Recommendations for the data management and structure for BGC extension within the Argo system at EU level (including recommendations of task 4.2.1)

Ref.: D4.11_V1.0

Date: 21/12/2022

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	**D4.11**
Deliverable title	**Recommendations for the data management and structure for BGC extension within the Argo system at EU level (including recommendations of task 4.2.1)**
Description	**Report providing recommendations for the data management and structure for the BGC extension at the European level**
Work Package number	**4**
Work Package title	**Extension to biogeochemical parameters**
Lead Institute	**Sorbonne Université**
Lead authors	**Catherine Schmechtig and Fabrizio D'Ortenzio**
Contributors	**Thierry Carval, Hervé Claustre, Giorgio Dall'Olmo, Claire Gourcuff, Birgit Klein, Violetta Paba, Sylvie Pouliquen, Virginie Racapé, Raphaëlle Sauzède, Kamila Walicka**
Submission date	**21.12.2022**
Due date	**31.12.2022**
Comments	
Accepted by	**Fabrizio D'Ortenzio**

Document History

Version	Issue Date	Author	Comments
0.1	10/05/2022	Catherine Schmechtig	Initial version
0.2	31/08/2022	Catherine Schmechtig & Fabrizio D'Ortenzio	Refined version
1.0	21/12/2022	Catherine Schmechtig & Fabrizio D'Ortenzio	Final version

EXECUTIVE SUMMARY

This deliverable aims to give an overview of the present Biogeochemical-Argo delayed mode status (procedures, organisation) to give some ways of thinking to the Euro-Argo management board to decide and organise (structures, workflow) efficiently the BGC-Argo Delayed Mode at the European level.

TABLE OF CONTENT

1. Acronyms

ADMT	Argo Data Management Team
BBP	Particulate Backscattering coefficient
BGC	Biogeochemical
CHLA	Chlorophyll concentration
DAC	Data Assembly Centre
DM	Delayed Mode
DOXY	Dissolved oxygen concentration
GDAC	Global Data Assembly Centre
NITRATE	Nitrate concentration
NN	Neural Network
OMZ	Oxygen Minimum Zone
QC	Quality Control
RT	Real Time

2. Introduction

In the framework of OneArgo, biogeochemical (BGC) parameters were officially included in the Argo network. This is the result of more than 10 years of technological development, pilot tests, scientific results (https://biogeochemical-argo.org/peer-review-articles.php). The key to the success of Argo (and then of OneArgo) is the data management and distribution system, which must deliver observations after a specific and "as-best-as-possible" Quality Control (QC) procedure. This constrain imposed to the BGC component of OneArgo a clear roadmap, which was for the most achieved in the last years. Presently, real time (RT) QC procedures (https://biogeochemical-argo.org/data-management.php) are established for the 6 endorsed BGC variables, for the most thanks to a huge international effort, coordinated by the BGC task team (https://biogeochemical-argo.org/data-management-task-team.php).

At the beginning of the Euro-Argo RISE project, the last segment of the BGC QC procedures , the Delayed Mode (DM, which is supposed to provide the "as-best-as-possible" data for each parameter) was not yet completely defined. This was the consequence of several concomitant factors: the different technological degree of maturity of the sensors, the availability (or not) of existing alternative data sets (i.e. climatology), the number of operational floats with several sensors configurations (i.e. only DOXY; only optics; DOXY + NITRATE etc), the size (and the degree of involvement) of the user's community (i.e. data experts, modellers, etc).

The first objective of the WP4 of EuroArgo-RISE was then to develop (when required, for example for chlorophyll concentration, CHLA) or consolidate (for example for pH) the methods for the DM QC of the six BGC variables. The results of this activity were reported in the deliverables D4.2, D4.4, D4.5, D4.6 and D4.7 (D4.3 reports RTQC procedures), available here . Moreover, most of the Euro-Argo-RISE propositions have been presented at the ADMT 22, some of them have already been endorsed (e.g., RTQC for BBP, DM for Radiometry) and some of them are still being tested (ex: DM for CHLA).

These results lay the scientific foundations of the future operational system to process the six BGC variables of OneArgo. There is, however, still one critical point. Although QC methods are now available, the organisation of the operational teams in charge of the processing and distribution of the data requires an additional effort. The identification of an efficient organisation among all the teams and the expertise involved will be critical for the next year's success of the BGC component of OneArgo.

This deliverable is devoted to a general review of the QC for the BGC parameters and, on the basis of this review, to some propositions for the organisation and the coordination of the European infrastructure devoted to the BGC data qualification.

On the basis of the methods described and tested in the first three years of Euro-Argo-RISE, that are recalled and synthesised in the first part of the deliverable, different propositions are suggested as an organisation (in terms of data flow, but, also, in terms of human and financial costs and partners' distribution of roles) for the QC for BGC variables. Note that the proposition of organisation is specific to the European context, and probably will not be adaptable to other situations.

The deliverable is organised in three sections:

1) BGC Data Workflow. Presently, documentation and methods have been developed and described separately for each parameter. In this section, an overview of steps required to process all the six parameters is described. This accounts for the level of maturity of each QC method, the required temporal hierarchy of the processing and the required inputs of ancillary data. Moreover, an analysis of the automatic, semi-automatic and human required processing is exposed. This point will be used at the end, as the cost analysis of the processing is dependent on the ratio between these three methods.
2) Organisation. In this section some alternative organisations for BGC processing in Europe are proposed. For each organisation, and on the basis of the results of the previous section, pros and cons are listed. A cost analysis (financial and human) is also attempted.
3) FInal Remarks. Finally, the participants to the deliverable will draw some conclusions, proposing recommendations for the evolution of the BGC component of the data Euro-Argo infrastructure. The final thoughts of the deliverable capture the present status of the BGC data quality organisation and are certainly not definitive. They will represent a starting point for further discussions within the Euro-Argo ERIC, and they have to include also teams not participating to Euro-Argo RISE project

3. BGC Data workflow

In this section, we will describe the QC workflow for BGC, considering all the parameters (and not separately as in the deliverables 4.2-7), to take into account their interdependence. This description should give a general overview of the processing and consequently should facilitate the identification of the general organisation. At the end of this section, an first attempt is proposed to quantify the time needed for each step of the BGC data flow. Although preliminary, and certainly biased by some assumptions (see section 3.3), this estimation is critical to identify a future BGC-Argo data organisation, which must be scientifically "as-best-as-possible" but also sustainable in terms of human and financial resources.

3.1. Timing of the data management processing

The timing of the different steps of QC is analysed and described in Bittig et al. (2019). It is recalled here to give a temporal context of the BGC QC processing. In the rest of the document, we focused specifically on the "final" situation (i.e. the dark green in Figure 1). Intermediate steps will also be briefly discussed.

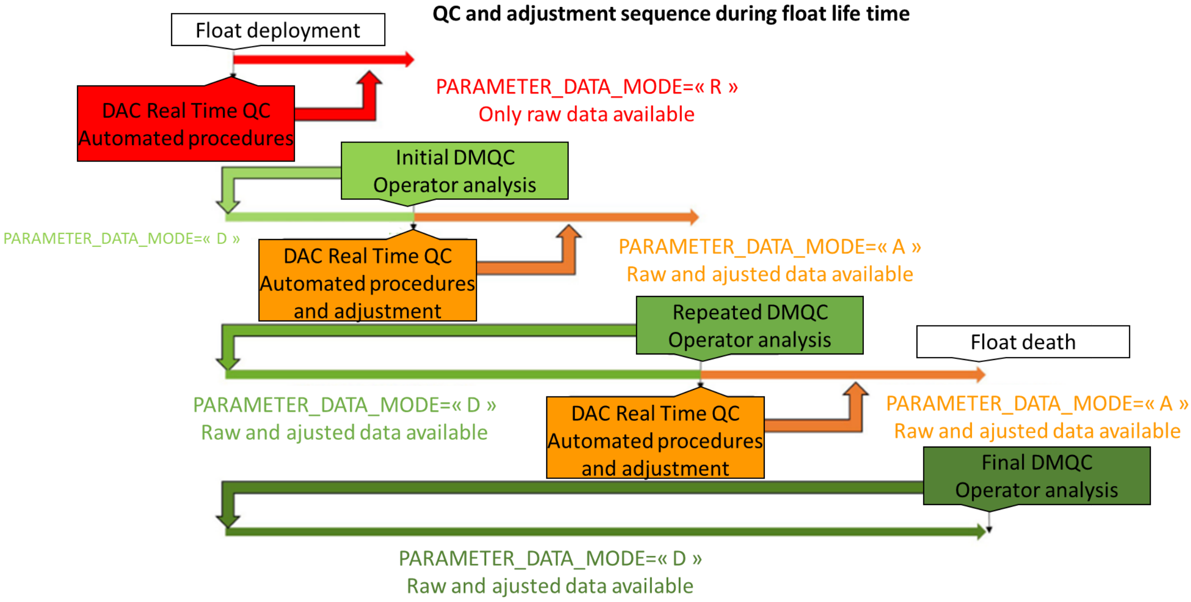

QC and adjustment sequence during float life time
Float deployment
DAC Real Time QC
Automated procedures
PARAMETER_DATA_MODE=« R »
Only raw data available
Initial DMQC
Operator analysis
PARAMETER_DATA_MODE=« D »
DAC Real Time QC
Automated procedures
and adjustment
PARAMETER_DATA_MODE=« A »
Raw and ajusted data available
Repeated DMQC
Operator analysis
Float death
PARAMETER_DATA_MODE=« D »
Raw and ajusted data available
DAC Real Time QC
Automated procedures
and adjustment
PARAMETER_DATA_MODE=« A »
Raw and ajusted data available
Final DMQC
Operator analysis
PARAMETER_DATA_MODE=« D »
Raw and ajusted data available

Figure 1: Sequence for QC (as obtained from Bittig et al., 2019). Sequence of quality control and adjustment steps during the lifetime of a float from float deployment to depth. The color shading indicates the data mode: “R” real-time data in red, ”A” real-time adjusted data in orange and “D” delayed-mode data in green. Initial DMQC should be performed soon after deployment (typically after 5-10 cycles, for DOXY, pH and NITRATE, and ~6 months for Radiometry, CHLA and BBP). With subsequent revisits (on an annual basis), adjustments become more reliable (indicated by the green shading).

3.2. BGC data workflow steps

When the schematic representation of Figure 1 is broken down into the 6 BGC core parameters (Figure 2, where the contours of the boxes have the same colour code as in Figure 1 for red=R, orange=A and green=D PARAMETER_DATA_MODE), we obtain an overall scheme of the data processing, which is based on the results described in the Euro-Argo-RISE deliverables (https://www.euro-argo.eu/EU-Projects/Euro-Argo-RISE-2019-2022/Deliverables D4.2, D4.3, D4.4, D4.5, D4.6, D4.7) and on the ADMT documentation (https://biogeochemical-argo.org/data-management.php).

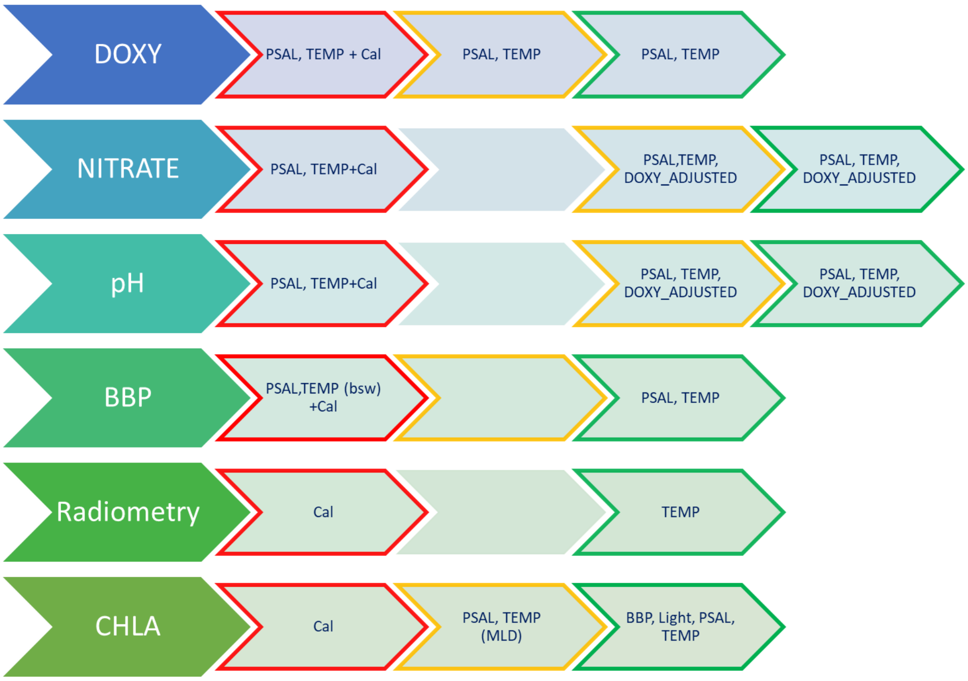


Figure 2: Scheme of the data workflow of the 6 core BGC parameters with the requested inputs (“Cal” stands for factory calibration): computation of the raw PARAMETER (PARAMETER_DATA_MODE=”R”, in red), automatic computation of the PARAMETER_ADJUSTED (PARAMETER_DATA_MODE=”A”, in orange) and computation of the PARAMETER_ADJUSTED in delayed-mode (PARAMETER_DATA_MODE=”D”, in green). The red and the orange parts are performed by the DAC, while the green part is performed by a DM operator.

Note that:

1. RT processing (in red) is performed simultaneously for all the parameters at the DAC.
2. Two parameters (bbp and radiometry), as it is not needed, don’t have an Adjusted step performed automatically in real time (see deliverables D4.3 and D4.4, and Jutard et al. 2021) but the community has agreed to fill the BBP_ADJUSTED field after the RTQC has been applied.
3. Two parameters (NITRATE and pH) depend on the prior computation of DOXY.
4. For the CHLA parameter, DMQC method depends on the combination of sensors equipping the float (see deliverable D4.2). In some cases, this implies the use of bbp and/or radiometry data. For this reason, the CHLA parameter must be processed after bbp and radiometry profiles.
5. All the parameters require (at different moments of the processing) PSAL and TEMP. Temperatures in the Argo profiles are accurate to ± 0.002°C and uncorrected salinities are usually accurate to ± 0.01 psu (Alert set at ± 0.05 psu). These relatively small uncertainties in PSAL and TEMP (in Real time) have little impact on the estimation of the BGC parameters in delayed time and are adapted to the “routine” mode (light green shadings in Figure 1). Then, we don’t need to wait for the CTD to be qualified in delayed mode for the “Initial DMQC Operator Analysis” of BGC parameters. The specific case of strong salinity drift will be addressed in the Annex 5.3.

When we focus on the DM step, and after the operational end of the float functioning (dark green line in Figure 1), the definitive sequence of the operations is (Flowchart on Figure 3) :

- Perform the DM for the CTD data
    - use the DM of TEMP to perform the DM of the radiometry data
    - use the DM of PSAL, TEMP to perform the DM of DOXY
        - use the DM of PSAL, TEMP, DOXY to perform the DM of NITRATE
        - use the DM of PSAL, TEMP, DOXY to perform the DM of pH
    - use the DM of PSAL to perform the DM of BBP
        - use the DM of PSAL, TEMP, BBP, radiometry to perform the DM of CHLA (MLD, Light penetration, ratio BBP/CHLA in the MLD)

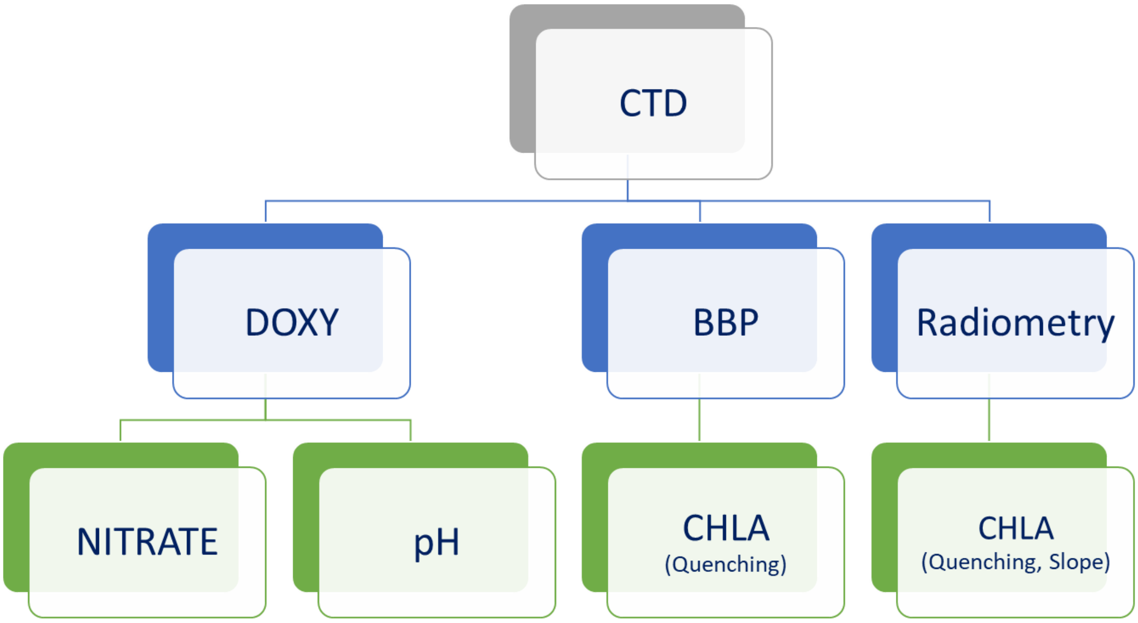


Figure 3 : Definitive sequence of the Delayed Mode for a dead float

Regarding “chemical” data, NITRATE and pH adjustment procedures rely on a neural network approach (Bittig et al., 2018, deliverables D4.3 and D4.7) by comparing float data at depth with a reference value. The reference value is estimated with the PSAL, TEMP, DOXY and location of the profile. While PSAL and TEMP often don’t require any adjustment, the DOXY parameter requires, likely always, an adjustment (with in-air measurements or compared to a climatology). So one **step that can not be neglected in the timing of DM operations for the chemical data, is the adjustment of DOXY prior to NITRATE and pH adjustments**.

Figure 2 has been then redrawn as Figure 4 to distinguish the steps of the workflow that are automatic (DAC PROCESSING), the steps that can be implemented by data scientists (OPERATOR

PROCESSING which means without deep know-how on the parameters) from those that require strong expertise of the parameters (EXPERT DECISION) and, in some specific cases, of the geographical region.

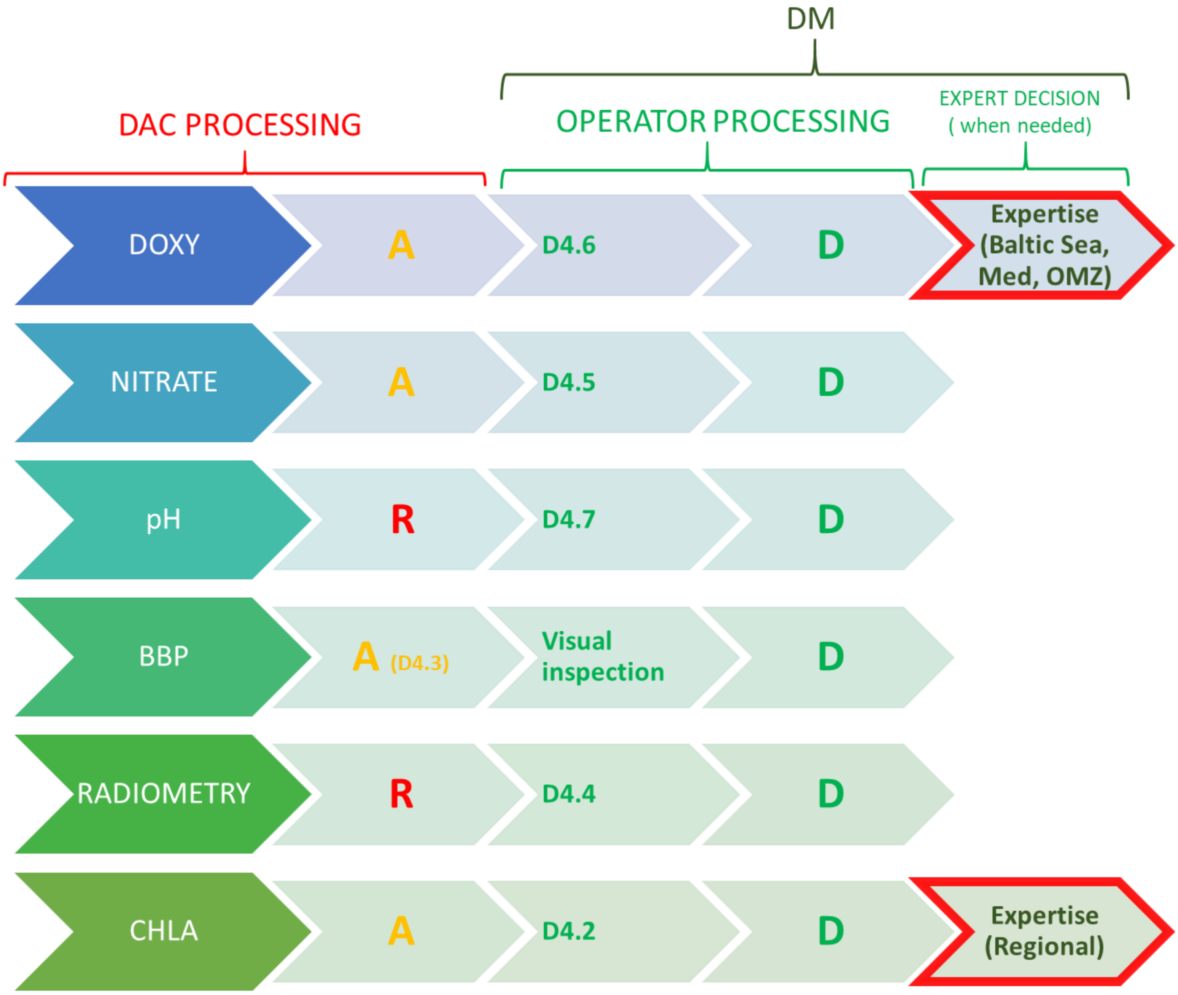


Figure 4 : Illustrations of the status of the QC procedures. The red arrows illustrate some situations encountered by an operational DM operator (Data scientist) that need scientific expertise to make the final decision.

One of the most important results of Euro-Argo-RISE was to develop and improve a series of algorithms/tools to perform DM for all the 6 core BGC variables, which could be operated by a data scientist (OPERATOR PROCESSING "zone" in Figure 4, (https://www.euro-argo.eu/EU-Projects/Euro-Argo-RISE-2019-2022/Deliverables D4.2, D4.3, D4.4, D4.5, D4.6, D4.7) without requiring strong know-how on the parameter. In other words, methods are now sufficiently robust to provide QC without requiring strong scientific expertise to validate the results.

However, for the DOXY and CHLA parameters, questionable qualification could still occur, mainly because the variability of these two parameters is still not completely characterised (thus not inventoried in the existing databases used as reference) or strongly anomalous compared to the general behaviour of the parameter in the rest of the ocean (as for example the DOXY in the OMZ).

Generally, these cases occurs in specific regions of the ocean (i.e. Black Sea, Mediterranean and Baltic Seas, OMZ areas).

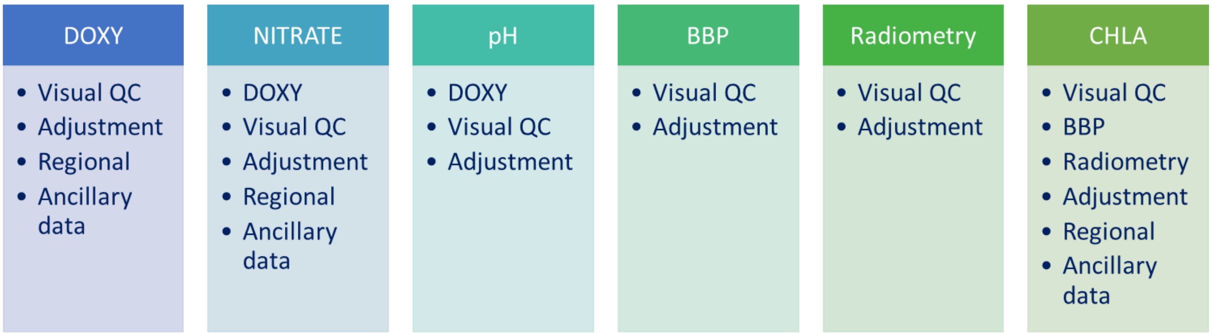
Presently, the final decision of these cases is submitted to a thematic “Expert” (or a group of “Experts”), who have recognized knowledge of the regions and of the parameter. Experts validate (or not) the OPERATOR PROCESSING and, potentially, propose an alternative correction(red arrows in Figure 4).

On the one hand, the role of the experts could be less critical in the next few years, when reference databases will grow, or thanks to an evolution of the QC methods. On the other hand, when the number of floats will increase, with deployments in regions weakly sampled in the past (and thus poorly represented in the reference database), the experts’ intervention could be still required.

3.3. Time allotted to DM

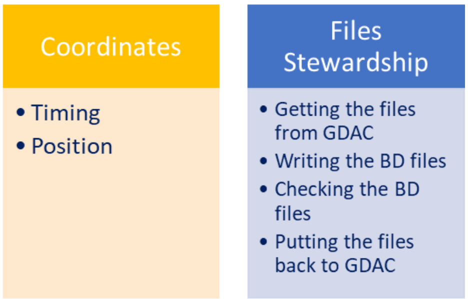
On the basis of the results presented in the previous sections, a rough estimation of the time required to perform the different Delayed Mode QC steps is presented. We based our calculation on the legal working time in France, which is 1607 hours for a Full Time Equivalent (FTE) and assuming that the DM is performed by a data scientist, expert on Argo data format. Note that the time presented here concerns only the OPERATOR PROCESSING step (i.e. without the red arrows of figure 4) and can be adjusted according to the scenario that will be discussed in section 4. The expert time is not included, as presently it is really difficult to evaluate and strongly dependent on the organisation. Moreover, these estimates are arithmetic estimates (sum of all the allotted times), it doesn’t take into account the extra time needed to address unplanned situations which may be linked to many potential issues (sensor failure, float operating in an area presently unknown where more investigations are needed, weakness of the present procedures…). The experience with Temperature and Salinity shows that time to process those floats is multiplied by a factor of 2 to 5. Finally, due to the potential different actors that may be involved in the processing, additional time to coordinate the DM for all the variables including submission to the GDAC with all the DM report should be added. Therefore our estimation of the workload is estimated to at least be 4 (FTE) which is coherent with the recommendations from the EuroGOOS DATAMEQ working group to allocate 10% of the cost of an equipment for its data management all along its lifecycle.

The delayed mode process is supposed to be performed at least once a year. All the estimated time associated with one parameter includes the visual inspection of each parameter profile. This visual inspection allows to remove obvious outliers and doesn’t require strong expertise.



DOXY	NITRATE	pH	BBP	Radiometry	CHLA
• Visual QC • Adjustment • Regional • Ancillary data	• DOXY • Visual QC • Adjustment • Regional • Ancillary data	• DOXY • Visual QC • Adjustment	• Visual QC • Adjustment	• Visual QC • Adjustment	• Visual QC • BBP • Radiometry • Adjustment • Regional • Ancillary data

Figure 5: Checklist of each step related to the DM of a specific BGC parameter. (Check https://www.euro-argo.eu/EU-Projects/Euro-Argo-RISE-2019-2022/Deliverables D4.2, D4.4, D4.5, D4.6, D4.7 for details). Regarding BBP, DM procedures are still pending, few floats were corrected into DM after visual inspection and an OFFSET application.



Coordinates

- Timing
- Position

Files Stewardship

- Getting the files from GDAC
- Writing the BD files
- Checking the BD files
- Putting the files back to GDAC

Figure 6: Checklist of generic steps that needs to be checked and done during the DM process

In Figure 6, we illustrate what are the common steps to be performed that are not relative to a specific parameter. These steps are relative to data management and not to scientific expertise.

Table 1: Estimated time to perform the DM of a float by a data scientist. The target of the BGC-Argo float array is 1000 floats at sea, and the European part represents one quarter of the global array which is 250 floats.

Note that only OPERATOR PROCESSING is considered here (i.e. without red arrows of Figure 4).

Parameter	Time alloted for one float (hours per year)	Time for the European array (hours per year)
CHLA, BBP, Radiometry	8	~2000
DOXY, pH, NITRATE	8	~2000
Coordinates/ Files stewardship	4	~1000
Total	20	~5000

4. Organisation

4.1. The actors

For the BGC data management, two main actors are identified:

1. Operational DM operators (data scientist)

   "Operational DM operators" are data scientists familiar with BGC-Argo format and BGC-Argo procedures. They contribute to the improvement of the methods by interacting with the scientific community, end users and with the deploying teams. They regularly provide statistics on the database they are responsible for (i.e. audit).

2. Experts

   "Experts" are scientific experts on a specific location and/or a parameter and they are able to provide recommendations and feedback on the correction proposed, but don't necessarily need to be expert of the method used or familiar with Argo format. They have a recognized role and they accept to support the QC systems.

4.2. The different scenarios

The structure of BGC data management at the European level should then comprise an efficient and cost-effective arrangement of the two actors cited above.

Three scenarios are identified:

Scenario 1 (“every man for himself”): Each country implements its “DM processing capabilities ” and its “experts” pool, who will apply the recommended procedures (endorsed at ADMT).

Scenario 2 (“long-lived centralisation”): A centralised entity performs DM for all the floats (the pink circle). The centralised entity is helped by a pool of “Experts”, scientific experts on a parameter and/or on an area identified in the community, involved or not in the deployment of floats. The “centralised entity” must be understood as “a well identified team” and not with a geographical meaning. “Experts” and “operators” could be geographically distant from the centralised data centre.

Scenario 3 (“happy to be in solidarity”): Some countries implement their “DM processing capabilities” which will perform DM for their own fleet, plus some of the other European floats deployed in the same areas or for a specific parameter (DOXY for example as illustrated, then once the DM is done, it can be sent directly to the GDAC or to another team to perform DM for pH and NITRATE). This scenario can be also arranged in groupings of parameters.

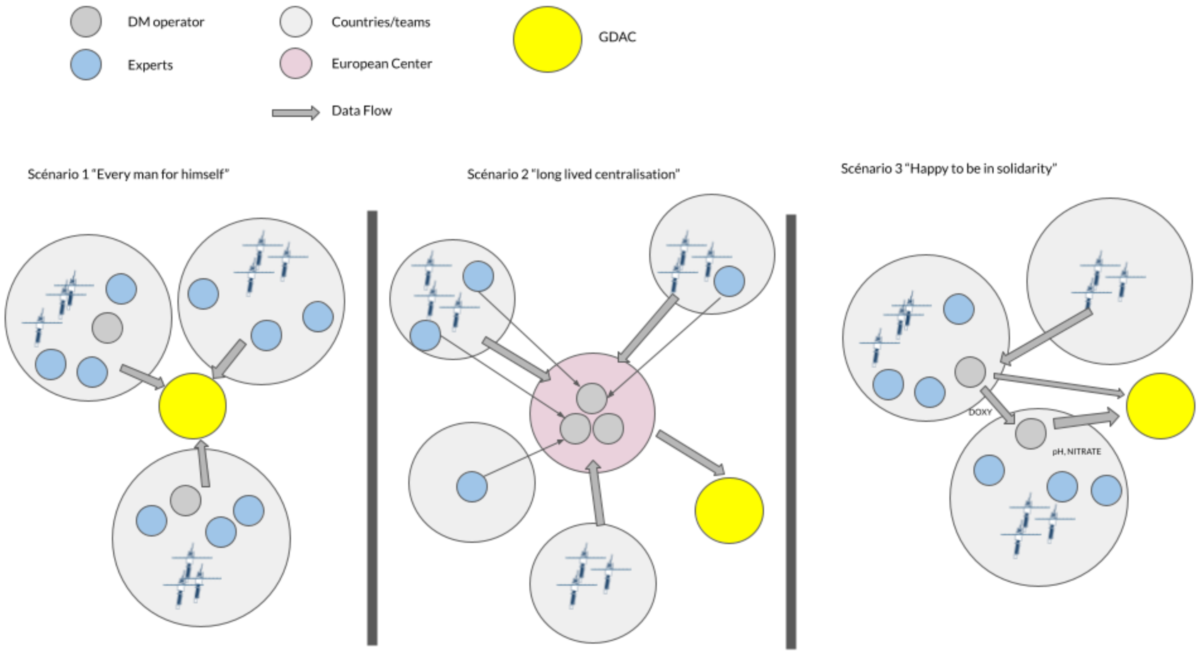

DM operator
Experts
Countries/teams
European Center
Data Flow
GDAC
Scénario 1 "Every man for himself"
Scénario 2 "long lived centralisation"
Scénario 3 "Happy to be in solidarity"
DOXY
pH, NITRATE

Figure 7: Illustrations of the different proposed scenarii

4.3. Pros and cons

Table 2: Pros and cons for each scenario identified

Scenarios	Pros	Cons
1	- Each country is autonomous in the BGC processing - Proximity between deploying teams and data managers, favoring interactions - The system could be built up gradually, without huge investments	- Not efficient for the Country that are deploying few floats (time consuming) - Discouraging for countries seeking to join Argo and wishing to only deploy a few floats - Poor Consistency of the dataset threatened by the application of the agreed procedures from one team to another - Poor Consistency of the dataset can be threatened by the independence of the team on the agreed procedures for DM - Not cost effective (multiplication of experts and operational DM operator) - Low sustainability (depends on independent team fundings) - Potentially difficult to find experts for all the situations
2	- Economy of scale for the members who really rely on a professional team - Consistency of the methods and on the databases - Cost effective (number of experts and operators reduced) - Sustainability - Coordination at the european scale - Unique interlocutor for the end users, which minimise errors and wrong information	- Funding scheme needs to be defined - Require some investments (although mainly at the beginning) - Have to deal with a relative large number of experts (as required to cover all the potential situations) - It should have an “international” status, to be directly included in the ERIC
3	- This organisation has proven for T&S to be efficient as the DM-Operators are processing enough floats to have strong expertise in DMQC - Cost effective - Sustainability	- Volunteers that will perform DM for others countries should assure engagement on the long term - If the scenario is organised on the PARAMETER basis, bottlenecks could appear (Figure 3, ex : DOXY prior to NITRATE, pH) - Funding schema needs to be defined - Risk of complicating the interaction between experts and data management (i.e. experts for a given region will likely come from interested countries, even if these countries have no DM capabilities) - Poor readiness for end-users and deploying teams

5. Final remarks

5.1. General Synthesis on QC

Presently, and thanks to the effort done in Euro-Argo RISE, QC of BGC floats is reaching a level which could be considered robust and consolidated. Data flows are now quite finalised. The pH and the bbp parameters will still require an additional effort: the first because the number of floats is still too low to have robust statistics, the second because of the delay in the finalisation of RT[1].

As a general remark, we would like to stress the difficulty of "adapting" the model of the T&S DM to the BGC DM. The 6 BGC parameters, compared to the 2 parameters of Argo (mainly PSAL), implies that there are numerous different adjustment and control quality procedures that need to be mastered. The main consequence of this is that the "hierarchy" of the processing is critical. There is a specific order to follow in the DM workflow (i.e. PSAL, DOXY, NITRATE and pH), which consequently requires a strong coordination between the involved teams. This is particularly relevant if the DM is shared over several teams for the same float/parameter.

Moreover, the QC procedures on "positions" and "time" parameters should deserve a greater attention than in the present configuration. With an increasing complexity due to the presence of this piece of information in the Core files, the B files and the S files, a special care must be taken to also verify this information in the DM process.

As a matter of fact, the main issue of the QC for the BGC floats is the lack of a robust, consolidated and global data set of reference. This implies that a large range of methods and procedures are developed, which, although often supported by scientific results (i.e. papers), were often hard to apply in a general system.

The development, and the consecutive integration in the QC system, of the new Neural Networks (NN) methods is dramatically changing the situation. Integrating independent sources of data (i.e. remote sensing, rosette bottle), which alone would be useless for a global world validation, in a unique frame able to generate a modelled profile for each measured profile, is now providing a reference data set to verify the quality of the QC processing. Moreover, the evolutive intrinsic nature of the NN techniques provides a method to adjust, and possibly improve, the existing procedure. Finally, they are completely automatic, minimising the need for expert intervention.

At the present day, the QC processing could be operated without a specific expertise (see section 3.2). Experts are, however, still required. Presently, this is the case only in some regions, although it could be required somewhere else, when the OneArgo network coverage will be achieved.

Without considering the experts, our evaluation of the number of people required to operationally process the DMQC BGC European fleet is 2 (see table 1), which should be at least doubled to maintain a 24h/365 days operational structure. This estimation is, of course, really raw and could be applied only in the case of a central entity managing all the floats of the European fleet (i.e. scenario 2 of section 4.3). In the case of a distributed system (i.e. scenarios 1 and 3), the number of floats decreases for each country, consequently decreasing the total number of hours required for the QC. However, for these scenarios, to the time required to process a float, it should be added the time (a) to update the procedures, (b) to identify the experts and (c) to constantly check the consistency with other data centres (see discussion in the next section).

[1] A publication providing the recommendations of the international community on RT QC for bbp has just been submitted in Autumn 2022: https://doi.org/10.12688/openreseurope.15047.1

5.2. Recommendations for a BGC QC Organisation at European level

On the basis of the analysis presented before (and in particular section 4.2), some recommendations could be provided by the Euro-Argo RISE WP4 participants, in order to initiate a discussion around the European organisation of the BGC QC.

1. **The consistency of the data and of the methods must be considered the first preoccupation for a BGC QC organisation.** We consider that we have to absolutely avoid data in different data centres that are processed in a different way. Consequently, and under this aspect, we consider a centralised entity (scenario 2) the most suitable organisation for the QC. For the other scenarios, a tight coordination must be organised, to regularly verify the processing chains and the intermediate and final data in the different data centres. Centralisation would also strongly facilitate potential re-processing of data, which could be hard to coordinate if data centres are multiple.
2. **The interaction with the scientific community is the second main point to consider.** For the scientific community we intend: the experts, the deploying teams, scientists working on the QC methods, end-users (however see later for a specific point on experts). The three scenarios have all pros and cons concerning this point, and it is then hard for us to identify the best model. For example, a centralised system (i.e. scenario 2) could be more suitable to interact with end-users, while it could be complex for the coordination of the large experts pool required to deal with all the potential conditions requiring an expert intervention. On the other hand, experts could be more easily managed in scenario 3 (with a sort of repartition of parameters/regions between 2-3 data centres), although contact with the end-users could be more complicated in this case. A complete distributed scenario (i.e. number 1), will certainly help deploying teams, as each country could have its supporting QC team tightly in contact with deploying teams. On the other hand, in the case of scenario 1, determining all the required experts for each country would be hard. Additionally, several pools of experts will require a top-level coordination, to ensure consistency of processing (see points 1 and 4).
3. **The sustainability of the system is the third point to be accounted for.** In terms of the manpower required to operationally process the European fleet (i.e. the goal of European contribution to OneArgo, 250 floats with the 6 BGC parameters), our estimation gives 4 full time equivalent (FTE) for scenario 2 for an operational processing of the European fleet. For the two other scenarios, the computation is hard to perform, as the number of floats for each country could be highly variable and, also, evolving with time. We anticipate, however, that, even with a relatively low number of floats, a data centre should, at least, employ one person full time to be operational and to achieve the tasks required. This could be then highly expensive for scenario 1, and also for scenario 3. In terms of the financial investment for facilities, a BGC data centre is relatively less demanding than, for example, hardware installations for sensor and platform support. High-speed internet connection and medium-level computational and storage capacities are required, as it is possible to find in 90% of the research laboratories in Europe. Note, however, that a multiplication of data centres (as in scenario 1 and 3) would also multiply the cost of facilities.
4. **We would also suggest a specific point on the "Experts"**. At the present day (see section 3), processing of BGC requires episodically an expert's contribution to manage "exceptional" data (i.e. poorly represented in climatologies or obtained under rarely or never observed conditions or in regions strongly undersampled). In these cases, data centres require then a validation of an expert, who could accept the final result of the automatic QC or, alternatively, suggest modifications to improve correction or, in the worst case scenario, reject the

observations definitively. It is hard, at this stage of development of BGC QC, to evaluate the timing that experts spend to support data centres, although we consider that their activity is critical and it should be organised if an efficient system is implemented. Consequently, for the specific aspect of the experts, it is hard to recommend the best scenario among the three proposed. We have however some recommendations:

a. **Expert activity has to be recognized.** Experts are presently working on a volunteer basis, which cannot be sustainable in the long term. A sort of acknowledgement (even financial) needs to be found.
b. **Expert activity has to be coordinated.** General criteria and practices have to be defined by dedicated working groups and they should be used systematically during the operational process of expert validation.
c. **Expert intervention needs to be regularly revisited.** Development of automatic methods is a continuous process, which regularly decreases the number of profiles demanding expert intervention. On the other hand, BGC floats are continuously deployed, also in regions strongly under-sampled or currently unexplored, thus demanding an expert know-how which was unnecessary before. Moreover, new parameters (derived by existing parameters or directly acquired by new sensors) could be endorsed in the future and expert intervention could be required to perform QC. Overall, a regular and coordinated review of experts intervention is then mandatory and should be organised.
d. **Experts have to work closely with the data centre.** This point would be considered obvious, although it is, for us, particularly critical and not easy to manage. Ideally, experts and data operators should work under the same entity and they must maintain tight and not only virtual interactions. Alternatively, systematic and recurrent exchanges should be organised, for the most virtually, and, as much as possible, with in-person meetings. The expert-data operator exchanges should follow general protocols (see point b), although adapted by the specific configuration of the data centre (i.e. country dedicated, parameter dedicated, region dedicated).

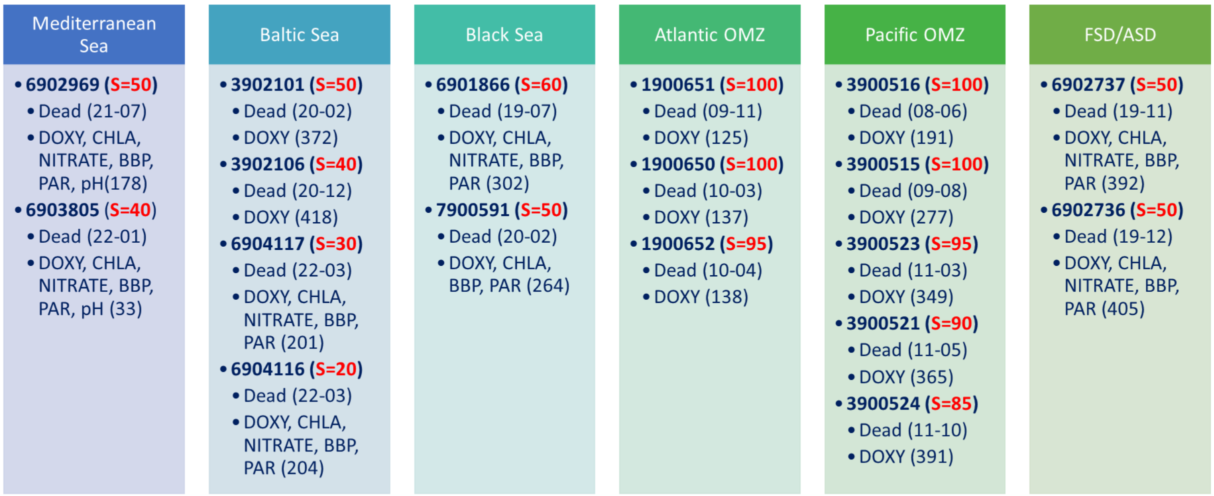
The previous recommendations are, obviously, the results of the WP4 Euro-Argo RISE consortium and therefore they are, by definition, partial. The final decision on the BGC-Argo structure will be decided by high-level coordination entity of the Euro-Argo infrastructure (i.e. Euro-Argo ERIC Management Board) and by extending the discussion outside the perimeter of the project. All the participants of the WP4 are, however, strongly implicated in the BGC Argo activity at the European level. They will thus continue to contribute also after the end of the project.

6. Annexes

6.1. Setting up alerts

One of the first challenges that a DM operator is facing is to know what floats should be processed first. To solve this issue, we imagine setting a global priority list which will be the gathering of several priority lists. These priority lists will be calculated at the DAC.

To sum up, these priority lists will be set up with a score that will be calculated with a combination of the argo-bio profile index , the greylist and the PROFILE_PARAM_QC and SCIENTIFIC_CALIB_xxx fields (in Bfiles). It will take into account : the number of profiles, the number of parameters, the issues present in the greylist, the DATE_UPDATE, the status of the float (Dead, Alive), the date of the last transmission, with parameters that are affected by drift (Salinity, Doxy) ….



Mediterranean Sea

- **6902969 (S=50)**
  - Dead (21-07)
  - DOXY, CHLA, NITRATE, BBP, PAR, pH(178)
- **6903805 (S=40)**
  - Dead (22-01)
  - DOXY, CHLA, NITRATE, BBP, PAR, pH (33)

Baltic Sea

- **3902101 (S=50)**
  - Dead (20-02)
  - DOXY (372)
- **3902106 (S=40)**
  - Dead (20-12)
  - DOXY (418)
- **6904117 (S=30)**
  - Dead (22-03)
  - DOXY, CHLA, NITRATE, BBP, PAR (201)
- **6904116 (S=20)**
  - Dead (22-03)
  - DOXY, CHLA, NITRATE, BBP, PAR (204)

Black Sea

- **6901866 (S=60)**
  - Dead (19-07)
  - DOXY, CHLA, NITRATE, BBP, PAR (302)
- **7900591 (S=50)**
  - Dead (20-02)
  - DOXY, CHLA, BBP, PAR (264)

Atlantic OMZ

- **1900651 (S=100)**
  - Dead (09-11)
  - DOXY (125)
- **1900650 (S=100)**
  - Dead (10-03)
  - DOXY (137)
- **1900652 (S=95)**
  - Dead (10-04)
  - DOXY (138)

Pacific OMZ

- **3900516 (S=100)**
  - Dead (08-06)
  - DOXY (191)
- **3900515 (S=100)**
  - Dead (09-08)
  - DOXY (277)
- **3900523 (S=95)**
  - Dead (11-03)
  - DOXY (349)
- **3900521 (S=90)**
  - Dead (11-05)
  - DOXY (365)
- **3900524 (S=85)**
  - Dead (11-10)
  - DOXY (391)

FSD/ASD

- **6902737 (S=50)**
  - Dead (19-11)
  - DOXY, CHLA, NITRATE, BBP, PAR (392)
- **6902736 (S=50)**
  - Dead (19-12)
  - DOXY, CHLA, NITRATE, BBP, PAR (405)

S is estimated with the Argo bio profile index (parameter, (R,A,D) + date_update + location) + Greylist (sensors, salinity) + nb profiles
The specifications to estimate the score to establish the priority lists are still under development (Schmechtig C., Racapé V.)

Figure 8 : Example of how could be designed a priority list. (On-going work)

6.2. Checking positions and Time

In 2015, in order to ease the insertion of BGC data in the ARGO data system, it was decided to split the profile parameters into two main files. First, the so-called “c” file or “core file” contains all the parameters relative to the CTD, that are PRES, TEMP and PSAL. Then, the so-called “b” file contains PRES and all parameters relative to biogeochemical data (for example, CHLA, BBP700, DOXY, NITRATE ….). This choice aims to keep together both information while they were at different stages of maturation.

This said, both “c”file and “b”file share some pieces of information that need to be checked and aligned, for example the parameters relative to the timing and position.

LATITUDE is the latitude of the profile

LONGITUDE is the longitude of the profile

POSITION_QC is the quality flag of the position (1,2,3,4, 8 (interpolated), 9 (undefined))

JULD is the julian day of the profile

JULD_QC is the quality flag of the date of the profile

JULD_LOCATION is the julian day of the location information (it can be the julian day of the GPS point).

Dealing with time and position is not a strictly a “BGC” topic, but, in the BGC community, synthetic profiles (https://doi.org/10.13155/55637) are widely used and their generation requires to have “c” files and “b”files aligned regarding time and position.

6.3. DMQC analysis in the particular case of salinity drift

Over the last three years, there has been an upsurge of CTDs with a drifting salinity. As mentioned previously, most of the time, the PSAL is “not bad enough” to prevent the BGC argo data QC to be performed, but in some cases identified as FSD, (Fast Salinity Drift) or ASD (Abrupt Salinity Drift), this drift should be accounted for while it generates aberrant values mostly for DOXY, NITRATE and pH, but can be also used to recompute the contribution of pure sea water for the BBP parameter.

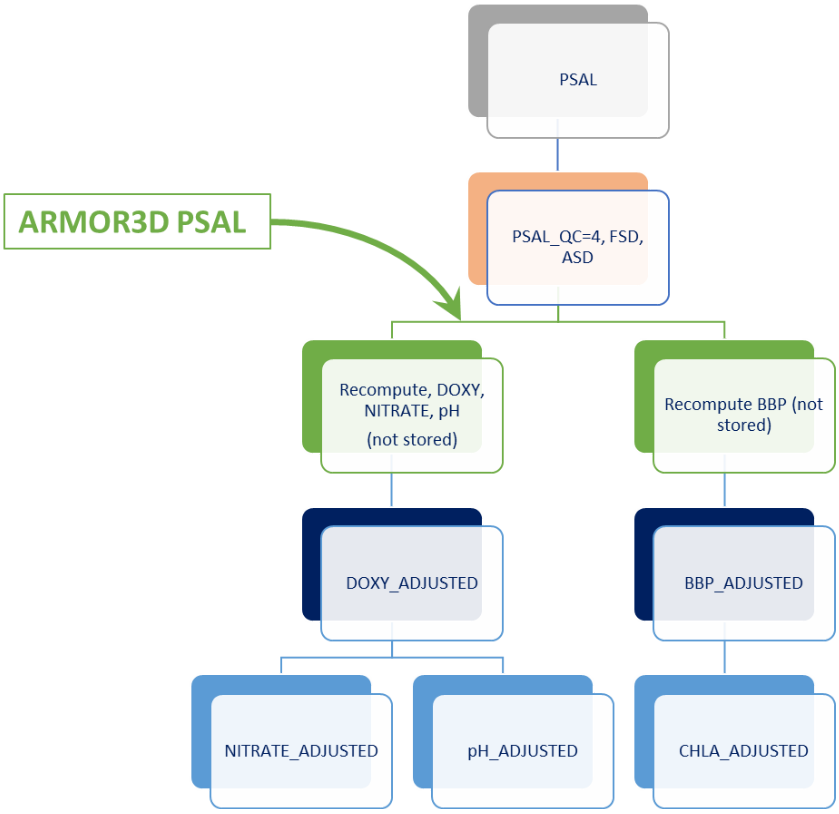


Figure 9 : Illustration of the sequence of the DM procedures in case of PSAL_QC=4 caused by a salinity drift. The “not stored” mention illustrates the fact that the raw parameters are recomputed with a PSAL proxy (but these “raw” parameters are not stored in the files) and are used to perform and estimate the final stored PARAMETER_ADJUSTED.

7. References

Bittig Henry C., Maurer Tanya L., Plant Joshua N., Schmechtig Catherine, Wong Annie P. S., Claustre Hervé, Trull Thomas W., Udaya Bhaskar T. V. S., Boss Emmanuel, Dall'Olmo Giorgio, Organelli Emanuele, Poteau Antoine, Johnson Kenneth S., Hanstein Craig, Leymarie Edouard, Le Reste Serge, Riser Stephen C., Rupan A. Rick, Taillandier Vincent, Thierry Virginie, Xing Xiaogang, 2019 : A BGC-Argo Guide: Planning, Deployment, Data Handling and Usage. Frontiers in Marine Science, 6. https://doi.org/10.3389/fmars.2019.00502

Bittig, H., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N.L., Sauzede, R. Körtzinger, A. and J.-P. Gattuso (2018). An alternative to static climatologies: Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S and O2 data using Bayesian neural networks, 5 (328), Frontiers in Marine Science, 10.3389/fmars.2018.00328

Dall'Olmo, G., Bhaskar TVS, U., Bittig, H., Boss, E., Brewster, J., Claustre, H., Donnelly, M., Maurer, T., Nicholson, D., Paba, V., Plant, J., Poteau, A., Sauzède, R., Schallenberg, C., and C. Schmechtig (2022) Real-time quality control of optical backscattering data from Biogeochemical-Argo floats, [version 1; peer review: awaiting peer review]. Open Res Europe 2022, 2:118 (https://doi.org/10.12688/openreseurope.15047.1)

Jutard, Q.; Organelli, E.; Briggs, N.; Xing, X.; Schmechtig, C.; Boss, E.; Poteau, A.; Leymarie, E.; Cornec, M.; D'Ortenzio, F.; Claustre, H. Correction of Biogeochemical-Argo Radiometry for Sensor Temperature-Dependence and Drift: Protocols for a Delayed-Mode Quality Control. Sensors 2021, 21, 6217. https://doi.org/10.3390/s21186217