Further characterization of MinMax QC approach

Technical note - MinMax sensitivity study - version 2

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

February 14, 2020

This technical note follows a previous one, "Adapting MinMax QC approach to NRT constraints at CORIOLIS for CMEMS INSTAC products", published in February 2019, and used for a first operational implementation of MinMax Quality Control (QC) for ARGO NRT data at Coriolis data center. In this work, we used an image of the CORIOLIS database at the earliest diffusion step to quantify the changes in amounts of good and bad alerts raised by the MinMax check while varying artificially and homogeneously the validity intervals using a P widening factor, see the corresponding note for further details. The alerts where counted by profile rather than by individual observation, as we wanted to focus on the amount of profiles to be visualized by a QC operator. The reference fields used in that study were some preliminary ones used in delayed-mode QC at Coriolis; they had been smoothed out horizontally over 3 layers of ISEA grid cells and vertically over the immediately above and below layers (with a regular 20 db layer thickness).

For this updated study, a specific database has been setup at OceanScope; its initial state consists of 1) an ARGO GDAC image built on january 2019, 2) several datasets of historical CTDs and 3) the MEOP Sea Mammals dataset. In a second step, dedicated QC focused on the extreme values of the local parameter distribution is performed. From the resulting dataset, we build several different versions of the reference MinMax fields, using the profiles acquired over time periods of increasing extent (up to October 2015, as in the previous study, up to december 2016/2017/2018). Evaluation of the robustness is always performed only with profiles that are posterior to the time period considered when building the reference fields. New features in this study:

- the datasets used to build the reference fields and to perform the sensitivity study come from the same database, and the same effort on a dedicated QC. It was not the case in the previous study, as the flags of the evaluation profiles were set to mimic a specific state (the CMEMS state at the earliest distribution stage)
- the programs used to perform the QC are now built by OceanScope, and run over the OceanScope database. The results might differ somehow from what would be obtained with the Coriolis validation database.
- the smoothing over 3 layers of neighbours performed in the MinMax reference fields for the previous study reduces the spatial resolution of the methodology to more than 700 km. Here we aim at coming back to smoothing over only 1 layer of neighours (which was shown to be the minimum required to have sufficiently robust statistics, see gourrion et al. 2019), allowing a resolution close to 300 km. All extra widening of the validity intervals is performed using the P factor.
- in this new database, 3 additional years of ARGO data are available; we can evaluate the impact of the time period used to build the reference Min/Max fields.

- in addition to statistics in terms of profiles, we will also present statistics in terms of observations. A profile alert is named good if it contains any good observation alert; it is named bad if it does not contain any good observation alert but some bad ones.
- it will be possible to distinguish results in terms of depth layer, so that optimal P values can be defined as a function of depth.
- while the previous work had focused on salinity, this study also presents results for temperature, which appears to behave somehow differently.

In Section 2, we check that our new data system provides results that are close to the Coriolis system used in the previous study. In Section 3, we will show how the results degrade when we come back from 3 layers of neighbours to only 1. The improvement caused by increasing the time period used to build the reference Min/Max fields is described in Section 4. the impact of the above mentioned widening factor P is described in Section 5. In Section 6, depth/amplitude histograms of alerts are presented. Section 7 provides statistics for the specific case of the cells which characteristic bathymetry is smaller than 1800 m. Similar results for temperature are summarized in Section 8, and operational values of the P factor are proposed in Section 9.

1 Background - The MinMax approach

The MinMax method has been developped for Delayed-Time (DT) QC activities, resulting in a significant reduction of the number of profiles unnecessarily visualized by the DT operator. Nevertheless, for direct use in Real-Time or near-Real-Time (RT) production, the total number of alerts per day is still too large to be treated by the RT operator. This is understood as an imperfection of the reference fields due to a persistent lack of geophysical variability as observed in the reference dataset; the reference dataset does not describe the entire ocean variability, resulting in a significant number of erroneous detections. As a solution, the present study is proposing to simulate the missing variability through artificially extending the validity interval by a factor P as follows:

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

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

Sensitivity of the method robustness to different factors is now evaluated through statistics of good and bad detections. Optimal P values are proposed as a trade-off to best meet the NRT constraints.

2 Preliminary results

In this section, we first make the link with the previous study. The results obtained with the old Coriolis tools are compared with those from our new OceanScope tools for the same 3-month period (July to September 2018), see Figure 1. We present statistics of good (dashed line) and bad (full line) alerts for profiles in order to ease the comparison with the results from the previous study.

First, the black and blue curves correspond to the same configuration but from the two different systems. Clearly, statistics from the new system are better in terms of ratio of good to bad alerts. The reason for that is the "age" of the July/August/September 2018 data in both systems: the old system is in a state corresponding to the earliest distribution stage, while the new system contains the same profiles a few months later, i.e. in January 2019. Within a few month, about 10 % of profiles may have been processed in delayed-mode and have seen their quality improved, see section 4.1 on the impact of RT/DT processing.

Why do bad alerts reduce faster than good ones when increasing P ?

In Figure 1, we can see that, when increasing the P factor, the reduction in bad alerts occurs faster than for good alerts. If this is an intrinsic characteristic that led us to imagine that widening the intervals might improve the robustness at some point, it may not be so obvious why this happens. Here we provide some description that should help illustrate why the chosen strategy makes sense.

For all alerts, we compute the minimum required P value beyond which the method would stop raising an alert for that observation. Figure 2 displays the probability density function of such P values, distinguishing good alerts from bad ones. It appears that, for salinity, the distribution of P values associated with bad alerts is significantly much narrower than that corresponding to good alerts. Therefore, it is now clear why the proposed strategy makes sense.

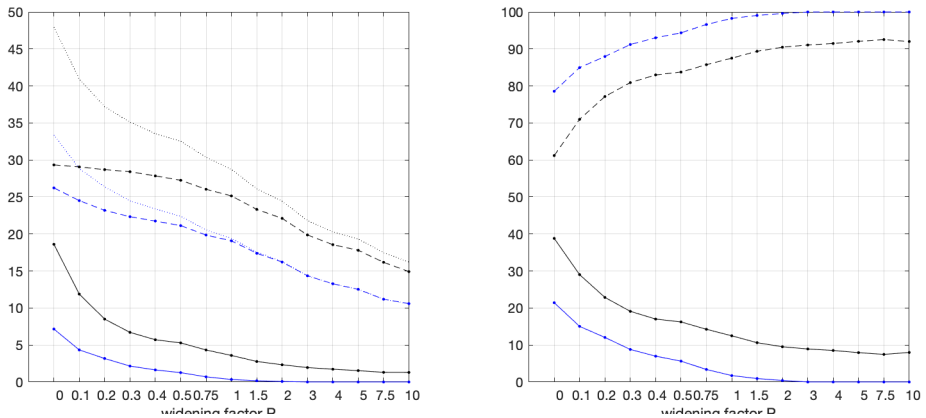

Figure 1: Left - average daily number of all alerts (dotted), good alerts (dash-dotted) and bad alerts (full line) as a function of the widening factor P. The black lines refer to the results presented in the previous study. For the blue lines, the Min/Max fields are smoothed in the same way as in the previous study, ie. vertically and horizontally over 3 layers of neighbours. Alerts are counted per profile.

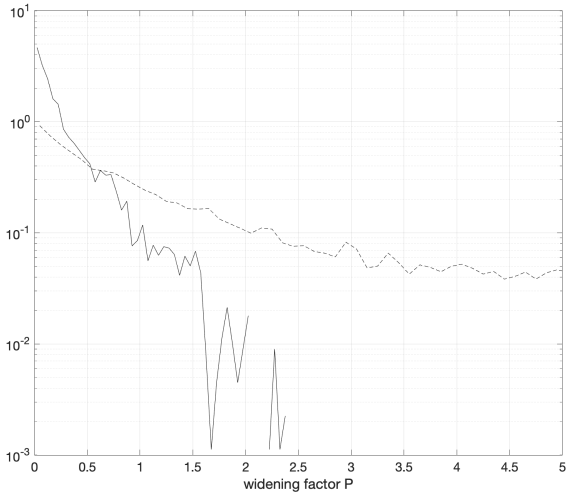

Figure 2: Histogram of P values for bad alerts (full line) and good alerts (dashed line). The dataset and Min/Max fields are the same as those used in Figure 1 for the blue curve.

It is also interesting to note that, as our dedicated QC work preliminary to the estimation of the Min/Max fields focused on the extreme values, the final quality of our dataset flags is expected to increase with the distance between the detected observation and the corresponding validity interval bound. As a consequence, we should expect the overall quality of the bad alert statistics to be somehow poorer than that of the good alert statistics.

3 Impact of increasing the spatial resolution of the reference fields

This experience aims to show the impact of spatial smoothing on the detection, see Figure 3. Spatial smoothing consists in using observations from the cell of interest together with those from its neighbors at computing the statistics, allowing to increase the statistical robustness to the price of a degraded resolution. In the first step, the original setup (blue lines, vertical smoothing + horizontal over 3 layers of neighbours i.e. max scale close to 700 km, same as green lines from Figure 1) is modified removing the vertical smoothing (red lines). We further reduce the number of layers of neighbours to 2 (green lines, 500 km) and 1 (pink lines, 300 km). As expected, increasing the resolution results in an increasing number of good detections, to the price of a degraded robustness.

In this study, we made the choice of coming back to smoothing over 1 layer of neighbours in order to keep the effective resolution close to 300 km. If such a choice leads to well increased number of bad alerts, it also allows to detect an increased number of erroneous data (good alerts). The robustness of the method is temporarily

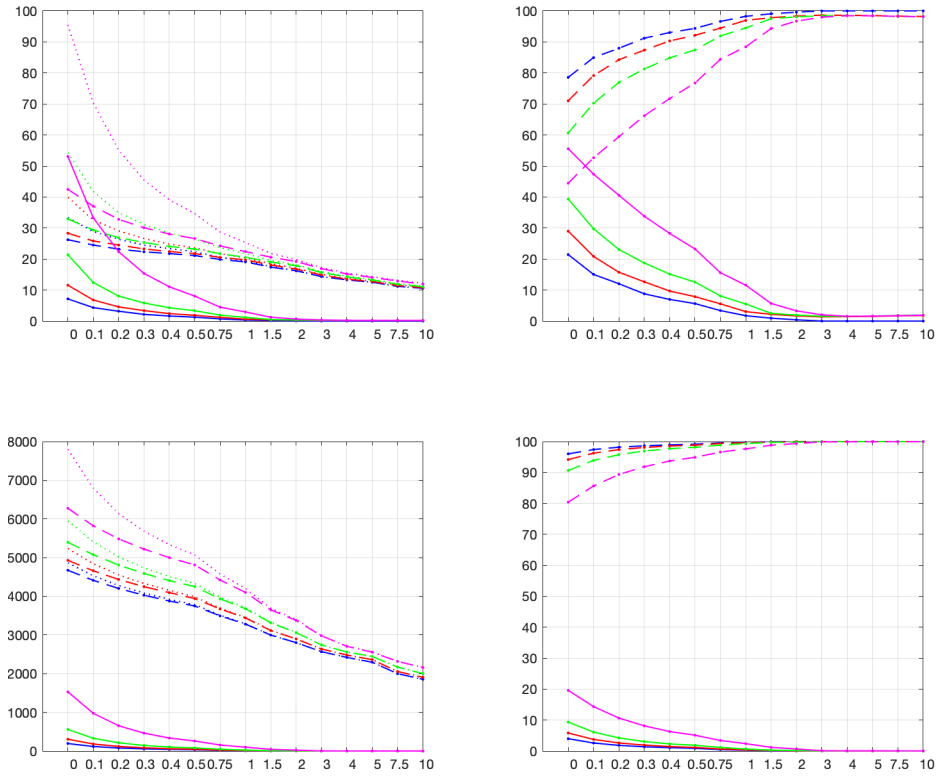

Figure 3: Left - Daily number of all alerts (dotted), good alerts (dash-dotted) and bad alerts (full line) as a function of the widening factor P. The blue lines refer to the blue lines from Figure 1, i.e. the case with smoothing over 3 layers of neighbours + vertical smoothing. For the red curves, no vertical smoothing is applied. For the green curves, horizontal smoothing is performed over 2 layers of neighbours, while for pink curves, only 1 layer of neighbours is used for smoothing. Right: same thing in terms of percentage rather than number of daily alerts. Top panels correspond to profile alerts while bottom panels refer to observation alerts.

degraded, but, in the following, we will show how the number of bad alerts can be reduced: through 1) extending the time period of the dataset used to derive the Min/Max reference fields, and 2) using positive values of the widening factor P. First, we start with distinguishing the statistics between the real-time observations and the delayed-time ones.

4 Impact of the temporal extension of the reference dataset

We decided to increase the resolution of the Min/Max reference fields to 300 km. Such a decision leads an increase of both the number of good alerts (about 50 % more errors detected) and the fraction of bad alerts (performance reduced by a factor 3 to 4), see Figure 3, left panels.

In this section, we characterize the increase of performance obtained from a set of Min/Max reference fields derived from datasets for which the upper temporal bound is successively extended by 6 months, from december 2015 to june 2018. For each dataset, all the data available since that upper temporal bound and up to december 2018 are used to evaluate the methodology robustness.

It appears that good and bad alerts statistics differ significantly over different subsets of observations: with RT or DT processing, and for negative ($<$ Min) or positive ($>$ Max) detections. Therefore, the results about the impact of the temporal extension of the reference dataset are presented separately for these categories of observations.

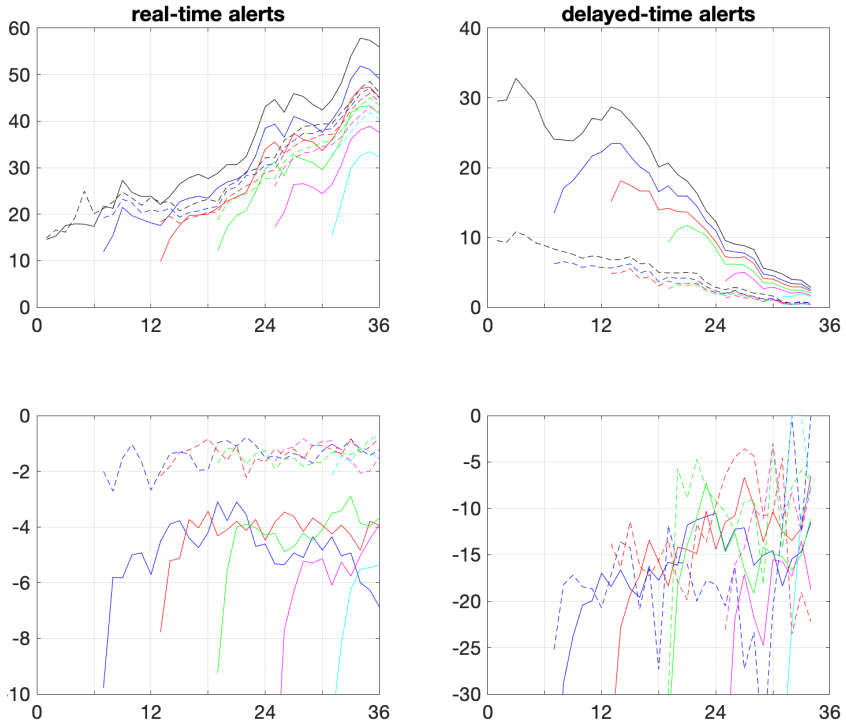

Figure 4: Top left: monthly average number of good (dashed line) and bad (full line) RT alerts for the same 36 month period. The black curves refer to the reference dataset with upper temporal bound by the end of 2015. Increasing the temporal extent by steps of 6 months, we obtain the blue (June 2016), red (Dec. 2016), green (June 2017), pink (Dec. 2017) and cyan (June 2018) curves. Bottom left: increment of the monthly average number of good (dashed line) and bad (full line) alerts at increasing the time extent of the reference dataset by the same 6-month steps. Right panels are similar to the left ones but for DT alerts, except that the increment is now multiplicative i.e. a percentage.

4.1 Differencing statistics between observations with RT or DT processing

The results are presented Figure 4 in terms of average daily number of alerts and separating the statistics obtained from the profiles with only real-time (RT, left panels) from those with delayed-time (DT, right panels) processing. We only use alerts raised without widening factor ($P = 0$).

The RT alerts increase over time from 15 to 50 alerts per day over 3 years (upper left panel). They also reduce additively with the extension of the reference dataset. The lower left panel shows that while, each day, 2 good RT alerts are lost at each 6-month extension of the reference dataset, 4-5 bad RT alerts disappear in the same time, making the method robustness increase with a larger reference dataset, as expected. The DT alerts reduce with time (in relation with the fraction of DT processing, see below); at each 6-month extension of the reference dataset, good and bad alerts are both reduced by about 15%.

We note that, as a recurrent feature, the first 3 months of all bad alert time series are characterized by significantly lower values. Such feature is understood as the signature of average decorrelation time in the upper ocean as sensed by ARGO; during these initial months, the profilers are statistically often in the same oceanic structure as a few months earlier, leading to anomalies already observed in the reference dataset. We believe that it is reasonable to discard such features before interpreting the results in the general context of real time QC.

We now investigate whether the increase/decrease of RT/DT alerts with time is consistent with the time evolving fraction of RT/DT observtions. The results are presented in Figure 5 with blue lines; full (dashed) lines correspond to DT (RT) QC.

Bad alerts

First, the percentage of bad alerts (see Figure 5, right panel) is quite stable over observation time (no tendency is visible), while the dataset consists of an increasing part of RT observations (see left panel). As bad alerts are associated to method imperfections, it is a nice result to show that the method robustness is quite homogeneous

Figure 5: Left: monthly percentage of observations with RT (dashed line) or DT (full line) processing over the 36 months from Jan.2016 to Dec.2018. Middle: monthly percentage of good RT (dashed line) and DT (full line) alerts for the same period. Right: monthly percentage of bad RT (dashed line) and DT (full line) alerts for the same period. For all panels, the blue curves refer to the whole dataset while the red curves refer to the same dataset reduced by the set of platforms having SBE sensors with serial number in the ranges specified in the text.

over time, at least at first order. Second, the percentage of bad alerts appears to be systematically higher by about 50% for observations with RT processing relatively to those with DT processing. It is useful to recall that bad alerts have two possible origins: 1) imperfection of the Min/Max reference fields, leading good data to be improperly qualified as erroneous, and 2) imperfect QC in the database, leading detections to be improperly qualified as bad alerts instead of good ones. Once this recalled, it is clear that if the first above-cited reason is not expected to vary with the RT/DT processing, the quality of the database QC is expected to improve after DT processing. Accordingly, we consider that the systematical lower percentage of bad alerts associated to DT processing is an understandable and robust result.

Good alerts

First, the percentage of good RT alerts is about one order of magnitude larger than that for DT processing. It is an expected result, with the ARGO DT dataset showing a smaller amount of erroneous data while the ARGO DT processing aims to improve the overall quality of the dataset through correction of decoding errors, obvious offsets or estimated sensor drifts. This result illustrates the improved quality of the DT dataset.

Second, the percentage (and not the number) of RT good alerts appears to systematically increase by a factor 3 along the period 2016-2018 (or decrease with the "age" in the database of the corresponding observations). This is understandable only in terms of a slow degradation of the overall dataset quality since early 2016. Indeed, recent results among the ARGO community suggest a higher than usual occurrence of drift problems with SBE sensors having serial number in the ranges 6000-7100 and 8100-9200. Following, the same statistics are recomputed once discarded all observations from the corresponding sensors. Results are presented in Figure 5, red lines. Obviously, the set of suspected platforms allows to explain the observed increase in good alerts, while the bad alerts statistics are not affected. Variations in the number of detections by the present method appear to be a reasonable proxy of the overall quality of the dataset.

As the objective of this study is the near-real time quality control, in the following we will discard DT alerts from our analysis.

4.2 Differencing statistics from negative (MIN) and positive (MAX) alerts

Here, we propose to distinguish statistics by their sign, i.e. lower than the MIN or greater than the MAX; the left and right panels of Figure 6 provide the same statistics as Figure 4, left panels, but specific to the MIN and MAX alerts. MAX alerts are more numerous and increase by a factor 3 over the observed 3-years period, while the MIN ones only by a factor 2. This is certainly the effect of the recent problem with some SBE sensor series which induces majoritarily positive errors and impacts essentially the MAX alerts statistics;

As in the previous section, increase of the temporal extent of the reference dataset induces a regular reduction of all alerts; the bottom panels indicate that the good MIN (resp. MAX) alerts reduce by 0.5 (resp. 1) at each 6-month extension while the bad alerts reduce by 2 (resp. 3) events for same extension.

In the following, we are going to keep this MIN/MAX alerts distinction by default.

Figure 6: Top left: monthly average number of good (dashed line) and bad (full line) RT MIN alerts for the same 36 month period. The black curves refer to the reference dataset with upper temporal bound by the end of 2015. Increasing the temporal extent by steps of 6 months, we obtain the blue (June 2016), red (Dec. 2016), green (June 2017), pink (Dec. 2017) and cyan (June 2018) curves. Bottom left: increment of the monthly average number of good (dashed line) and bad (full line) alerts at increasing the time extent of the reference dataset by the same 6-month steps. Right panels are similar to the left ones but for RT MAX alerts.

5 Impact of widening the validity intervals

In the previous sections, the main factors controlling the statistics of good and bad alerts have been identified. It has also been shown that a multiplicative widening of the validity interval makes sense. Here, we present how the good and bad alerts statistics are modified through widening of the validity interval using a multiplicative factor P.

The results are presented in Figure 7. The left (right) panels refer to MIN (MAX) alerts. The 4 top panels provide statistics in terms of profile alerts, which are adequate for estimation of operator time spent in visual QC tasks; the 4 bottom panels concern individual observation alerts, which are more relevant for method performance evaluation. For each type of alerts, statistics are presented in terms of 1) mean daily number of alerts and 2) percentage among all alerts (good+bad).

First, the statistics clearly show a poorer performance (higher bad alert ratio) for the profile alerts than for the observation alerts. This simply results from the fact that a single observation bad alert within a profile leads the whole profile to be qualified as a bad alert, unless some good alert is present.

As already observed, the MAX alerts are more numerous than the MIN ones and display a larger impact of the temporal extent of the reference dataset. The results corresponding to the cyan curves (i.e. the upper temporal bound is June 2018) show an increased performance relatively to the other ones, which is likely associated to some higher correlation of the July/August/September 2018 observations with ones included in the corresponding reference dataset. The performance is certainly exagerated in comparison to the case where these specific reference fields are used with data posterior to September 2018.

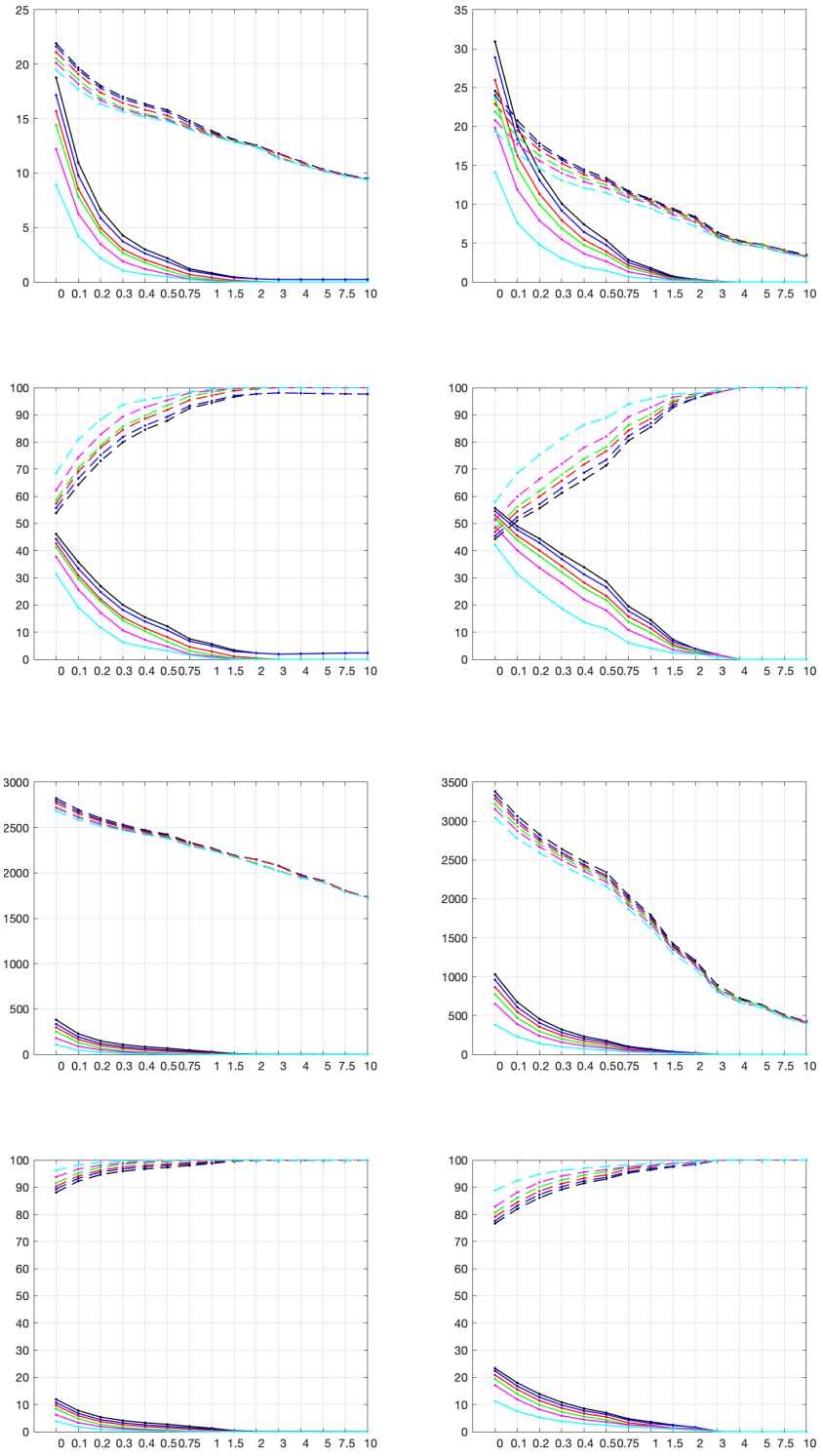

Figure 7: Top left: average dailly number of good (dashed line) and bad (full line) RT MIN profile alerts for July/August/September 2018. The black curves refer to the reference dataset with upper temporal bound by the end of 2015. Increasing the temporal extent by steps of 6 months, we obtain the blue (June 2016), red (Dec. 2016), green (June 2017), pink (Dec. 2017) and cyan (June 2018) curves. Middle top left: same as top panel but for percentages instead of number of profiles with detections. Top right and middle top right panels are similar to the left ones but for RT MAX profile alerts. The four bottom panels are similar but for individual observation alerts.

The percentage graphs are expected to be the adequate frame in order to make decisions about the choice of operational P values. We propose that, starting from an expected performance in terms of good to bad alerts ratio (e.g. 90 %), such graphs be used to obtain the P value that should allow to fullfill such a constraint on average.

6 Depth/amplitude distributions of alerts

In this section, we present the joint histograms of amplitude and observation depth of alerts. The amplitude can either be described through an additive (Q) or multiplicative (P) formulation:

$$Q_{min} = \mathcal{S}_{min} - \mathcal{S} \tag{3}$$

$$Q_{max} = \mathcal{S} - \mathcal{S}_{max} \tag{4}$$

$$P_{min} = \frac{\mathcal{S} - \mathcal{S}_{med}}{\mathcal{S}_{min} - \mathcal{S}_{med}} - 1 \tag{5}$$

$$P_{max} = \frac{\mathcal{S} - \mathcal{S}_{med}}{\mathcal{S}_{max} - \mathcal{S}_{med}} - 1 \tag{6}$$

where P_{min} (P_{max}) is the P value (from Eq. 2) that would be required to cancel the alert, i.e. to let the $\mathcal{S}_{min}$ ($\mathcal{S}_{max}$) values reach the alert value $\mathcal{S}$. The first (third) row of Figure 8 refers to Q (P) as obtained using the reference dataset up to end 2015. The second (fourth) row is obtained using the reference dataset up to end 2017. The first (second) column concerns the good MIN (MAX) alerts while the third (fourth) concerns the bad ones. In each panel, the 2-D histogram is represented with colours while the white line shows 1-D histogram of observation depths. The 2-D histogram is corrected by a factor accounting for the average number of observations in the corresponding depth layer.

One can summarize observations as follows:

- the MAX alerts are more numerous than the MIN ones (see the extent of the brown/red saturated patches near 0)
- good MAX alerts occurrence increases linearly with depth (contrarily to good MIN); this is consistent with drift detections being more effective where the validity intervals are thinnest
- the good alerts are much more widely spread in the P or Q dimensions than the bad ones, consistently with our Figure 2
- the good alerts are weakly affected by a 2-year increase of the temporal extent of the reference dataset while the bad ones are reduced by about a factor 2 (see e.g. the 1-D histograms)
- with the Q formulation, the distribution of bad alert amplitudes varies strongly with depth; the P formulation provides a distribution of bad alert amplitudes much more homogeneous with depth, supporting the multiplicative choice of Eq. 2 for interval widening.

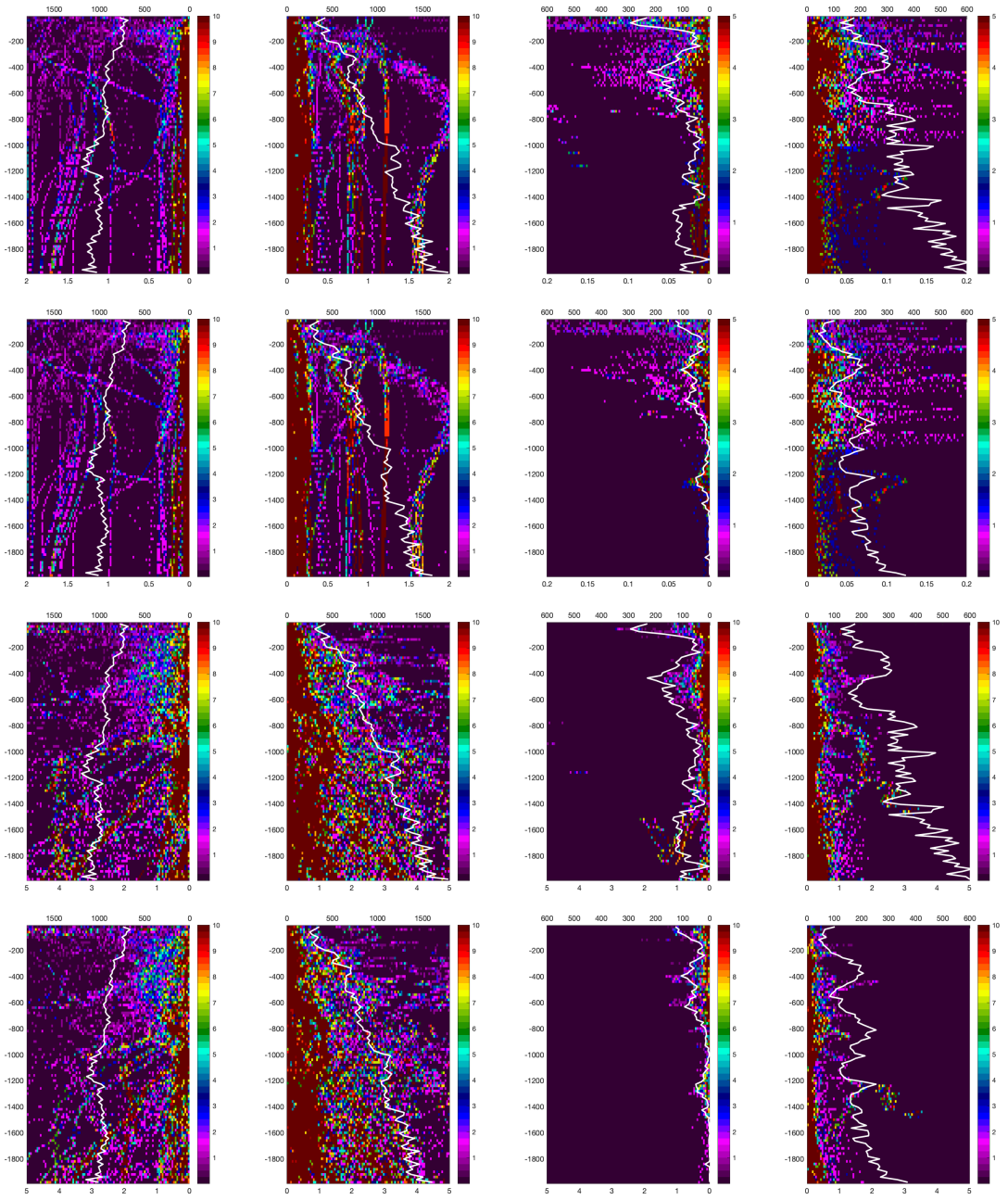

Figure 8: Top row, left panel: depth/amplitude histogram of good MIN alerts with the amplitude characterized through the additive Q factor. The histogram is normalized by a factor accounting for the variations of the average number of observations per layer. The alerts are computed using the reference dataset with upper temporal bound as the end of 2015. The white line is the one-dimensional depth histogram (top labels). Top row, middle left panel : same for good MAX alerts. Top row, middle right panel: same for bad MIN alerts. Top row, right panel: same for bad MAX alerts. Middle top row: same as top row but the upper temporal bound of the reference dataset is end of 2017. Middle bottom and bottom rows: same as the two first rows but using a multiplicative P factor instead of an additive one.

Some surprising features appear in the bad alerts distributions:

- MAX between 1200 and 1400 m, whatever the time extent of the reference dataset and with amplitudes higher than usual at such a depth; looking more in details, it appears that the corresponding observations come from profiler 3901156, which displays a salinity drift undetected during our QC of the reference dataset; the QC values are improperly set to 1; such alerts should be classified as good.
- MAX between 800 and 1200 m, only visible with the shortest time extent of the reference dataset; 5901015, equatorial Pacific
- MIN between 800 and 1200 m; the alerts come from 2 profilers (2902744 and 2902745) while they are in the South China Sea, where the reference dataset is poor of data; likely due to imperfect reference fields.

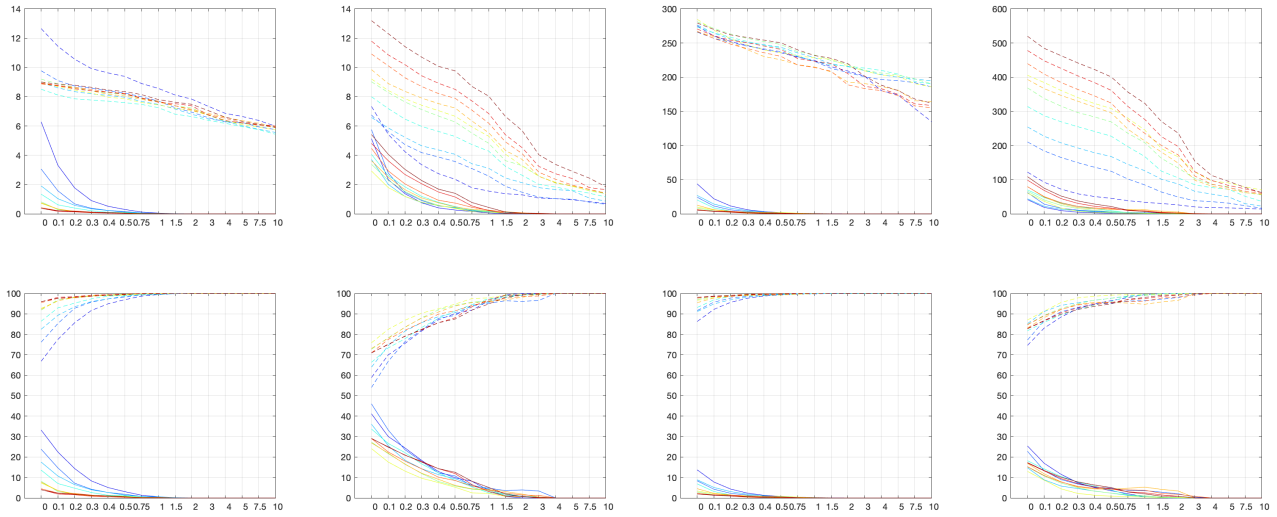

Figure 9: Top: average daily number of good (bad) detections, dashed (full) line for 10 depth layers from 0-200 m (blue) to 1800-2000 m (red). Bottom: same for percentages instead of number of detections. The four left panels refer to profile alerts while the four right ones to individual observation alerts. In each set of 4 panels, the first column corresponds to RT MIN alerts while the second one to RT MAX alerts.

- MIN above 200 m; the alerts come from 3 profilers at the time where they are located in the Indonesia Sea where again the reference dataset is poor of data.
- MIN between 1500 and 1900 m; Caribbean Sea

We note that the present 2-D distribution may help improve the reference dataset QC.

For the reference dataset ending in Dec. 2017 (pink curves in Figure 7), Figure 9 shows the variation with depth of the detection statistics. The numbers of detections are corrected from the variations of the average number of observations per layer. Clearly, the good MIN alerts are randomly distributed with depth, while the good MAX ones are preferentially found at depth; this is consistent with drift problem of conductivity sensors which mostly result into positive errors and are better detected at depth where the variability is generally weaker.

7 Impact of grid cells with bathymetry shallower than 1800 m

All the results presented in the previous sections adress the case of profiles acquired within geographical grid cells having some characteristic bathymetry beyond 1800 m; this is the way used in the previous studies to focus on deep waters as a first objective. Here, we provide statistics for all other grid cells where the reference fields are defined, i.e. where some profiles were present in the reference dataset. As we use ARGO data for the method evaluation, it should be noted that our alert statistics do not account for the cells located in shallow water (typically with bathymetry smaller than 250 m) mostly not sampled by the ARGO network.

Figure 10 provides such results for MIN and MAX alerts. The comparison with Figure 6 allows to claim that the fractions of good and bad alerts are similar for these shallow grid cells than for the other ones. The method can be applied to all cells having available statistics, without significative modification of its robustness.

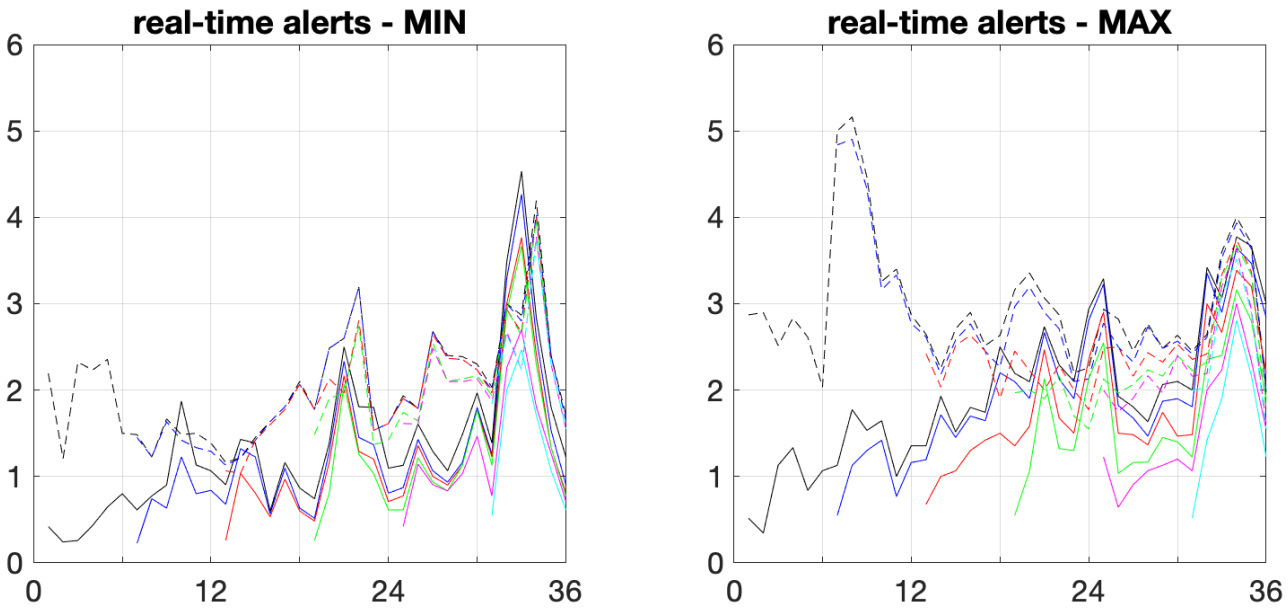

Figure 10: Same as upper panels of Figure 6 but for grid cells with characteristic bathymetry shallower than 1800 m.

For the sake of completeness, Figure 11 shows, in red, the geographic cells where Min/Max statistics are provided. In the blue/green cells, the reference dataset does not contain any observation. Among these, the blue cells correspond to "deep" ocean (bathy $>$ 250 m) while the green ones to shallow waters (bathy $<$ 250 m).

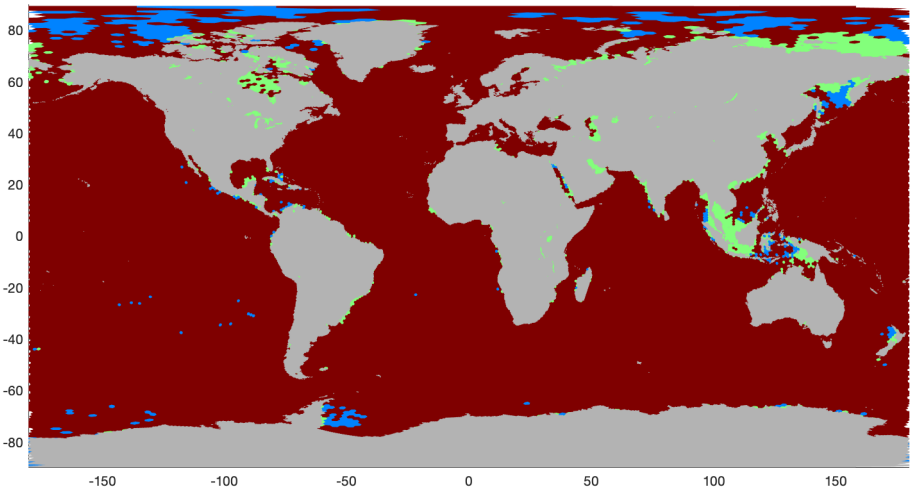

Figure 11: Map of grid cells with available statistics for the reference dataset extended up to December 2018, without horizontal smoothing. Red indicates cells with Min/Max values available; blue (green) indicates cells without Min/Max values available where the characteristic bathymetry is larger (smaller) than 250 m.

8 Similar statistics for temperature

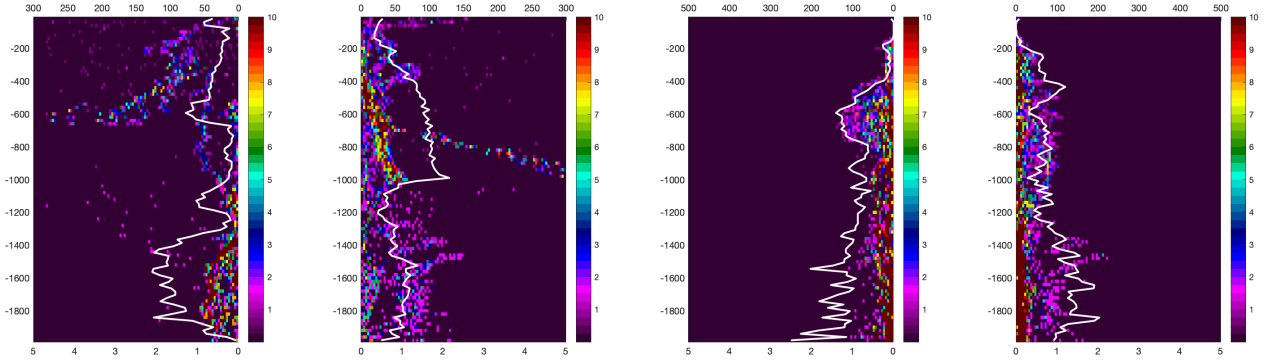

Figure 12: Left: depth/amplitude histogram of good MIN temperature alerts with the amplitude characterized through the multiplicative P factor. The alerts are computed using the reference dataset with upper temporal bound as the end of 2017. The white line is the one-dimensional depth histogram (top labels). Middle left: same for good MAX temperature alerts. Middle right panel: same for bad MIN temperature alerts. Right panel: same for bad MAX temperature alerts.

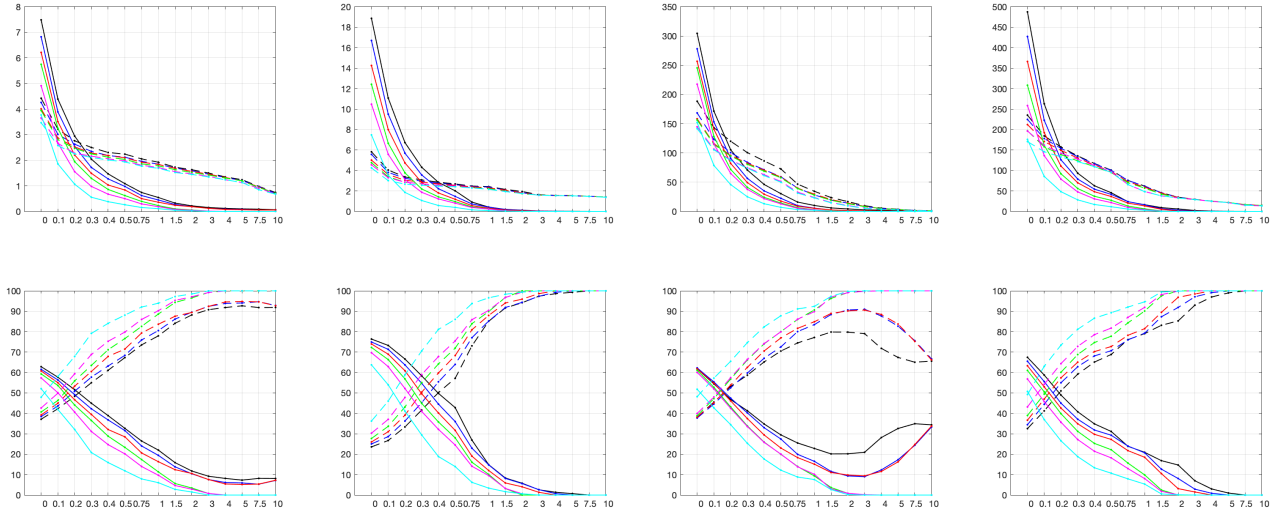

Figure 13: Same as Figure 7 but for temperature.

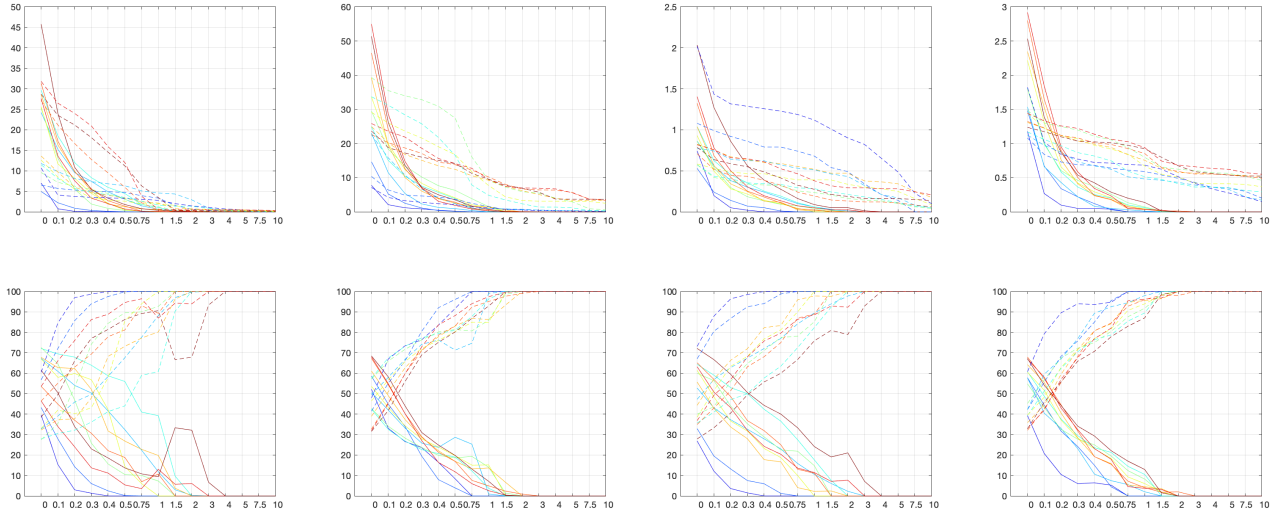

Figure 14: Same as Figure 9 but for temperature.

9 Proposal of operational P values

For salinity

Following the previous analysis of salinity detection statistics, we propose the following P values for implementation in the operational processing chain:

- for MIN detections: $P = 0.5 - 0.3 * min(Z, 1000)/1000$
- for MAX detections: $P = 0.2$ for all depths

For temperature

Following the previous analysis of temperature detection statistics, we propose the following P values for implementation in the operational processing chain:

- for MIN detections: $P = 0$ for $Z < 400$, $P = 0.5$ for $Z > 400$
- for MAX detections: $P = 0$ for $Z < 200$, $P = 0.5$ for $Z > 200$

Bad alert locations

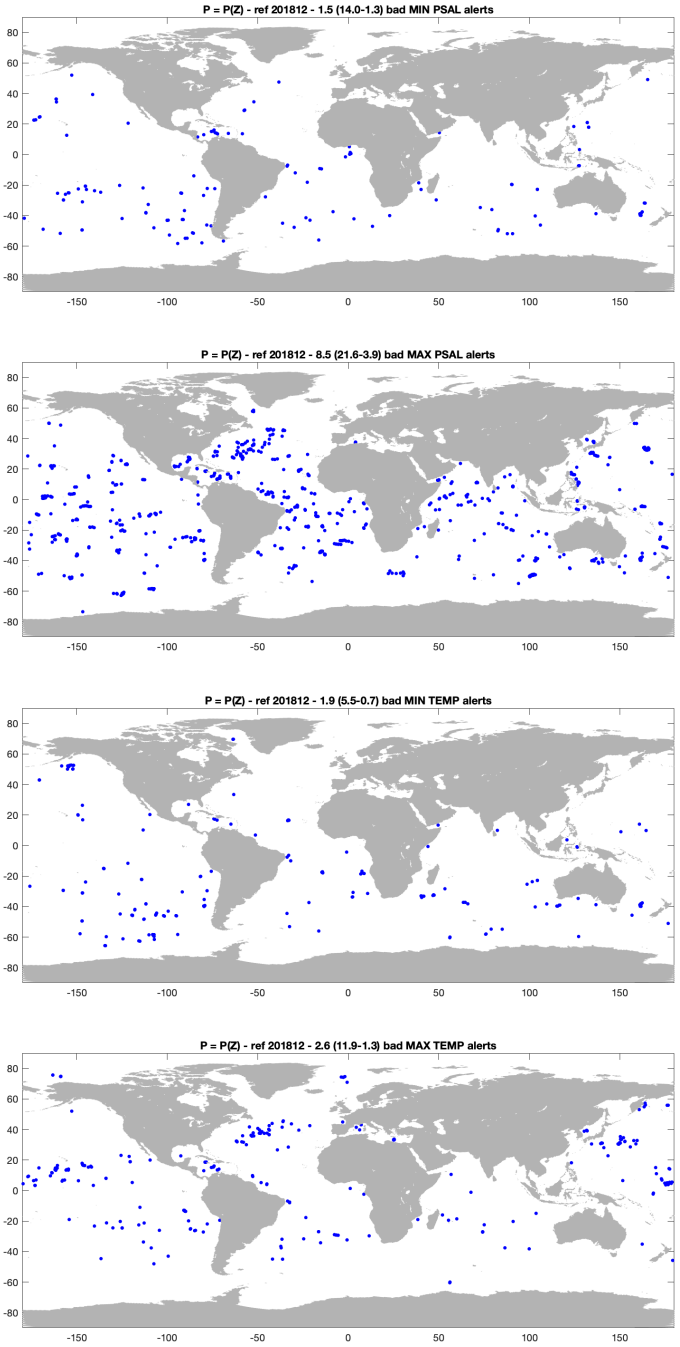

Figure 15: Maps with locations of bad alerts raised with the reference dataset extended up to December 2018. Top: MIN salinity; middle top: MAX salinity; middle bottom: MIN temperature; bottom: MAX temperature.