Filtering method based on cluster analysis to avoid salinity drifts and recover Argo data in less time

Emmanuel Romero[1], Leonardo Tenorio-Fernandez[2], Iliana Castro[3], and Marco Castro[4]

[1,3,4]Tecnológico Nacional de México/Instituto Tecnológico de La Paz

[2]CONACyT-Instituto Politécnico Nacional–Centro Interdisciplinario de Ciencias Marinas

Correspondence: Leonardo Tenorio-Fernandez (ltenoriof@ipn.mx)

Abstract. Currently there is a huge amount of freely available hydrographic data and it is increasingly important to have access to it efficiently and easily provided with as much information as possible. Argo is a global collection of around 4000 active autonomous hydrographic profilers. Argo data ~~goes~~ go through two quality processes, real time and delayed mode. This work shows a methodology to filter profiles within a given polygon using the odd-even algorithm, this allows analysis of a study area, regardless of size, shape or location ~~. Also, gives~~ and aims to offer two filtering methods to discard only the real time quality control data that present salinity drifts, thus taking advantage of the largest possible amount of valid data within a given polygon. In the study area selected as an example, it was possible to recover around 80% in the case of the first filter that uses cluster analysis and 30% in the case of the second, which discards profilers with salinity drifts, of the total real time quality control data that are usually discarded by the users due to problems such as salinity drifts, this allows researchers to use any of the filters or a combination of both to have a greater amount of data within the study area of their interest in a matter of minutes, unlike waiting for the delayed mode quality control that takes up to 12 months to be completed. This methodology has been tested in five selected areas around the world to test its replicability, obtaining good results.

1 Introduction

Autonomous oceanographic instruments have become very important tools in observational oceanography. Hydrographic Autonomous Profilers (HAPs) are tools that reduce the costs of ~~in situ~~ *in situ* oceanographic observations, obtaining a large number of hydrographic profiles in time and space, at a lower cost compared to those carried out on oceanographic cruises. An example of these HAPs is those belonging to the Argo program, each measured profile is processed by its Data Assembly Center (DAC) in a quality control system, before being published (Argo Data Management Team, 2019).

HAPs have the ability to continuously measure hydrographic parameters in the water column. Since the beginning of the program and up to the present, there are ~~currently~~ data records collected from around 15 300 core HAPs and around 1300 biogeochemical HAPs belonging to the global Argo group around in the world ~~'s~~ oceans, which have measured temperature, salinity and biogeochemical parameters in most cases from 2000 m depth to the sea surface or vice versa, from which around 4000 are currently active (Argo). However, around 75% of the total profiles have completed the quality control process and therefore it is considered that the rest are not of such good quality, in areas with a low concentration of profiles, ~~the amount of~~

good quality data is scarce, this percentage is more significant and it is important to obtain as much data as possible to support scientific research.

The data of each HAP has have to be validated , verified and processed by a quality control system, before being used or published. This The Argo quality control system consists of two stages, Real Time Quality Control (RTQC) and Delayed Mode Quality Control (DMQC). The RTQC's goal is to make data tests performed by the RTQC are automated and limited, due to the requirement to be available within the first 24 hours of transmission, and these tests in real time are therefore automated and limitedafter transmission. These data are free of serious errors in each of the variables measured by the profiler and must be consistent with the hydrography of the area where the profile was madetheir variables (e.g. impossible data in dates and coordinates) and it must be within the global and regional ranges. In the case of having adjusted parameters available, these are placed in the same variables, but named with the suffix "_ADJUSTED", in this way the data are preserved without adjustments in the variables without this suffix. The second quality control process is the DMQC, the data from adjusted by this quality control replaces the data obtained adjusted by the RTQC, since, during this process, it is subjected to detailed scrutiny by oceanographic experts, DMQC data can take a year to be published (Argo)(Wong et al., 2021). Normally, due to the problems presented by the RTQC data, such as the salinity drifts presented in this work, users of the Argo program data decide not to use the data from are advised to only use DMQC data for scientific analysis, or to perform it by themselves this quality control is explained in the manuals, for this reason many users decide not to use the RTQC data.

The objective of this work is to present a methodology based on cluster analysis to admit the data in RQTC that conforms to the same hydrography patterns as the DMQC data and thus increase the amount of data available for scientific research, avoiding the complete discard of the RTQC data. To do this , the data to be evaluated is determined by a study area carry out this methodology, first, the data must be delimited by a polygon, the one that represents the study area of interest. Using a point in polygon Point in Polygon (PIP) algorithm the profiles that were measured within the study area of interest are determined. In addition, a web application was developed to show results of applying this methodology in a study area of scientific interest, but with little concentration of profilesthe results of the application of this methodology and the usefulness that it can have if it were integrated into the data access platforms, such as statistics and graphs of study areas defined by the user.

2 Data collection and methods

To achieve the objectives of this research in any study area given by a polygon, irregular or not, and since the selection of the data can be of interest both at a global and regional level, it was decided to filter the data using the geographic coordinates where the profiles were measured. By establishing a polygon, we can determine if a point is inside or outside of it, this is a Point in Polygon (PIP) problem, derived from computational geometry and it was approached using the even-odd algorithm (J. D. Foley and Hughes, 1990), which draws a line from a point in a fixed direction and intersects the edges of the polygon. If the point is on the outside of the polygon, the ray will cross the edge an even number of times, if it is on the inside it will intersect the edge an odd number of times.

~~Although this algorithm may fail when the point is at the perimeter of the polygon, it is sufficient to delimit study areas in the ocean. To filter them, the polygon that will be the study area was established, the maximums and minimums of the latitude and longitude of its points were extracted, to set them as a range, the profiles outside this range were automatically discarded. Because it is necessary to evaluate each of the points not discarded with the polygon using the chosen PIP algorithm , parallel programming was used so that each processor core evaluated a certain number of points equally and accelerate this process. Once the measured profiles are obtained~~ geographical coordinates of the profiles stored on the servers of Argo were used as points for the PIP problem, and thus determine if they were measured within the study area. To solve it, the even-odd algorithm (J. D. Foley and Hughes, 1990) was used and once the profiles within the polygon are obtained, the profile data ~~is~~ are downloaded.

For the purposes of testing the methods of this work, a study area was selected (Fig. 1), it is located at 25° and 19° north and 113° and 105° west. In this area it is known that ~~the interactions betweencurrents produce~~ there are current interactions between: the California Current, transported southwards by the tropical branch of the California Current; the Gulf of California current, resident of the Gulf of California with a varying southward extension; and to the north the Mexican Coastal Current (Lavín and Marinone, 2003; Lavín et al., 2009; Godínez et al., 2010; Portela et al., 2016). These interactions produce intense mesoscale activity and a high complexity in the circulation (Kessler, 2006)~~, in this area the~~ . Two mesoscale structures such as cyclonic and anticyclonic eddies, interact and play an important role in circulation (Zamudio et al., 2001; Lavín et al., 2006; Zamudio et al., 2007; Pantoja et al., 2012)~~and~~ . Besides, this area is part of the minimum oxygen zone (Fiedler and Talley, 2006; Stramma et al., 2008). The study area encompasses parts of the California Current System, the Gulf of California, the Transition area and the ~~tropical Pacific off central Mexico (Portela et al., 2016) hereafter~~ Tropical Pacific off Central Mexico (Portela et al., 2016) hereinafter this study area will be named TPCM.

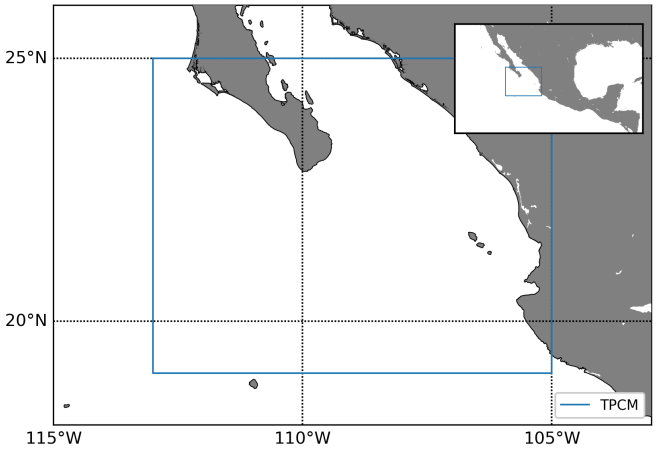

Figure 1. Study area. The upper right corner shows the location of the study area composed of parts of the California Current System, the Gulf of California, the Transition area, and the tropical Pacific off central Mexico (TPCM), shown in the foreground.

One of the great benefits of using a PIP algorithm to filter locations is that it can be used with data from other geo-referenced databases. To demonstrate this, tests were carried out with the World Ocean Atlas 2018 (WOA18) database, which provides quality controlled data to calculate the climatology of temperature, salinity, dissolved oxygen and dissolved inorganic nutrients derived from profiling floats, OSD (Ocean Station Data), CTD (Conductivity, Temperature and Depth) and many others contents in the NCEI World Ocean Database 2018 (WOD18). The monthly data of temperature (Locarnini et al., 2018) and salinity (Zweng et al., 2018) of the statistical mean of each quarter degree (1/4) from 2005 to 2017 were downloaded and the PIP algorithm described in this work was applied to the polygon that delimits the TPCM. ~~Corrections of the data according to~~ Conversions from *in situ* temperature to conservative temperature and practical salinity to absolute salinity were carried out according to the Thermodynamic Equation of SeaWater 2010 (TEOS-10~~were applied and compared with the data from the DMQC of the Argo HAPs to~~). The monthly TS (temperature and salinity) diagrams of the WOA18 data and the Argo DMQC data were compared to corroborate that they are located in the same water masses and review the quality of the DMQC data in the area.

The data ~~and the number of hydrographic profiles~~ measured within the TPCM were ~~also~~ statistically analyzed, it was found that there are few profiles within the area and that around only 30% of the ~~data are part of~~ total data have been evaluated by the RTQC. The Argo manual (Argo Data Management Team, 2019) indicates that ~~there are flags that establish the quality of the adjusted data in both quality controls, one being the best and the fourth being the worst~~the quality control flags establish how good or bad the data are, with 1 being good data and, as the value increases, the quality deteriorates. These flags are determined in the RTQC and when the DMQC arrives it creates flags for the adjusted data, in the event that data have been adjusted in the RTQC they are replaced, this because during the RTQC it is not possible to detect all the bad data or good data could be detected as bad during the real-time evaluations (Wong et al., 2021).Tests were performed by graphing the TS diagrams using these flags, adding the density isoline and the water masses according to Portela et al. (2016), although only the ~~data with the best RTQC quality~~ flagged RTQC data with 1 were used, salinity drifts were shown, so it is not feasible to use these indicators to filter the data in RTQC. To increase the amount of available data, cluster analysis was applied to the data, since two groups of data can be visually located in the TS diagrams; those that form the same patterns as those of the DMQC and those that do not. This analysis, groups a set of objects in such a way that the characteristics of the objects of the same group are more similar to each other than to the other groups (Everitt et al., 2011). In this case, the aim is to separate the RTQC data into groups, a group that contains data with characteristics similar to DMQC data and other groups with salinity drift problems.

To perform the cluster analysis, the unsupervised K-means classification algorithm was chosen, this algorithm groups the data into k groups, minimizing the distance between the data and the centroid of its group (Hartigan and Wong, 1979). The algorithm starts by setting the k centroids in the data space, regardless of where the data were obtained, and assigning the data to its closest centroid. Then, it updates the position of the centroid of each group, calculating the position of the average of the data belonging to each group, and the data is reassigned to its closest centroid. This process is repeated until the centroids do not change position. An algorithm based on distances was selected because it seeks to obtain only the RTQC data closest to the DMQC data.

Since it is necessary to indicate the number of k centroids when we use K-means, a manual enumeration of the groups to be searched is required. To automate this process ~~to~~ and avoid the user having to indicate the exact number of centroids needed to retrieve RTQC data for each month and for each study area chosen, Algorithm 1 was programmed.

ALGORITHM 1

RTQC data filtering

```
dataset ← FilterByMonth(dataset)
for i ← 0 to 11 do
    for j ← 0 to 10 do
        data ← GetDataWithDepthHigherThan(dataset[i], depth ← 1500)
        mid_ranges ← GetMidRangeOfDMQCandRTQC(data)
        groups ← kmeans(data, k ← 2, init ← mid_ranges)
        if groups[0] have DMQC data and groups[1] do not then
            dataset[i] ← MatchDataByProfilerAndProfule(dataset[i], groups[0])
        else
            break
        end if
    end for
end for
return dataset
```

The Algorithm 1 receives the adjusted data from the DMQC and the RTQC, ~~separates it~~ in the case of profiles that have not been adjusted, the data are received without adjustment. The algorithm separates these data by month in an array, and iterates it. Within each iteration, it calculates the ~~mid-ranges~~ salinity mid-range of each quality control and divides the data into two groups (using the mid-ranges as the starting position of the centroids), up to a maximum of ten iterations, each time verifying if there are DMQC data in both groups, if so, the algorithm stops and returns the data without grouping them, on the contrary, if only a group contains the data in DMQC, it associates the data of that group with the data at depths less than 1500 m as default, taking into consideration the month, the profiler code and the profile number and replaces the group data with the associated data. The mid-ranges are used as the initial position of the centroids to prevent them from being generated randomly. The procedure described above is the first filter of the RTQC data, in each iteration the algorithm discarded the groups that presented salinity drifts and kept only the group where the DMQC data were found, thus, when the execution of the algorithm ends, RTQC data within the group with DMQC data are considered to contain no salinity drifts. To increase the reliability of the filtering, a second filter was created. In the second filter, the algorithm stores in memory the profilers that presented salinity drifts ~~. Thus~~ during the execution of the Algorithm 1. Thus, the second filter not only discards the profiles with drift problems, it also directly discards all the ~~filtering, in addition to being carried out by cluster analysis, now discards the~~ profiles of the profilers that ~~presented problems~~they have at least one profile with salinity drifts.

A library was developed that contains all the procedures described in this work. Using it, as an extra example, five study areas were delimited with different extension, location, density of profiles and hydrographic characteristics. The first area was the Alboran Sea, this area was selected because the data were measured in shallow water (0 to 1200 m), the Antarctic area was selected in the high altitude (cold water), the third area was the Bermuda Triangle, it was selected because it is located in the Atlantic transition between the tropical and subtropical area, the fourth, the tropical zone of the Pacific that surrounds an archipelago of the central Pacific and the last one is in the tropical sea of Indonesia. All data from these areas were downloaded from the snapshot of December 2020 (Argo, 2020) and evaluated by Algorithm 1.

To test the above methods in a more extensive and irregular polygon area, a web application was developed. The study area for this web application was delimited by the Exclusive Economic Zone (EEZ) of Mexico as example (Fig. 2) and the geographical location of the profiles from around the world are filtered by the PIP algorithm, to automatically download the data every 24 hours within this irregular polygon through the IFREMER synchronization service~~(?)~~. In Figure 2, the ~~blue~~ orange line delimits the EEZ of Mexico and the ~~yellow~~ blue box delimits the TPCM.

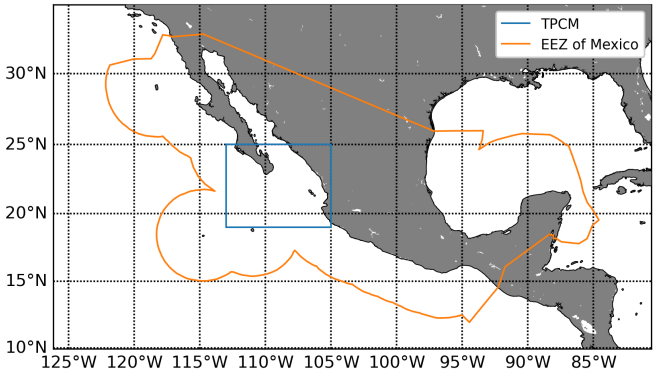

Figure 2. Comparison of study areas. The irregular polygon that delimits the Mexican EEZ and TPCM that was used as an example for the use of the proposed methodology are shown.

Every time that new data from HAPs is downloaded, they go through a processing phase, the data ~~is~~ are cleaned and transformed to be integrated into the web application. For example, the variables of temperature and salinity are converted to conservative temperature and absolute salinity, as ~~the Thermodynamic Equation of SeaWater 2010 (~~defined by TEOS-10~~),~~ , the current description of the properties of seawater defines it. Afterwards, graphs and useful files are generated to show information about the HAPs and their profile data.

The web application was developed on a satellite map, to which tools were added for data management and visualization, such as drawing irregular polygons to define study areas within the main polygon, filtering data to display statistical and graphical data according to the selected filter, trajectory tracing, among others. Also, RTQC data filtering was implemented in the web application, the same irregular polygons with which statistical data are obtained, can be used to indicate a study area in which it is sought to obtain as much data as possible without salinity drifts.

3 Results

The use of the chosen PIP algorithm to filter the measured profiles within the polygon (Fig. **??**) worked correctly, in addition to establishing the range of maximums and minimums of the latitude and longitude of the polygon to discard the profiles measured outside it, allowed the PIP algorithmto filter only the profiles made near or inside the polygon. In Figure **??**a, the geographical locations of the profiles from HAPs that were made within the polygon filtered by the even-odd algorithm are shown, in the same way, in Fig. **??**b, the location of the filtered data belonging to WOA18 is shown. The blue line represents the given polygon and the locations of the filtered profiles inside and outside the polygon are represented by dots in red and black respectively.

Filtered geographic locations within the defined polygon, the upper right corners show the location of the TPCM shown in the foreground. **(a)** Location of the HAP profiles, each point represents a hydrographic profile. **(b)** WOA18 data locations, each point represents the statistical mean of each quarter degree of hydrographic profiles from different instruments.

data used to obtain the following results was downloaded from the Coriolis GDAC FTP server in 2019 and the "Profile directory file of the Argo Global Data Assembly Center" file was used as input for the chosen PIP algorithm, the algorithm filtered the measured profiles inside the polygon correctly. Figure 3 shows the result of the TS (temperature and salinity) diagram comparison between the DMQC data and the WOA18 data. The DMQC and WOA18 data are located in the same water masses, and the data is spliced at depths greater than 1500 m, which validates that the DMQC data following the same patterns as the data from other international DBsdatabases. According to Portela et al. (2016), this region is made up of the California Current Water (CCW), Tropical Surface Water (TSW), Gulf of California Water (GCW), Subtropical Subsurface (SS) and the Pacific Intermediate Water (PIW).

On the contrary, the data in RTQC with the best quality flag present drifts in salinity. The RTQC and DMQC data were plotted in the TS diagrams together per month of the TPCM, some of the data in RTQC were the cause of salinity drifts in almost all the months (Fig. 4).

In Figure 4 it is clear that the salinity drift drifts in the RTQC data is important and therefore they are labeled as erroneousare important, however it is also shown that certain data follow the structure (shape) of the DMQC data. To avoid discarding the entire RTQC data, it is proposed to use cluster analysis. By applying cluster analysis to all data in RTQC with the K-means algorithm and with different values in k, the resulting groups mix data that show salinity drifts, with data that follow the same patterns as the DMQC data at 1500 meters, this is because, at depths less than 1500 meters, salinity data is are more dispersed than at greater depths.

Taking into consideration that at depths greater than 1500 m, the variations in salinity and temperatures are imperceptible, the cluster analysis was performed with the salinity data measured at depths greater than 1500 m. The resulting groups are shown in Fig. 5a and b, in the figure it is observed that one of the resulting groups contains the data that follow the same patterns as the DMQC data and the rest of the groups contain data with salinity drifts, therefore the next step was to associate the data of these groups with the rest of the data, taking into consideration the profiler code and the profile number and thus obtaining complete groups (Fig. 5c and d).

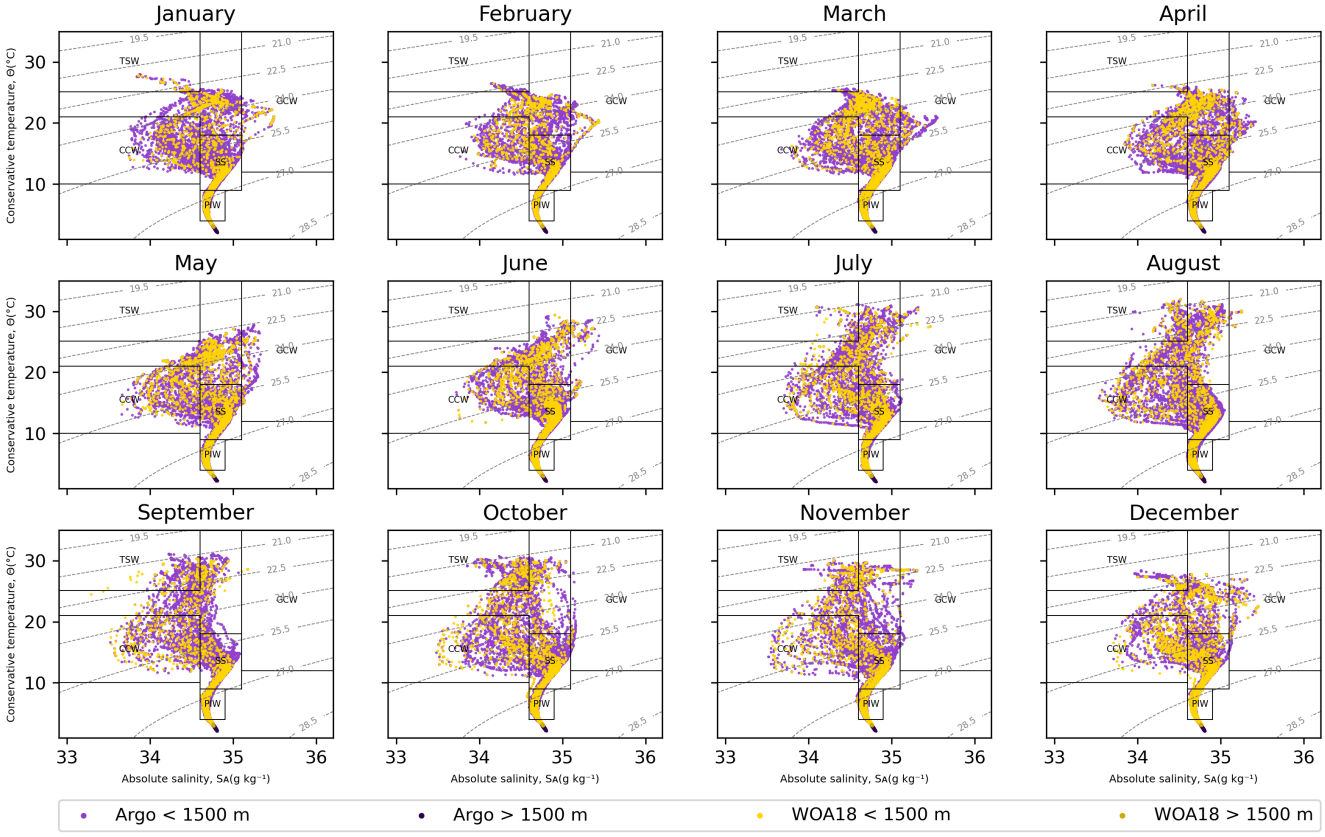

Figure 3. Monthly comparison of TS diagrams of data from DMQC and WOA18. The black boxes delimit the limits of the water masses in the region and the gray isolines the density (kgm^{-3}).

Figure 5 shows how the groups are separated with the chosen algorithm. In the months of January and December, DMQC data ~~is~~ are displayed as yellow dots and the orange groups contain the RTQC data that follow the patterns of the data in DMQC. The blue, green and red groups contain the data showing salinity drifts.

To manually avoid indicating the number of k centroids, Algorithm 1 was developed. Figure 6 shows the first three iterations of the month of ~~January~~ December as an example. In Figure 6a and b blue data ~~represents~~ represent the group that contains DMQC data and the orange color group represents the group of the RTCQ data. The data contained in the orange groups are discarded by the algorithm. The Figure 6c is the third iteration, both groups contain data in DMQC~~therefore the algorithm~~ , this because the data are so close to each other that the K-means algorithm (which is based on distances to separate the groups) divides the DMQC data into two different groups, so in this iteration the algorithm stops.

The results of the first filtering of the proposed algorithm are shown in Fig. 7a, the filtered data from the RTQC show the same patterns as the DMQC data, except for the months of July, August and September. In July and August, the salinity drifts are found at depths less than 1500 m, while in September, the drifts present values very close to the DMQC data and this

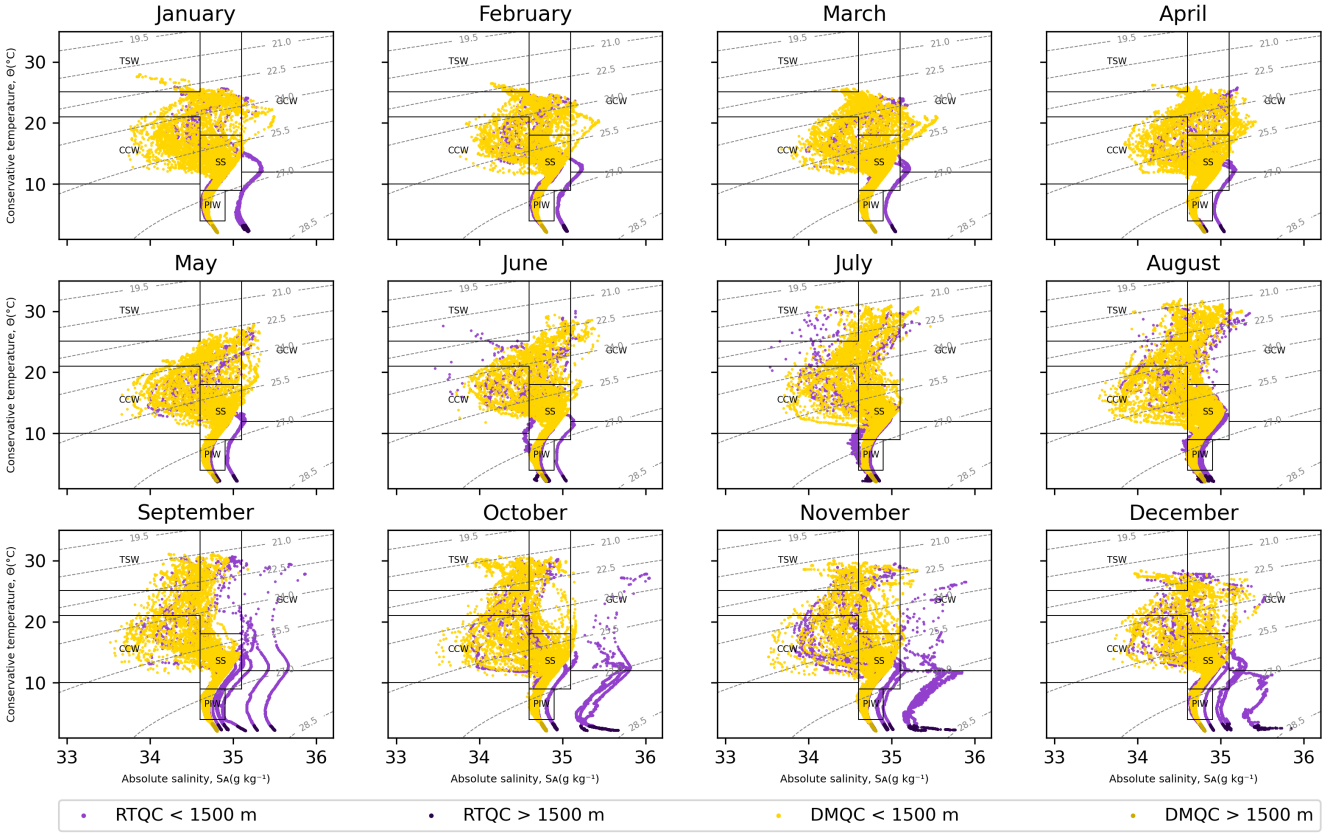

Figure 4. Monthly comparison of TS diagrams of data from RTQC and DMQC. The black boxes delimit the limits of the water masses in the region and the gray isolines the density (kgm^{-3}).

prevents the algorithm from being able to separate them. This filter allows obtaining a greater amount of admitted RTQC data, but as seen in the figure, it still shows salinity drifts in some cases. For this reason, the second filter was incorporated, Fig. 7b shows the results of it, since it considers those profilers that have presented salinity drifts and removes their profiles completely, a significant reduction in admitted data from the RTQC is observed, but these no longer show salinity drifts.

Table 1 shows the total measurements (meas.) made in the TPCM area and the measurements filtered by the aforementioned algorithms.

The total usable data in the TPCM due to the first and second filters represent ∼95% and ∼80% of the data, compared to the ∼70% that would be obtained by automatically discarding the data in RTQC. By presenting this option to the researcher and filtering the data from the RTQC, instead of discarding ∼30% of the total, only ∼5% would be discarded in the case of the first filter and ∼20% in the case of the second, which would mean a considerable increase in the data available for use, after all, the admitted data ~~presents~~ present similar characteristics to the data that were already evaluated with the DMQC, they have a high

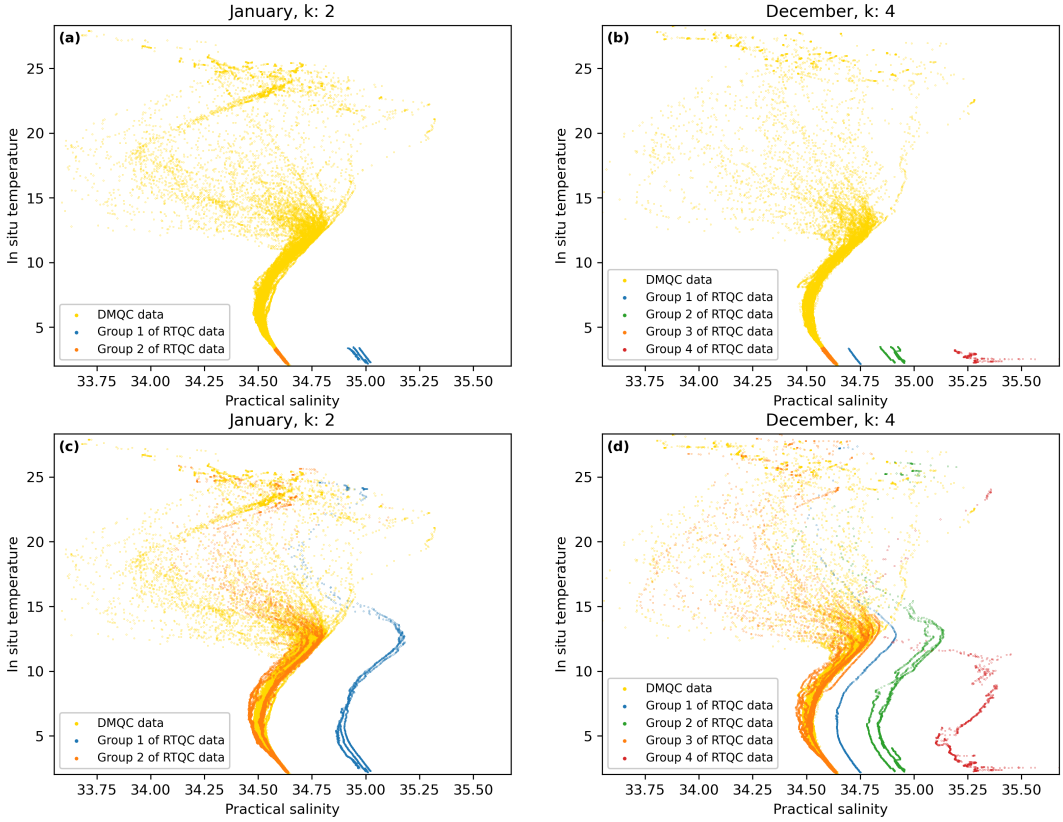

Figure 5. Cluster analysis results. **(a)** and **(b)** show the groups formed with the RTQC data measured at depths greater than 1500 m. **(c)** and **(d)** show these same grouped data but matched data with the rest of their profile data.

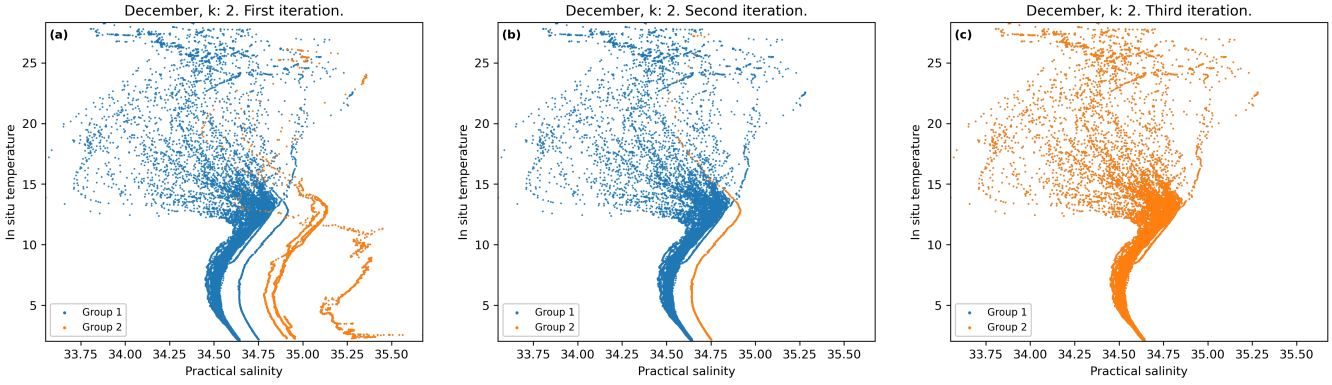

Figure 6. First three iterations of the proposed algorithm using the data for the month of December. **(a)**, **(b)** and **(c)** are the first, second and third interactions.

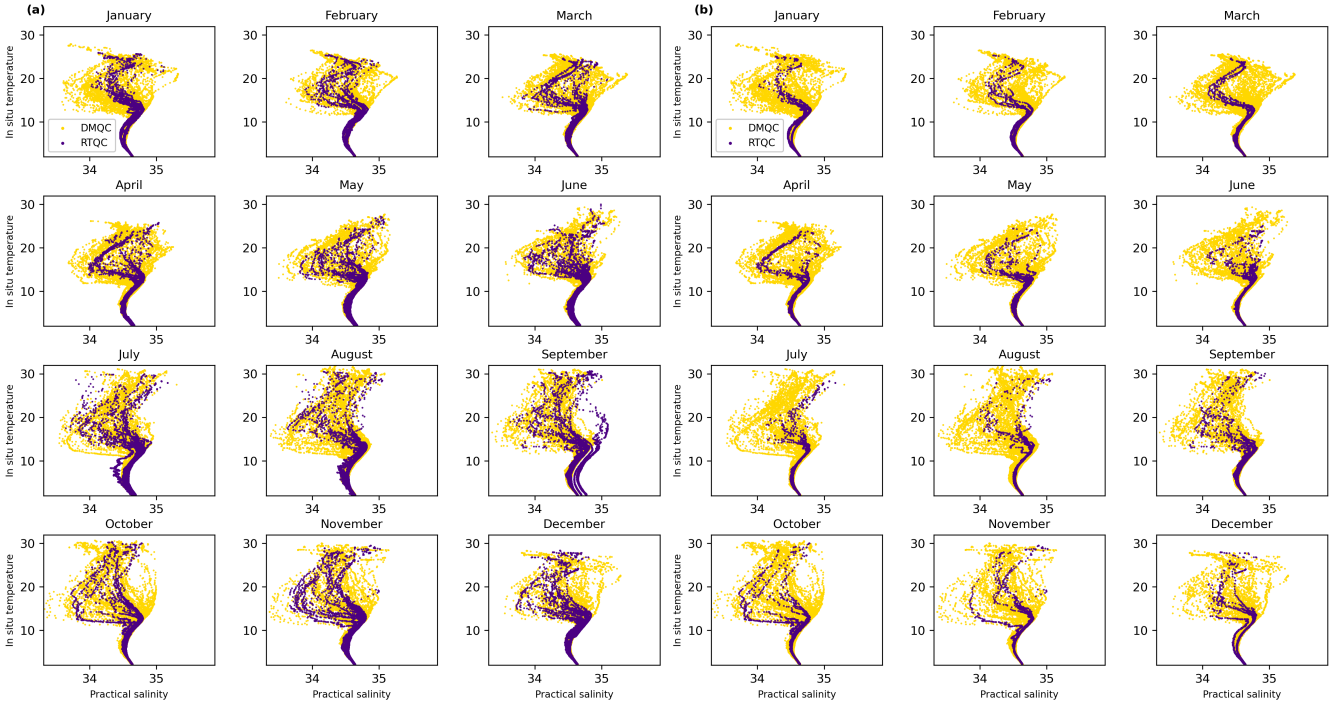

Figure 7. Monthly comparison of the TS diagrams of RTQC and DMQC. **(a)** First filtering of RTQC. **(b)** Second filtering of RTQC.

Table 1. Percentages of DMQC and RQTC data admitted and discarded normally and by the two proposed filters.

	Without filter		First filter		Second filter	
Data	Meas.	%	Meas.	%	Meas.	%
DMQC	594 385	69.96%	594 385	69.96%	594 385	69.96%
Admitted RTQC	0	0.00%	209 392	24.64%	82 196	9.67%
Discarded RTQC	255 184	30.03%	45 792	5.39%	172 988	20.36%
Total	849 569	100.00%	849 569	100.00%	849 569	100.00%

probability of not needing adjustments and therefore could be used in research before waiting for the DMQC to be applied to them.

Despite the fact that in the first filter some months were not filtered in the desired way in the study area, the researcher may simply not use the data from those months or use the second filter if the researcher wishes to use only the most reliable data. Also, the possibility of using a combination of both filters is not ruled out, if the researcher uses the months of the first filter that no longer present salinity drifts and uses the data of the second filter in which they present drifts, the largest possible amount of admissible data would be used in any study area. The results of the algorithm will change depending on the extension

and the hydrographic characteristics of the study area that the user selects, selecting which filter to use or whether to make a combination of them is the responsibility of the user and it is recommended to have knowledge of the study area.

A library for python 3.7 named *cluster_qc* was developed alongside this work, contains all the procedures described in it and is available under the Creative Commons Attribution 4.0 International License, latest package version is v1.0.2 (Romero et al., 2021). Using this library, five study areas were delimited with different extension, location, profile density and hydrographic characteristics, the data were downloaded from the snapshot of December 2020 (Argo, 2020) and evaluated by Algorithm 1, the results are shown in Table 2.

Table 2. Results of Algorithm 1 in five study areas.

	DMQC		RTQC - Original		RTQC - First filter		RTQC - Second filter	
	Meas.	%	Meas.	%	Meas.	%	Meas.	%
Alboran Sea	49 401	54.96%	40 481	45.04%	40 481	45.04%	40 481	45.04%
Antarctica	1 117 571	92.14%	95 346	7.86%	93 647	7.72%	92 204	7.60%
Bermuda Triangle	2 060 348	70.49%	862 455	29.51%	468 483	16.03%	243 752	8.34%
Hawaii	3 252 097	70.81%	1 340 462	29.19%	1 308 773	28.50%	1 259 247	27.42%
Indonesia	5 260 566	86.86%	795 900	13.14%	780 727	12.89%	771 874	12.74%

In the results of the Alboran Sea, the westernmost part of the Mediterranean Sea and there are no data deeper than 1500 m or salinity drifts in this study area, so the algorithm directly returns the data set without modification. The algorithm receives the depth of 1500 m by default, sending a greater depth could eliminate salinity variations if there were any. In the case of Antarctica, we found that the months of February and April contain salinity drifts, which could not be completely eliminated with the first filter. For this case, it is recommended to use the RTQC data supported by the second filter. On the other hand, in the Bermuda Triangle, salinity drifts are shown in the months of June to October, in addition to atypical values in the rest of the months. The first filter already eliminates salinity drifts, so in this case it is recommended to use this filter and eliminate outliers. The fourth study area, the one that surrounds a central Pacific archipelago, there are many outliers in all the months, however, the first filter managed to rule out the salinity drifts present in the months of September to December. In this case it is recommended to reduce the study area into smaller areas to apply the filters and treat the outliers separately. Finally, the large study area located next to Indonesia shows salinity drifts in the months of March and July to December. The first filter was able to filter the salinity drifts except for the month of December, because the deviations are above 1500 m, in this case it is recommended to use the data from the first filter for the months of January to November, or use only the months with no outliers.

3.1 Web application

The web application got interesting results and its access is through the *cluster_qc* library repository. In Figure 8, it is observed that the PIP algorithm filters the profiles that were made within the EEZ of Mexico correctly, even when the irregular polygon

that comprises the study area is defined by more than 350 vertices. The blue line represents the given polygon and the locations of the filtered profiles inside and outside the polygon are represented by dots in red and black respectively.

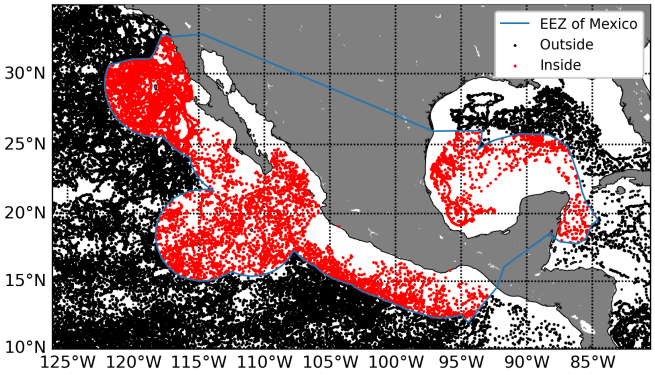

Figure 8. Filtered geographic locations within the EEZ of Mexico. The irregular polygon that delimits the EEZ of Mexico and the profiles measured inside and outside of it are shown.

Once the data ~~has~~ have been downloaded and transformed, statistical data specific to the EEZ of Mexico can be obtained, such as the number of profilers within the polygon, the number of profiles or profilers per year, the DACs to which these profilers belong, among others. Table ~~2~~ 3 shows the profilers that have carried out measurements within the polygon given in the month of November 2019. We can see from the table that there is a shortage of biogeochemical profilers within the polygon. These 4 biogeochemical HAPs are capable of measuring oxygen in addition to temperature and salinity, but none of their oxygen data satisfactorily finish the quality control process, so they are not available. So we can conclude that within the Mexican EEZ there are no good quality biogeochemical data from ~~PHAs~~ HAPs Argo.

Table 3. Profilers and profiles present in the Mexican EEZ.

	Core		Biogeochemical		
DAC	Actives	Inactives	Actives	Inactives	Profiles
AO: AOML	51	114	0	3	32 998
IF: CORIOLIS	6	3	0	1	1 098
ME: MEDS	1	1	0	0	201
Total	58	118	0	4	34 297

For each of these profilers their profiles of temperature (Fig. 9a) and salinity (Fig. 9b), the Temperature-Salinity (TS) diagram (Fig. 9c), the estimation of the profiler trajectory (Fig. 9d) and the profiles of temperature (Fig. 9e) and salinity (Fig. 9f) with respect to time were generated, these diagrams are basic for analysis in scientific ocean research, the profiler 4901635 is shown as an illustrative example in Fig. 9.

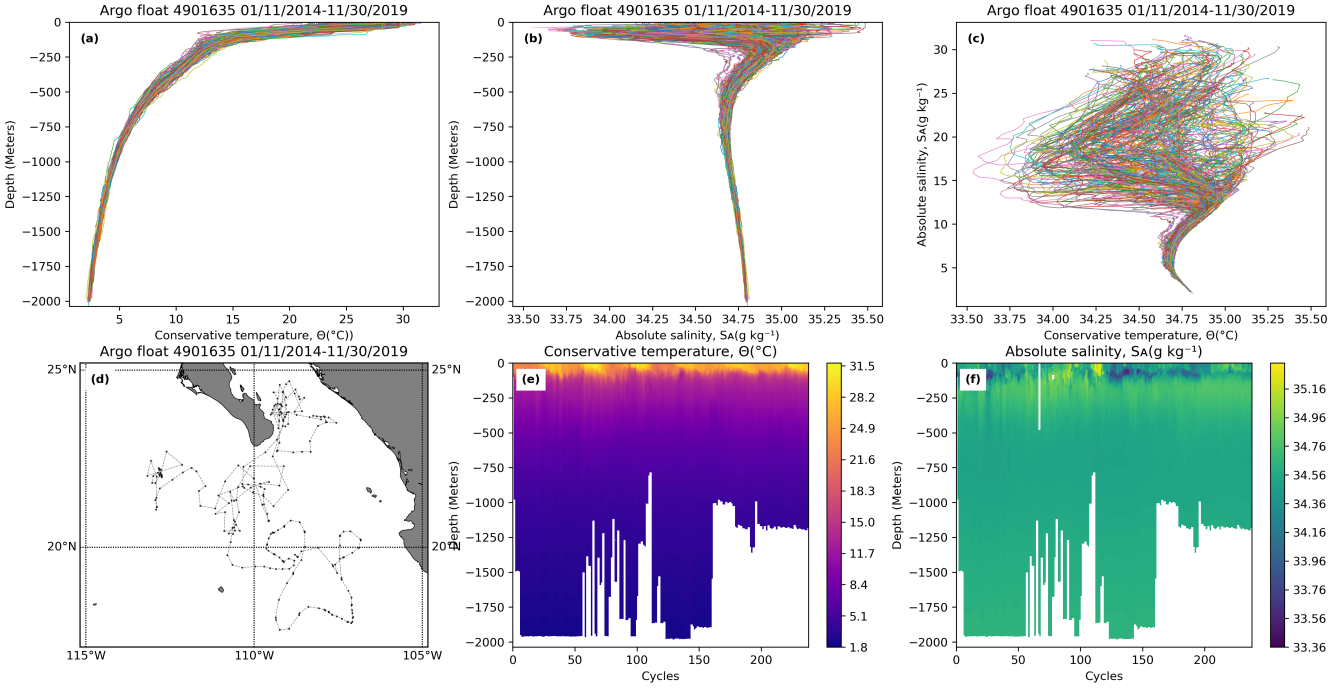

Figure 9. Diagrams produced by the web application. a) Profile of conservative temperature. b) Profile of absolute salinity. c) TS diagram. d) HAP trajectory. e) Profile of conservative temperature with respect to time. f) Profile of absolute salinity with respect to time.

The satellite map of the web application is interactive, it shows the active and inactive HAPs, filters the data, shows statistics, trajectories, diagrams (Fig. 10a) and has other tools to facilitate the visualization and management of the data, such as displaying statistics of a given study area within the main polygon (Fig. 10b and c).

Finally, the filtering of RTQC data that have patterns similar to DMQC data are offered in the web application, which allows that to filter the data in a study area within the EEZ of Mexico, it is not necessary to have programming knowledge. Access to the web application is through the *cluster_qc* library repository.

4 Discussion

Despite the existence of reports on salinity drifts such as the one announced by Argo Data Management on September 25, 2018, the quality control processes in real time are not yet robust enough to identify them, since these processes are automatic and ~~search for data that is impossible or~~ mainly look for impossible data, for example, latitudes greater than 90 degrees or temperatures and salinities outside the global and regional ranges. Therefore, the quality established by the flags does not take these drifts into account. A possible solution is for the interested researcher to apply the DMQC on their own (Wong et al., 2021), this process can be long and tedious, ~~after all, according to the manual (Argo Data Management Team, 2019)~~

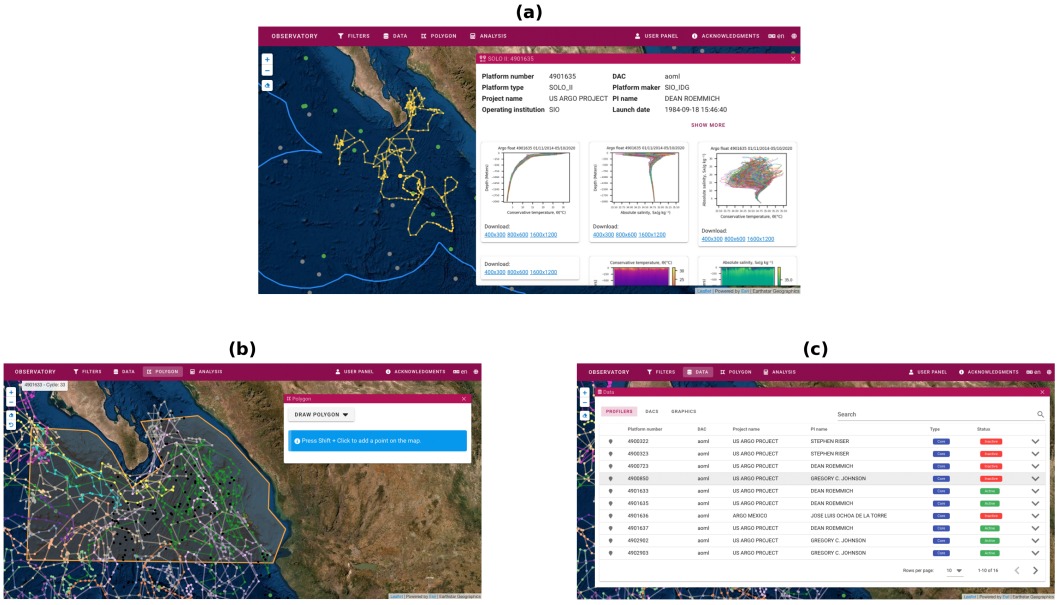

Figure 10. Web application. a) Data, charts and trajectory of a HAP. b) HAP trajectories filtered by a drawn polygon. c) Profiler data within the polygon.

for this reason and under Argo's recommendation to use only DMQC for scientific research, a large number of users who use the data from ~~PHAs~~ HAPs Argo prefer to ~~automatically~~ directly discard the RTQC data and only use the DMQC data.

The data in DMQC ~~is~~ are consistent with other international databases such as WOA18 within the study area delimited by the irregular polygon, which validates this process, however, too much data ~~has~~ have to be discarded due to the drifts present in RTQC. The filtering proposed in this work is based on using the patterns followed by the DMQC data to filter the RTQC data, especially useful for areas where there are few profiles. This process is carried out by zone and by month, in this way it does not matter if the study area is close to the arctic or the tropics, the filtering of the RTQC data ~~is~~ are carried out based on the characteristics of the area reflected in the DMQC data. In addition, when separating the data by month, their seasonal changes are taken into account. This means that the resulting RTQC data will have a high probability of being accepted when the DMQC is applied to them.

The time it takes for a modern computer to do cluster analysis is relatively short compared to the 12 months it can take to perform the DMQC, this will help researchers interested in recent data from ~~PHAs~~ HAPs to have greater reliability when using RTQC data. Two filters are proposed, the first is the result of using cluster analysis on the data and the second discards the ~~PHAs~~ HAPs that have presented salinity drifts in the result of the first filter. Therefore the second filter is more reliable but contains a smaller amount of data. ~~The~~ As seen in the TPCM example, the researcher is free to use either one or a combination of both. For example, as seen in Fig. 7a and b, where around 80% and 30% of the total discarded data are admitted, the months of July to September continued with salinity drifts after applying the first filter, to take advantage of more data the researcher

can use the data from the months of July to September of the second filter and the rest of the months use the data from the first. However, as shown by the five study areas used as an extra example, the percentage of data admitted by the filters depends on the study area and its characteristics. It is the responsibility of the researchers to make the decision based on their knowledge of the study area, which filter to use, if the study area should be resized or if the default depth value should be changed.

There are platforms to access data from HAPs Argo, such as Argo Data Management, Coriolis and Euro-Argo in addition to other options such as FTP or snapshots. The current platforms already provide graphics and data from the profilers, as well as filters to display or download the data, however, the geographical filter they use is by maximum and minimum coordinates, so it is only possible to filter by polygons in rectangle or square shape without rotation.

Another platform called ~~?~~ OceanOPS (Joint Centre for Oceanography and Marine Meteorology in situ Observations Programmes Support) does perform statistical ~~analyzes~~ analyses on the data, nevertheless this one performs them globally and it is not possible to choose a smaller area, for example, only the EEZ of Mexico or the tropical Pacific off central Mexico and surrounding areas to obtain statistical information on it. It is worth mentioning that said platform has a large number of statistics for each variable registered within the source files, however being able to generate graphs and tables in real time using an irregular polygon defined by the user (as shown in this work with the PIP algorithm), would be a great tool for studying these data.

The web application described in this document tries to cover some of the problems that the aforementioned have and include some of their characteristics, in addition to proposing unpublished options such as filtering by irregular polygons, statistics adaptable to filters, generation of graphs according to user needs and RTQC data filtering. However, the web application is in its initial phase, there are still many tools and ~~DBs~~ databases that can be integrated to offer an even more complete experience.

5 Conclusions

~~The data in DMQC is consistent with other international databases such as WOA18 within the study area delimited by the irregular polygon, which validates this process, however, too much data has to be discarded due to the drifts present in RTQC.~~ This work gives two filtering methods to discard only the RTQC data that present salinity drifts and with it to take advantage of the largest amount of data within a given polygon. In the TPCM, of the total RTQC data it was possible to recover around 80% in the case of the first filter and 30% in the case of the second~~of the total RTQC data that~~ , which are usually discarded due to problems such as salinity drifts, this allows researchers to use any of the filters or a combination of both to have a greater amount of data within the study area of their interest in a matter of minutes, unlike waiting for the DMQC that takes up to 12 months to be completed.

The result of this work provides useful tools to increase productivity in scientific investigations that use data from the water column. The PIP algorithm turns out to be an efficient method to directly filter the data from any georeferenced database using geographic locations, while the algorithms proposed for filtering RTQC data allows the separation of the data not yet adjusted by the DMQC into data with salinity drifts and data that show patterns similar to those of the DMQC data, in order to increase the amount of data in study areas with scarce data from HAPs. Finally, the web app demonstrates one of the applications in which these proposals can be used.

Code availability. *cluster_qc* was developed in python 3.7 and is licensed under a Creative Commons Attribution 4.0 International License. Source code is available at doi.org/10.5281/zenodo.4595802. Latest package version is v1.0.2.

Data availability. These data were collected and made freely available by the International Argo Program and the national programs that contribute to it. (https://argo.ucsd.edu, https://www.ocean-ops.org). The Argo Program is part of the Global Ocean Observing System. The data was downloaded from the Coriolis GDAC FTP server in 2019 and the snapshot from December 2020 (Argo, 2020) was also used. The data used from the NCEI World Ocean Database 2018 are the monthly data of temperature (Locarnini et al., 2018) and salinity (Zweng et al., 2018) of the statistical mean of each quarter of a degree (1/4) from 2005 to 2017.

Author contributions. Emmanuel Romero: Writing - original draft, Methodology, Software. Leonardo Tenorio-Fernandez: Writing - review & editing, Conceptualization, Supervision. Iliana Castro: Writing - review & editing. Marco Castro: Writing - review & editing, Conceptualization.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are grateful to CONACYT for granting scholarship No. 780669 to Emmanuel Romero. We appreciate that these data were collected and made freely available by the International Argo Program and the national programs that contribute to it. (https://argo.ucsd.edu, https://www.ocean-ops.org). The Argo Program is part of the Global Ocean Observing System. We also thank to the Instituto Tecnológico de La Paz (ITLP) and to the Centro Interdiciplinario de Ciencias Marinas (CICIMAR) for their institutional support. We also acknowledge the critical comments from the reviewers.

References

Argo: Argo, https://argo.ucsd.edu/.

Argo: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of December 10st 2020, https://doi.org/https://doi.org/10.17882/42182#79118, 2020.

Argo Data Management: Argo Data Management, http://www.argodatamgt.org/.

Argo Data Management Team: Argo user's manual V3.3, Report, https://doi.org/10.13155/29825, 2019.

Coriolis: Coriolis: In situ data for operational oceanography, http://www.coriolis.eu.org/.

Euro-Argo: Argo Fleet Monitoring - Euro-Argo, https://fleetmonitoring.euro-argo.eu/.

Everitt, B., Landau, S., Leese, M., and Stahl, D.: Cluster Analysis, Wiley Series in Probability and Statistics, Wiley, 2011.

Fiedler, P. and Talley, L.: Hydrography of the Eastern Tropical Pacific: a review, Progress In Oceanography, 69, 143–180, https://doi.org/10.1016/j.pocean.2006.03.008, 2006.

Godínez, V. M., Beier, E., Lavín, M. F., and Kurczyn, J. A.: Circulation at the entrance of the Gulf of California from satellite altimeter and hydrographic observations, Journal of Geophysical Research: Oceans, 115, https://doi.org/10.1029/2009JC005705, 2010.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100–108, https://doi.org/10.2307/2346830, 1979.

J. D. Foley, A. van Dam, S. K. F. and Hughes, J. F.: Computer Graphics: Principles and Practice. The Systems Programming Series, Addison-Wesley, 2nd edn., 1990.

Kessler, W. S.: The circulation of the eastern tropical Pacific: A review, Progress in Oceanography, 69, 181–217, https://doi.org/https://doi.org/10.1016/j.pocean.2006.03.009, a Review of Eastern Tropical Pacific Oceanography, 2006.

Lavín, M., Beier, E., Gomez-Valdes, J., Godínez, V., and García, J.: On the summer poleward coastal current off SW México, Geophys. Res. Lett, 33, https://doi.org/10.1029/2005GL024686, 2006.

Lavín, M. F. and Marinone, S. G.: An Overview of the Physical Oceanography of the Gulf of California, pp. 173–204, Springer Netherlands, https://doi.org/10.1007/978-94-010-0074-1_11, 2003.

Lavín, M. F., Castro, R., Beier, E., Godínez, V. M., Amador, A., and Guest, P.: SST, thermohaline structure, and circulation in the southern Gulf of California in June 2004 during the North American Monsoon Experiment, Journal of Geophysical Research: Oceans, 114, https://doi.org/10.1029/2008JC004896, 2009.

Locarnini, R., Mishonov, A., Baranova, O., Boyer, T., Zweng, M., Garcia, H., Reagan, J., Seidov, D., Weathers, K., Paver, C., Smolyar, I., and Locarnini, R.: World Ocean Atlas 2018, Volume 1: Temperature, A. Mishonov Technical Ed.; NOAA Atlas NESDIS, 1, 81, 52 pp., 2018.

OceanOPS: OceanOPS, https://www.ocean-ops.org/board.

Pantoja, D., Marinone, S., Pares-Sierra, A., and Gomez-Valdivia, F.: Numerical modeling of seasonal and mesoscale hydrography and circulation in the Mexican Central Pacific, Ciencias Marinas, 38, 363–379, https://doi.org/10.7773/cm.v38i2.2007, 2012.

Portela, E., Beier, E., Barton, E., Castro Valdez, R., Godínez, V., Palacios-Hernández, E., Fiedler, P., Sánchez-Velasco, L., and Trasviña-Castro, A.: Water Masses and Circulation in the Tropical Pacific off Central Mexico and Surrounding Areas, Journal of Physical Oceanography, 46, https://doi.org/10.1175/JPO-D-16-0068.1, 2016.

Romero, E., Tenorio-Fernandez, L., Castro, I., and Castro, M.: romeroqe/cluster_qc: Filtering Methods based on cluster analysis for Argo Data, https://doi.org/10.5281/zenodo.4595802, 2021.

Stramma, L., Johnson, G. C., Sprintall, J., and Mohrholz, V.: Expanding Oxygen-Minimum Zones in the Tropical Oceans, Science, 320, 655–658, https://doi.org/10.1126/science.1153847, 2008.

Wong, A., Keeley, R., and Carval, T.: Argo Quality Control Manual for CTD and Trajectory Data, Report, USA, FRANCE, https://doi.org/https://doi.org/10.13155/33951, 2021.

Zamudio, L., Leonardi, A., Meyers, S., and O'Brien, J.: ENSO and Eddies on the Southwest cost of Mexico, Geophysical Research Letters - GEOPHYS RES LETT, 28, https://doi.org/10.1029/2000GL011814, 2001.

Zamudio, L., Hurlburt, H., Metzger, E., and Tilburg, C.: Tropical Wave-Induced Oceanic Eddies at Cabo Corrientes and the Maria Islands, Mexico, J. Geophys. Res., 112, 18, https://doi.org/10.1029/2006JC004018, 2007.

Zweng, M., Reagan, J., Seidov, D., Boyer, T., Locarnini, R., Garcia, H., Mishonov, A., Baranova, O., Paver, C., and Smolyar, I.: World Ocean Atlas 2018, Volume 2: Salinity, A. Mishonov Technical Ed.; NOAA Atlas NESDIS, pp. 82, 50 pp., 2018.