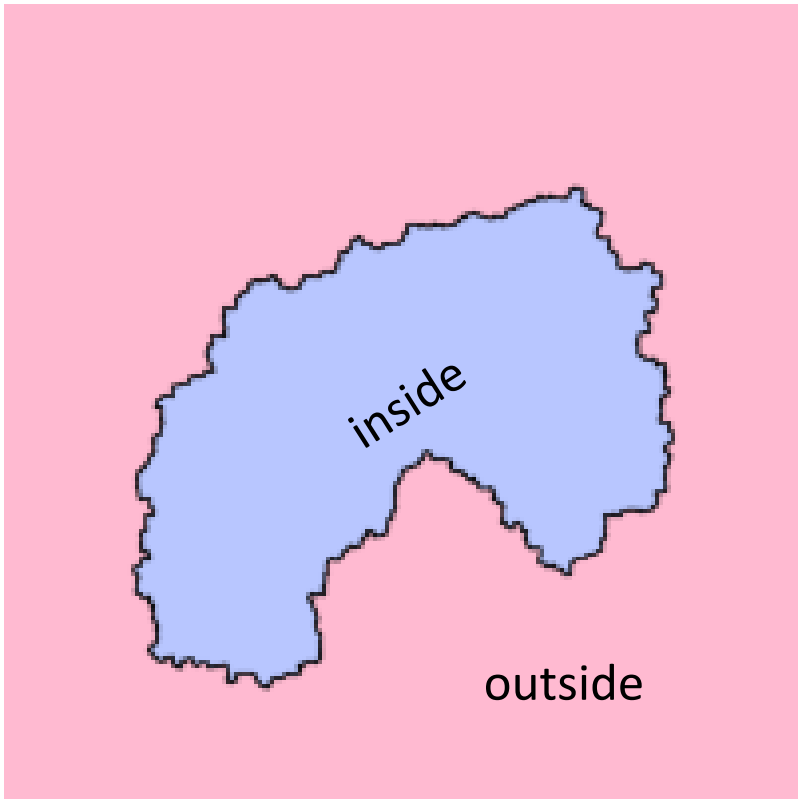




Patterns of salinity in the Indian Ocean when plotted against (a) longitude and (b) latitude corresponding to the depth 0 mts obtained from World Ocean Atlas 2013. Black, blue and red dots represent the mean, (mean - 2*SD), (mean + 2*SD) respectively

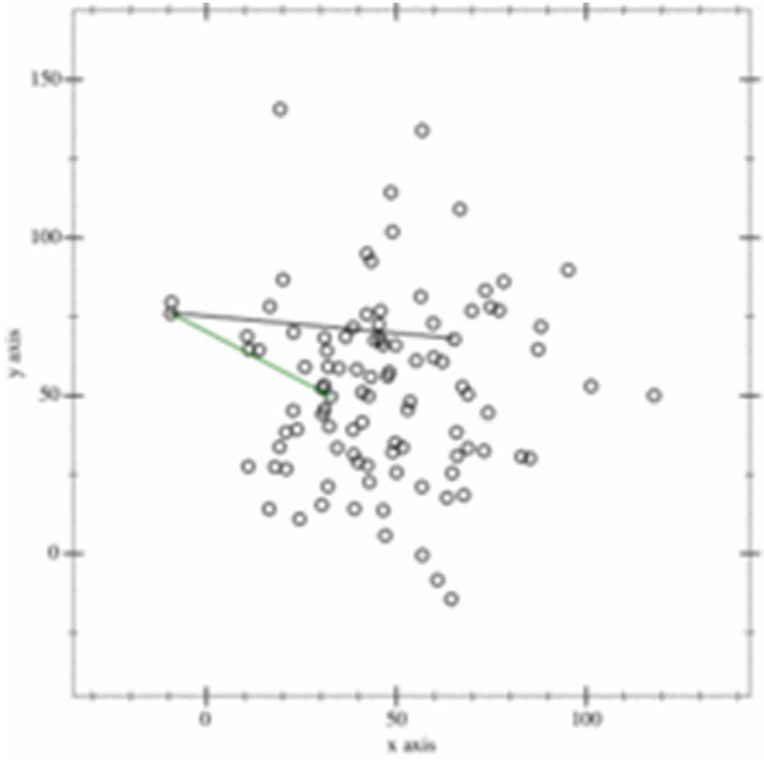
Jordan Curve Theorm

- If **J** is a simple closed curve in $\mathbf{R^2}$, then the Jordan curve theorem, also called the Jordan-Brouwer theorem states that $\mathbf{R^2}$ - **J** has two components (an "inside" and "outside"), with J the boundary of each



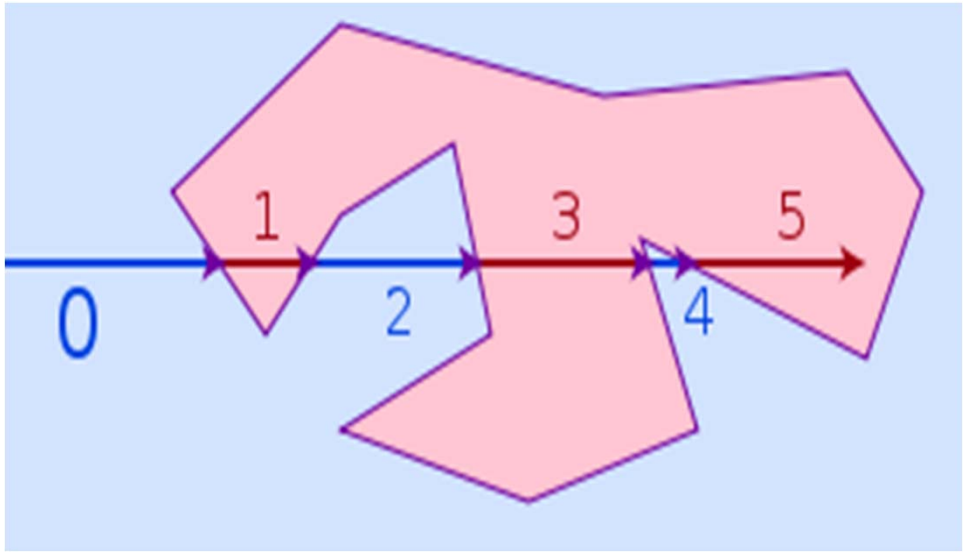
Building polygons (Convex hulls)

- Once the patterns are observed we used Jarvish March algorithm to build the polygon.
- This polygon is the one with least area encompassing all the given points.
- Works with all the edge points in building the polygon.

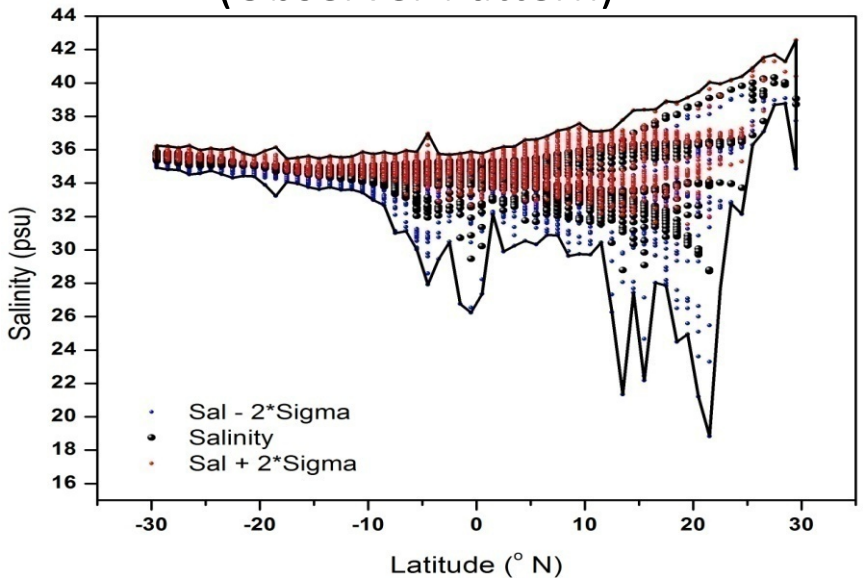


Point In Polygon (PIP) Algorithm

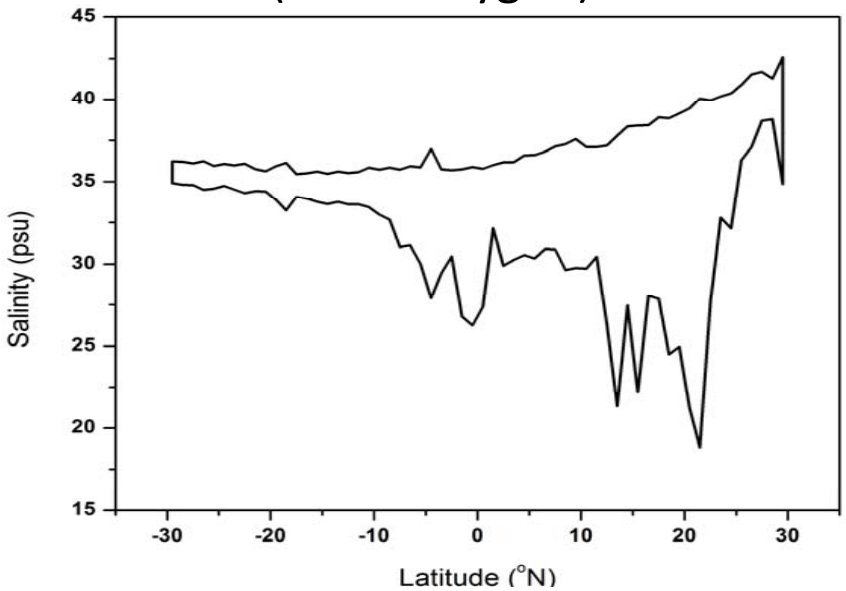
- PIP algorithm is used to check if the point is inside or outside the polygon.
- Ray casting method is used for implementing this algorithm.
- Any ray to this point if cross the polygon
  - even number of times (outside)
  - Odd number of times (inside)

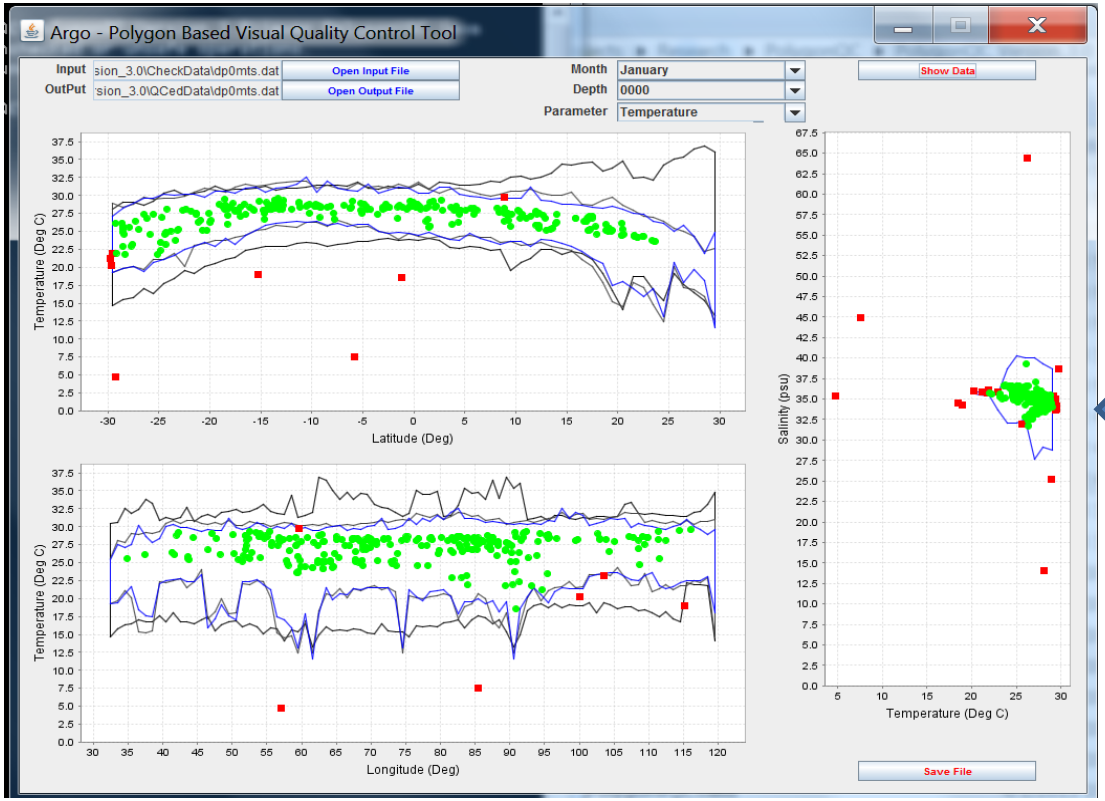


(Observer Pattern)

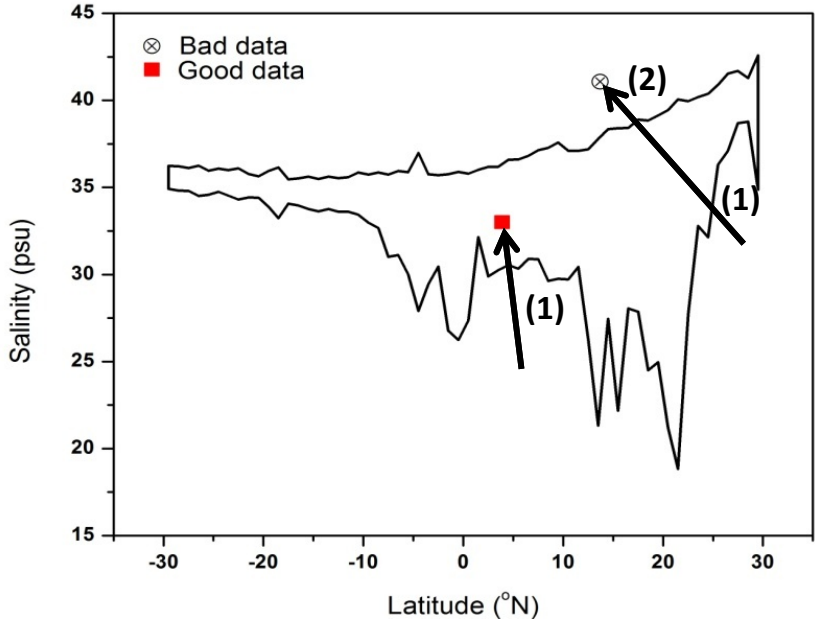


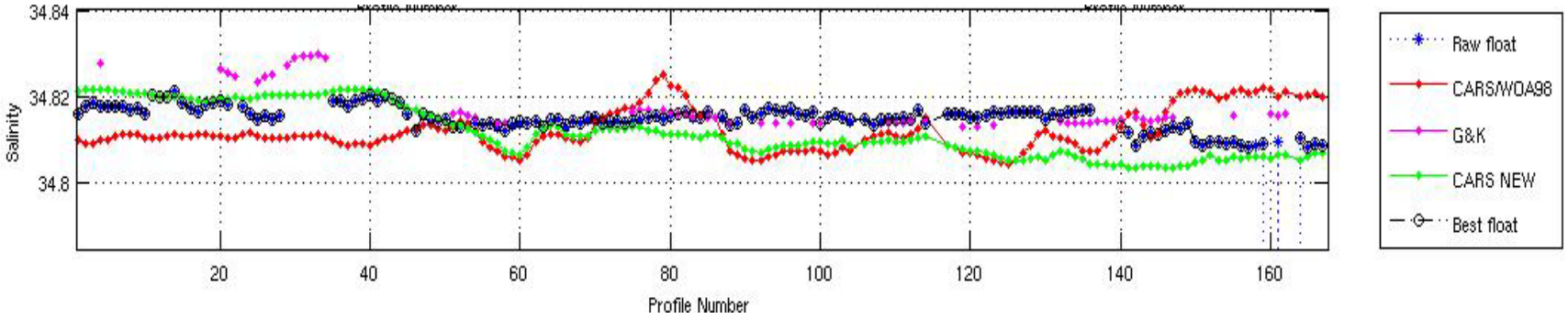
(Build Polygon)

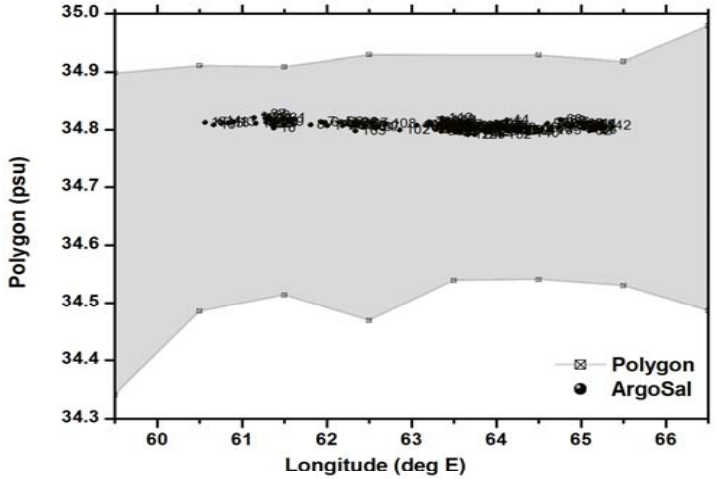


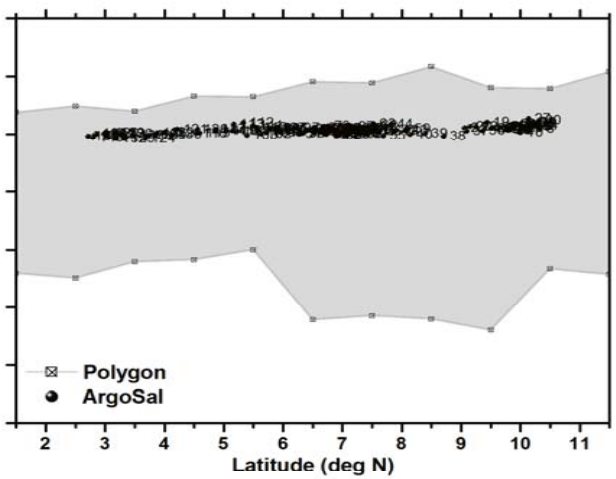


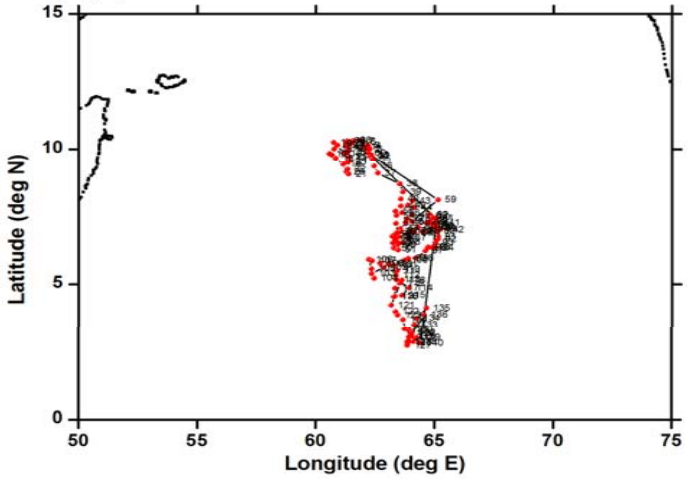
Check points inside outside



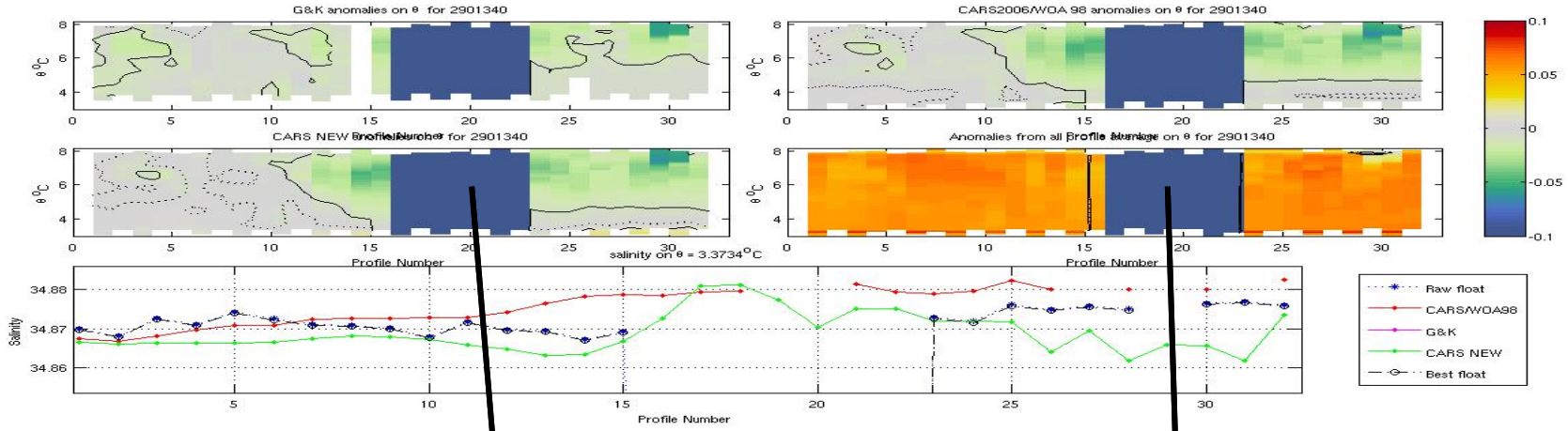


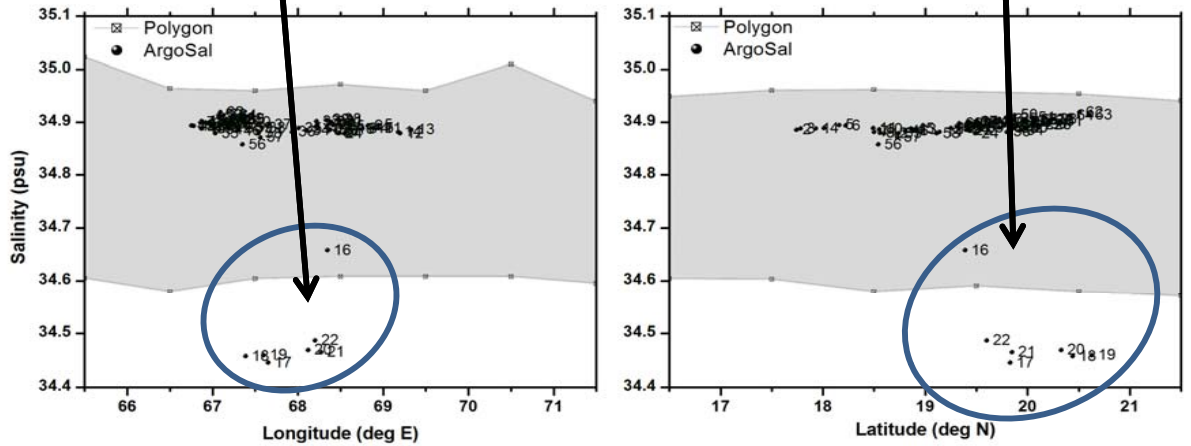


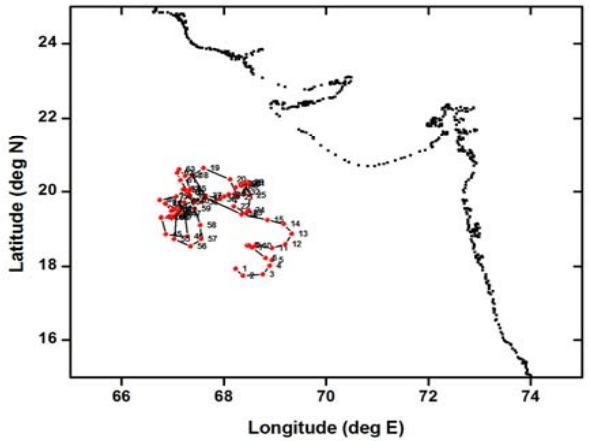




Good float data

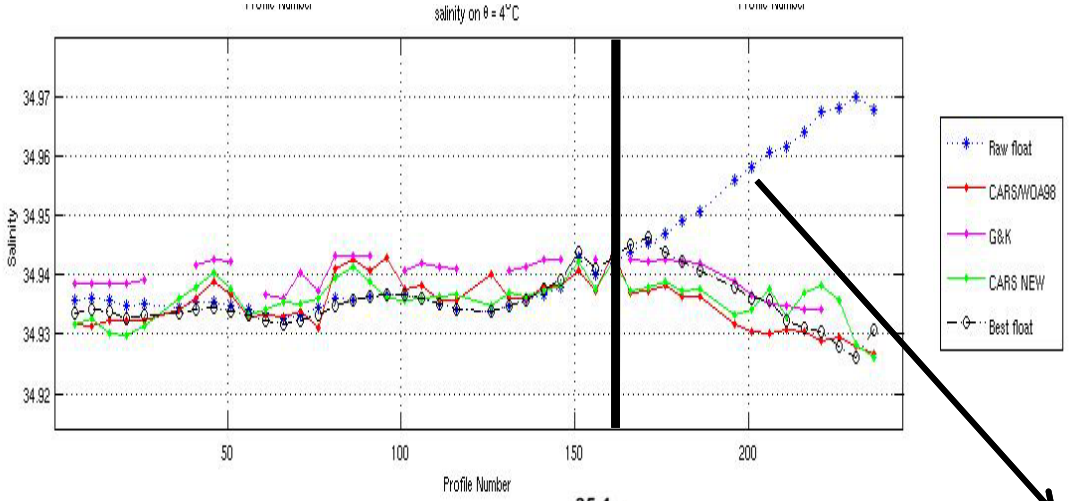


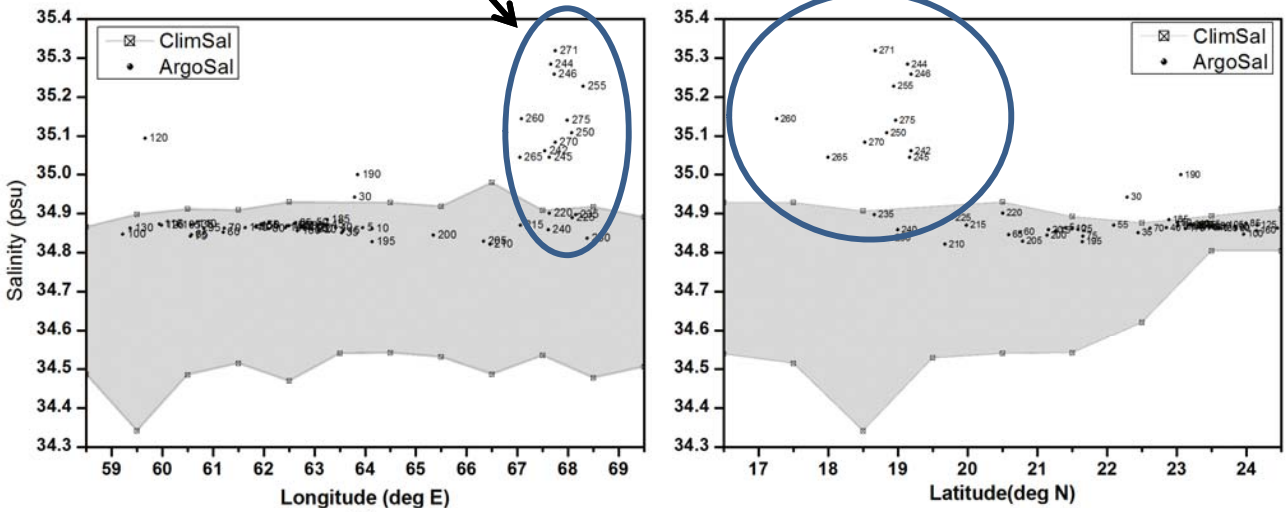


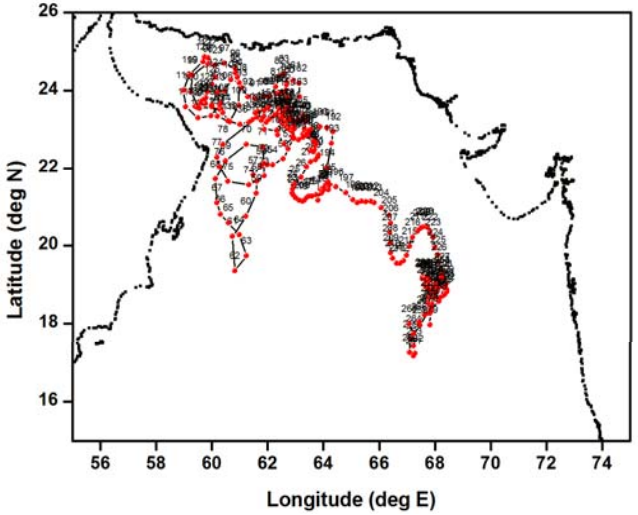


Sensor offset

Sensor Degradation







Observations

- A good climatology is very much required for the proposed method to work perfectly.
- All the pointed outliers need not be outliers. There can also areas where there is no good data coverage in climatology.
- The possible outliers need to rechecked if they are due to some extreme events like ENSO, IOD, cyclones etc.
- Polygon need to be update with addition of high quality data from all sources.

Thank you