Technical Note 202312-MinMax-01

MinMax QC approach: Characterization of alert statistics and configuration for CMS Nov. 2023 EiS Case of ARGO T/S dataset

J.Gourrion, D.Leroy
(POKaPOK)

C.Coatanoan
(Ifremer/SISMER)

version 1

Dec. 26th, 2023

contact: jerome.gourrion@pokapok.org

Abstract

Since 2019, the NRT operations at CORIOLIS data center apply the MinMax approach as a *local range* test to the Temperature and Salinity ARGO profiles. The test detects suspicious profiles that are visualized by the NRT operator in order to take a quality decision. While updated by early 2020, the method has demonstrated its capacity to early detect ARGO platforms suffering from Salty drift (see "Abrupt Salty Drift (ASD) in SeaBird CTDs" section in `https://argo.ucsd.edu/data/data-faq/#fsd`). Since early 2023, the method is also applied to all non-ARGO profiles distributed in the CMS product.

In this technical note, we document the recent developments performed at POKaPOK in order to extend the domain of applicability and improve the method robustness without degrading its performance. Main updates concern 1) a new reference dataset with 4,5 years of additional ARGO profiles, new historical CTD dataset (from ARGO DMQC operations) and latest MEOP (Sea Mammals) dataset, 2) a vertical extension of the reference fields from 0-2000 dbar to 0-5500 dbar, 3) revisiting the validity interval widening strategy that aims at improving the robustness through accounting for the variability not sampled in the reference dataset.

In a last step, from the results of the widening study mentioned above, we present the configuration choice made for the CORIOLIS operational chain update.

Acknowledgement

J.Gourrion and D.Leroy have realized all the R&D tasks behind this work. J.Gourrion has been in charge of the reference dataset update; D.Leroy has implemented the vertical extension of the reference fields; they jointly designed an evaluation strategy that answers the questions critical for NRT operations such as impact on the dataset quality or NRT operator workload. C.Coatanoan carries the daily role of the NRT operator; she provides regularly operations feedback that helps understanding the strengths and weaknesses of the method; she also contributed to the discussions that lead to the decision on the new configuration for the CMS Nov. 2023 Entry into Service (EiS). She also provided corrections and suggestions on the present manuscript. The authors are grateful to D.Dobler for her past contribution to a robust implementation of the MinMax prototype into the CORIOLIS NRT operational chain. Finally, they acknowledge the long term confidence and support from the CMS In Situ TAC coordination team at Ifremer, especially S.Pouliquen in the past and, currently, S.Tarot, D.Obaton and L.Drouineau.

Contents

List of Figures

1 Introduction

Within the Copernicus Marine Service (CMS) In Situ TAC project, POKaPOK is in charge of the R&D activities aiming at improving the quality of the NRT Temperature and Salinity dataset. In addition to quality check tests dedicated to specific errors (spikes, density inversion, ...), the MinMax approach [1] is a *local range* test checking whether a given observation lays inside or outside of a validity interval defined from local historical distributions.

The validity interval bounds are initially inferred from minimum/maximum samples from the reference dataset. Next, they are modified to account for the fraction of variability that is not sampled in the reference dataset. Presently, such a model consists in a linear widening of the validity interval around the local median value:

$$Min_W = Median + (1 + W) \cdot (Min - Median) \tag{1}$$

$$Max_W = Median + (1 + W) \cdot (Max - Median) \tag{2}$$

where the linear widening factor W is allowed to vary with the parameter (T or S), the sign of the anomaly (lower than minimum or higher than maximum) and the observation depth; apart from this, W is taken as geographically homogeneous; the optimization has essentially focused on the adjustment of W since the first deployment in NRT operations.

Possible remaining errors in the current method implementation are caused by

- remaining QC imperfections in the reference dataset (QC = 1 while should be 4)
- and the geographical homogeneity of W.

If the amount of good detections is primarily driven by the overall dataset quality (and on the method performance to a lesser extent), the erroneous detections are the result of the above-mentioned method imperfections and are responsible of wasted operator time.

In the context of NRT operations, the available operator energy is limited, and the optimal configuration is a compromise between 1) maximizing good detections (better dataset quality, increased method performance) and 2) minimizing the erroneous ones (lower operator time waste, increased method robustness). As a reminder (as well as a medium-term objective), the lower the amount of erroneous detections, the closer the point where full automation (without operator check) may become conceivable.

In this study, both objectives 1) and 2) are addressed.

Concerning objective 1), a methodology to extend the applicability of the method below 2000 dbar is proposed; it should allow to increase the overall amount of detected erroneous data.

Objective 2 is addressed through 2 different approaches. In (2a), we rebuild the reference dataset from which the raw local validity intervals are inferred, adding additional ARGO data from 2019-2023, and new platforms from the latest MEOP repository; more robust parameter distributions are expected as a consequence of an increased sampling of the local variability. In (2b), we revisit the widening strategy in order to account for the increased part of sampled variability. The classification of detections into good or bad ones is improved through a consistency analysis of the input dataset quality flags considering the expected behaviour of bad detections with increasing widening.

The present technical note is organized as follows:

1. Chapter 1: New version of the reference dataset and reporting on the additional detection of suspicious profiles.
2. Chapter 2: Method for Min/Max statistics extension down-to ocean layers below 2000 dbar where the reference dataset is much sparser than in the 0-2000 dbar one.

3. Chapter 3: Definition of the strategy to properly estimate statistics of good/bad detections for decision about an optimal widening configuration for NRT operations at CORIOLIS

4. Chapter 4: Results of the widening study and decision

2 The 2023 version of the MinMax reference dataset

As mentioned in the introduction, the robustness of the MinMax QC approach relies primarily on the fraction of the natural viariability actually sampled by the reference dataset.

2.1 Reference dataset content

The 2019 and 2020 versions of the operational MinMax reference fields were inferred from a version of the reference dataset with the following characteristics:

- ARGO GDAC image from January 2019
- MEOP dataset version 2018-04-10
- Selection of historical CTDs from CORIOLIS database
- Only data down to 2050 dbar are used

In 2023, POKaPOK has been working on updating the MinMax reference dataset as follows:

- ARGO GDAC image from July 2023
- Apply QC information derived from the Salty Drift spreadsheet as provided by ARGO on July 11^{th} 2023
- MEOP dataset version 2021-11-26
- Historical ship-based CTDs provided by ARGO for its DMQC activities
- All observation depths are used

2.2 Dedicated QC procedure for extreme values

Once all the data loaded in a dedicated database, the tedious task comes: it is necessary to perform a dedicated QC analysis in order to improve the robustness of the Minima and Maxima estimates, ensuring that all possible outliers are identified. The following strategy is implemented and run:

1. Bin the latitude/longitude/pressure measurements in order to classify the observations into discrete horizontal and vertical cells. The horizontal grid consists in hexagonal cells of equal area with different possible resolutions (1, 1/2, 1/4, 1/8 degree), while the vertical is an adhoc one, with 20 dbar thick layers from surface to 2000 dbar, and 100 dbar thick layers below.
2. Scan all layers one by one and
   (a) Extract and visualize maps of layer statistics (Minimum, Maximum, Maximum - Minimum, Median - Minimum, Maximum - Median, ...)
   (b) Scrutinize such maps in order to detect spatial discontinuities between neighbour cells
   (c) For each selected discontinuity, identify the profile responsible of it and draw new figures to help decision: the Temperature, Salinity and Density profile of interest in 2 different figures, one with all the other profiles from the same platform, and one with all profiles found in the same horizontal grid cell independently of time; may also use time series of the measurement departures from the local median as a function of depth
   (d) Make a decision on whether the observation responsible of the discontinuity is an erroneous data or an infrequent extreme event. Making a decision on the observation from the current depth layer, the operator may see an opportunity to flag other similar data from the same profile or from any profile from the same platform
   (e) update the statistics maps in order to check that any flag modification is correctly accounted for.
   (f) keep on scrutinizing the maps !

2.3 Summary of flag modifications in comparison to original dataset

Here we report some statistics on the profiles which have been flagged during the dedicated QC process. A feedback to the concerned data centres will be provided in early 2024.

2.3.1 ARGO GDAC image from July 2023

```
----------------------------------------------------------------------
Statistics on ARGO profiles flagged during MinMax manual QC
----------------------------------------------------------------------
Number of platforms concerned: 1452  corresponding to 19764 profiles
Number of platforms with initial drift problems: 31 corresponding to 711 profiles
Number of platforms with problems after some cycle: 185 corresponding to 8074 profiles
Number of platforms having profiles with spikes: 667 corresponding to 4062 profiles

Number of remaining profiles: 6917 from 741 platforms
```

2.3.2 MEOP dataset v20211126

```
-----------------------------------------------------------------
Statistics on SEA MAMMAL profiles flagged during MinMax manual QC
-----------------------------------------------------------------
Number of platforms concerned: 135
Number of platforms where all profiles have been entirely flagged: 97
Number of other platforms: 38   corresponding to 115 profiles
        75 (115) profiles have    1 obs flagged
        17 (115) profiles have  all obs flagged
        23 (115) profiles have 2-11 obs flagged
```

2.3.3 ARGO DMQC reference dataset

```
----------------------------------------------------------------------
Statistics on ARGO DMQC CTD profiles flagged during MinMax manual QC
----------------------------------------------------------------------
Number of profiles concerned: 675
    CTD-DMQC-COR : 422 profiles among which 415 are entirely flagged
    CTD-DMQC-OCL : 193 profiles among which 183 are entirely flagged
    CTD-DMQC-CCH :  35 profiles among which  34 are entirely flagged
    CTD-DMQC-ICE :  10 profiles among which   9 are entirely flagged
    CTD-DMQC-GSD :   6 profiles among which   6 are entirely flagged
    CTD-DMQC-OGS :   6 profiles among which   6 are entirely flagged
    CTD-DMQC-DPY :   2 profiles among which   0 are entirely flagged
    CTD-DMQC-UDA :   1 profiles among which   0 are entirely flagged
```

3 Strategy for Vertical Extension of MinMax reference fields

In this section, we describe the strategy proposed and implemented in order to extend the applicability of the MinMax approach below 2000 dbar.

Originally, the MinMax approach was an attempt to build validity intervals directly from minima/maxima estimates (ideal but noisy) rather than from first and second order statistical moments (robust but imperfect). Obviously, such an approach relies on a dataset that should be extensive in order to get extrema estimates with a reasonable noise level.

For that reason, the original MinMax version was focused on the 0-2000 dbar ocean layer, a layer sensed as a priority by the ARGO network where the spatio-temporal sampling is quite homogeneous and where the amount of observations is at least one order of magnitude larger than for deepest layers.

If the amount of observations is low, the local parameter distributions will be poorly described; in that case, we expect the [Min Max] intervals to be abnormously thin and responsible of a high rate of erroneous detections.

In order to extend the MinMax applicability domain to deepest ocean layers, we need robust enough validity intervals. To this end, we propose the following strategy.

1. Integrate in the reference dataset observations acquired deeper than 2000 dbar: we benefit from the DEEP-ARGO program as well as the deep casts from the historical CTD dataset provided by ARGO for its DM QC activities. With such a type of casts, we populate the deep layers of our reference dataset.

2. Define 135 vertical layers:
   - layer 1: [-5 20] dbar
   - layers 2-99: 20 dbar thick from 20 to 1980 dbar
   - layer 100: [1980 2050] dbar
   - layers 101-135: 100 dbar thick from 2050 to 5550 dbar

3. A first version of the parameter P interval validity bounds $[PMin_k^1\ PMax_k^1]$ is obtained with $k \in [1, 135]$.

4. Convert the WOA depth levels into pressure levels using the Gibbs Sea Water library that solves Eqn.(3.32.3) of IOC et al. (2010); only latitude is provided as input, no geopotential.

5. Define an adhoc version of the validity intervals taking advantage of some information from the analyzed fields of the World Ocean Atlas (2018):

   (a) Interval center location: use the vertical variations of WOA analysis relative to 2000 dbar to infer the deep reference median from 2000 dbar value:

$$PMed(p) = PMed(p = 2000) + WOA(p) - WOA(p = 2000) \tag{3}$$

   (b) Interval width: propagate downwards the 2000 dbar interval width i.e. $[Min_1\ Max_1]$ as follows:

$$PMin^2(p) = PMin(p = 2000) + WOA(p) - WOA(p = 2000) \tag{4}$$

$$PMax^2(p) = PMax(p = 2000) + WOA(p) - WOA(p = 2000) \tag{5}$$

   (c) For each layer k, with p_1 (p_2) the depth of the layer top (bottom),

$$PMin_k^2 = Min(PMin(p_1),\ PMin(p_2)) \tag{6}$$

$$PMax_k^2 = Max(PMax(p_1),\ PMax(p_2)) \tag{7}$$

$$PMed_k = PMed((p_1 + p_2)/2) \tag{8}$$

6. Merge the 2 validity intervals into a single one:

$$PMin(z) = Min(PMin_k^1,\ PMin_k^2) \tag{9}$$

$$PMax(z) = Min(PMax_k^1,\ PMax_k^2) \tag{10}$$

Clearly, in this approach,

- we trust the WOA analysis values at depth but we only use its vertical gradient in order to avoid some discontinuity near 2000 dbar
- we do not use the variability information provided by WOA as it is provided in terms of standard deviation and inconsistent with or minimum/maximum approach; as a tentative, we first propagate downwards the 2000 dbar interval width i.e. $[Min_1\ Max_1]$ and then modify it in all locations where the reference dataset contains more extreme values.

The above approach for deep ocean layers is implemented as described and will be evaluated together with the shallower layers in chapter 4.

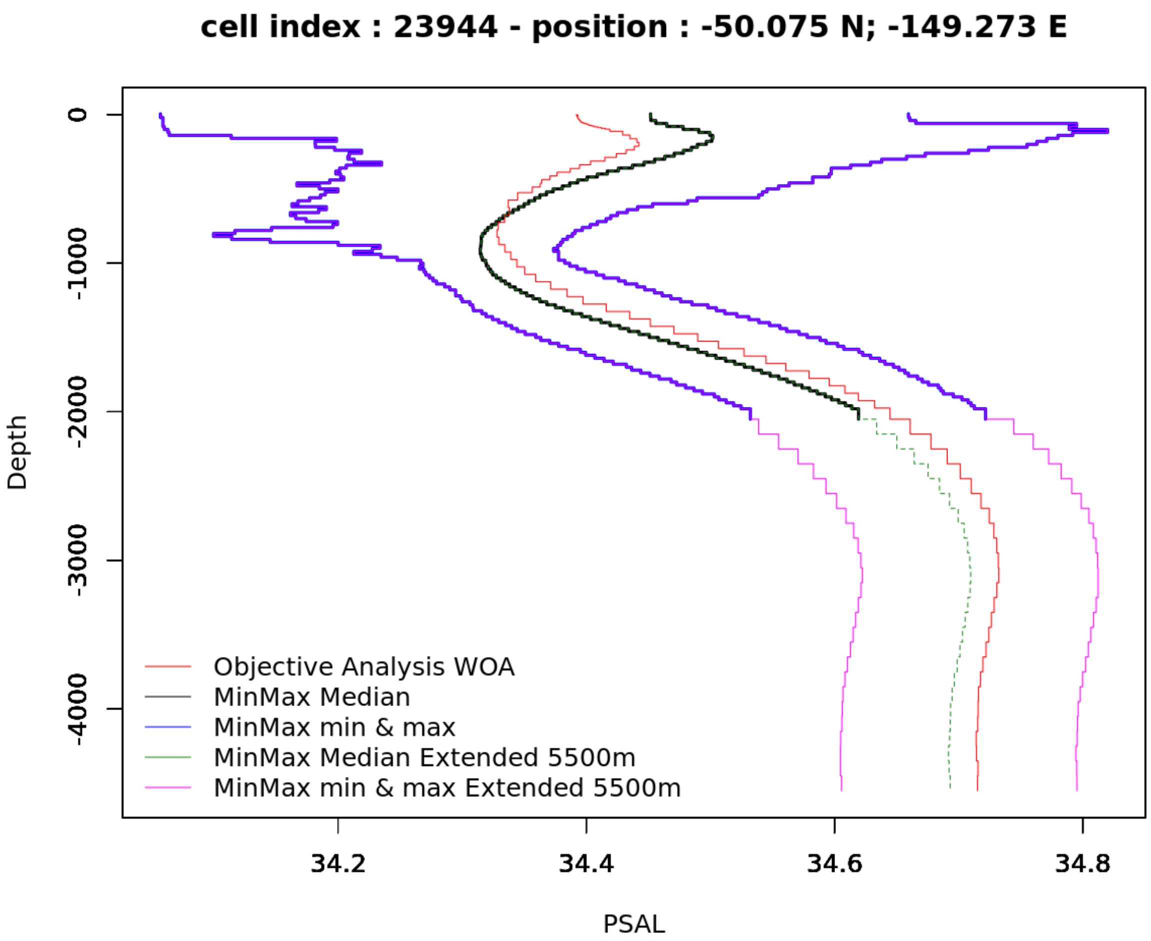

Figure 1: Illustration of the deep validity interval construction.

4 Performance evaluation strategy

Once the reference dataset QCed as described in chapter 3, the reference Min/Max fields can be generated following the deep extension procedure described in chapter 3.

It is now time to evaluate the degree of widening that brings a compromise between the detection performance, its robustness and the NRT operational constraints.

In this chapter, we present the strategy proposed to address such an objective and apply it to the CMS HISTORY PR_PF files containing the ARGO dataset.

4.1 Approach

First, we recall the metrics that will be used to evaluate the different criteria:

- detection performance: amount of detectable bad data i.e. of good detections or "True Positive"
- detection robustness: amount of detected good data i.e. of bad detections or "False Positive"
- operational constraint: amount of profiles that must be visualized to assess the detection i.e. of either good or bad detections

Then, we have to describe how the detections will be 1) obtained and 2) qualified as "good" or "bad". Qualifying a detection of "good" or "bad" means defining a reference "truth" that brings an answer to the question: are the detected data of good quality (bad detection) or of bad quality (good detection) ?

At this stage, a rational strategy would be to mimic the NRT operations conditions; this means 1) use the data exactly in the same state as for NRT operations and 2) perform a visual QC of all detected profiles in order to build the necessary "truth". Clearly some problems appear.

1. the CMS In Situ TAC archive is a rolling archive; it is instantaneous and permanently overwritten, making impossible to recover any past state. Since May 2022, POKaPOK has set up an incremental version of the full In Situ TAC archive which allows to recover the archive state at 2:00, 10:00 and 18:00 UTC everyday. Such an incremental version of the CMS archive would allow to recover the state of the data at the time where the detection was run; nevertheless, obtaining the quality "truth" over an amount of data sufficiently large to allow a robust estimate of NRT conditions means a huge (and always imperfect) work of visualization and decision. In 2018, in the study to define the first MinMax configuration, such a dedicated QC work was performed over a 3-month period of data. Here, an alternative approach has been set up.
2. The main leading idea is avoid performing a dedicated extensive visual QC work, and infer the quality "truth" from pre-existing information. Assuming that the HISTORY CMS archive does reach a Delayed-Mode quality level within several months, we a priori trust the quality flags provided together with the data. Whether the flag value is 3 or 4 (1 or 2), the detection is automatically qualified of "good" ("bad") detection. Three important remarks:
   (a) we are making the strong assumption that the HISTORY CMS quality flags are perfect; clearly, the assumption holds better than for the LATEST (NRT) files, but is probably still imperfect, and may bias our estimates. In the following, we will introduce a criterion - based on the expected behaviour of "bad" detections - that allows the detection of erroneous flags 1 that should be 4, allowing to better respect such an hypothesis.
   (b) the total amount of alerts will underestimate the reality of NRT conditions; indeed, HISTORY includes the quality improvement related to Delayed-Mode (DM) activities during which data may be adjusted from offsets, drifts, etc. Gourrion et al. (2020), in their figure 10, show that NRT ARGO data typically carry 2-2.5% of profiles with errors while for DM ones the ratio reduces to 1%. This seriously alters the approach potential in providing a realistic absolute number of daily alerts i.e. realistic estimate of the operator workload.

Consequently, the approach will be only used to provide relative numbers of alerts; using as reference a version already in operations and where such numbers are known from the past, our results will be presented in terms of increase / decrease of "good", "bad" and total alerts relatively to the reference configuration.

(c) For the Nov 2023 EiS, we want the operational reference fields to be built on the most extensive possible reference dataset, here we use the July 2023 ARGO GDAC image. Assessing the configuration choice with HISTORY PR_PF, i.e. with profiles that have been used in building the reference dataset might appear incestuous, with strong underestimation of the amount of "bad" detections; clearly, the absolute numbers of "bad" alerts will be low-biased as if the reference dataset would have sampled the whole variability.

Nevertheless, the authors claim that in the "relative" approach proposed above, the results will remain robust. Only one condition will have to be filled in: at comparing two versions of the reference fields, we will have to check that statistics from both configurations are computed over subsets of data that were fully included in (or absent from) both datasets. We will show in the following how different can be the results if such a condition is not satisfied.

Finally, the MinMax detection treats all observations of a given profile independently. Initially, we can build statistics of observation detections; unfortunately, they are not a good estimate of the operator workload: the operator visualizes entire profiles at once, not observations one by one. Further than raw "observation alert", we introduce the notion of "profile alert" and define it as follows:

- a "good" profile detection contains at least one "good" observation detection
- a "bad" profile detection contains only "bad" observation detection

In the following, the definition of the approach is based on the analysis of salinity detections uniquely. In the last part where the detection time series are analyzed to evaluate the operator workload and decide on the configuration, both results for temperature and salinity are presented.

4.2 Preliminary results and assumption assessment

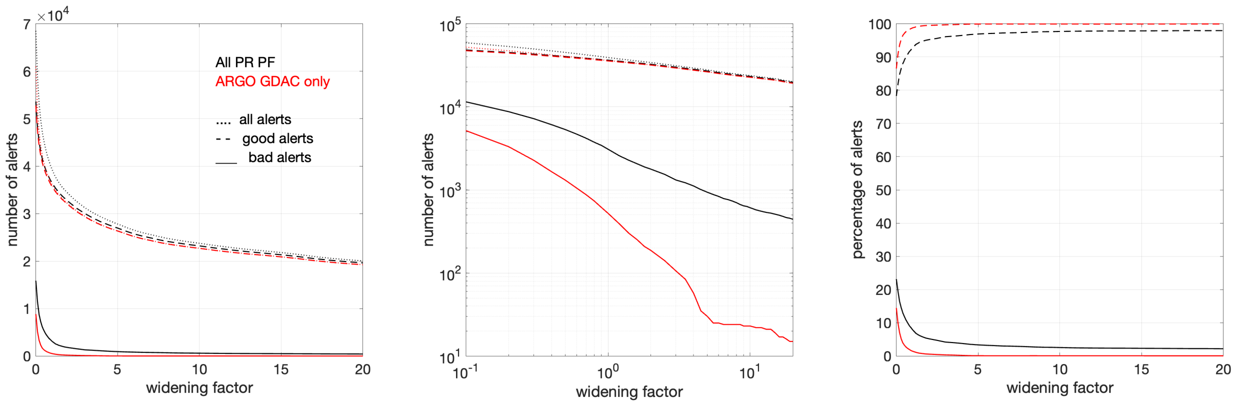

Figure 2: Profile alert statistics evolution with widening degree for all CMS PR_PF platforms (black) and for the subset of them coming from ARGO program (red). Left: total number of alerts (dotted), "good" ones (dashed), "bad" ones (full). Middle: same as previous but with logarithmic scale. Right: Percentage of "good" (dashed) and "bad" alerts (full).

Following the approach proposed in the previous section, we use the CMS In Situ TAC PR_PF HISTORY archive, as downloaded on Nov. 6^{th} 2023 from CMS repository, to run the detection method for a large set of widening factors from 0 to 20 (i.e. 0 to 2000%).

Figure 2 presents the results.

It clearly appears that, along increasing widening factor, the fraction of "bad" detections decreases faster than that of "good" ones. This illustrates and confirms the expected adequacy of the widening strategy to improve the robustness of the method: "good" detections (i.e. "bad" data) are spread more widely in the parameter space, while "bad" ones ("good" data erroneously detected) are accumulated closer to the validity interval bounds. This is consistent with the fact that we do not expect the variability observed while not sampled in the MinMax reference dataset to lay far from the MinMax envelop; and we expect that in general the number of "bad" detections should converge to zero with increasing widening.

We use now such a consideration as a postulate.

Nevertheless, we observe that the amount of "bad" detections does converge to a non zero finite value, close to 2-3 %. This suggests that some profiles are erroneously flagged with 1 instead of 4, a characteristic that violates our previously mentioned assumption that HISTORY flags would provide the quality truth.

Looking for some platforms responsible of that, we identify a list of about 600 PR_PF ones that are not referenced in the ARGO GDAC. We thus understand that such platforms do probably not benefit from the ARGO Real-Time and Delayed-Mode Quality Control procedures. In order to better respect the approach assumptions, we remove these platforms from our analysis, keeping only platforms referenced by the ARGO GDACs. The updated statistics are shown in Figure 2 as red lines; as expected, the amount and ration of false alerts are reduced, and their rate of reduction with increasing widening factor is increased, in a better agreement with our postulate.

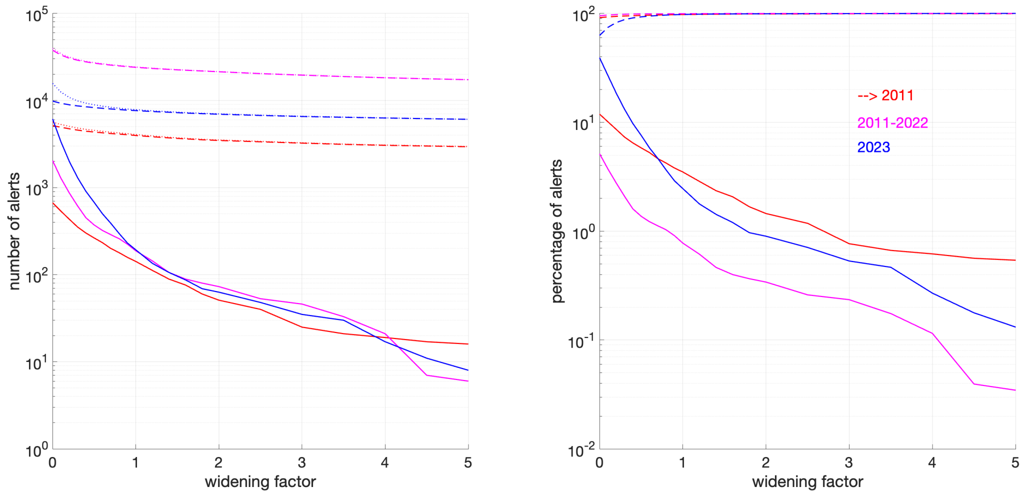

Figure 3: Profile alert statistics evolution with widening degree for the subset of PR_PF platforms coming from ARGO program. Left: Total number of alerts (dotted), "good" ones (dashed), "bad" ones (full); for the full ARGO dataset (black), only data before 2011 (red), from 2011 to 2022 (pink) and from 2023 (blue). Right: Same for percentage of alerts rather than for numbers.

Figure 3 presents the results for this subset of PR_PF platforms. In order to further investigate the consistency of the dataset, we now distinguish different time periods to compute our alert statistics. It appears that the gross part of the dataset corresponding to the 12 years 2011 to 2022 has the lowest amount of "bad" detections. Yet, two specific time periods (specifically chosen to maximize the discrepancy) present significantly different results.

The oldest data before 2011 display a "bad" detection ratio that struggles to vanish with increasing widening factor. After checking in details which platform and profiles are responsible of that, we find that 3 platforms, responsible of nearly 70 % of such cases, present some inconsistency between the ARGO GDAC and the CMS PR_PF: the PSAL_ADJUSTED variables are not present in the corresponding CMS files while they are in the GDAC ones; the result is that, at running the detection, the raw data are read instead of the ADJUSTED ones, and the flags being set to 1, the detections are classified as "bad" erroneously. The corresponding platform numbers are: 1900508 and 1900509 from bodc, 5901162 from csiro. The authors believe that most of these old "bad" detections must be associated to some profiles for which quality flags should be revisited, either on GDAC or CMS side. Such a task is left for future work.

The latest data from 2023 show poorer robustness results; as mentioned in the previous section of this chapter, this is understandable as these data have sampled some variability that is not accounted for in the MinMax reference dataset (here limited to data up to late 2022), making that subset probably more representative of realistic NRT conditions. We understand this different behaviour and claim that it is not a sign of discrepancy.

The gross part of the dataset, corresponding to the 12 years 2011 to 2022, has the lowest amount of "bad" detections, and, again, we understand that it is because all these data have been included in the MinMax reference dataset. Nevertheless, we still obtain curves somehow flat, suggesting that some of these "bad" detections might be explained by erroneous CMS flags.

We now check whether such "bad" detections could be explained by erroneous CMS flags.

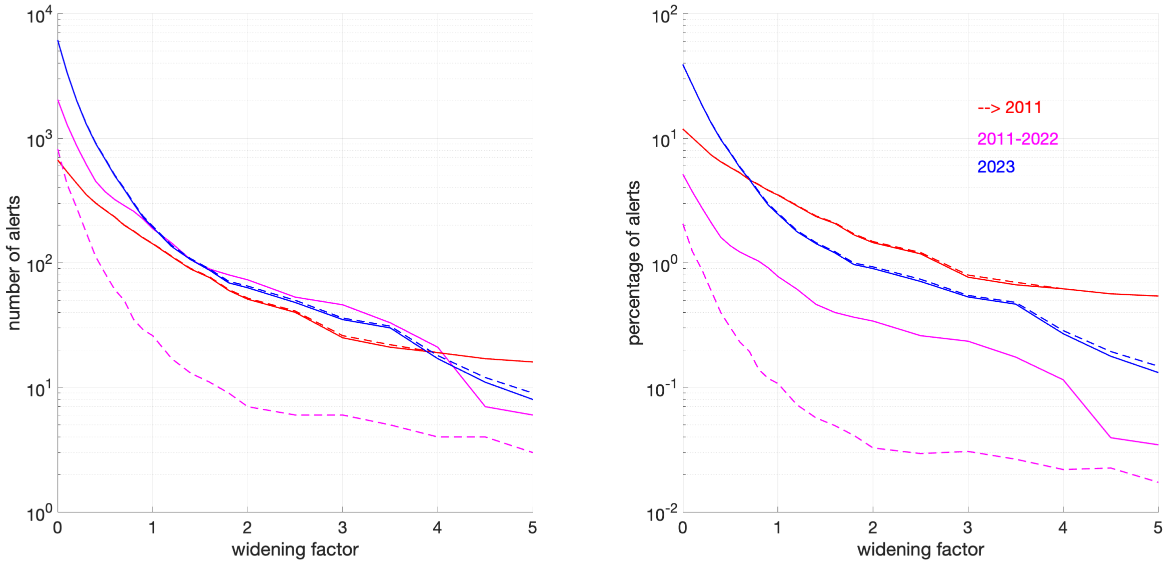

Figure 4: Bad profile alert statistics evolution with widening degree for the subset of PR_PF platforms coming from ARGO program for the original CMS dataset (full) and the modified one (dashed). Colours correspond to the full ARGO dataset (black), only data before 2011 (red), from 2011 to 2022 (pink) and from 2023 (blue). Left panel corresponds to number of alerts and right panel to percentages.

In order to check this, we create a modified CMS HISTORY dataset, in which we force the CMS flags to match with those from the MinMax reference dataset whenever they were modified along the MinMax dedicated QC step, see Chapter 3. The results are presented in Figure 4.

Clearly, reporting the MinMax QC flags to the CMS files has a large impact on the results, with a much better agreement with our above-mentioned postulate on the "bad" detection behaviour.

For the 2023 data, as expected, no impact is seen; as explained above, we understand that the corresponding level of "bad" detections is not associated to QC errors or discrepancies but rather to new variability not sampled in the reference dataset.

The old data curve is little affected; indeed, in the flag overwriting, we have copied the raw flags from one dataset onto the raw flags from the other, same for adjusted ones, and the discrepancy caused by the absence of ADJUSTED variables in the CMS files is still present.

At this stage, we consider that we have identified the main problems and discrepancies in the dataset in order to safely apply the approach described in the previous section.

4.3 Expected modifications of amount of alerts - Choice of a configuration

Here, as proposed earlier in this chapter, we present a comparison of the total number of alerts obtained when using a recent version of the operational MinMax reference fields (version from Nov 08^{th} 2023) or the new version for a set of widening factors. The dataset used for this comparison is the modified CMS HISTORY archive.

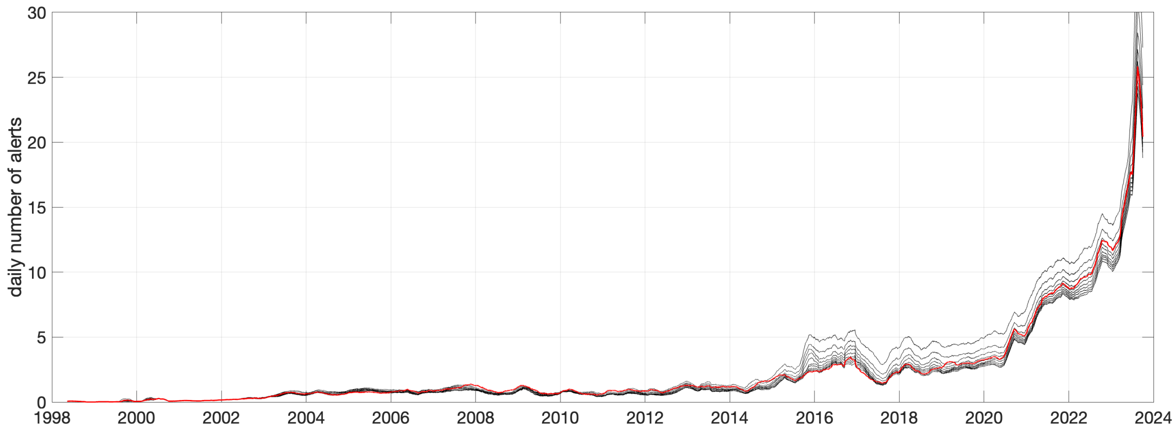

Figure 5: Daily number of salinity profile alert as a function of the observation date. The dataset used for this comparison is the modified CMS HISTORY archive. The black curves correspond to the new reference field for different widening values from 0 to 1 by step of 0.1. The red curve correspond to the results for the Nov 08^{th} 2023) version of the operational reference fields.

Figure 5 shows that up to 2015, the amount of detectable profiles with errors is quite constant; increasing the widening degree does not change significantly the count, suggesting that detected errors are mostly gross errors or typical outliers, laying far from the variability envelop characterized the MinMax intervals.

Starting from late 2015, the set of curves corresponding to increasing widening factor clearly separate; the detection starts to be sensitive to the widening degree, suggesting that the associated errors are statistically laying closer to the MinMax envelop; we understand that this transition corresponds to the onset of the Salty drift sensor problem, see `https://argo.ucsd.edu/data/data-faq/#fsd`.

By 2020, the number of daily alerts starts increasing severely; as already described in Gourrion et al. (2020), see their Figure 10, we understand that such an increase is related to an increasing fraction of profiles that are missing a Delayed-Mode (DM) quality control: indeed, a profile with correctable errors will raise an NRT alert until the moment where a correction is applied, usually during the DM operations.

In order to focus on the relative behaviour of the two versions of the reference fields, Figure 6 shows the increment in number of daily alerts at shifting from the current version of the reference fields to the proposed one. We shall mention that all time series presented here were filtered with a 3-month smoothing window. Several features are observed:

- a sharp increase of "good" detections by late 2015: for the lowest widening values (< 0.5), the spread between the different curves becomes much larger at that time; it indicates an increase of erroneous data located immediately outside the validity intervals. We interpret this jump as likely corresponding to the onset of the major "Salty drift" problem, see ARGO documentation (url / doc ??); we leave for future work to evaluate the realism of its suddenness.

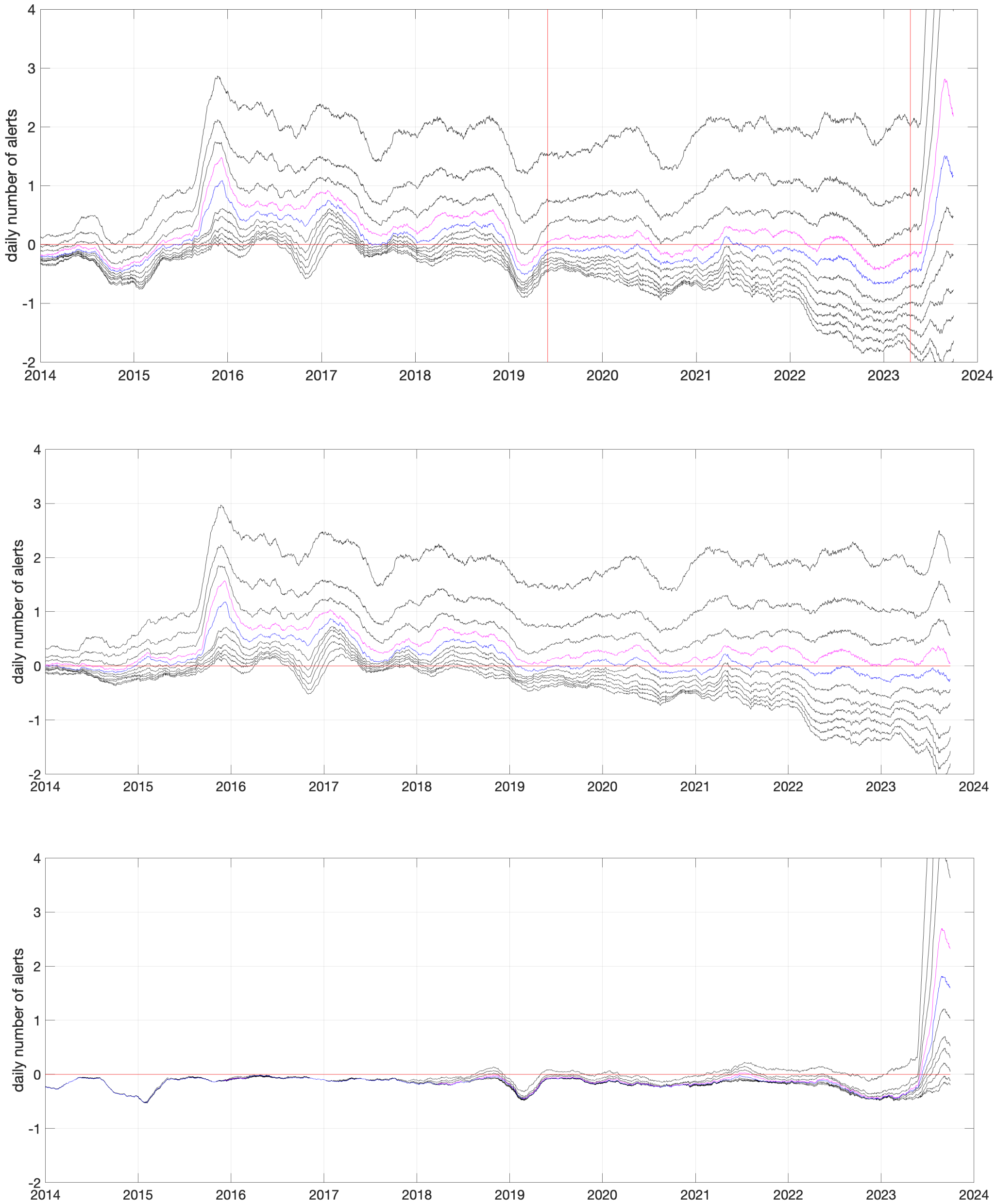

Figure 6: Same as Figure 5 but in increment mode i.e. new fields proposed in this study relative to the operational fields used at CORIOLIS by early Nov 2023. All alerts (top), "good" ones (middle) and "bad" ones (bottom). The horizontal red curve emphasizes the location of the no-increment value. The pink (blue) curve correspond to the widening value 0.3 (0.4). The vertical red lines define the time period over which characteristic values are estimated in the following.

- from 2016 to 2023, a slow decay of the number of "good" alerts, especially for the highest values of the widening factor. Again, we do not try here to explain it and leave it for future work.
- a time period from spring 2019 to spring 2023 where the behaviour is quite stable, and over which we focus our attention in the following.
- a sudden increase in the number of "bad" alerts by mid-2023 when the MinMax detection starts to see profiles that were not included in its reference dataset, while the operational one is updated at most false alerts encountered, see first section of the present Chapter.

For the seek of completeness, we present in Figure 7 similar time series for the amount of temperature detections. The results are far different from those for salinity:

- Without widening, the "bad" detections are more numerous than the good ones. The number of "good" alerts, related to the dataset quality, is rather low, while the number of "bad" ones, related to the method robustness is independent. A first explanation may come from the fact that the ARGO temperature dataset is much cleaner than the salinity one. Another one is likely related to the fact that the performance of the method depends on the typical size of the errors relatively to the typical size of the variability; the structure of the temperature variability is certainly related to its reduced performance.

- the large spread of the curves for the low values of the widening degree suggests that after 2016, the number of erroneous data laying immediately outside of the reference intervals does increase. The reason for such a change remains to be explored.

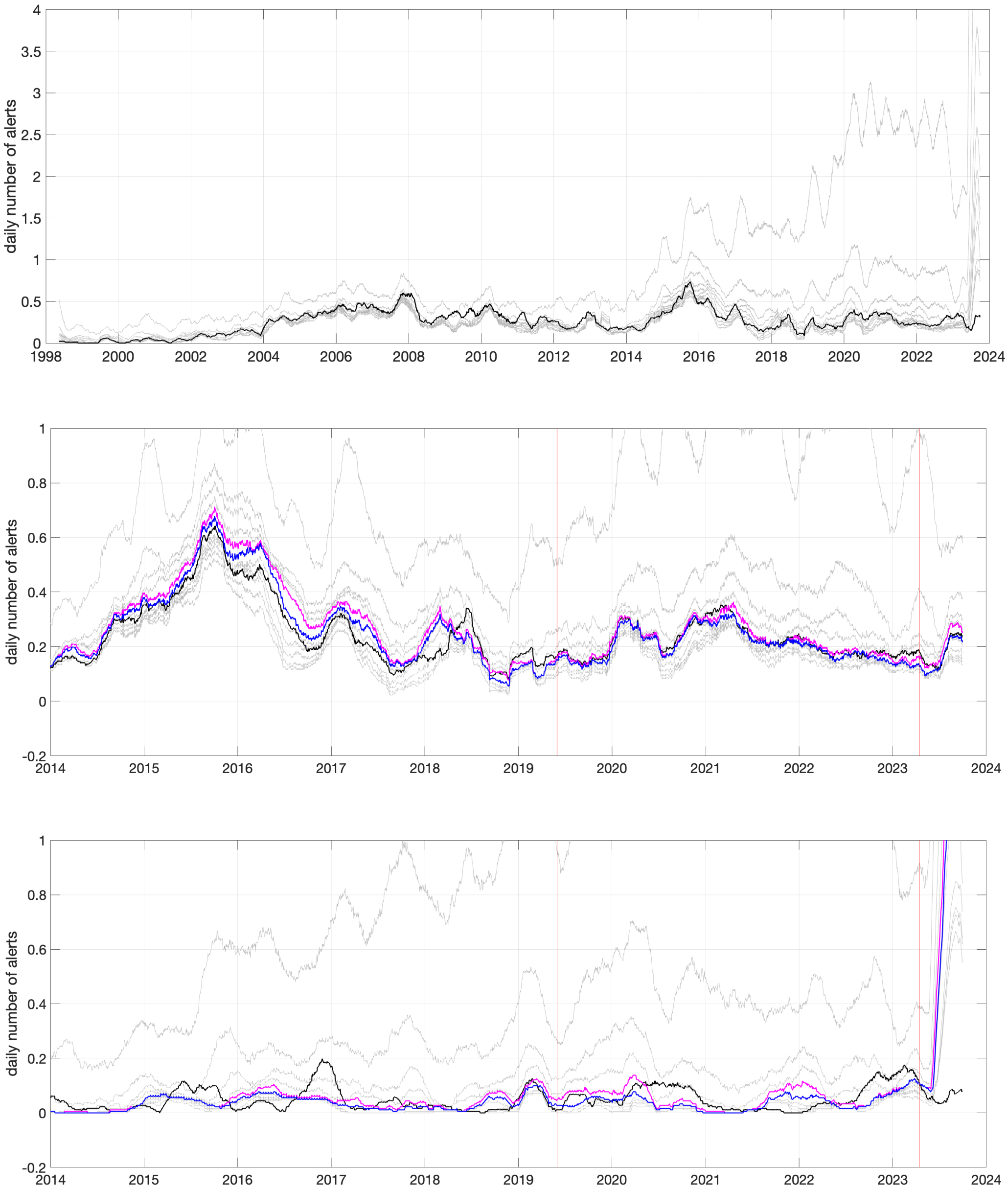

Figure 7: Daily number of temperature profile alert as a function of the observation date. The dataset used for this comparison is the modified CMS HISTORY archive. The grey curves correspond to the new reference field for different widening values from 0 to 2 by step of 0.2. The black curve correspond to the results for the Nov 08th 2023) version of the operational reference fields. Top: all alerts. Middle: good alerts Bottom: bad alerts. In middle and bottom panels, results for specific values of the widening degree are highlighted: 0.8 (pink) and 1.0 (blue).

As a final synthetic view, we now focus on the 2019-2023 period where some "plateau" is observed and materialized in Figures 6 and 7 with the two vertical red lines.

5 Synthesis and Operational choices

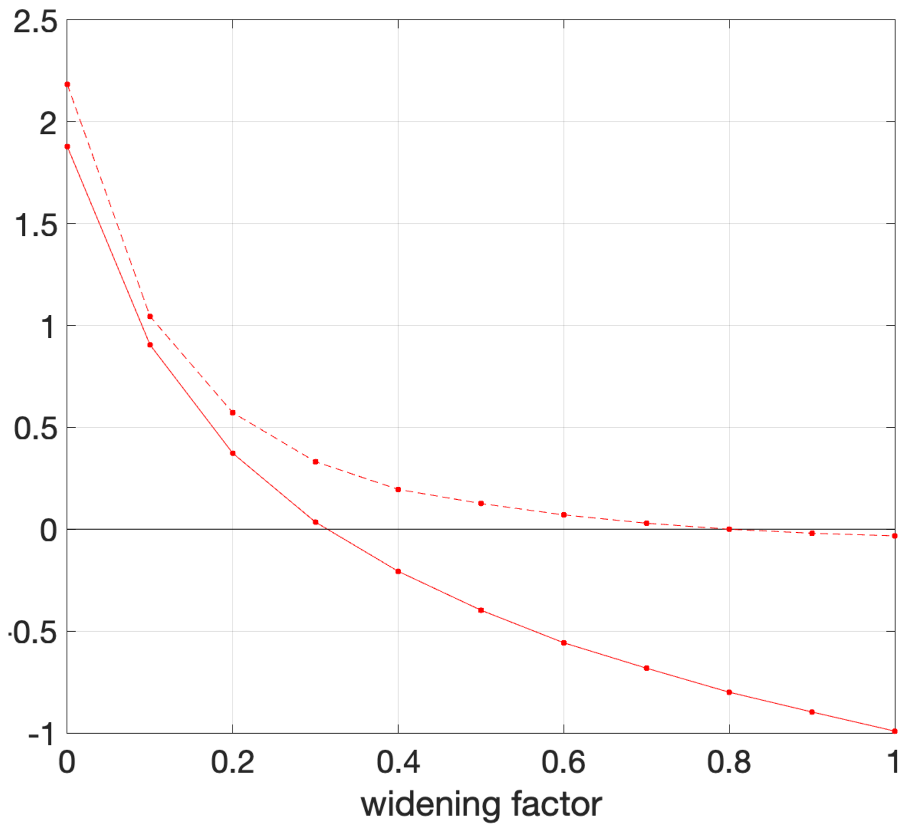

Figure 8: Average relative daily number of profile alerts as a function of the widening degree for Salinity (full line) and Temperature (dashed).

For each widening value i.e. each curve from Figures 6 and 7, we compute the average number of alerts over this "plateau". The results are shown in Figure 8 for salinity and temperature in a relative mode i.e. differences between the new fields and the ones used in operations by early November 2023. It allows to estimate the expected change in operator workload depending on the choice for the widening configuration.

After presentation of the results and discussions between the POKaPOK and SISMER teams, the following widening values have been retained for operations:

- Salinity: $W = 0.3$
- Temperature: $W = 0.5$

They will be used during several weeks after the November 29$^{\text{th}}$ 2023 EiS, the results will be analyzed and the widening possibly adjusted if necessary.

References

[1] Jérôme Gourrion, Tanguy Szekely, Rachel Killick, Breck Owens, Gilles Reverdin, and Bertrand Chapron. Improved statistical method for quality control of hydrographic observations. Journal of Atmospheric and Oceanic Technology, 37(5):789–806, 2020.