Quality control of large Argo datasets

F. Gaillard, E. Autret, V. Thierry
Laboratoire de Physique des Océan
UMR 6523: CNRS-IFREMER-IRD-UBO

Context:

- Argo: more than 9000 profiles per month
- The **utopia**: develop an automatic tool to remove all bad data and keep all good data !!!
- The data center **approximation**: apply automatic QC to separate the problem in 2 parts
  1. Real time assimilation: remove all bad data (ex: ECMWF wind measurements from ships) (QC: 1)
  2. Delayed mode use (re-analysis): keep all possibly good data, that will need later processing (QC: 1, 2,3?)
- The **fact**: the real world is not binary and frontiers are difficult to specify in the QC continuum 1-2-3

- Major reason to the failure of automatic testing: Qualification of **individual** profiles does not exploit all available knowledge

Using estimation theory for QC

- Estimation theory permits combination of informations from previous knowledge (climatology and statistics) and the full dataset (neighbors): **Optimal estimation (BLUE)**
- *Underlying assumption: zero mean and gaussian distribution of variables*
- This method, can be used in delayed mode and in real time to allow for early detection of errors.
- To discriminate between the error sources, it is better applied on the basic variables (temperature and salinity on depth levels), work on elaborate variables (isotherms or isopycnals reference, dynamical height) should be introduced at a later stage, as for example
  - Wong Method
  - Comparison with altimetry

Analysis system: Method

Objective analysis
(Optimal Interpolation)

- yo: data vector (observations)
- xa: state vector (T or S at grid points)
- Pa: covariance matrix of analyzed field
- R: data error covariance matrix
- Co and Cao: covariance matrices of obs-obs and analyzed-obs
- Xf: first guess for x at analyzed points
- Yf: first guess at data point

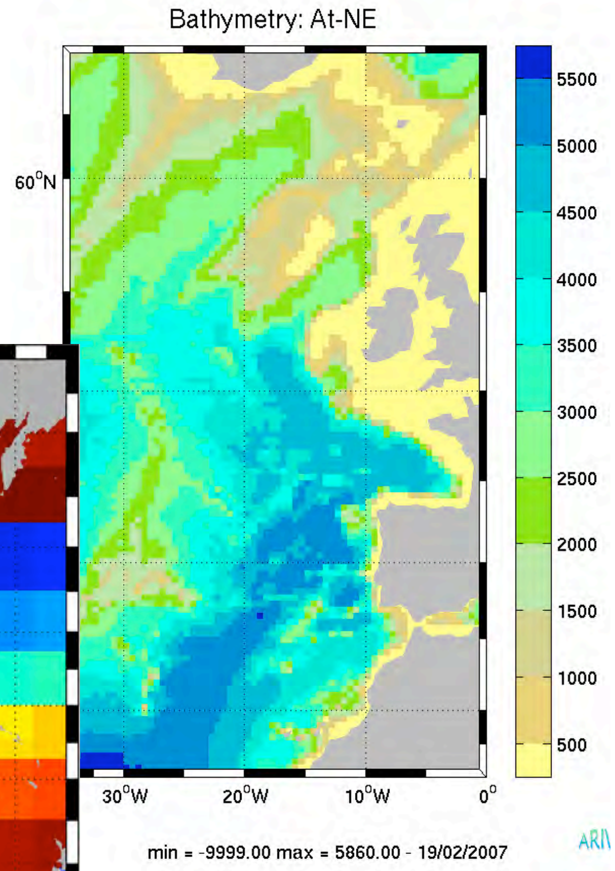
$$
\left\{
\begin{aligned}
d &= y^o - y^f \\
x^a &= x^f + C_{ao}\left(C_o + R\right)^{-1} d \\
P^a &= P^f - C_{ao}\left(C_o + R\right)^{-1} C_{ao}^T \\
y^o - y^{ao} &= R\left(C_o + R\right)^{-1} d
\end{aligned}
\right.
$$

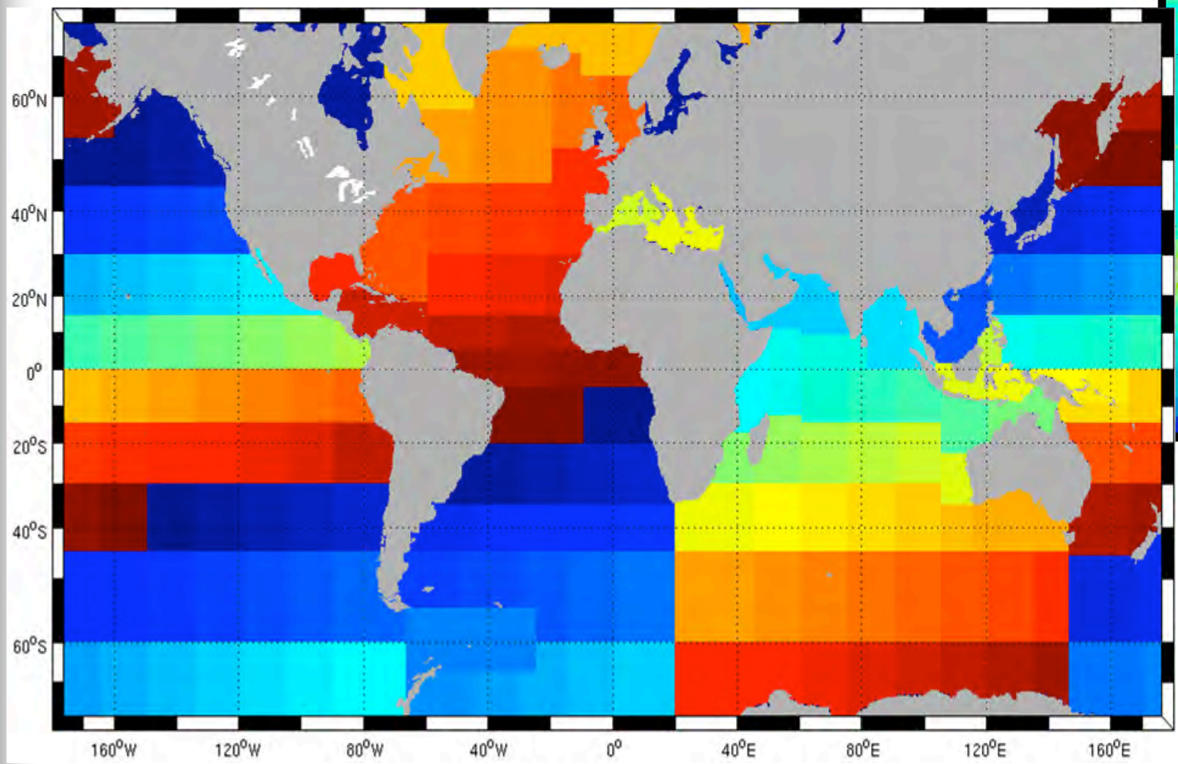
Equivalent observation/mapping matrix:

$$H^T = P^{f\,-1} C_{ao}$$

In Situ Analysis System : ISAS-V4 ARIVO

- 151 std-level: [[0 3] [5:5:100] [110:10:800] [820:20:2000]]
- 1/2° Mercator limits:





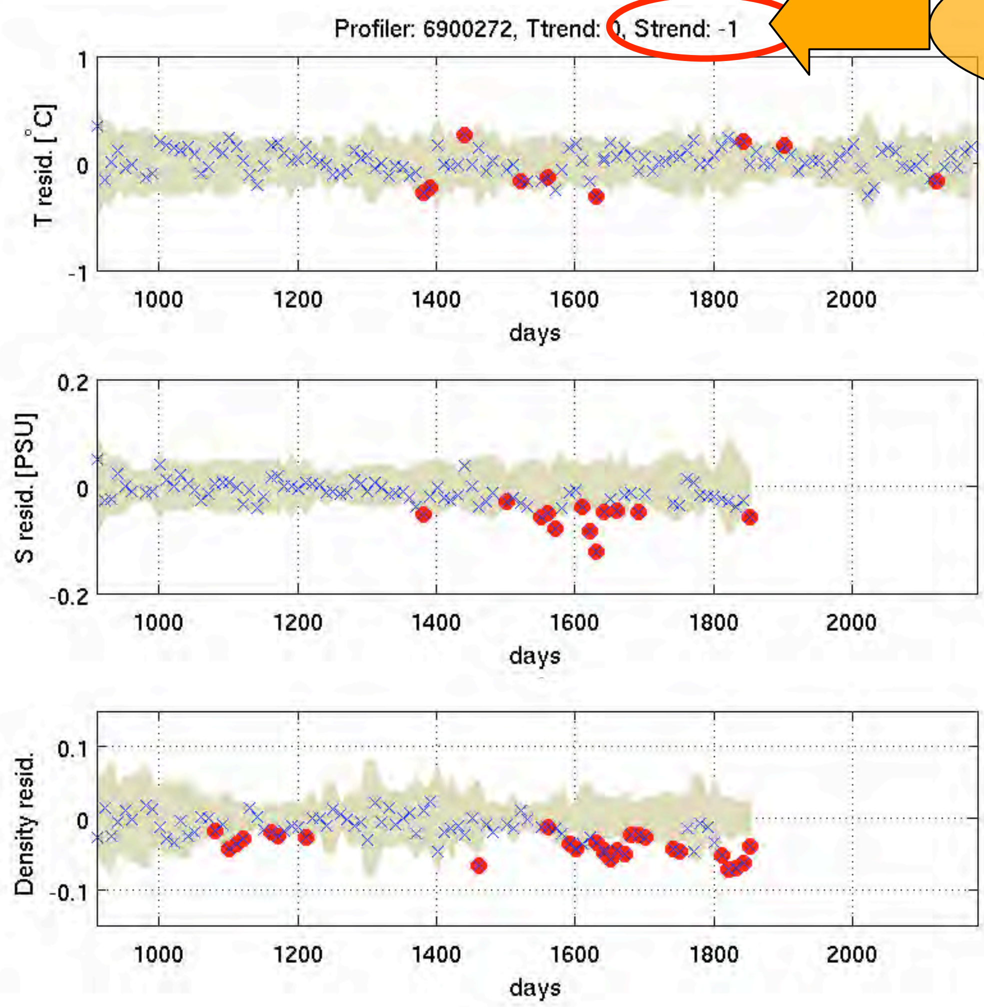
Data validation

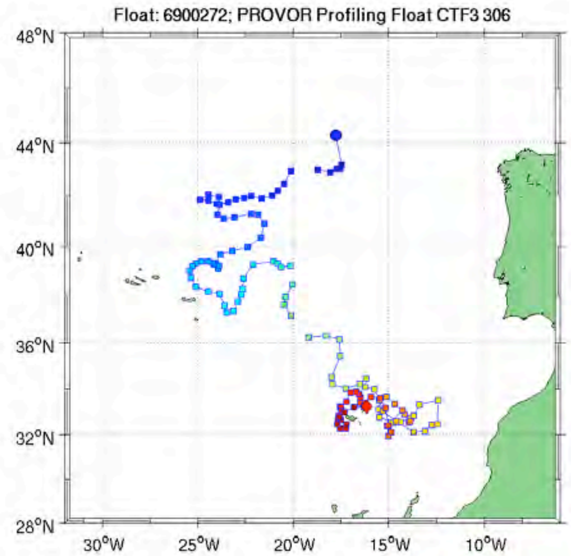
- Analysis of residuals:
  - $y_o - y_a = R\,(C_o + R)^{-1} d$

- Applied to ARGO floats:

Data validation: Profiler 6900272

Trend detected with reverse arrangement test.



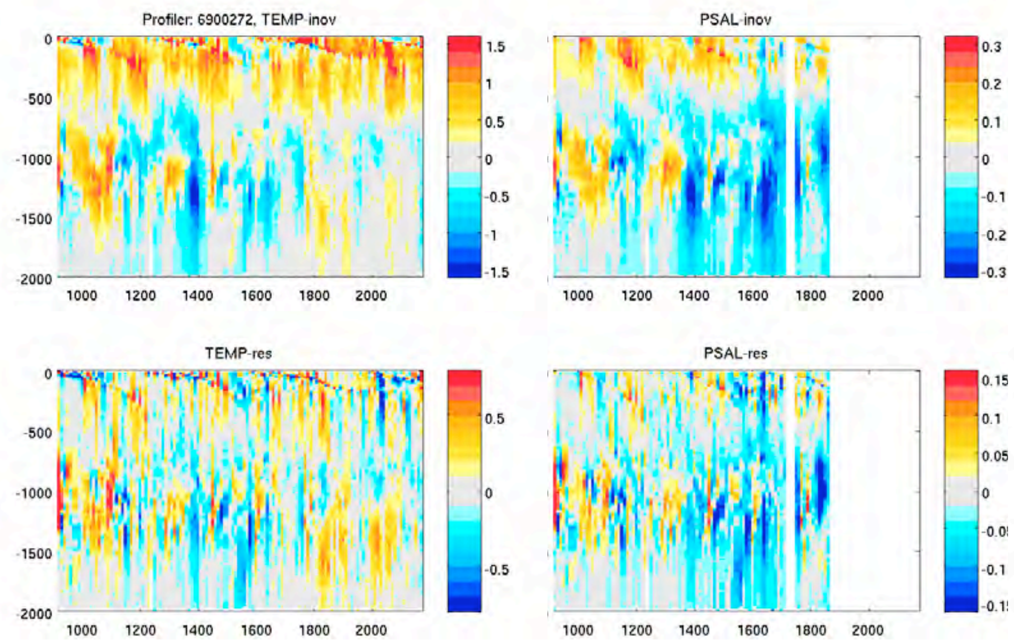


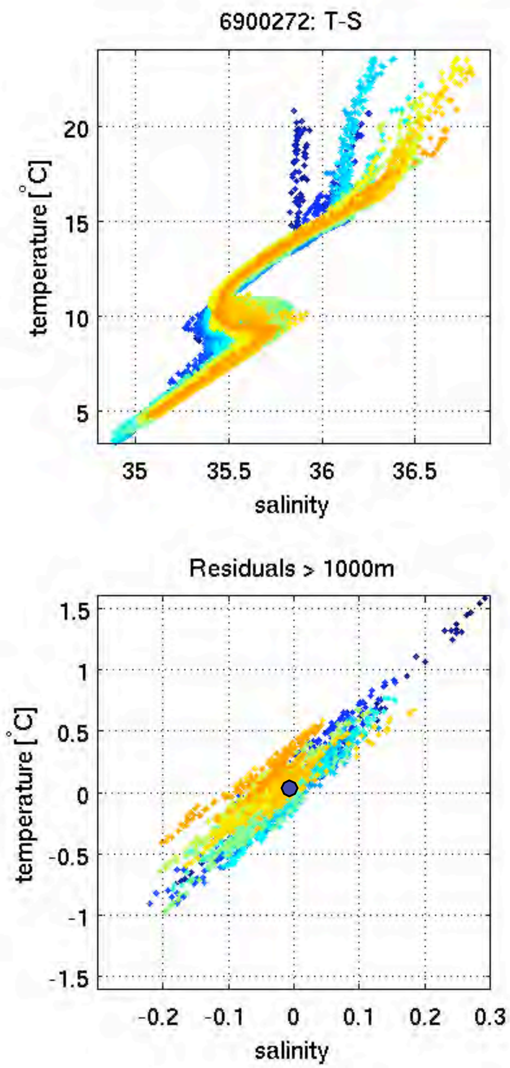
If trend = ± 1

- Screen the profiler series of average residuals (z> 400m) for confirmation
- see salinity and density anomaly,

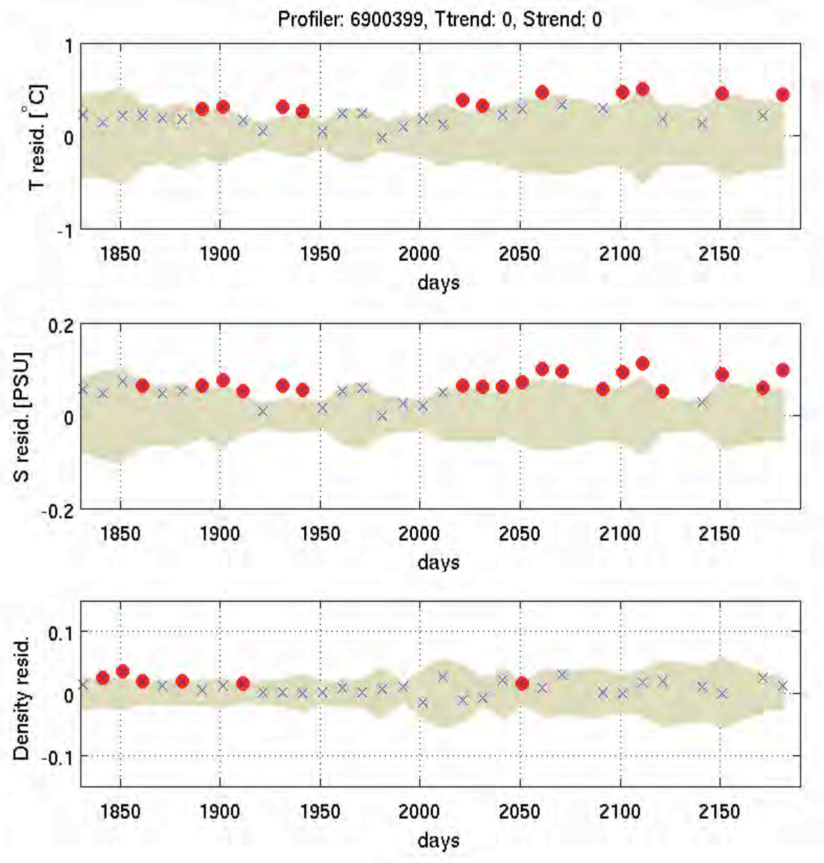
Data validation: salinity drift

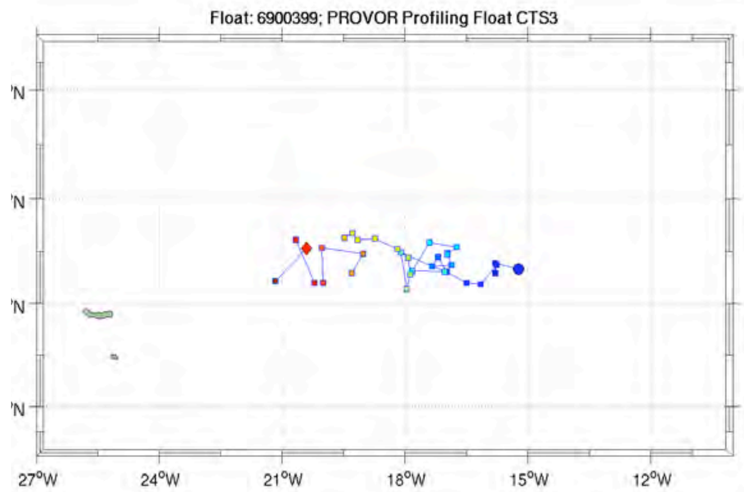
- T/S plots reveal the salinity drift





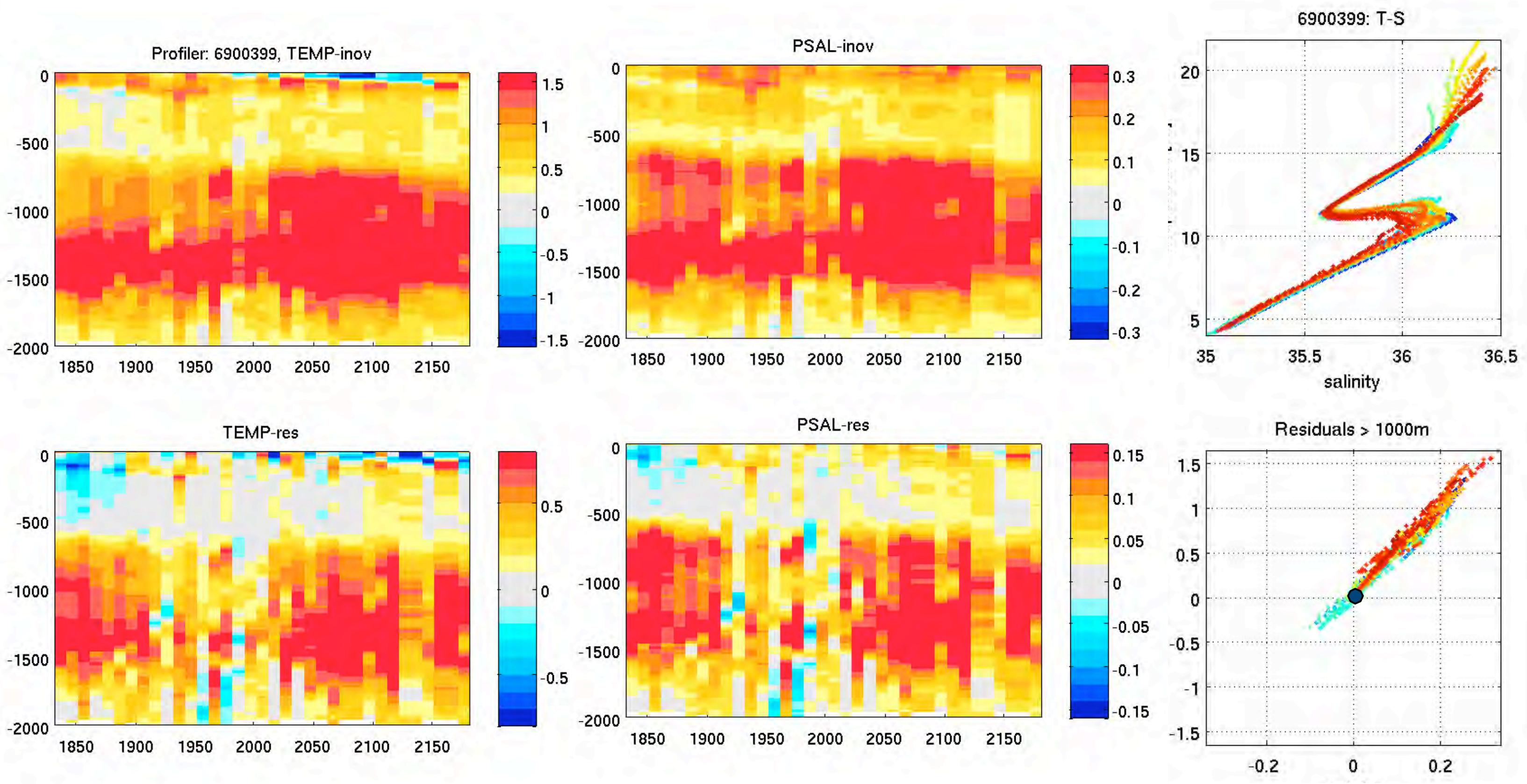
Data validation: PF6900299





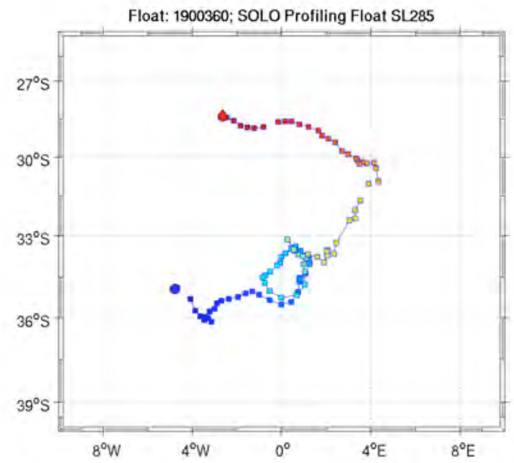
No trend detected with reverse arrangement test.

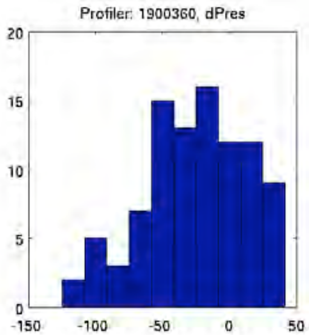
Profiler in a meddy

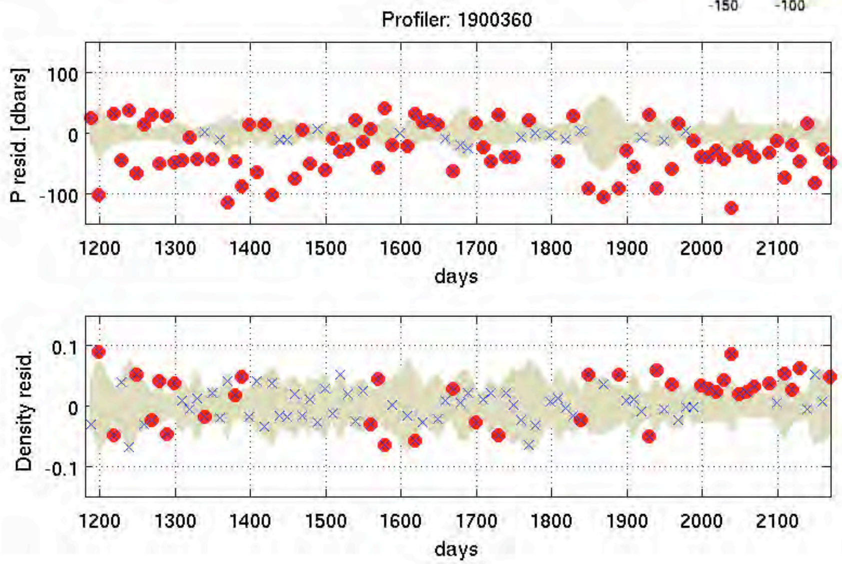


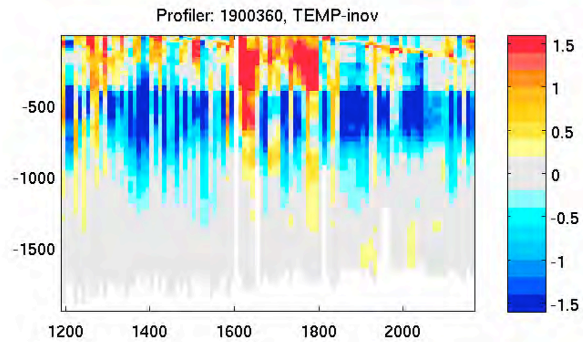
Data validation

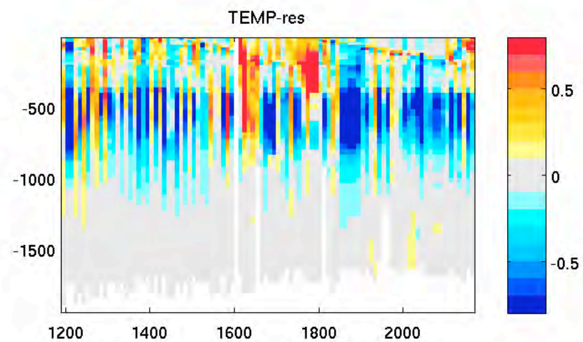
- Pressure sensor problem



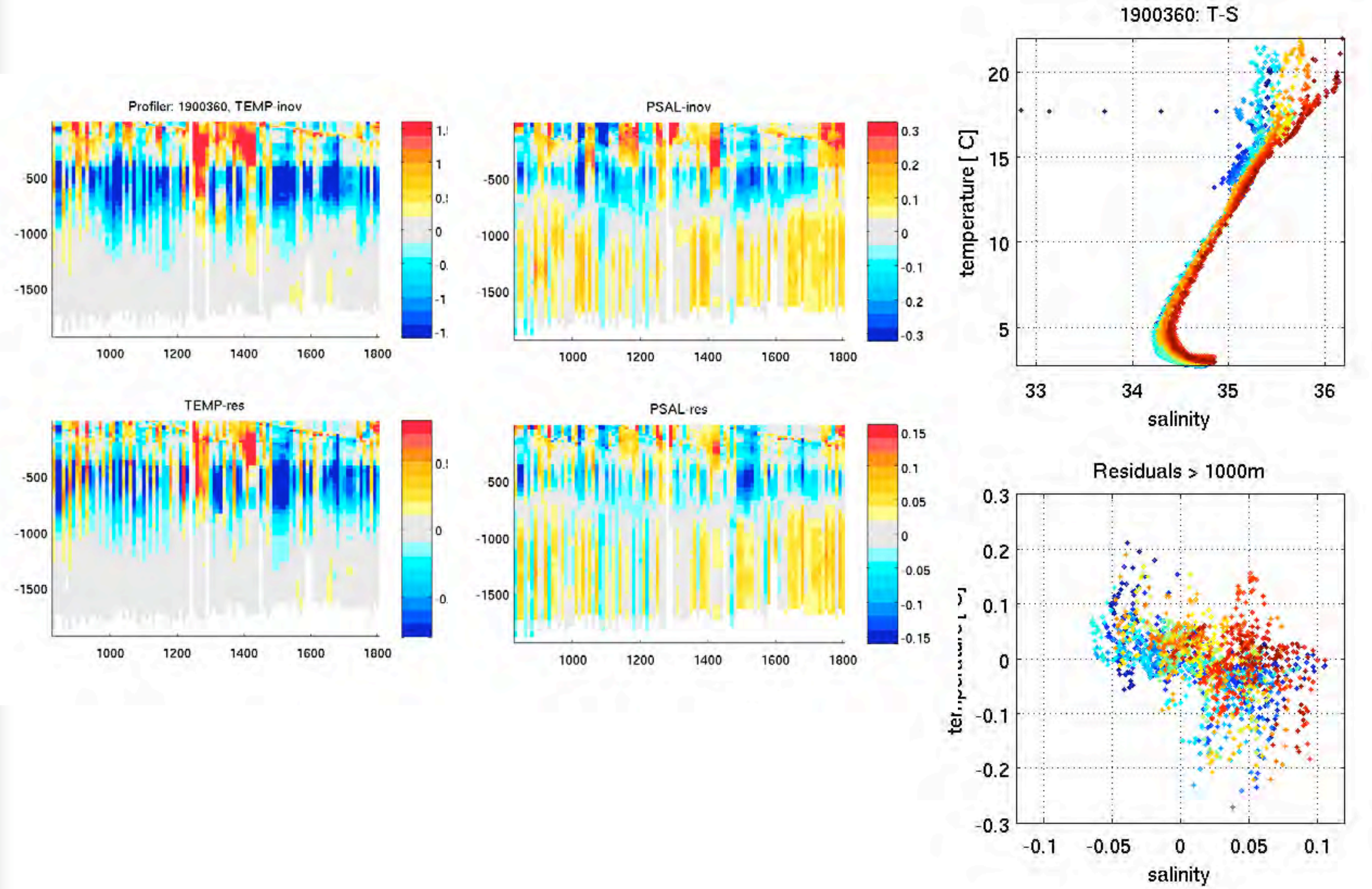








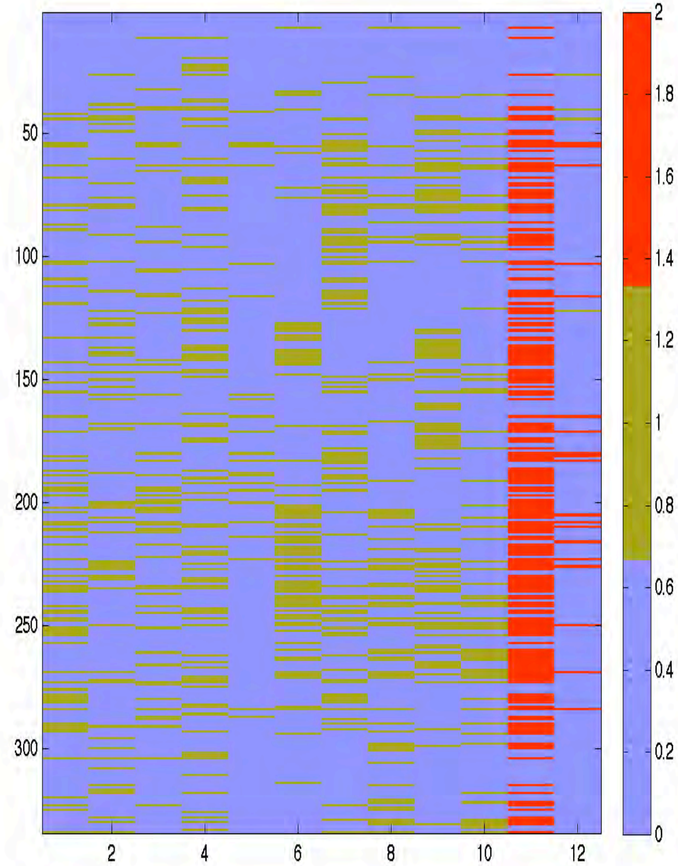
Data validation: Pressure problem



Converting this into numbers

- Define thresholds on various measures:
- 1 line = 1 profiler, 1 column = 1 test
  - Col 1, 2: T-offset, Tdrift
  - Col 3, 4: S Offset, Sdrift
  - Col: 5, 6: Density offset and STD
  - Col 7, 8, 9: Pressure offset, STD and Pout

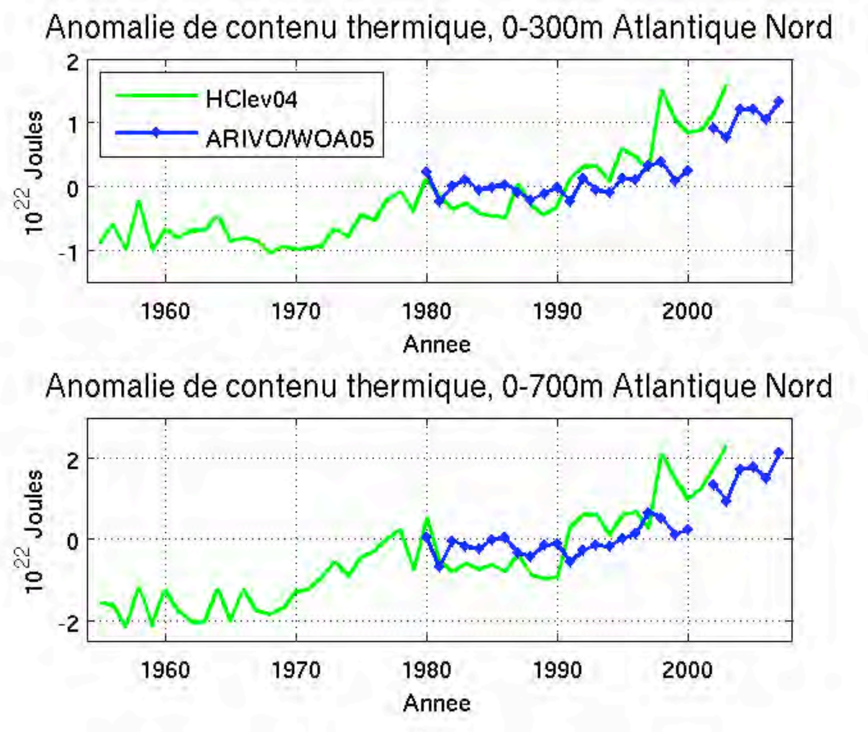
- Define combined indices:
  - Col 11: Pressure pb.
  - Col 12: opposite T-S drift or offset (expression of T sensor error)



Conclusion

- Obtaining a perfect dataset will need many iterations and requires to alternate simple automatic QC, global statistical coherence tests and single profiler screening
- A first pass on the 2002-2007 dataset already gives robust results

Heat content anomaly



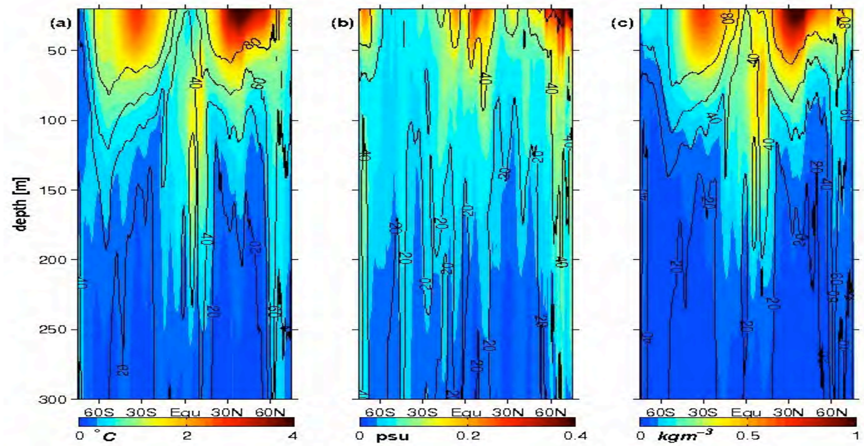
- Blue: anomaly of ARIVO analysis relative to WOA05
- Green:Levitus 2004 anomaly (fitted on ARIVO series over 1980-2000)

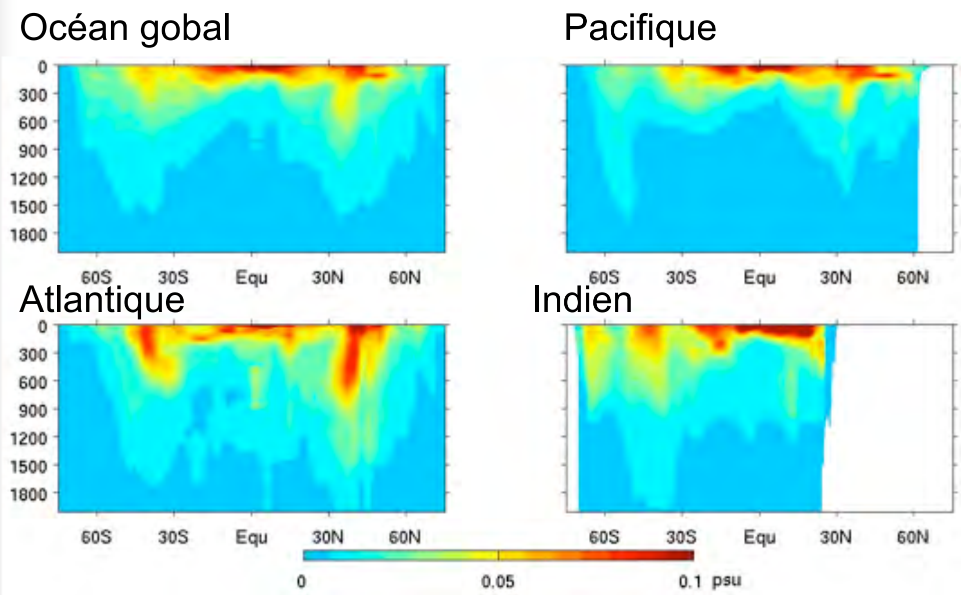
Global ocean variability (2003-2007)

(K. Von Schuckmann)

Seasonal cycle:

- The 12 month harmonic (extracted from the 5 years series) show penetration of the seasonal cycle deeper than 200m.





Intra-saisional to interannual variability:

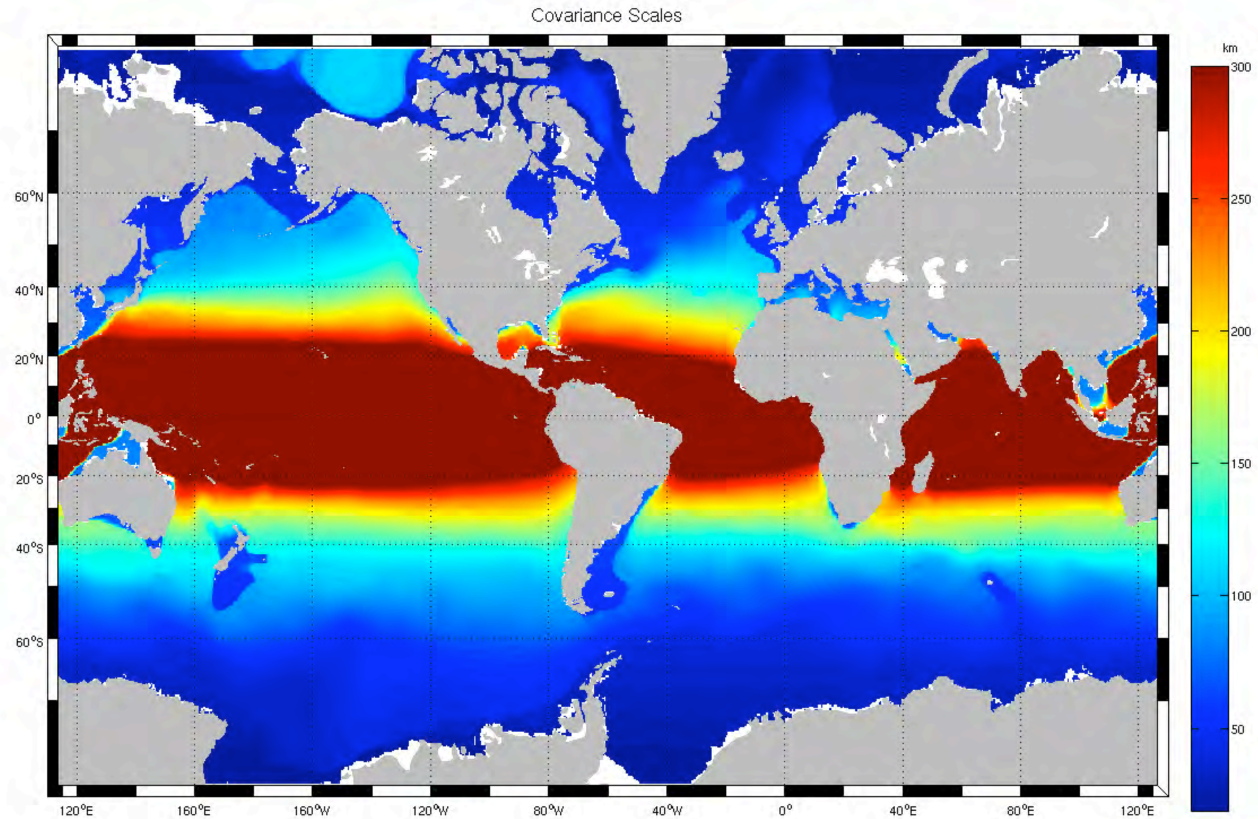
Zonal mean of the standard deviation show the strong deep variability of the North Atlantic (mid and high latitudes) and of the southern ocean (south of 30S)

Annex

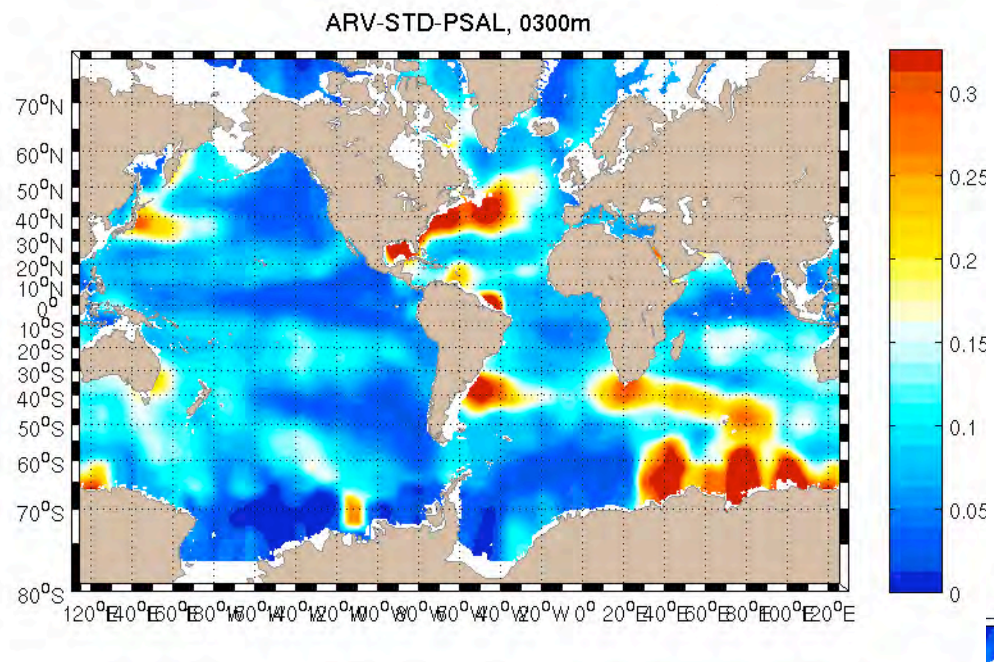
Covariances scales

$$C_o = C_o(dx_o, dy_o, dt_o) = \sum_i \sigma_i^2 e^{-\left[\frac{dx_o^2}{2L_x{}^{i2}} + \frac{dy_o^2}{2L_y{}^{i2}} + \frac{dt_o^2}{2L_t{}^{i2}}\right]},$$

i = 1; Lx=Ly=300km; Lt = 21 days
i = 2; Lx=Ly=2*resol<4*Rossby radius)<300km; Lt = 21 days



A priori variance



signal noise

$$Var_{tot} = var_{LS}+var_{ms} + var_{ur}+var_{me}$$

- $Var_{tot:\ deduced}$ from 2002-2006 database (S_{obs} - $S_{clim\text{-}month}$)
- Var_{me}: from instrument type
- Ratio var_{ms}/var_{ls} and var_{ur}/var_{ls} specified