Real-time Temperature and Salinity Quality Control based on minimum/maximum estimates from the known local variability

Jérôme Gourrion[1], *Caroline Paugam*[2], *Vincent Fachero*[3], *T. Szekely*[1], *D. Briand*[2], *C. Coatanoan*[2]

(1) OceanScope (2) Ifremer (3) Cap Gemini

Contact: jerome.gourrion@ocean-scope.com

Context

Coriolis is a French multi-institution facility to provide quality-controlled datasets for both scientific purposes and operational ocean state monitoring and prediction systems. In the Copernicus Marine Environment Monitoring Service context (CMEMS), Coriolis is coordinating the global In-Situ Thematic Assembly Center (INSTAC), feeding the assimilation system of the global Monitoring Facility Center (MFC) lead by Mercator-Océan with ***quality-controlled in-situ data***; here the attention is focused on temperature and salinity observations.

The Quality Control (QC) procedures include standard checks on station time, location and its metadata, as well as spikes or density inversion in the vertical profiles. Apart from such standards, specific procedures based on comparing data to the known local variability are implemented.

Unfortunately, such procedures, if efficient in detecting erroneous data, may 1) require a large computational time and 2) are responsible of a non-negligible amount of erroneous detections and are used combined with operator check. If very efficient in a delayed-time (DT) context, ***such procedures do not meet the short delay (hourly) constraint inherent to Near Real Time (NRT) context***. A significant amount of erroneous data are still not caught by the present QC system and are distributed while they should not be.

In this study***, a robust statistical procedure already validated for DT applications is adapted to meet the NRT distribution constraints***. Feedback from Mercator-Océan about examples of erroneous data going through the system (BlackList, BL) are used for evaluation purpose.

Why ?

Reduce the number of observations in BLs in the NRT processing chain

How ?

① Set up a test database to mimic the state of the operational one by the time of NRT diffusion
② Choose a test time period: → *Aug 1st to September 15th 2018*
③ Select and check all BLs over that period: *among 91, 5 rejected, 5 out-of −scope, 81 kept.*
④ Run the QC procedure for all data and P values from 0 to 10
⑤ Check visually all BLs
⑥ For each P value, get the number of a) detected BLs, b) all alarms and c) false ones.

Present test restrictions

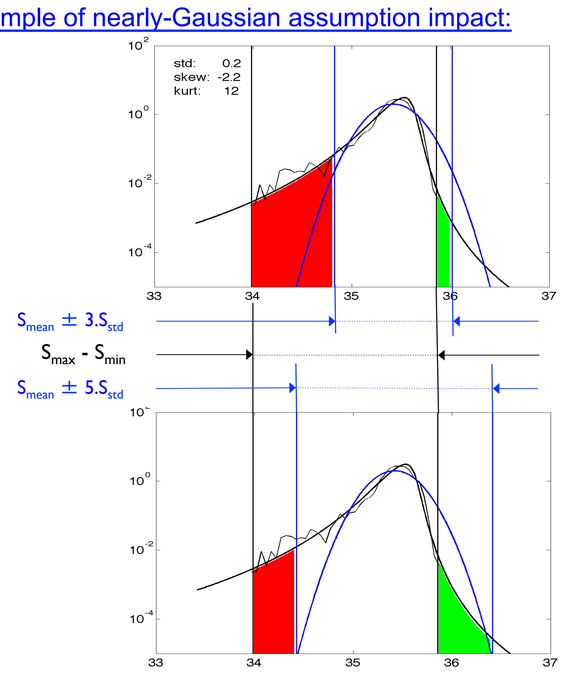
a. Only primary ARGO profiles
b. Present Min/Max validity domain : *bathymetry > 1800 m, Z < 2000 m*
c. Assumption on test BDD state: *RTQC results (not exactly NRT distribution state)*
d. Statistical simplification: *Median estimated from Mean*

Strategy

Use a method implemented and validated in Delayed-Time processing chains

Validity intervals are based on local Minimum/Maximum values inferred from historical datasets rather than from (Mean +/- N*Std) estimates

Example of nearly-Gaussian assumption impact:



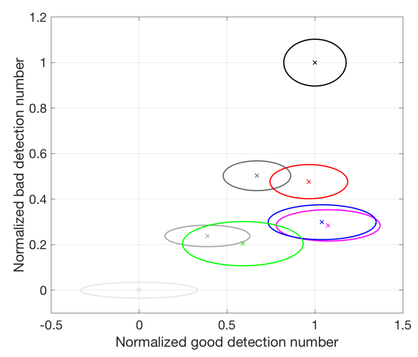
Gourrion et al., 2018: to be submitted to JAOT

True salinity distribution
Gaussian model - same variance
False alarm (FA) probability
Inefficient Detection (ID) probability

nearly-Gaussian case
N = 3 does a good job

Skewed distribution
N = 3 : FA ↗
N = 5 : FA ↘ but ID ↗

FA and ID optimization are contradictory !



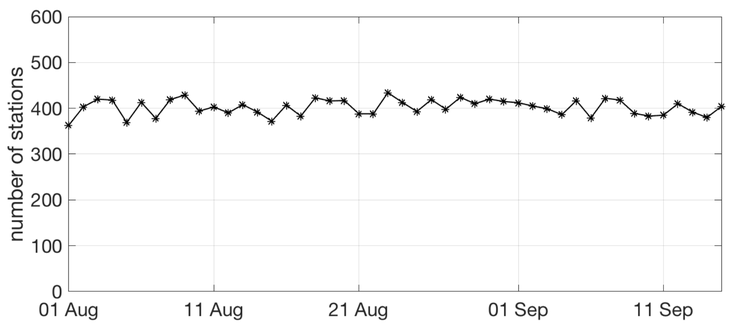
Severe improvement in comparison to Mean +/- N*Std intervals !

To meet NRT constraints,
Introduce an enlargement **P** factor :

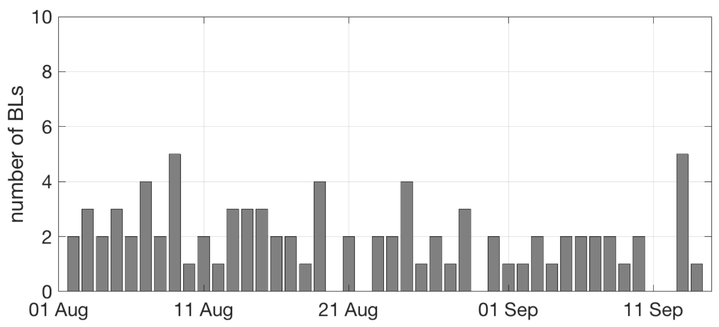
new Min = Median + (1 + **P**) . (Min − Median)
new Max = Median + (1 + **P**) . (Max − Median)

Results

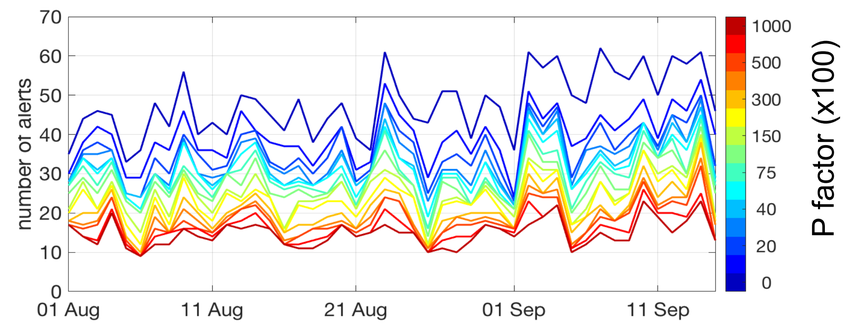
Number of stations available each day



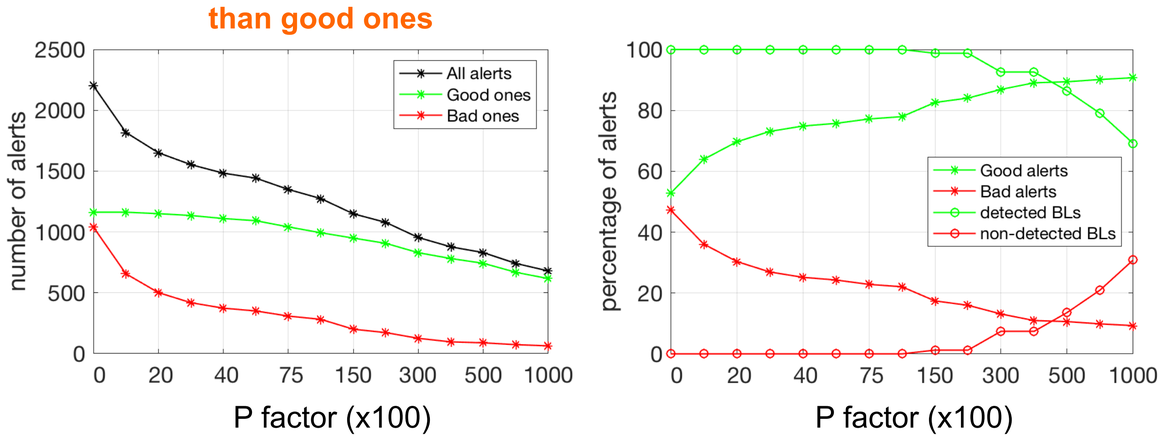
Number of observations in BL each day

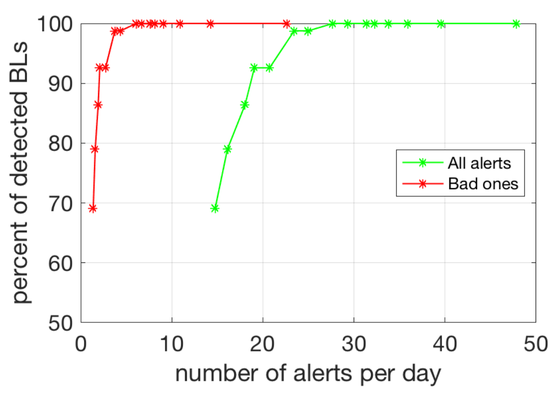


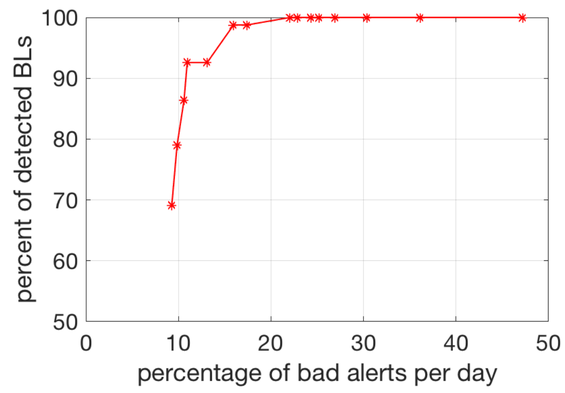
Decreasing number of alerts with increasing P factor



Bad detections decrease faster than good ones







Perspectives

- Improve test BDD state: *set all QC to actual NRT values using distribution backup*
- Revise operational NRT QC processing chain *some tests not run before NRT distribution*
- Improved median estimation
- Repeat study distinguishing depth layers *optimal P factor might change with depth or other*
- Tune a validity interval based on Mean/Std where Min/Max not available *bathy < 1800 m, Z > 2000 m*
- Extend Min/Max validity domain (shelf, regional seas)
- Extend to all non-ARGO profiles (potentially with different quality level)
- Extend to all observation types beyond vertical profiles

The P factor may be adjusted to NRT constraints as a trade-off between:

for hourly diffusion (fully automatic):
- **The percentage of observations in BLs detected by the system**
- **The number of false alerts (good data discarded)**

for daily diffusion (automatic + operator check):
- **The percentage of observations in BLs detected by the system**
- **A manageable number of alerts per day (visual inspection)**