Adapting MinMax QC approach to NRT constraints at CORIOLIS for CMEMS INSTAC products

Technical note - CORIOLIS MinMax NRT study - version 1

Jérôme Gourrion (OceanScope) and Delphine Dobler (SISMER/Ifremer)

February 19, 2019

This technical note is written by the above-cited authors to keep track of the work conducted at CORIOLIS centre in 2018 to improve CMEMS NRT products quality. The task was initiated in early 2018 by Claire Clénet during her contract at SISMER, with support from CORIOLIS R&D team (T.Szekely and J.Gourrion), and Vincent Fachero (Cap Gemini). Following, Caroline Paugam, J.Gourrion and V.Fachero took over the methodology development and implementation; they designed an evaluation study which results are presented in this document. Delphine Dobler helped a lot in finalizing the analysis of the study results and provided corrections to this document. Contributions from Yann-Etienne Prigent, Dominique Briand, Laure Fontaine, Christine Coatanoan, Anne Piron are kindly acknowledged, as well as the permanent support from CMEMS INSTAC coordinators, Sylvie Pouliquen and Loïc Petit de la Villéon. Compared to the version 0 of this technical note, here the state of the test database perfectly mimics that of the operational one at the earliest stage of the NRT data diffusion.

1 Context

The CMEMS Near-Real Time (NRT) global In-Situ data products are produced at CORIOLIS data center. Quality control (QC) for these products consist in 1) automatic procedures to detect and flag standard problems in the profile metadata, location, spikes, density inversion or regional ranges, operated in a very short term (hourly) after data collection, as well as 2) other automatic procedures complemented by operator visualization and decision (1-3 days delay). In the last years, Mercator-Océan and CLS, as main users of these products, have been providing feedback to highlight erroneous data going through the Quality Control. Since 2017, such feedback is operational through daily production of blacklist (BL) alerts managed by CLS (C.Boone). Unfortunately, such imperfections concern the hourly distributed products where procedures from 2) have not been applied yet. Within a few days, the overall quality clearly improves when such procedures are finally applied.

Seperately, since 2014, the R&D Coriolis team (J.Gourrion, T.Szekely) has developped and implemented in delayed-time (DT) processing a method consisting in comparing any observation to a local validity interval based on the minimum and maximum values ever observed as inferred from a reference historical database. The method has demonstrated its capacity to improve severely the automatic detection accuracy, allowing to save a significant amount of DT operator time spent in visualizing suspicious observations.

In this study, the DT MinMax approach is adapted in order to meet NRT quality and time constraints. BL statistics are used to evaluate the performance of such a new NRT procedure.

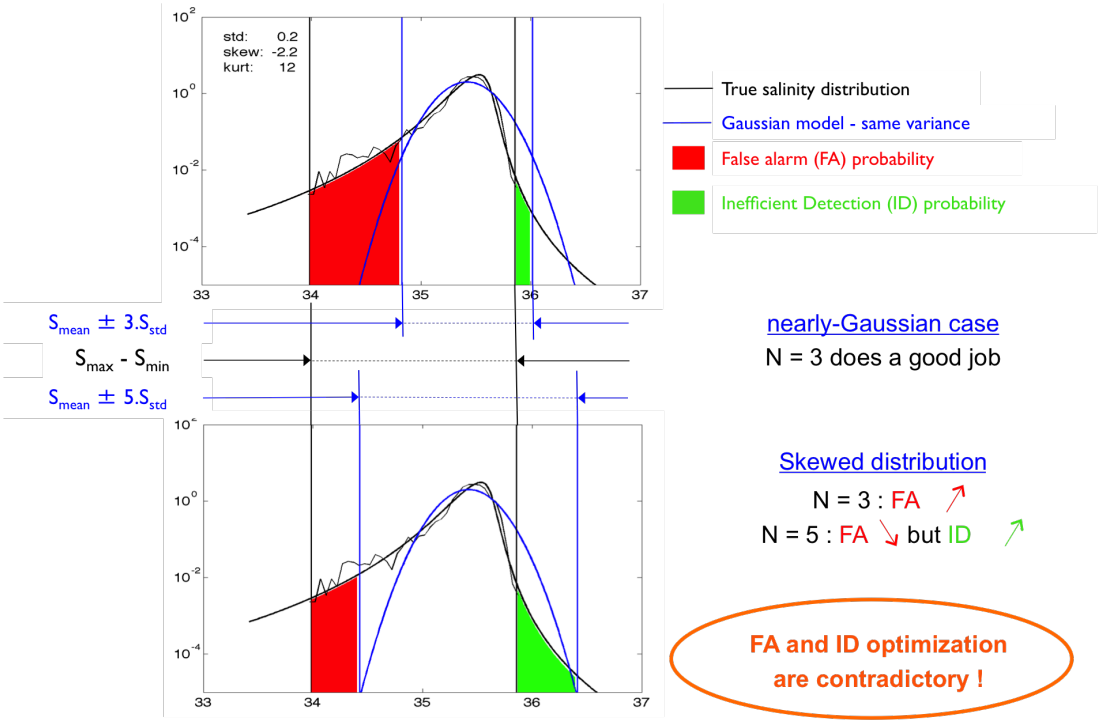

Figure 1: Scheme describing the impact of nearly-Gaussian assumptions on the quality control of a realistic salinity distribution. Thin black line: observed distribution and location of minimum and maximum values. Thick black line: skewed Student pdf model with same mean, variance, skewness and kurtosis. Blue line: Gaussian model with same mean and variance and location of the validity interval boundaries with N = 3 (upper panel) and N = 5 (lower panel).

2 Background - The MinMax approach

Comparison of in-situ observations to a validity interval based on the local knowledge of the geophysical variability is an old idea. Unfortunately, the interval bounds are usually inferred from first and second order statistical moments (mean and standard deviation), under strong underlying assumptions on the shape of the probability density function of the corresponding physical variable. The validity interval is centered on the mean value and half-wide as N*std, where N is an adjustable parameter. Such assumptions degrade the method accuracy, with a large number of false alerts and a significant amount of undetected erroneous data. This is graphically described in Figure 2. Consequently, such an approach is not activated operationnally at CORIOLIS for NRT production.

The MinMax approach simply proposes to modify the estimation method of the interval bounds i.e. the minimum and maximum values. Such values are derived from a reference dataset built with a specific attention on the quality control of its extreme values. The method validation has demonstrated its ability to reach an improved ratio of good to bad detections, see Figure 2. The method has ben used successfully for the last years in the production of the delayed-time INSTAC dataset, allowing an important improvement in the station selection to bbe visually controlled by the delayed-time operator.

Nevertheless, for direct use in NRT production, the total number of alerts per day is too large to be treated by the NRT operator. This is understood as a persistent lack of statistical robustness of the minimum and maximum estimates: if the description of the actual distribution tails is improved using such estimates, it is still not perfect; the reference dataset does not describe the entire ocean variability and false detections are still present. As a solution, the present study offers to artificially extend the validity interval by a factor P as follows:

$$newMin = Median + (1 + P) \cdot (Min - Median) \quad (1)$$

$$newMax = Median + (1 + P) \cdot (Max - Median) \quad (2)$$

Sensitivity of the method performance to varying P values is evaluated in terms of 1) its ability to catch the erroneous data listed in BLs, 2) the total number of alerts per day (proxy of the operator time spent in visualization) and 3) the number of false alerts (good data erroneously discarded, important for fully automatic procedure).

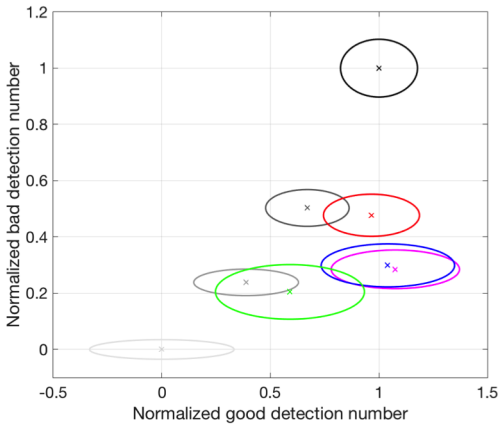

Figure 2: Efficiency diagram for the "classical" (mean/std) and MinMax approaches in terms of normalized relative variations of "good" (horizontally) and "bad" (vertically) detection statistics. Grey lines refer to the classical approach for N = 4, 4.5, 5, 6 with increasing line darkness; color curves refer to the Min/Max approach for different ocean depth layers: 0-200 m (red), 200-500 m (pink), 500-1000 m (blue) and 1000-2000 m (green). "Good" detection reference values are defined as the classical approach time series for N = 4 filtered with an 11 points rectangular window. "Bad" detection reference value are defined as the temporal average of the classical approach time series for N = 6. The presented statistics are first computed as the difference to the above defined reference levels and, second, normalized so that the classical approach results for N = 6 and N = 4 have respective coordinates (0,0) and (1,1) in the diagram frame.

An optimal P value can then be chosen to best meet the NRT constraints.

3 Study design and setup

3.1 Analysis strategy

In this study, we are interested in testing the impact of a new QC procedure on the data quality at the earliest production and diffusion stages. It is proposed to

1. simulate an adequate state of the database over a specific time period: July to September 2018.
2. load the operational blacklist information from CLS in the station history table.
3. run the new detection procedure for different P values. Store the alert results in the station history table.
4. define some truth about the quality of all observations in alert in order to characterize them as "good" or "bad" detections. Visualization with SCOOP software.
5. over the chosen time period, obtain statistics on the considered stations, the alerts for each P value, their good/bad alert status, the list of stations in BL.

3.2 Dedicated tools

In order to make the study results reproducible, it is necessary to work independently of the operational CORIOLIS database (DB). A specific test database must be setup, mimicking the earliest distribution state for which the blacklist feedback from CLS is available.

Similarly, all softwares available from SISMER to raise MinMax alerts (Java code) or visualize (SCOOP) have been designed for operational use. They often do not provide the flexibility required for development tasks. The design of the study must be adapted to the associated constraints; in cases of strong constraints limiting the scope of the study, specific development activities might be planned for the future.

3.3 Simplifications

1. In a preliminary step, the detection performance is only evaluated over the subset of all *primary* ARGO profiles (CORIOLIS PR_PF). It will allow to setup and refine the evaluation protocole. In a second step, it will be extended to other profiles (all PR_) and later to all observations (CORIOLIS time series, TS_).

2. The study is restricted to observations in areas where the minimum/maximum reference fields are available, and expected to have sufficient statistical robustness. Observations at depth below 2000 m and those in grid cells lying over the continental slope or shelf (roughly determined from the bathymetry as the ones with depth at cell center being shallower than 1800 m) are not considered.

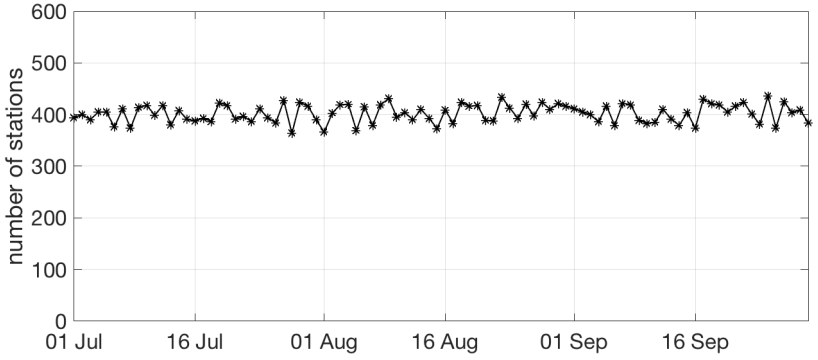

Figure 3: Number of stations per day analyzed in the present study over the study time period.

3. CORIOLIS does not maintain an archive of all the files distributed in NRT. The files may be distributed again at any time (either if additional QC procedure is run, typically with 1-3 days delay, or if the data provider updates the data) and only the latest state is archived at CORIOLIS. Both QC flags AND physical values may be modified in the days and weeks following their first distribution. Checking the quality at the earliest distribution stage for development purposes is not something feasible in the CORIOLIS system alone. Two solutions are proposed:

   (a) Get from CLS opportunity archive the earliest version of the distributed files. Not done in this preliminary study because of a missing software to load the files content into the database. *Caution: this solution creates a dependence on CLS archive whereas they are absolutely not committed to maintain it.*

   (b) Simulate the earliest distribution stage, starting from a latter state (setup the DB state for July to September 2018 from its state in November). This is what is done in the present preliminary study. From the physical values, automatic Real-Time Quality Control procedures are run again. This can be satisfactory, as long as only QC flags have been modified since the earliest distribution stage. *Fails if physical values have been modified in the meanwhile as RTQC will not run on the adequate data.* See Section 3.4 for 12 such cases.

4. In Eq. 2, the validity interval is extended around the median. Presently, the median from the reference dataset used to build Min/Max estimates is not available (because impossible to estimate iteratively). In this preliminary study, the median is approximated by the mean. *In a near future, sample median should be used instead.*

3.4 Blacklist selection

To build their blacklisting feedback, CLS runs different automatic tests operationnally and provides the information to identify the detected observations.

- a letter code is provided to trace which automatic test has raised a detection. Only observations with alerts from tests 'N' (local comparison to climatology based on mean/std based validity intervals) and 'O' (regional bounds) are considered in this study.
- as shown in section 2, the comparison to climatology may be imperfect. All stations in blacklists are visualized with SCOOP software and erroneous detections are discarded.

Over the period July to September 2018 and focusing on *primary* ARGO profiles, 337 stations (among 103 platforms) are present in CLS blacklists.

Keeping only alerts from tests 'N' or 'O' reduces to 305 stations (among 88 platforms).

Removing multiple detections for the same station reduces to 224 stations (still among 88 platforms).

Removing detections (see Section 3.3) only below 2000 m reduces to 213 stations (among 87 platforms).

Finally, removing stations on bathymetry criterion (see Section 3.3) reduces to 200 stations of interest (among 81 platforms).

Figure 3.4 displays the daily amount of stations of interest in BL over the concerned period.

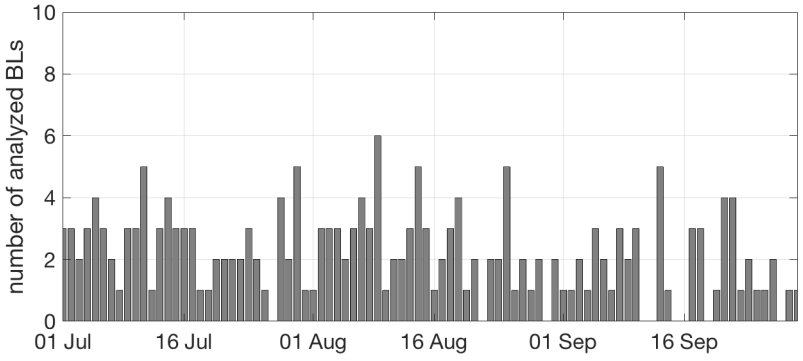

Figure 4: Number of stations per day present in the Blacklists from CLS over the study time period.

Contrarily to the work done in the version 0 of this technical note, here, all these 200 stations are analyzed.

After visually inspecting them, 6 of these stations (from 3 platforms) were judged to be actually good data, i.e. that the CLS detection, based on the comparison to climatology (letter code 'N') was erroneous. The list of corresponding station IDs follows:

corresponding CORIOLIS Station IDs: *60005464,60093228,60197838,60583183,60213877,61193747*

Thanks to setting the database state to the actual state of the operational database at the earliest diffusion time, the inconsistency about 21 stations observed in the version 0 of this technical note has now disappeared.

This preliminary study will use the 194 remaining stations that could be confirmed as erroneous data.

4 Results

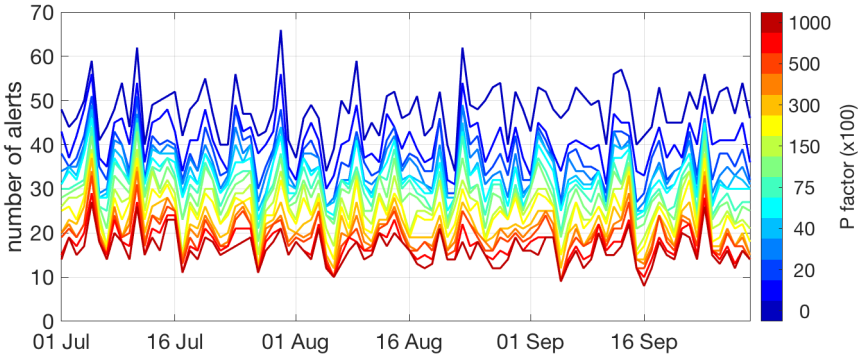

Figure 5: Number of alerts per day over the study time period. Color code indicates the corresponding P widening factor.

Figure 4 shows the number of alerts per day for a wide range of the widening factor P. Without widening (i.e. $P = 0$), the daily number of stations in alert lies between 40 and 50 (while the total number of alerts reaches 4176). Over the whole range of P values, this number is reduced approximately by a factor 3.

Figure 4 shows the average number of alerts per day as a function of P separately for the three month periods i.e. July, August and September.

As expected, the total number of alerts monotonically decreases with increasing P values. More interestingly, the number of "good" alerts decreases very slowly with P, while the number of "bad" alerts does decrease severely since the first widening stages. This clearly illustrates the interest of the Min/Max interval widening.

The quality of the detection, from the point of view of "good" to "bad" alerts ratio, appears to be significantly variable from one month to the other. The reason for this is investigated later on.

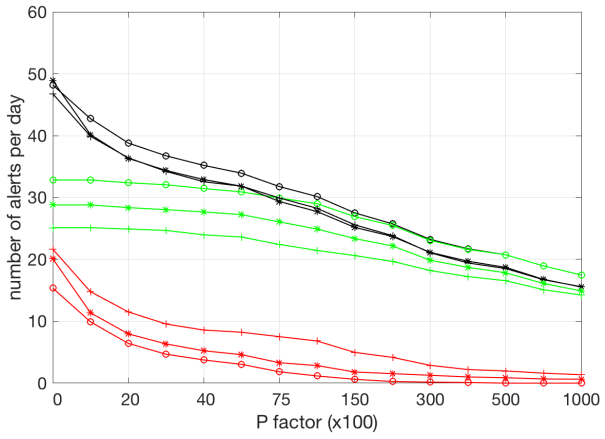

Figure 6: Number of alerts per day over the study time period as a function of the widening factor P. Black: all alerts. Green: "good" ones. Red: "bad" ones. Symbols correspond to different time periods: July 2018 (circles), August (pluses), September (stars).

Figure 7 provides the required information to make a decision on the optimal P value to be used in a specific context. On the left, it is convenient when the capacity to detect erroneous data both in general and specifically those listed in BL are of interest. Clearly, all stations in BL are detected up to doubling the original width of the Min/Max interval (i.e. $P = 1$ or $P * 100 = 100$). On the right panel, the diagram is built to help the decision when the statistics of stations in BL are not considered i.e. when the detection capacity needs to be evaluated in the general context of all alerts and not limited to that subset identified in the CLS blacklists.

Based on these diagrams, for the fully-automatic approach to be implemented in CMEMS NRT production and distribution chain, a P value of 0.4 has been chosen for operational testing. It is a good trade-off between a reduced number of false alerts and still a high number of good alerts. Let's remind that, in this context, "good" alerts characterize erroneous data that are discarded from the distributed data flow, while "bad" alerts characterize good data that are erroneously discarded.

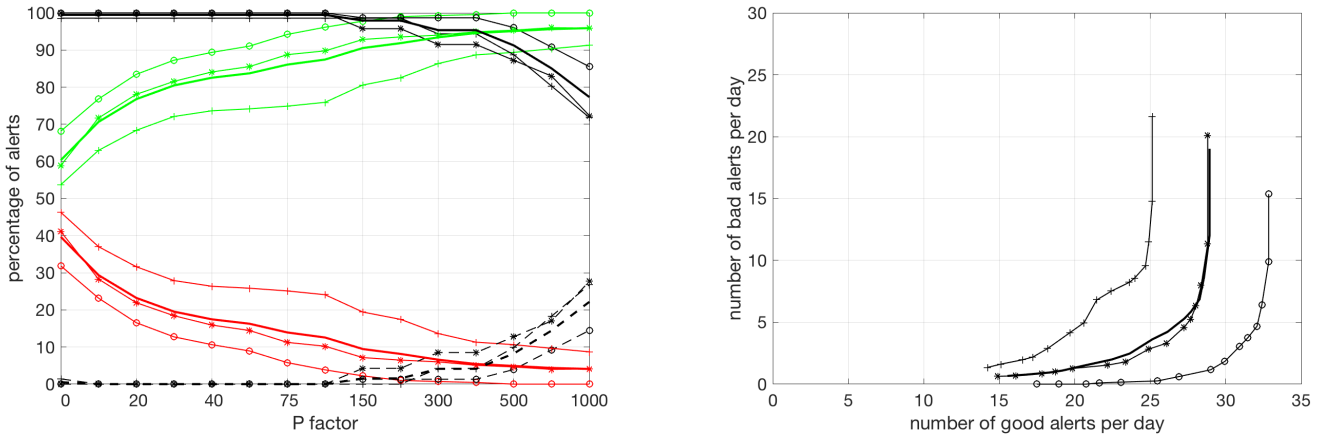

Figure 7: Left: percentage of "good" and "bad" alerts as a function of the widening factor P. Color codes are the same as for Figure 4 except that black color now stands for the percentage of detected stations in BL. Thicker lines without symbol correspond to the July to September statistics. Right: number of "bad" alerts as a function of the number of "good" ones. Each point corresponds to a different value of the widening factor P:

Now, we investigate further the origin of the statistics variability from one month to the other. There are two possible reasons: natural variability of the data quality at monthly scales or inaccuracy at some step of the evaluation procedure.

The accuracy of our quality "truth" obtained after visualization of the detected stations was questioned. Indeed, if this step may seem rational and reproducible, the repetivity of the operator decision for a set of similar anomalies is not a simple subject. Among different operators or even for a single operator, the decision made for similar anomalies may change significantly.

The impact of such an observation being not clear, a strategy to scrutinize the decisions was setup.

For all platforms having more than one alert over the study period, the corresponding stations are visualized again and the decisions checked. It has appeared that in a large number of cases, the same situation does not always lead to the same decision, even by the same operator. This check procedure has lead to 482 modifications of the operator decision (among 4411 alerts i.e. 10.9 %). Updated results accounting for such a consistency check are now presented in Figure 8.

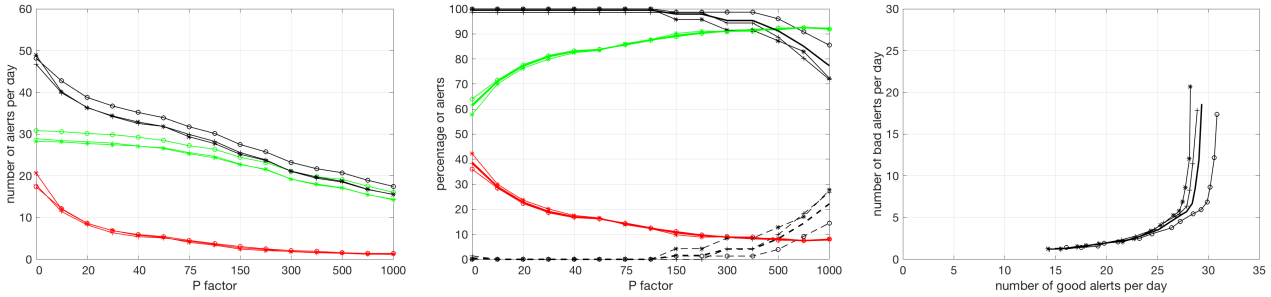

Figure 8: Same as Figure 4 and 7 for the updated operator decision.

Clearly, the statistics variability from one month to the other is highly reduced, especially for the highest P values. The operator decision is a noisy variable that requires post-processing steps to be homogenized.

Figure 9 displays the locations of the 1491 stations still with alert for $P = 1000\%$. Most of them have been confirmed as "good" alerts. Among them, 119 (i.e. 8%) have been identified as "bad" alerts, corresponding to erroneous detections, i.e. wrong minimum/maximum estimates. It is interesting to note that all these 83 stations are located in marginal seas (South China, Caribbean Seas, Gulf of Mexico). They correspond to specific locations where the bathymetry locally exceeds 1800 m while the area is mainly shallower. When refining the validity domain of the MinMax approach, such erroneous detections should disappear. Specific work on improving Min/Max fields in regional seas might also improve the detection quality in such cases.

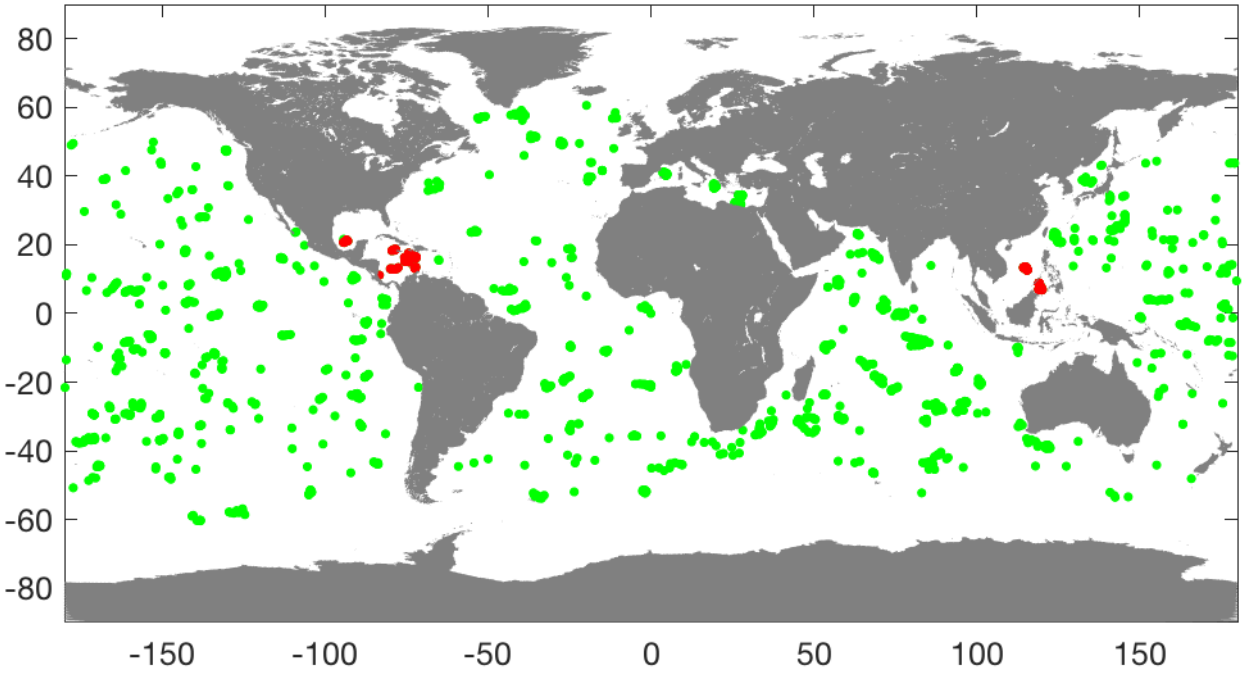

Figure 9: Locations of Min/Max alerts for $P = 1000\%$. Green (red) color indicates "good" ("bad") alerts.

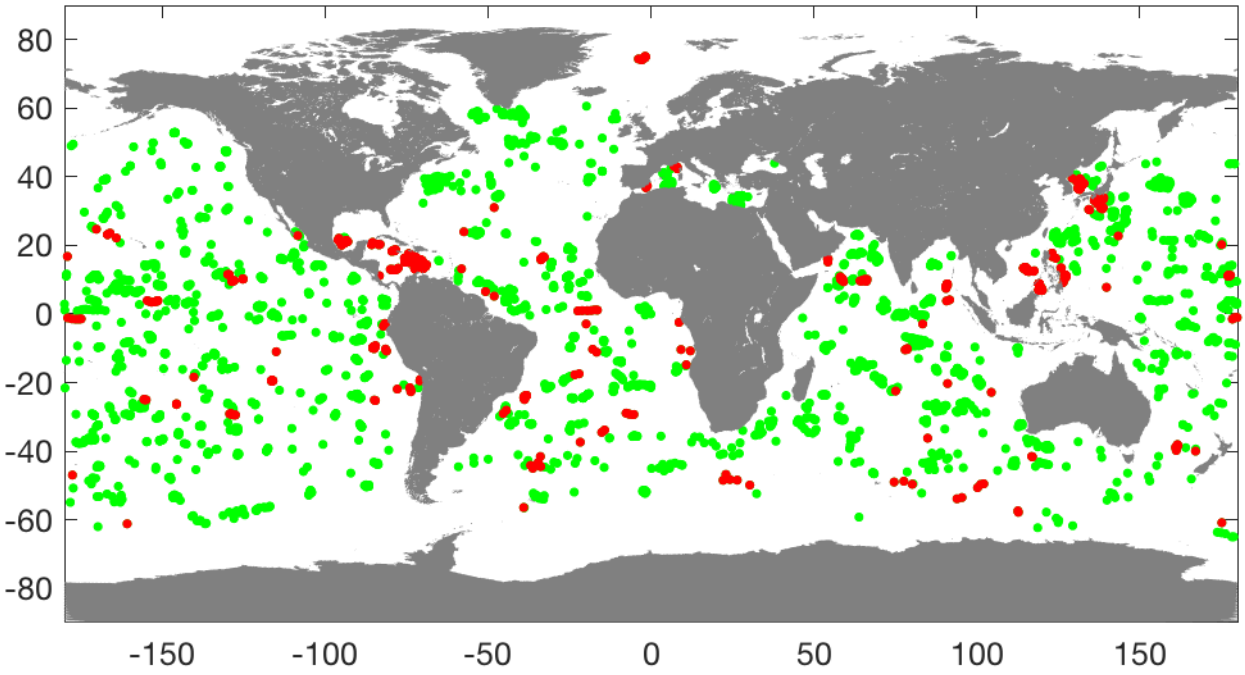

Figure 10: Locations of Min/Max alerts for $P = 40\%$. Green (red) color indicates "good" ("bad") alerts.

For the sake of completeness, we also show the location map for the 3087 stations with alert for $P = 40\%$, the value presently used at CORIOLIS for operational tests, see Figure 11. Among them, 525 (i.e. 17 %) are identified as "bad" alerts. Future steps of the study should focus on these erroneous detections and improve the method and auxiliary minimum/maximum fields in order to increase the method efficiency, and the amount of good data potentially blocked at the earliest diffusion stage.

Finally, we show a map of all stations in BL. Red dots indicate those stations considered, in this study, to be erroneous BL detections.

3 of them (all from platform 3901873, near 65^o N in the Atlantic, see Figure 12, top panels) correspond to conditions of extremely deep mixed layers that are probably not well accounted for in the climatology used by CLS for BL detections. The MinMax test identically fails for those cases.

For 3 other ones, the visualization led to infirm the BL detections. The conclusion is obvious for platforms 3901834 and 4901489 (middle top and middle bottom panels respectively). For platform 3901485, a low salinity surface value is not detected by the MinMax approach; indeed, even lower values are present in the MinMax reference dataset, but this particular case will be checked separately.

Among the 200 stations in BL, 194 are detected, 5 are clearly good data and 1 needs further check before confirmation.

5 Perspectives

In this section, we list the tasks to be conducted in the next steps of the study. Order is not priority.

- work on improved median estimator
- implement the possibility to build similar statistics for a set of ocean depth layers (e.g. 0-200 m, 200-500 m, 500-1000 m, 1000-2000 m). *The P factor might be optimized differently for each layer.*
- set up a validity interval based on climatological mean/std for observations acquired below 2000 m. *Important to minimize the discontinuity near 2000 m.*
- design a similar approach to the one proposed above for stations in grid cells with bathymetry shallower than 1800 m. *Similarly, important to minimize the discontinuity near isobath 1800 m. And to refine the 1800 m bathymetry criterion as the boundary of Min/Max validity.*

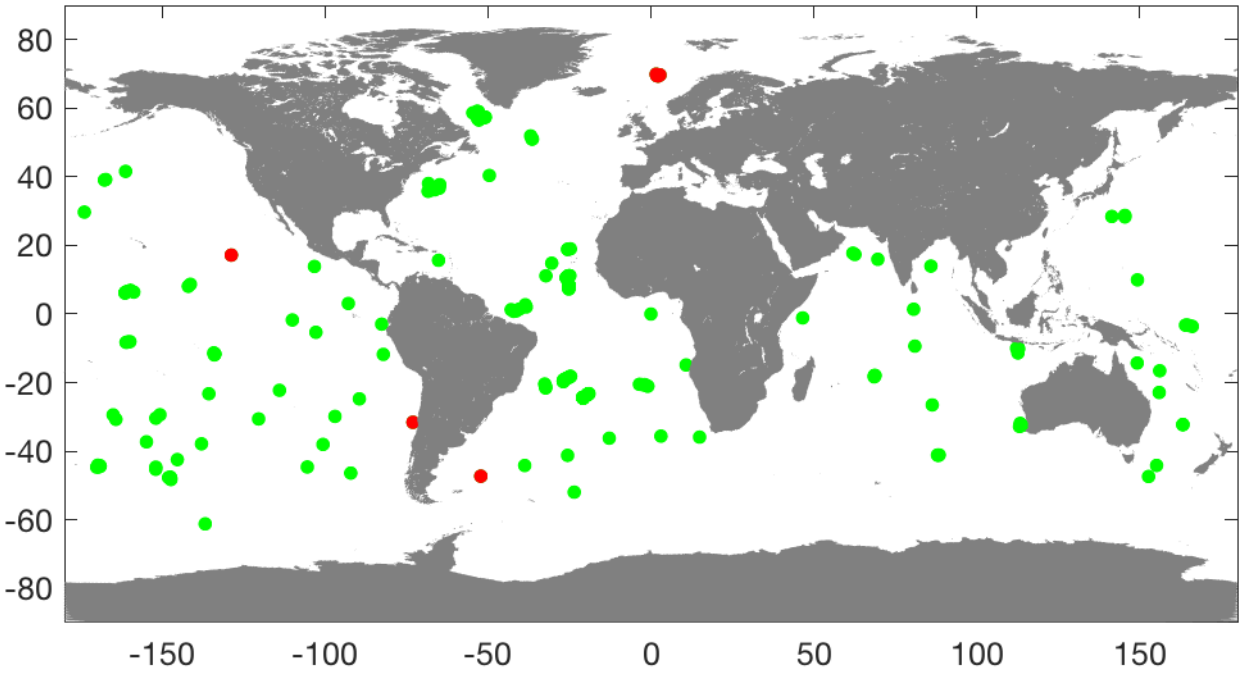

Figure 11: Locations of ARGO stations in CLS blacklists. Green color indicates "good" ("bad") alerts.

- further refine the procedure to homogenize the operator decision on the quality of the stations with alert. *Probably not so easy when adressing the case of non-ARGO stations.*
- run the analysis procedure on all PR_:
  1. run MinMax Java script
  2. run Blacklist insertion script
  3. SCOOP visualization
  4. operator decision check procedure
  5. alert information extraction from DB
  6. alert statistics computation
  7. diagnostic and decision graphs production

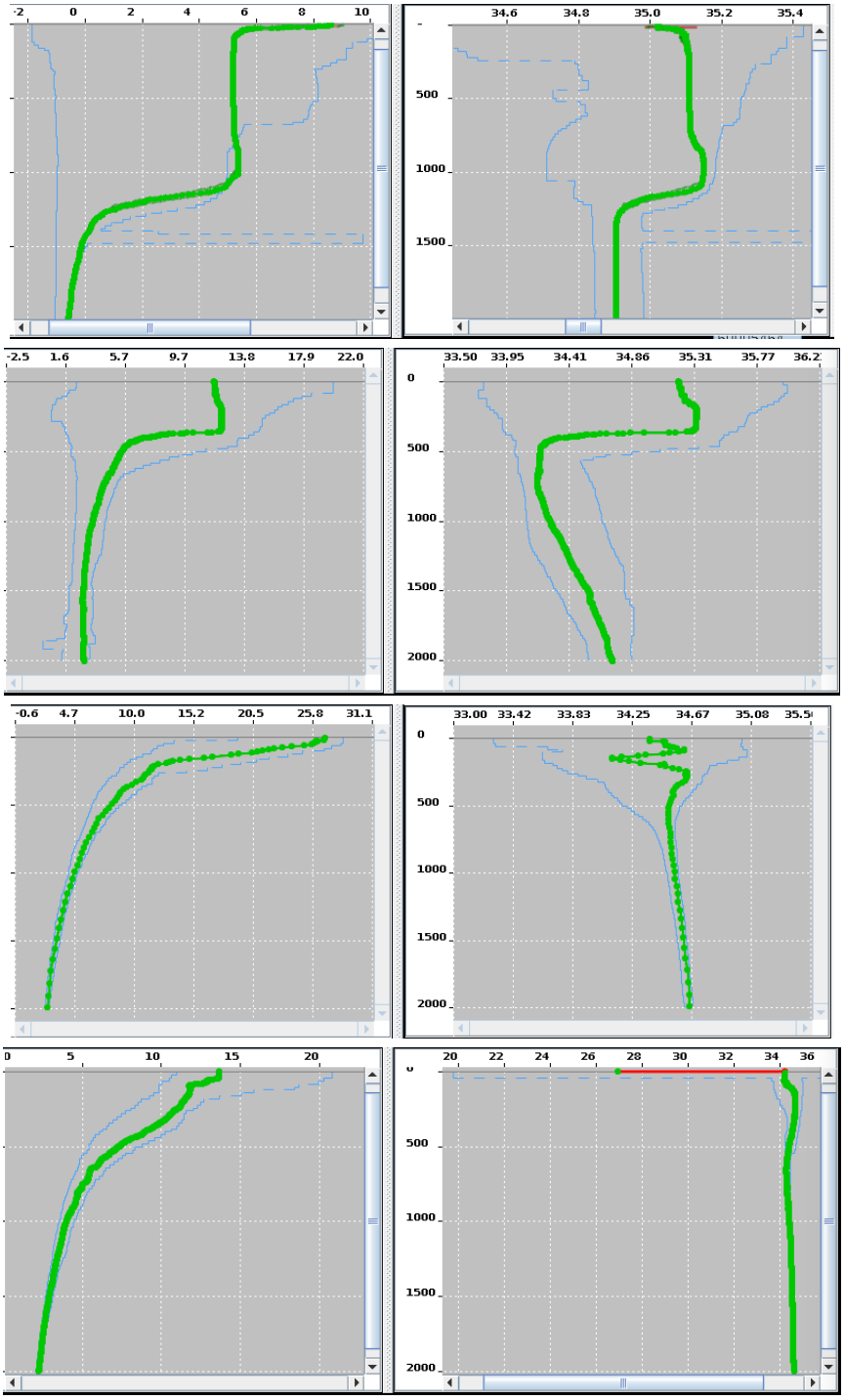

Figure 12: Screen captures of SCOOP software output for selected temperature (left) and salinity (right) profiles. Top: platform 3901873, station IDs 60005464, 60093228, 60197838. Middle top: platform 3901834, station ID 60213877. Middle bottom: 4901489, station ID 61193747. Bottom: 3901485, station ID 60049488.