PERFORMANCE ASSESSMENT AND IMPLEMENTATION PLAN FOR A NEW DMQC METHOD BASED ON MACHINE LEARNING (FOR TEMPERATURE AND SALINITY)

Ref.: D2.4_V1.2

Date: 14/12/2021

Euro-Argo Research Infrastructure Sustainability and Enhancement Project (EA RISE Project) - 824131

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 824131.
Call INFRADEV-03-2018-2019: Individual support to ESFRI and other world-class research infrastructures

Disclaimer:

This Deliverable reflects only the author's views and the European Commission is not responsible for any use that may be made of the information contained therein.

Document Reference

Project	**Euro-Argo RISE - 824131**
Deliverable number	D.2.4
Deliverable title	Performance assessment and implementation plan for a new DMQC method based on machine learning (for temperature and salinity)
Description	DMQC-PCM method uses a statistical classifier (a PCM: Profile Classification Model) to organise and select more appropriately reference data for the quality control of an Argo float.
Work Package number	WP2
Work Package title	Improvement of the Core Argo mission
Lead Institute	Ifremer
Lead authors	Andrea Garcia Juan
Contributors	Guillaume Maze, Kevin Balem, Cécile Cabanes, Birgit Klein, Romain Cancouët
Submission date	14/12/2021
Due date	[M36] - 31/12/2021
Comments	
Accepted by	Ingrid A. Benavides and Birgit Klein

Document History

Version	Issue Date	Author	Comments
1.0	06/12/2021	Andrea Garcia Juan	
1.1	13/12/2021	Andrea Garcia Juan	Integration of WP2 partners corrections
1.2	14/12/2021	Andrea Garcia Juan	Integration of Euro-Argo Office comments

EXECUTIVE SUMMARY

This deliverable provides a performance assessment and implementation plan for a new quality control method based on machine learning. This method uses a statistical classifier (a PCM: Profile Classification Model) to organise and select more appropriately reference data for the quality control of an Argo float. The PCM based selection is able to distinguish profiles from different dynamical regimes of the ocean (e.g. eddies, fronts, quiescent water masses). Thus, selecting reference data out of the same dynamical regimes as the Argo float data to be quality controlled ensures more robust and relevant reference statistics.

We show that this new DMQC-PCM method is able to improve the detection of salinity drift and temperature or salinity outliers. We further show that the new method, when combined with the standard salinity calibration method, is able to reduce the error on the correction while preserving confidence in this correction amplitude.

We further provide an implementation plan for the DMQC-PCM method. The goal of this implementation plan is to make the method easily accessible to Argo QC operators. It is based on a collection of fully documented Jupyter notebooks demonstrating the use of the method with only open source software already available online freely. We also provide guidelines for the full integration of the new method in existing salinity calibration software.

TABLE OF CONTENT

1 Introduction

The Argo quality control for temperature and salinity is based on quite a simple approach, but relies in practice on a complex implementation. This simple approach is the following: compare data to be validated with historical data already validated and considered correct. The implementation is complex because there are millions of such historical data from all over the world ocean and one has to sub-sample this collection to retain only the most relevant data (so that the comparison is meaningful). In other words, sub-sampling the reference dataset is mandatory to obtain useful statistics but complicated because the ocean dynamic creates a very wide spectrum of structures. We will refer to these varieties as ocean regimes. We understand here that *to retain only the most relevant reference data* is the cornerstone of a reliable quality control, and that relevance is impeded by the diversity of ocean regimes.

Standard quality control methods select reference data by sub-sampling the full reference dataset for profiles in a specified search radius and time periods from the float to control (Owens and Wong, 2009, Cabanes et al, 2016, Cabanes et al, 2021). In practice, this is the easiest implementation. But the issue with this distance criteria is that it can lead to the selection of reference profiles that, even if close to the profile to validate, are not part of the same dynamical regime.

In a frontal region, where a strong current takes place, this means that a space/time distance selection will not distinguish profiles from one side of the front (e.g. with warm and salty water masses) from profiles from the other side of the fronts (e.g. with cold and fresh water masses). On such small (meso) scales, typically a few tens of km, profiles can also be located or surrounded by eddies with very specific water masses. One should distinguish these profiles in the reference dataset and ensure that they match with the dynamical regime of the profiles to validate, a requirement that distance-based selection cannot handle.

Even in larger regions, it can be difficult to select the appropriate reference data. For instance, this is the case where a prominent flow like the Antarctic Circumpolar Current encounters high topographic features, islands, like the Kerguelen plateau. The flow is diverted meridionally over large distances and water masses start to depart from the upstream characteristics. But the Argo salinity calibration method (Cabanes et al, 2021) is based on a certain reference data selection distance. Thus, in this case the reference data selection will not distinguish profiles from the different dynamical regimes created by the split of the ocean flow around high topographic features (Rosso et al, 2020).

Mixing multiple ocean regimes ultimately leads to the statistics from the reference dataset to be biased and/or with a large spread that jeopardise the quality assessment of an Argo profile. Maze et al, 2017 introduced a new methodology to distinguish ocean profiles from different dynamical regimes and Maze et al, 2017 illustrated how it could be used in Argo quality control (see Figure 4 in there). This method is coined “Profile Classification Model” (PCM). It is based on the unsupervised classification of temperature and salinity profiles with a Gaussian mixture model. A PCM allows the user to automatically assemble ocean profiles in clusters according to their vertical structure similarities. It provides an unsupervised, i.e. automatic, method to distinguish profiles from different ocean regimes, precisely what needs to be done to improve existing Argo procedures.

Within task 2.4 of the Euro-Argo RISE project, we analysed this methodology and we provide here a complete performance assessment and implementation plan in Argo DMQC procedures.

2 Methodology: a PCM in DMQC procedures

In the current implementation of Argo DMQC procedures, reference data selection is based on a space, and also time, distance criteria. So, for a profile to quality control that was sampled at (x_0, y_0), one will compute the following metric for all profiles in the reference dataset:

$$D(x,y) = \sqrt{(x - x_0)^2 + (y - y_0)^2}$$

and select only the subset where $D(x,y) \leq D_{max}$ with D_{max} usually on the order of 10 or 20 degrees of latitude/longitude. This criteria can be refined to distinguish longitudinal from meridional distances.

We lay out in the introduction the possible irrelevance of such a selection where different ocean structures emerge from the complex ocean dynamic. This is ultimately due to the fact that the distance $D(x,y)$ does not know anything about the ocean dynamic.

One way to consider ocean regimes is to use a distance that is not based on the position of profiles in the ocean but that is rather based on the similarities of profiles. Indeed, ocean regimes are precisely defined by a specific vertical stack of alternating water masses and sharp transition layers (e.g. thermoclines). Maze et al, 2017 have shown that to use such a distance to classify profiles leads to the association of profile classes with ocean regimes, which then naturally emerge as coherent structures in space and time. The reader is invited to look at Maze et al, 2017 for the technical details of the PCM methodology.

Here, we take advantage of this method to improve the selection of reference data. We propose to combine the two approaches with the following sequential method:

- sub-sampling reference data using a distance criteria;
- classifying this subset of reference data, i.e. create a PCM of reference data, PCM_{ref};
- classifying the profile to quality control with the PCM_{ref};
- sub-sampling again the subset of reference data to profiles from the same ocean regime (class) as the profile to quality control.

Note that we make it obvious here that the improved relevance of reference data with the new selection method will come at the expense of a smaller collection, since we're adding a second sub-sampling step to the standard one.

2.1 Preliminary implementation

A preliminary implementation of the PCM profile selection method has been developed, including a classification notebook and a new option added to OWC software. The implementation workflow is presented in figure 1.

The reference profile classification is done using a Jupyter notebook (*Classif_ArgoReferenceDatabase.ipynb*), which works with the OWC configuration file (*ow_config.txt*) and the argo reference database as inputs. From the *ow_config.txt* we obtain the path to the argo reference database and the distance scales needed to select the reference profiles used for the classification. The classification notebook provides detailed descriptions for each step, so the user only needs to follow the instructions.

The path to the OWC configuration file, the float reference number, the max depth and the number of classes are the information needed to run the notebook.

Max depth. The PCM can not deal with NaN values, so the reference dataset is interpolated on standard depth levels and the profiles shallower than the max depth are dropped out. A max depth of 1000 m can be enough, however the user should find a compromise between keeping a sufficient number of reference profiles and having a comprehensive representation of the oceanography in the region. The user should also consider the depth of the float profiles: if they are shallower than the max depth, they will be dropped out, and they will not be classified by the PCM. In such cases, a lower value of max depth is recommended.

Number of classes. The user is invited to try different numbers of classes and to evaluate if the results are coherent with her/his knowledge of possible ocean regimes in the region. It is also possible to use the BIC notebook (*BIC_plot.ipynb*), which can help to choose the optimal number of classes K. It computes the BIC (Bayesian Information Criteria), fitting the model to the training dataset for a range of K values from 0 to 20. A minimum in the BIC curve (see exemple in figure 2) will indicate the optimal number of classes to be used.

Once the above input parameters are set, the reference dataset is sub-sampled using ellipses shaped with the longitude and latitude scales provided in the *ow_config.txt* file, in the same way as OWC software does. The PCM is trained using this reference data and the prediction of classes is also applied to the float profiles.

Notebook's outputs include the classification figures, the trained model (in a netCDF file for other user sharing and allows for the reproducibility of the analysis) and a txt file containing the classification labels for each profile. This txt file will be used by the OWC software to apply the additional reference profile selection.

In order to use the classification in OWC:

- the USE_PCM option should be set to 1 in the *ow_config.txt* file and
- the path to the classification labels txt file should be provided in the PCM_FILE variable.

For each float profile, OWC chooses reference profiles as usual: using the spatial and temporal scales provided in the configuration file. After that, the classification labels txt is loaded, and only the reference profiles that are of the same class as the float profile are selected. Using this method implies that the number of profiles used to calculate the correction decreases. If a float profile is not classified (because it is shallower than the max depth, for example), all profiles chosen by OWC are selected.

Finally, the OWC salinity correction computation method is not modified and can be run as usual (only the content of the reference dataset has been modified to consider PCM information). The OWC output figures do not change.

3 Performance assessment

An analysis has been carried out to determine the performance of the DMQC-PCM method. The method has been tested with floats evolving through two frontal regions (Gulf Stream and Southern Ocean) and through a large-scale turbulent region (Agulhas Current), where we expect the method to be the most instructive. Two categories of impact have been analysed in detail: the impact on the reference profiles selection and the impact on the final calibration computation.

3.1 Impact on reference profile selection: suitability of selected profiles

Front crossing

Float 4900136 has been chosen as an example to illustrate how reference profile selection is impacted by the new method when a float crosses a front. This float had been deployed north to the Gulf Stream and later crossed the front to continue its trajectory eastwards. The Gulf Stream current constitutes a sharp separation from the cold and fresh water masses in the north and the warm and salty water masses in the subtropical south. It is also a region with an important presence of eddies.

A PCM using 4 classes was found appropriate for this region (see figure 3). For both argo and CTD reference data, the water masses in the north and in the south of the front are well differentiated and a frontal class is identified (class 2). In the case of Argo reference data, two seasonal classes are found south of the front: the summer class 0 and the winter class 3 (see figure 4). Northern reference CTD profiles are separated into 2 classes, finding a colder class in the north east.

With both the Argo and CTD reference data classifiers, we found a change in class for cycle 78 (see figure 3.b). Cycle 77 belongs to the northern class (orange), cycle 78 belongs to the frontal class (pink) and cycle 79 belongs to the southern classes (green). This can be interpreted as the Argo float crossing the front in cycle 78. A more detailed analysis was performed for these transitional profiles.

The usual reference profiles chosen by OWC for float cycles 77, 78 and 79 are shown in figures 5.a, 5.b and 5.c. In these figures, three categories (i.e. ocean regimes) of profiles can be distinguished by eyes: a group of saltier profiles, corresponding to the south of the front, a group of fresher profiles, corresponding to the north of the front, and a transitional group, corresponding to frontal profiles. The issue from this usual profiles selection is clearly apparent here: whatever the float position (cycle 77, 78, 79) all ocean regimes are represented despite the obvious fact that each of these float profiles is clearly more similar to a specific regime than another. However, all of them are used to compute the correction in OWC.

The modified reference profiles selections, based on a 4-classes PCM additional sub-sampling are shown in figures 5.d, 5e and 5.f. For each float profile, only the reference profiles with the same class as the studied profile are chosen. For float profile 77, reference profiles correspond to the saltier southern group; for profile 78, only transitional or frontal profiles are selected; and for profile 79, reference profiles came from the fresher northern group.

As OWC only uses a spatial criterion to choose the profiles, it can not distinguish which side of the front they are. The classification is able to differentiate profiles from the north, the south or inside the front, so the class selection method is able to choose more appropriate profiles. We found the same results using CTD data.

Comparison with SAF option

Float 3901928 drifted in the Southern Ocean, eastward, passing near Crozet and Kerguelen islands and crossing some important fronts. It has been chosen as an example to compare the reference profile selection performed by the PCM option with the selection performed by the SAF option available in OWC.

For each float cycle, the SAF (South Antarctic Front) option allows to select reference profiles that are located on the same side of the SAF (north or south), using a climatology of the front position (Cabanes et al 2021). It implies a reduction in reference profile number, as the PCM option does.

A PCM with 4 classes was made to perform the class selection method (see figure 6). For both argo and CTD data, the classification reveals three different fronts: the South Antarctic Circumpolar Current Front (SACCF) in the south (frontier between classes 0 and 3 in figure 6.a), the South Antarctic front (SAF) in the middle (frontier between classes 3 and 2 in figure 6.a), and the South Tropical Front (STF) in the north (frontier between classes 2 and 1 in figure 6.a)(Kim and Orsi, 2014, Pauthenet et al, 2018). As shown in figure 6.b, CTD data is scarce and presents an inhomogeneous spatial distribution, but we can also find the same fronts.

The reduction in the number of reference profiles per float cycle obtained as a result of the PCM option and the SAF option is shown in figure 7. From float cycles 0 to 60, the reduction of reference profiles is more important using the PCM method in both argo and CTD datasets. From float cycles 90 to 140, the number of reference profiles taken into account by the PCM option and the SAF option is very similar. Therefore, two profiles corresponding to these two intervals have been chosen for a more detailed analysis: profile 30 and profile 120 (see figures 8 and 9).

The reference profile selection for float profile 30 is more accurate using the PCM option, compared to distance-based and SAF criteria (see figures 8 and 9). But, for CTD data, it implies a big decrease in the number of reference profiles used to compute the correction. The envelopes of reference profiles selected for float profile 120 are very similar using PCM or SAF option.

Using the PCM option, it is possible to differentiate several water masses and to take into account other fronts than the SAF. During cycles 0 to 60, the float drifted following the South Antarctic Circumpolar Current Front (SACCF), which is not taken into account by the SAF option. Only the reference profiles located south of the SACCF are selected by the PCM option, which implies a bigger reduction in the number of reference profiles. However, the reference profiles chosen by the PCM option are more similar to the float profile, as it is shown for profile 30.

When the float drifts near to the SAF (cycles 90-140), the number of reference profiles selected by the PCM and the SAF option is very similar, and the reference envelopes for profile 120 as well. Thus, for float cycles located near the South Antarctic Front, the PCM option is as good as the SAF option in the selection of reference profiles.

Turbulent region

Float 3901915 has been chosen to evaluate the effect of the PCM profile selection in a region particularly rich in meso-scale features over a very large scale. This float was deployed in the Indian Ocean, near the South African coast and it has drifted westward following the continental shelf for finally crossing to the South Atlantic Ocean.

In this region, the Agulhas Current (AC) retroflexion takes place, shedding warm-core rings to the South Atlantic Ocean. When the AC shifts direction back to the Indian ocean, current meanders can become closed loops, rings, trapping warm and salty Indian Ocean waters and moving independently westward (Lutjeharms & Ansorge, 2001).

A PCM has been made in this region using 6 classes (see figure 10). Two classes are predominant in the Indian Ocean: class 2 and 4 for argo data (classes 0 and 1 for CTD data), and they are characterized by high values of salinity and temperature. Class 2 (class 1 for CTD) seems to follow the Agulhas Current pathway. Some profiles belonging to these Indian Ocean classes appear in the South Atlantic, even in farther western longitudes.

The region between longitudes 11E and 20E has been chosen to make a more detailed analysis, as the Agulhas rings are formed and start a westward trajectory in this place (Feron et al, 1992). Also, it seems to be the most variable zone in terms of PCM classes: at least three different classes are present, not showing a clear stationary spatial coherence (see figure 11).

In this region, the float trajectory changes direction constantly. Profile 72 belongs to class 2 (with argo data, class 1 with CTD data), one of the Indian Ocean classes. Reference profiles selected for float profile 72 are shown in figure 12 and figure 13.

For Argo data, the PCM option only takes into account some reference profiles located near profile 72, and a few located in the east. The variability of the reference profiles decreases significantly, compared with the profiles selected by OWC, which present an important spread. The same results are found for CTD data.

In this turbulent region, where very different water masses are located nearby, the PCM option performs a very suitable reference profile selection, decreasing significantly the reference profiles variability, hence reducing uncertainties in the float profile quality evaluation.

3.2 Impact on calibration properties

This section is dedicated to present how the PCM reference profile selection impacts the final calibration computed by OWC. We analysed both the calibration value and its associated spread, a measurement of the calibration error.

Corrections have been computed using the spatial and time scales proposed by the DMQC cookbook (Cabanes et al, 2021) for the Subpolar North Atlantic (float 4900136) and the Southern Ocean (floats 3901928 and 3901915). Theta levels are chosen below 600 dbar (use_pres_gt = [600] in set_calseries.m) for floats 4900136 and 3901928.

Table 1 summarizes the results for the three tested floats.

Using Argo reference data		
Float WMO	**Correction value**	**Correction spread**
4900136	Decrease for last cycles	Change for last cycles
3901928	No changes	No changes
3901915	No changes	Decrease for turbulent zone

Using ships CTD reference data		
Float WMO	**Correction value**	**Correction spread**
4900136	No changes	No changes
3901928	Changes for first cycles	Decrease for 0-80 and last cycles
3901915	Decrease for turbulent zone	Decrease for turbulent zone

Table 1: Qualitative summary of the PCM reference profile selection impact on OWC calibration estimates.

Drift correction

There are no changes in the correction value for float 4900136 using CTD and for floats 3901928 and 3901915 using argo data. In these cases, even if the PCM method implies an important decrease in the number of reference profiles used to compute the correction (see figure 7), the calibration value is not changing . The DMQC-PCM method reduces the number of reference profiles but decreases its variability, selecting more appropriate profiles.

For the float located in the Gulf Stream, 4900136, a difference in the calibration results with and without the PCM option is observed for the last cycles (see figure 14). Between cycles 86 and 120, the float drifts south of the Gulf Stream front, a region that presents two seasonal patterns within the argo classification (see figures 3 and 4).

The introduction of this seasonality in the classification improves the reference profile selection, using profiles measured in the same season as the float profile and leading to a reduction of the correction values for the last cycles. Using the standard reference profile selection based only on distance to the float profiles, DMQC operators may think that the float is starting to drift, but here, we show that it is not the case.

If profiles are different enough, seasonality can be captured using the PCM classification and taken into account in the reference profiles selection.

An important difference is also found in the correction computed with CTD data for the float 3901915 (Agulhas Current) (see figure 15). This difference takes place between cycles 0 and 80, where the float drifts in the Indian ocean and crosses to the South Atlantic, passing through the turbulent zone

(latitudes 10E to 20E). If we look at the peak around cycle 40, the value of the correction decreases from -0.8 to -0.4 using the PCM option. The correction spread also decreases in this zone.

Having a good profile selection in this area is very important as the variability of reference profiles is very high (see figures 12 and 13). The PCM option decreases the variability of reference profiles, having an important impact on the calibration. Correction value using PCM is more reliable because it is based on a better selection of the reference profiles.

Calibration errors

A difference in the calibration spread for the first 80 cycles is observed for float 3901920 (Southern Ocean), using CTD reference data (see figure 16). For these cycles the float drifts near the limit between classes 1 and 2 (see figure 6), corresponding to the SACCF. The SAF option does not provoque any difference in the correction.

The number of reference profiles is significantly reduced for the first 80 cycles (see figure 7), but it leads to an important reduction in the correction spread, as the variability of the reference profiles has been reduced. The DMQC operator would have no doubt that it is not necessary to apply a correction for the first 80 cycles using the PCM option.

A clear change in the correction spread is observed for float 3901915 (Agulhas Current) (see figure 17) between cycles 30 to 70, corresponding to the floating navigating through the most turbulent region of the area. As in the exemple in the Southern Ocean, even if the number of reference profiles is reduced, the correction spread decreases because the reference profile selection is improved using the PCM. And it works especially well in a turbulent zone where very different water masses are close in space.

3.3 Short summary

The DMQC-PCM method has been tested with floats evolving through two frontal regions (Gulf Stream Extension and Antarctic Circumpolar Current) and through a large-scale turbulent region (Agulhas Current). In all cases, the reference profile selection is improved using the classification, which leads to a more reliable quality control.

In frontal regions, profiles from one side of the front are differentiated from profiles on the other side of the front, so the PCM selection method is able to choose reference profiles belonging to the same oceanographic regime. As a consequence, the variability in reference profiles is reduced. It is also the case in the turbulent region, where very different water masses are located nearby. The PCM option performs a very suitable reference profile selection, decreasing, again, the variability in reference profiles.

We also showed that the PCM reference profile selection is consistent with the profiles selected by the OWC SAF option when the float is near the South Antarctic Front. In addition, the reference profile selection is improved when the float navigates through other ocean regimes because the classification is able to distinguish other fronts than the SAF (Rosso et al, 2020, Jones et al, 2019).

In some cases, the PCM profile selection has no impact on the OWC final calibration. OWC computes the calibration using the 10 theta levels that present the minimum variability. These levels are usually in the deepest part of the profiles. So, sometimes even if the global reference profiles variability is reduced, it may have no impact on the choice of these theta levels. In such cases, the PCM profile selection does not have a great impact on the final correction but it increases the confidence in the reference profiles used for computing the correction. And we are able to get the same results reducing the number of reference profiles.

In other cases, the correction value has been reduced. We think that, in such cases, OWC is over correcting because of the unsuitability of the distance-based reference profiles selection. Correction value using PCM is more reliable because it is based on a better selection of the reference profiles.

4 Implementation plan

The preliminary implementation of the DMQC-PCM method is available for the project partners in the github repository (https://github.com/euroargodev/DMQC-PCM). It includes the jupyter notebooks used to make the classification of the reference profiles and the version of OWC matlab containing the new PCM option.

The implementation plan is organised in different steps:

- test of the preliminary implementation by Euro-Argo RISE partners and DMQC operators in specific regions (link with WP5 task 5.3, Regional data quality assessment in the Southern Ocean);
- analysis and inclusion of the partners feedback in the final implementation;
- presentation to and validation of the method by the ADMT (2022 meeting);
- distribution of the final code including the new PCM option.

The preliminary implementation performance assessment included in this deliverable gives us a good idea of the potentialities of this method. However, more tests carried out by specialised DMQC operators can give us a more precise understanding of method performance and limits. Testing the PCM method with floats or regions where DMQC operators are used to encounter difficulties to choose the appropriate correction, will help us to identify in which cases it is the most useful.

The DMQC operators feedback will be analysed and integrated to the DMQC-PCM methodology and code if necessary.

After the testing and correction phase, the DMQC-PCM method will be ready to be presented in the ADMT 23 meeting (Argo Data Management Team meeting), for validation within the international argo community. After this validation, the code will be publicly distributed as a github release.

5 Conclusions and future work

The DMQC-PCM method improves the reference profile selection in OWC, selecting reference profiles that are in the same oceanographic regime as the float profile we want to qualify. It leads to a reduction in the variability of reference profiles that can impact the final correction computed by

OWC. The correction obtained using the PCM option is more relevant because it is based on a more coherent selection of the reference profiles.

The DMQC-PCM method performance assessment carried out in this deliverable provides the detailed reasons for its opportunity. Obviously, more tests should be carried to better understand the performance and limits of the method in other regions and cases, as it is proposed in the implementation plan.

Future work

The preliminary implementation of the DMQC-PCM method is based only on discret class labels, selecting profiles in the same class and reducing the number of reference profiles used for the correction computation. Another interesting possibility we would like to explore is the use of the probability of a profile to belong to a class. As the PCM is based on a GMM (Gaussian Mixture Models, see Maze et al, 2017), it is possible to calculate the probability of a profile to be in each of the classes. With this information, we could calculate a PCM-based weight for each reference profile linked to the probability of belonging to the float profile class. Hence, the number of profiles used to calculate the correction would not be reduced but rather their importance weighted according to similarities with each ocean regimes the float profiles would be encountering.

Figures

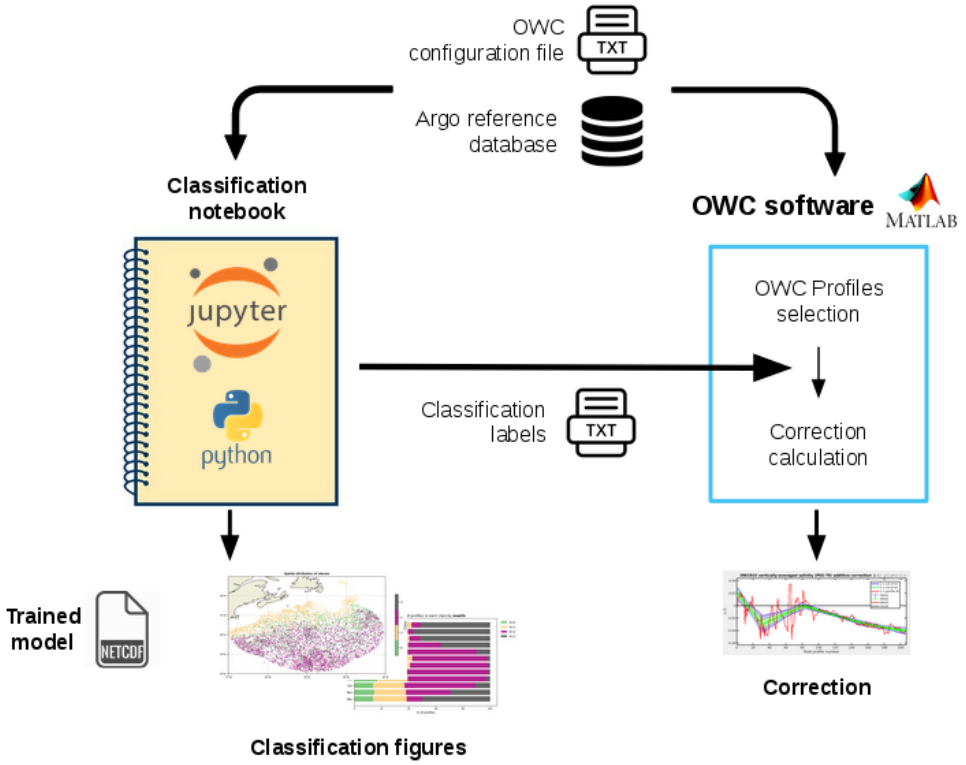

Figure 1. Workflow of the DMQC-PCM preliminary implementation.

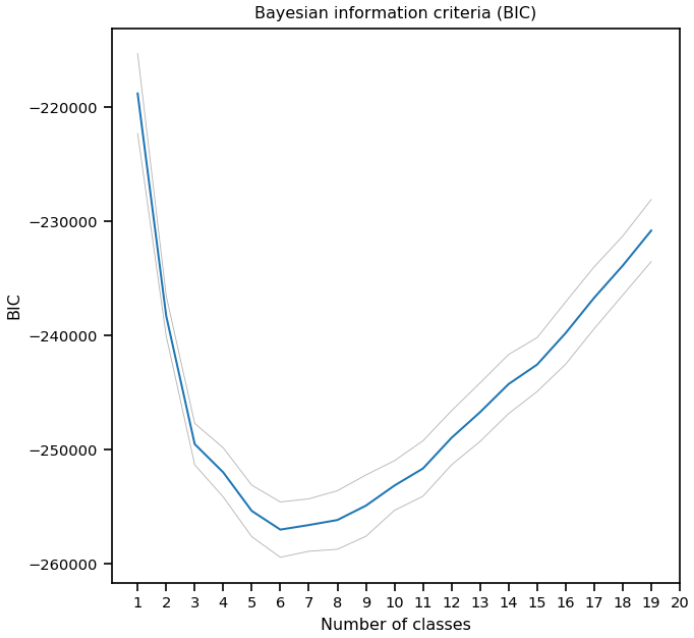

Figure 2. Example of BIC plot.

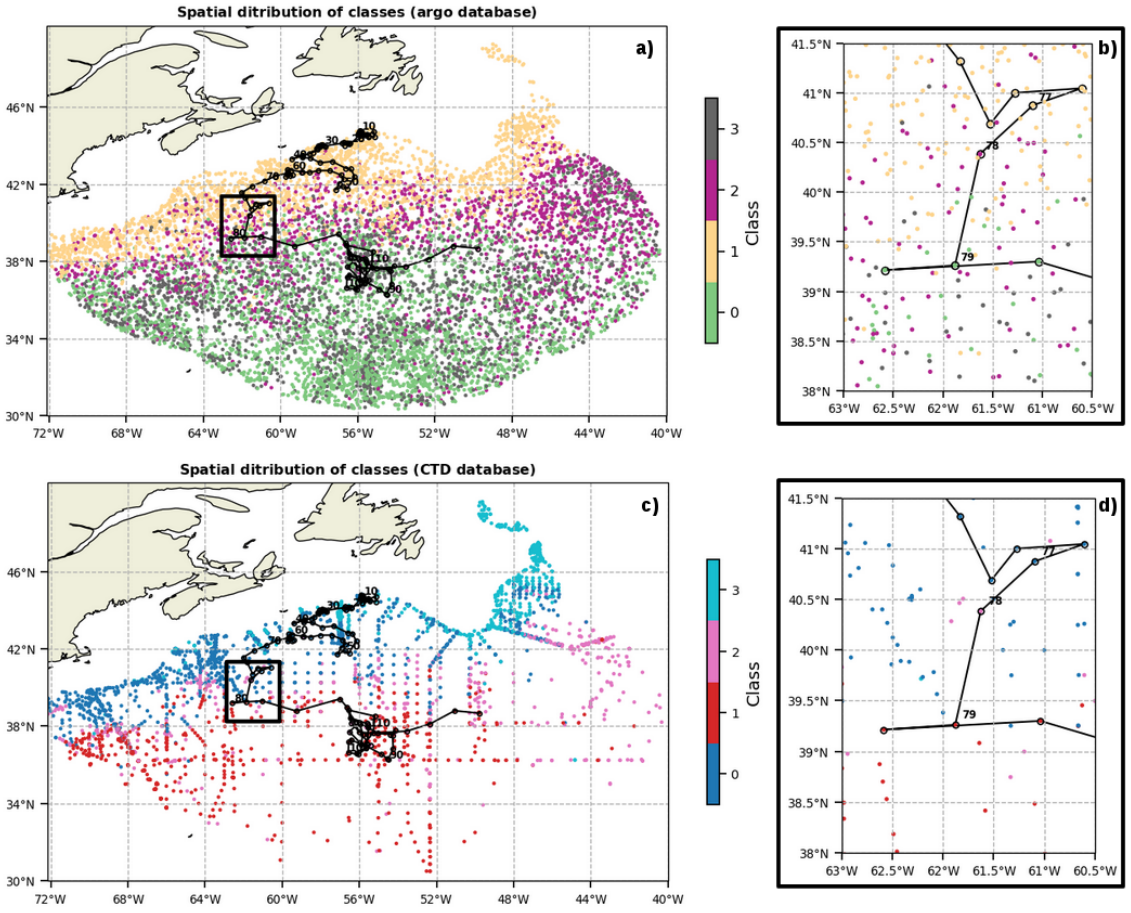

Figure 3. Reference data classification for float 4900136 using 4 classes. Reference argo database in a) and zoom to the turbulent zone in b). Reference CTD database in c) and zoom to the turbulent zone in d).

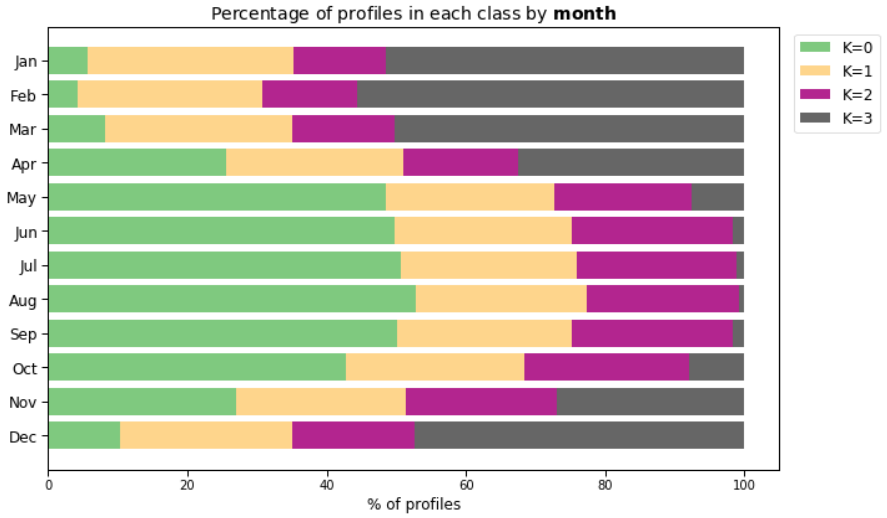

Figure 4. Temporal distribution of classes using Argo reference data for float 4900136.

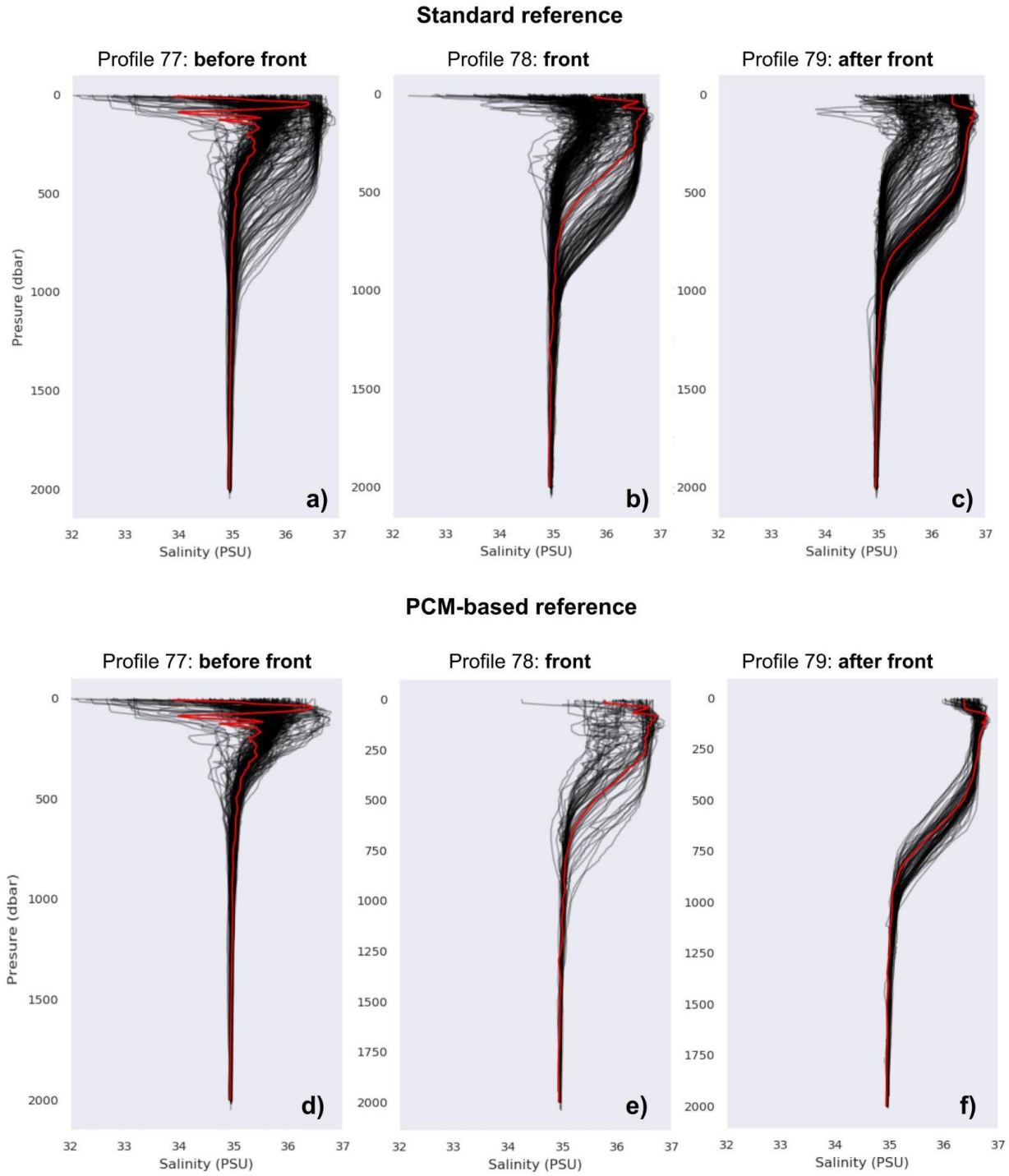

Figure 5. Reference profiles (in black) selected for float profiles 77, 78, 79 (in red) using OWC standard selection [a), b) and c)] and using PCM based selection [d), e) and f)], for float 4900136.

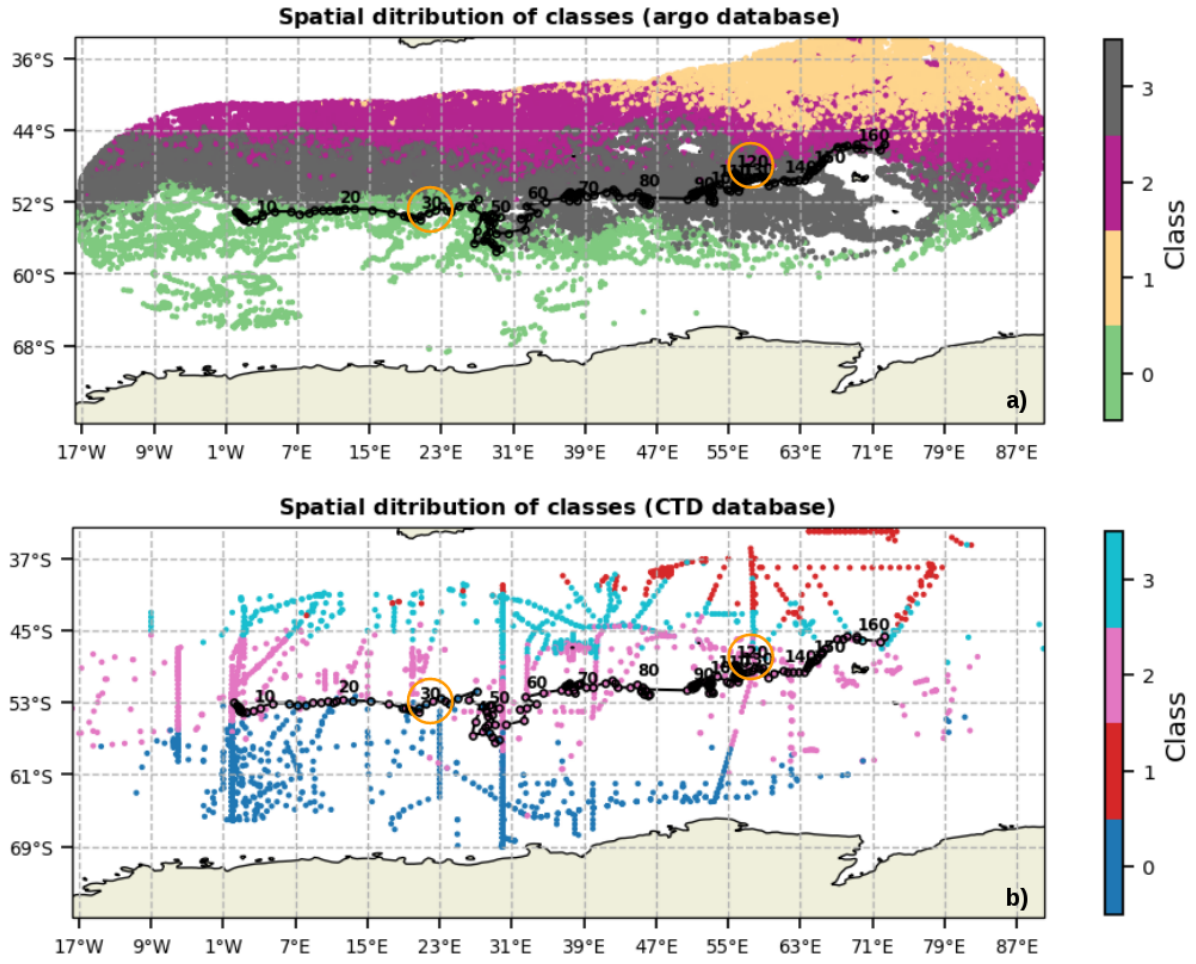

Figure 6. Reference data classification for float 3901928 using 4 classes, with reference argo database a) and reference CTD database b).

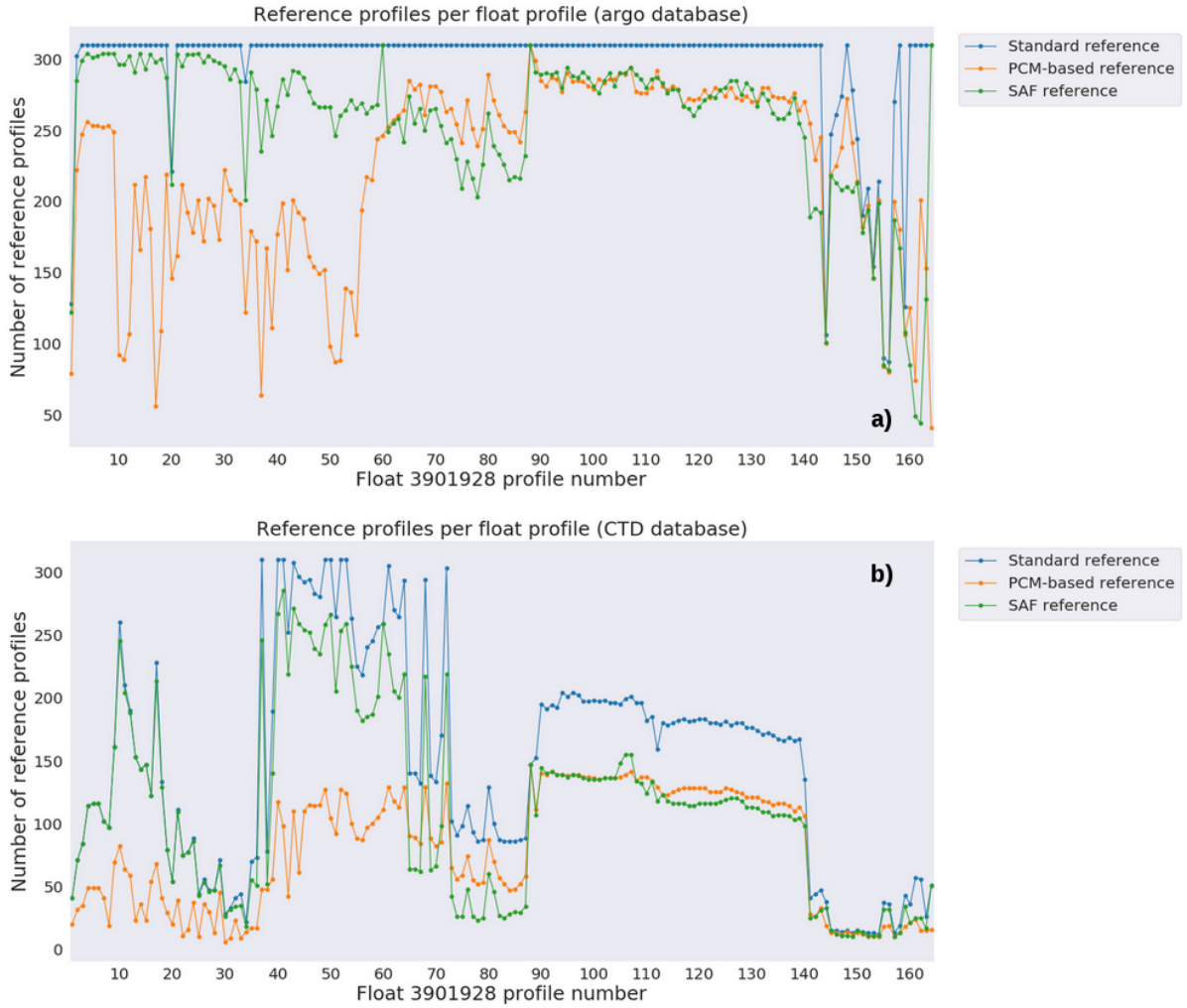

Figure 7. Number of reference profiles per float cycle in float 3901928 correction, using Argo reference database a) and CTD reference database b).

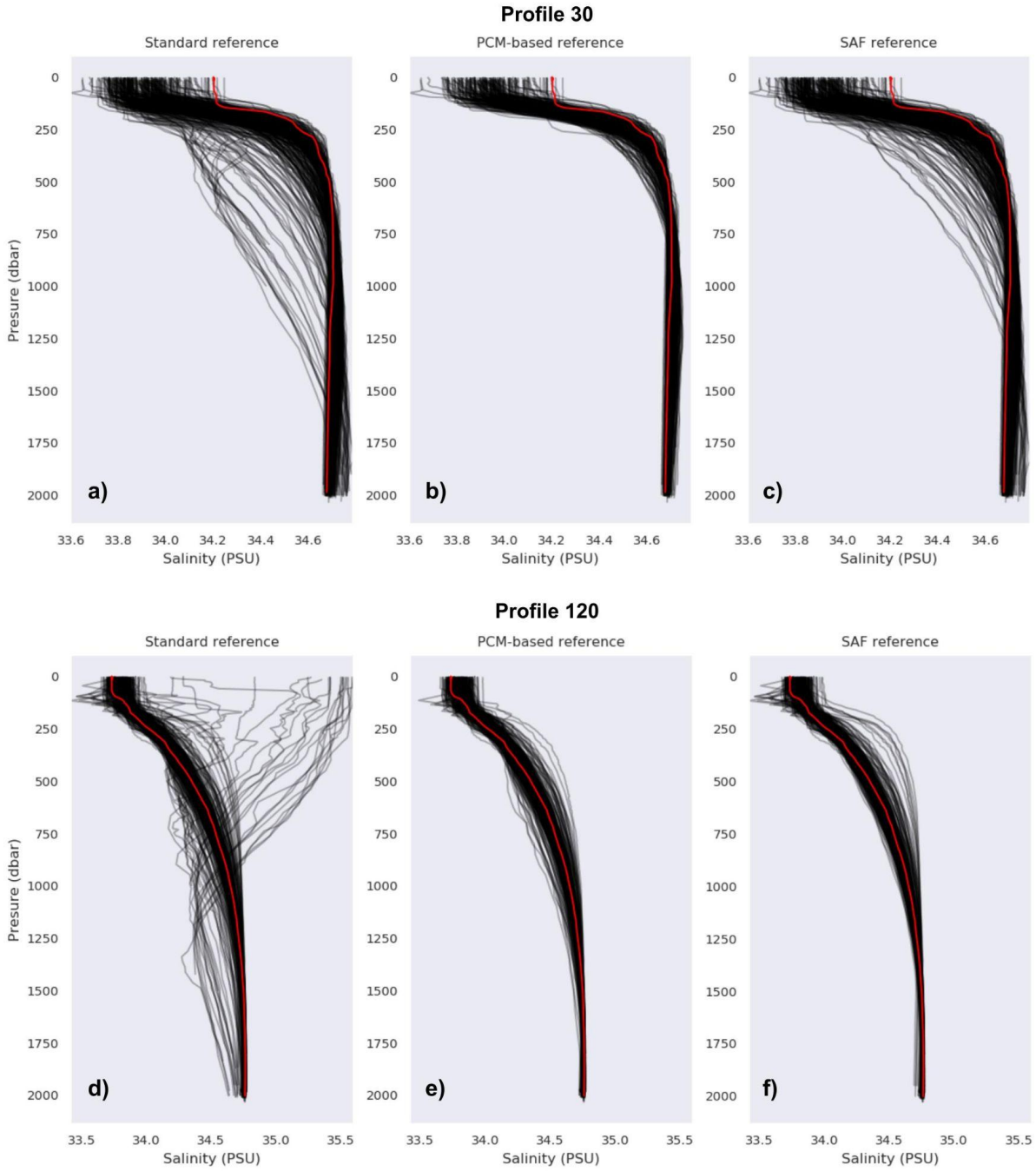

Figure 8. Argo database reference profiles (in black) selected for float 3901928 profiles 30 [a), b) and c)] and 120 [d), e) and f)], using standard, PCM-based and SAF references.

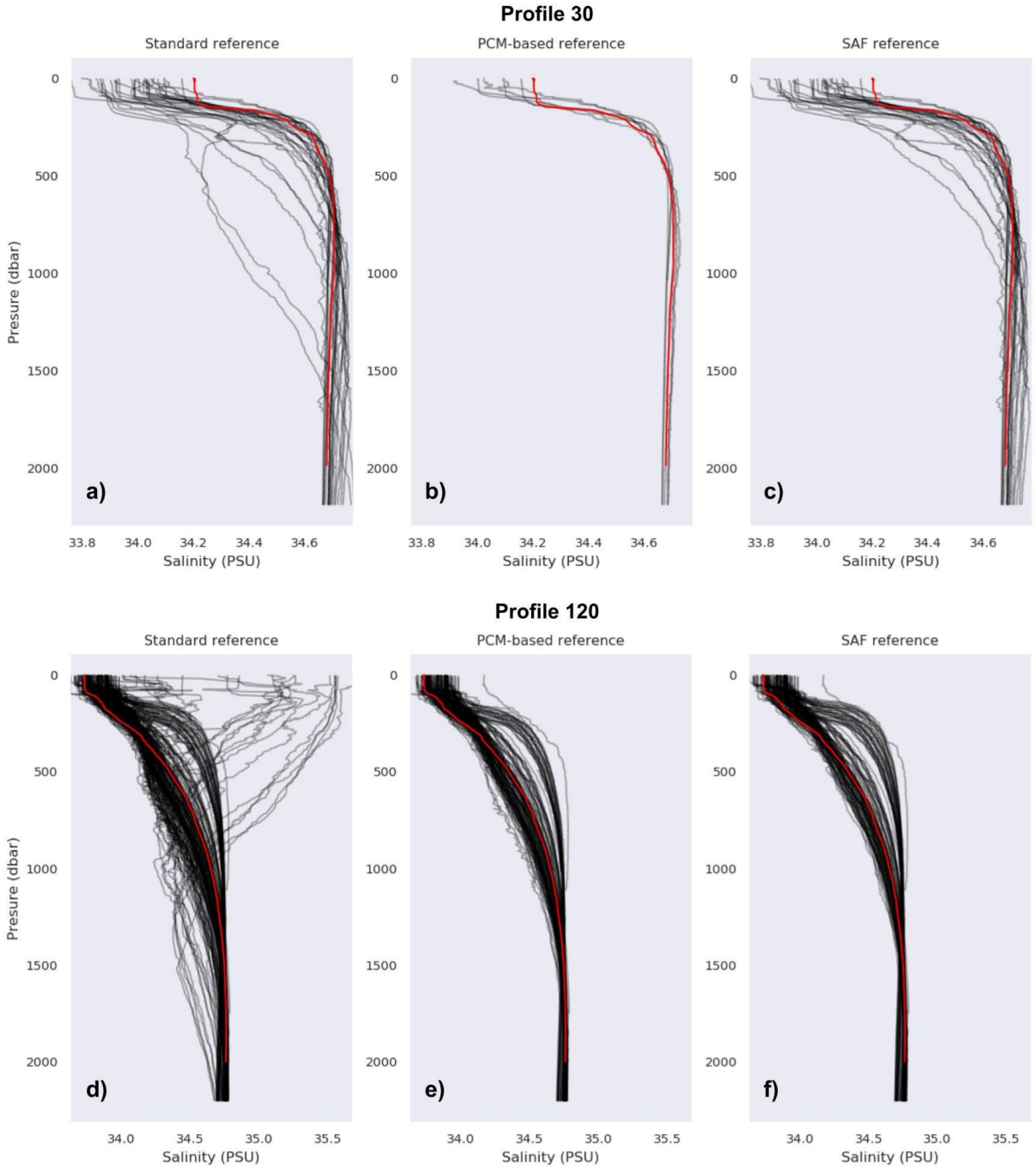

Figure 9. CTD database reference profiles (in black) selected for float 3901928 profiles 30 [a), b) and c)] and 120 [d), e) and f)], using standard, PCM-based and SAF references.

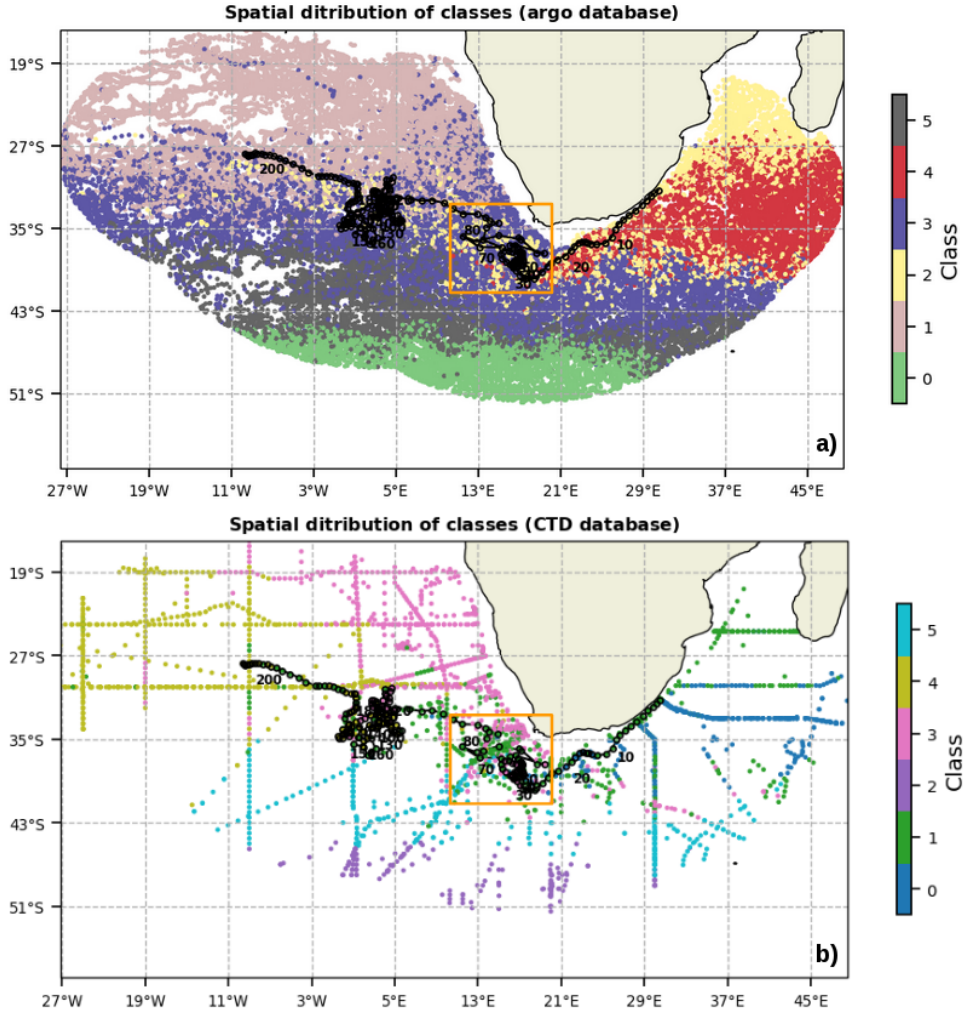

Figure 10. Reference data classification for float 3901915 using 6 classes, with argo reference database a) and CTD reference database b).

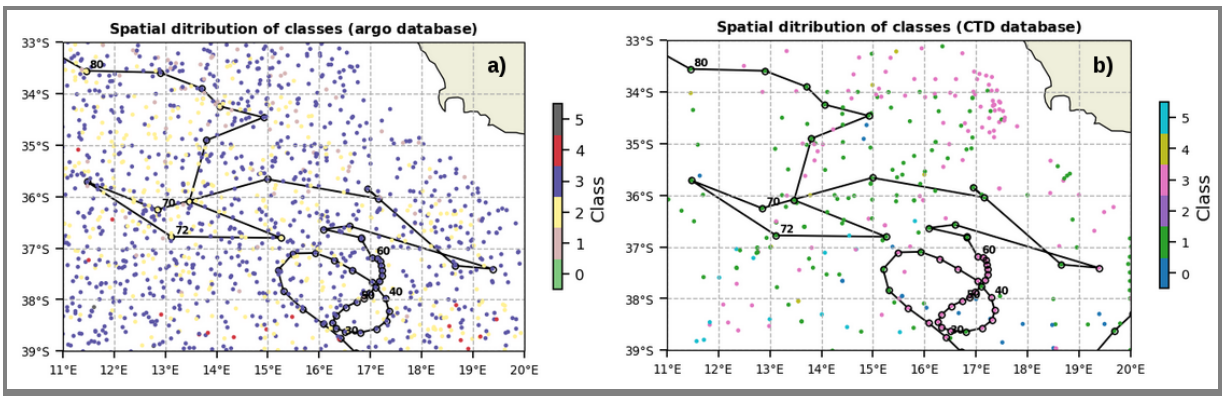

Figure 11. Zoom to the turbulent region in reference data classification for float 3901915, using argo reference database a) and CTD reference database b).

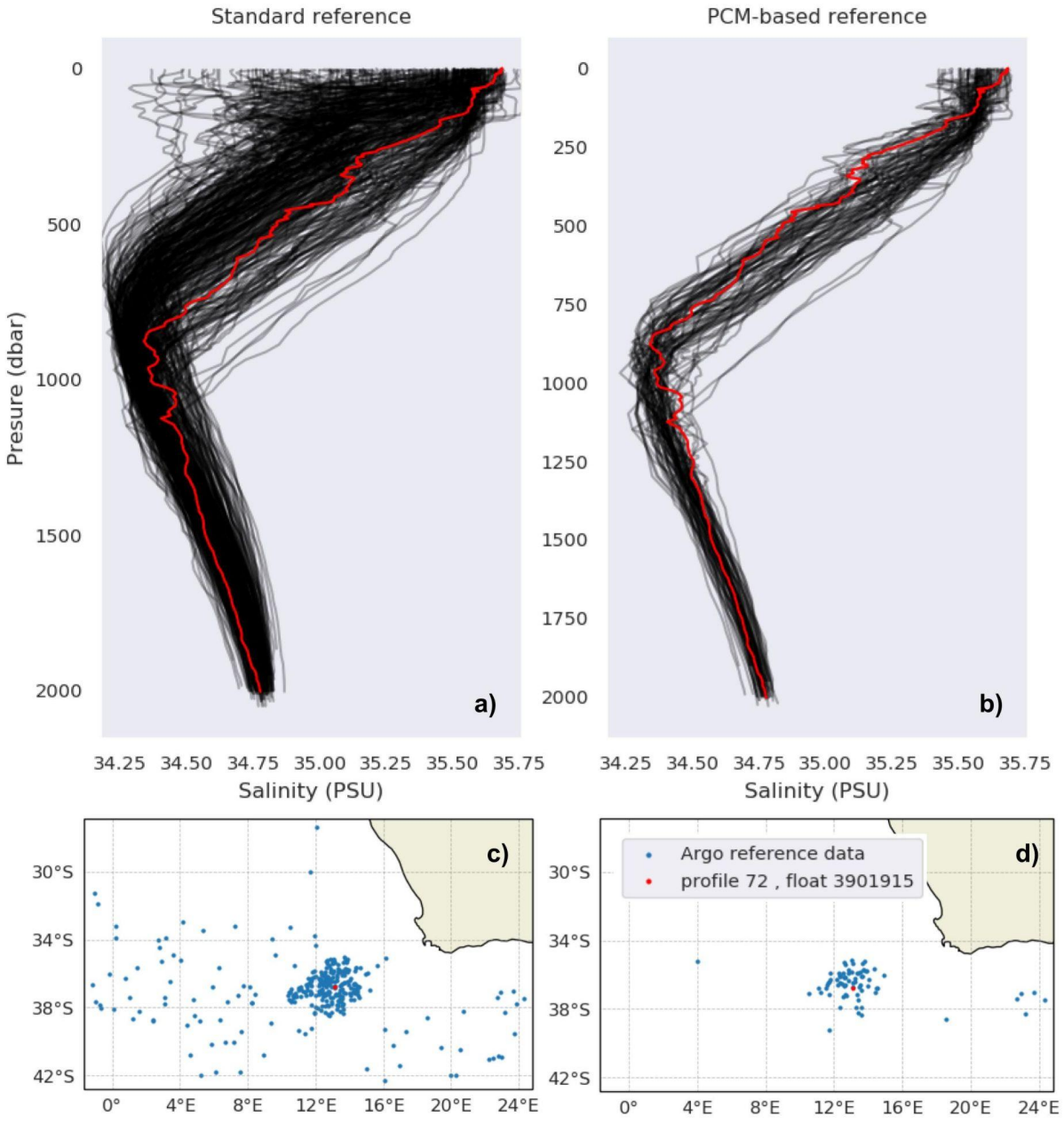

Figure 12. Argo database reference profiles (in black) selected for float 3901915 profile 72 using distance-based selection a) and PCM-based selection b), and its spatial distribution c) and d).

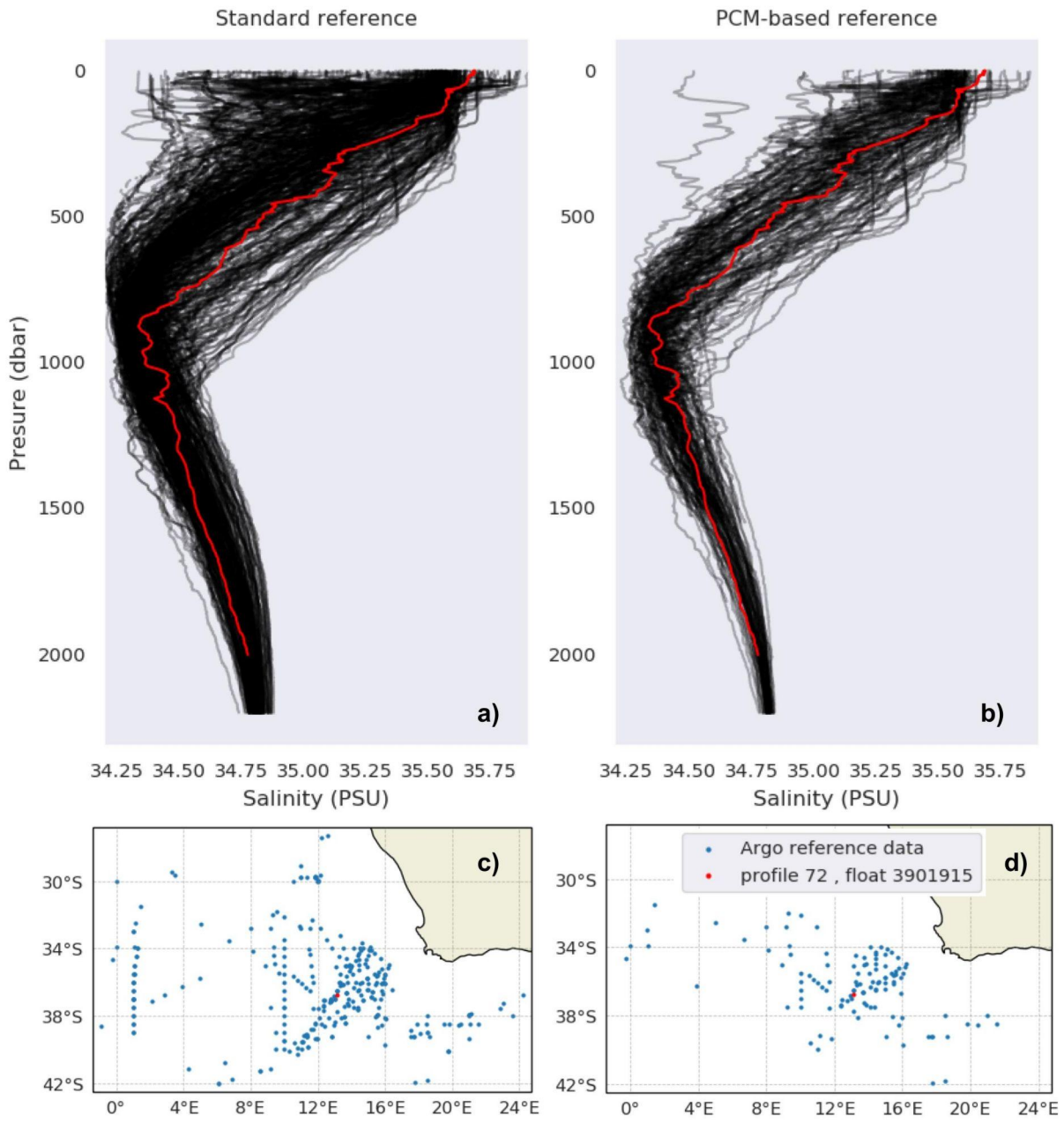

Figure 13. CTD database reference profiles (in black) selected for float 3901915 profile 72 using distance-based selection a) and PCM-based selection b), and its spatial distribution c) and d).

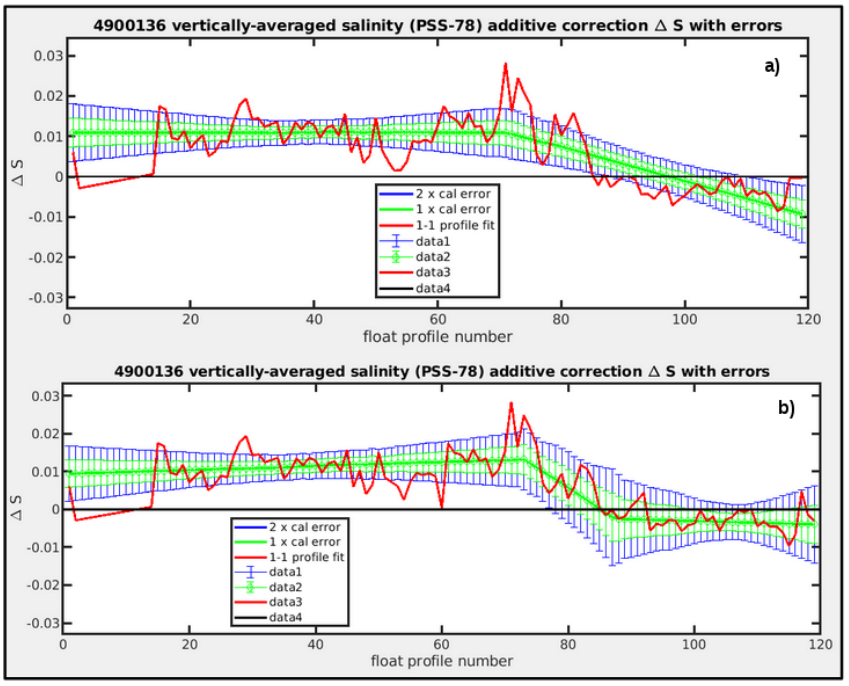

Figure 14. Correction computed for float 4900136 with argo reference data, using OWC distance-based reference profiles selection a) and DMQC-PCM method b).

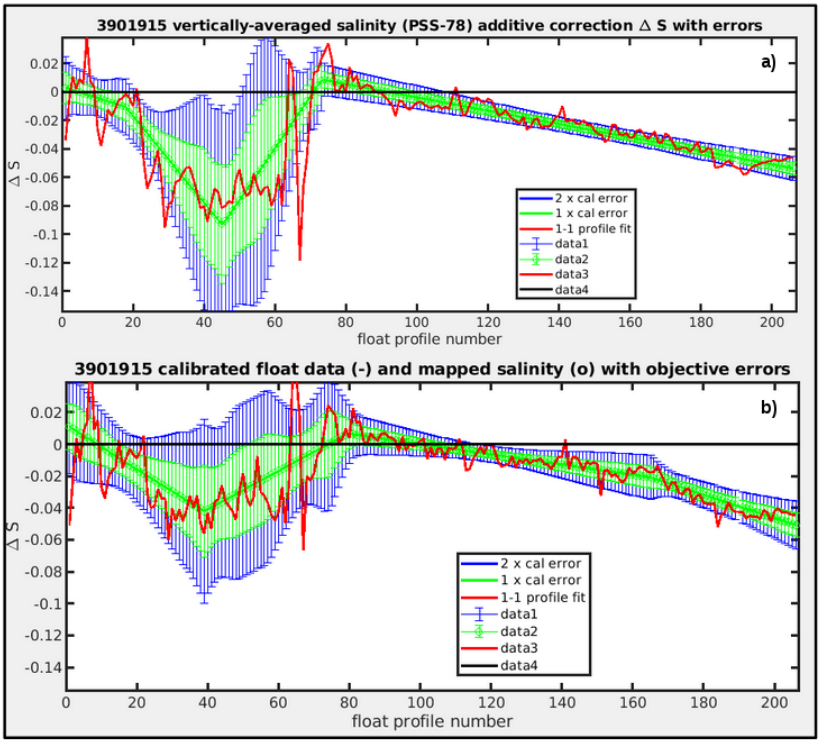

Figure 15. Correction computed for float 3901915 with CTD reference data, using OWC distance-based reference profiles selection a) and DMQC-PCM method b).

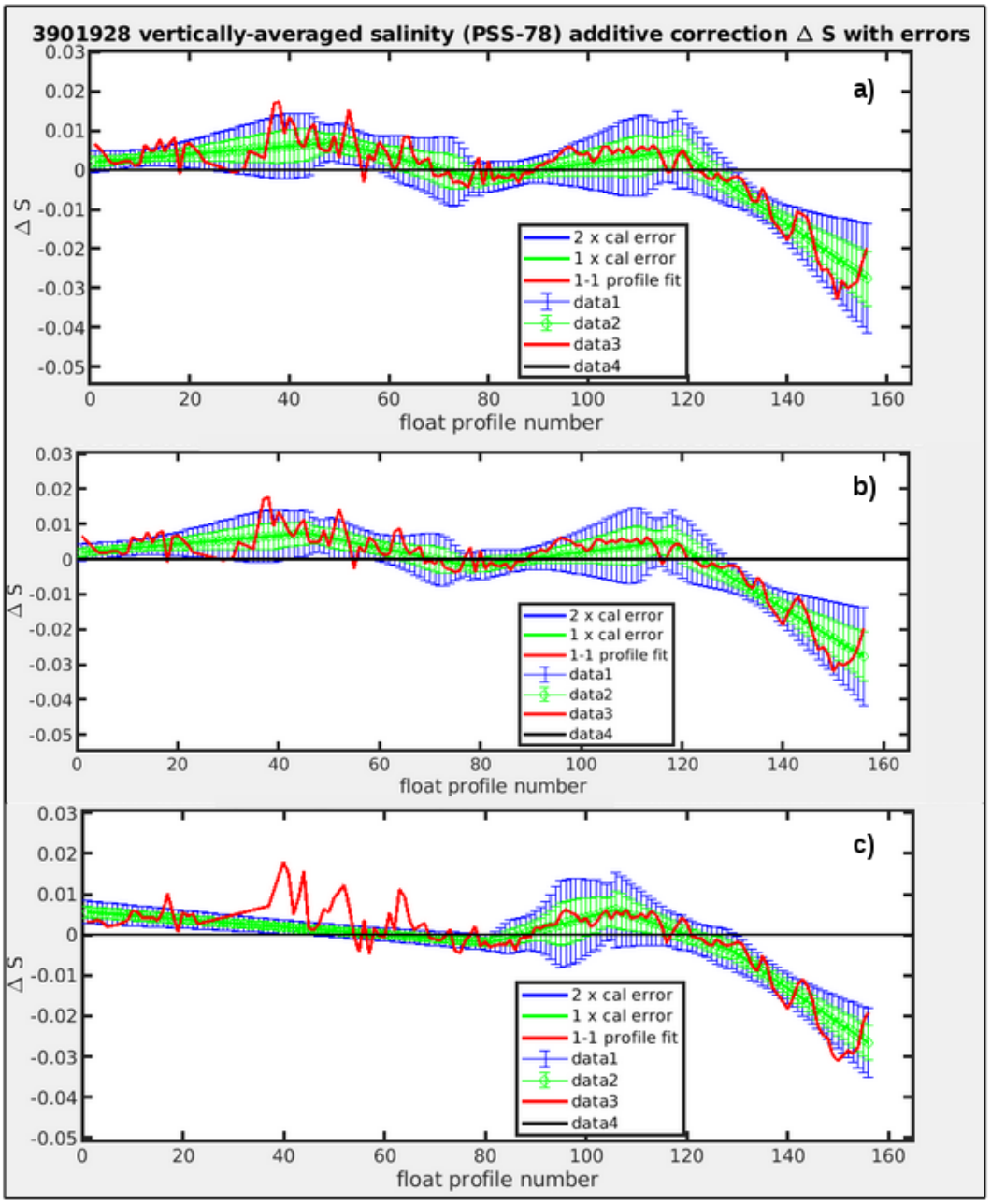

Figure 16. Correction computed for float 3901928 with CTD reference data, using OWC distance-based reference profiles selection a) SAF option b) and DMQC-PCM method c).

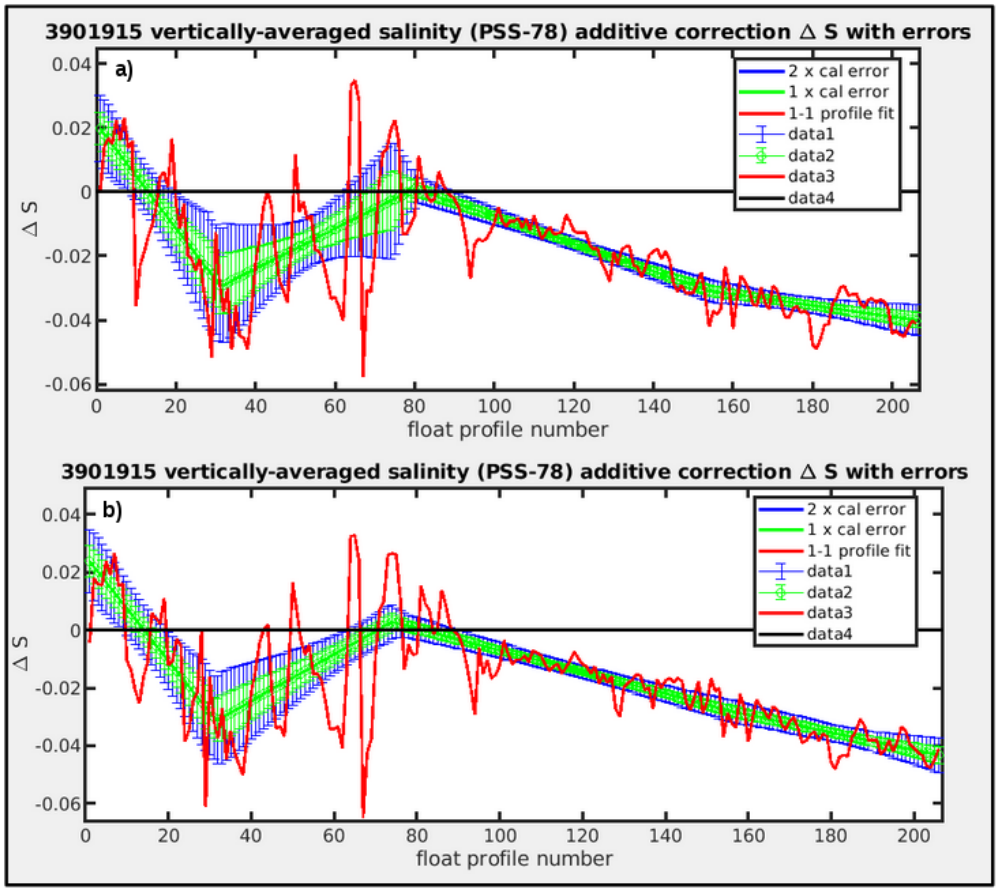

Figure 17. Correction computed for float 3901915 with Argo reference data, using OWC distance-based reference profiles selection a) and DMQC-PCM method b).

References

Cabanes C., I. Angel-Benavides, J. Buck, C. Coatanoan, D. Dobler, G. Herbert, B. Klein, G. Maze, G. Notarstefano, B. Owens, V. Thierry, K. Walicka, A. Wong, 2021: **DMQC Cookbook for Core Argo parameters**. https://doi.org/10.13155/78994

Owens W.B., Wong A.P.S., 2009: **An improved calibration method for the drift of the conductivity sensor on autonomous CTD profiling floats by θ–S climatology**, Deep Sea Research Part I: Oceanographic Research Papers, 56(3), 450-457. https://doi.org/10.1016/j.dsr.2008.09.008

Cabanes C., V. Thierry, C. Lagadec, 2016: **Improvement of bias detection in Argo float conductivity sensors and its application in the North Atlantic**, Deep Sea Research Part I: Oceanographic Research Papers, 114, 128-136, https://doi.org/10.1016/j.dsr.2016.05.007

Rosso, I., M.R. Mazloff, L.D. Talley, et al, 2020: **Water mass and biogeochemical variability in the Kerguelen sector of the Southern Ocean: A machine learning approach for a mixing hot spot**. Journal of Geophysical Research: Oceans, vol. 125, no 3, https://doi.org/10.1029/2019JC015877

Maze G., H. Mercier, R. Fablet, P. Tandeo, M. Lopez Radcenco, P. Lenca, C. Feucher, C. Le Goff, 2017: **Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean**. Progress In Oceanography , 151, 275-292 . Publisher's official version : https://doi.org/10.1016/j.pocean.2016.12.008 , Open Access version : https://archimer.ifremer.fr/doc/00363/47431/

Maze G., H. Mercier, C. Cabanes, 2017: **Profile Classification Models** . Mercator Ocean Journal , (55), 48-56 . Open Access version : https://archimer.ifremer.fr/doc/00387/49816/

Kim, Y. S., & Orsi, A. H., 2014: **On the Variability of Antarctic Circumpolar Current Fronts Inferred from 1992–2011 Altimetry**, Journal of Physical Oceanography, 44(12), 3054-3071. Retrieved Dec 6, 2021, https://doi.org/10.1175/JPO-D-13-0217.1

Pauthenet, E., F. Roquet, G. Madec, C. Guinet, M. Hindell, C. R. Mcmahon, ... & D. Nerini, 2018: **Seasonal meandering of the Polar Front upstream of the Kerguelen Plateau.** Geophysical Research Letters, 45(18), 9774-9781, https://doi.org/10.1029/2018GL079614

Lutjeharms, J.R.E. & Ansorge, I.J. , 2001: **The Agulhas Return Current**, Journal of Marine Systems, 30(1–2), 115-138, https://doi.org/10.1016/S0924-7963(01)00041-0

Feron, R. C., W. P. De Ruijter, & D. Oskam, 1992: **Ring shedding in the Agulhas Current system**. Journal of Geophysical Research: Oceans, 97(C6), 9467-9477. https://doi.org/10.1029/92JC00736

Jones, D. C., H. J. Holt, A. J. Meijers, & E. Shuckburgh, 2019: **Unsupervised clustering of Southern Ocean Argo float temperature profiles**. Journal of Geophysical Research: Oceans, 124(1), 390-402, https://doi.org/10.1029/2018JC014629

Thomas, S. D., D. C. Jones, A. Faul, E. Mackie, & E. Pauthenet, 2021: **Defining Southern Ocean fronts using unsupervised classification**. Ocean Science Discussions, 1-29, https://doi.org/10.5194/os-17-1545-2021