EuroGO-SHIP
Enhancing ocean observations

Defining a framework for estimating uncertainty as part of a secondary quality control (2QC) procedure

Deliverable 4.3

26 November 2024 / Version 1.0

Co-funded by the European Union
UK Research and Innovation

This work was funded by the European Union under grant agreement no. 101094690 (EuroGO-SHIP) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10051458, 10068242, 10068528]. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

About this document

Title	D2.4. Report on defining a framework for estimating uncertainty as part of a 2QC procedure
Work Package	WP2 Concept Development trough co-design
Lead Partner	NORCE
Lead Author (Org)	Marta Alvarez (IEO-CSIC)
Contributing Author(s)	Siv K. Lauvset & Elaine McDonagh (NORCE), Yvonne Firing & Fatma Jebri (NOC), Malek Belgacem (CNR)
Reviewers	Ryan Weber (NORCE)
Due Date	30.11.2024, M24
Submission Date	26.11.2024
Version	**1.0**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

EuroGO-SHIP: Developing a Research Infrastructure Concept to Support European Hydrography is a Research and Innovation action (RIA) funded by the Horizon Europe Work programme topics addressed: HORIZON-INFRA-2022-DEV-01-01– Research infrastructure concept development. Start date: 01 December 2022. End date: 30 November 2025.

EuroGO-SHIP is co-funded by the European Union, Horizon Europe Funding Programme for research and innovation under grant agreement No. 101094690 and by UK Research and Innovation

Table of Contents

1. Introduction 5
   1.1. Background and motivation of the deliverable 5
   1.2. Objectives and scope of the deliverable 7
   1.3. Previous concepts and vocabulary regarding uncertainty 7
      1.3.1. Metrology and Oceanography 7
      1.3.2. Metrology terms used in this deliverable 8
2. Secondary quality control (2QC) procedures: crossover analysis 11
   2.1. Introduction 11
   2.2. GLODAP 2QC crossover analysis description 11
   2.3. 2QC crossover analysis flowchart and software tool 14
3. Framework to assess the uncertainty in a 2QC crossover analysis 15
   3.1. Applicability of metrological methods to assess uncertainty in a 2QC crossover analysis 16
      3.1.1. Bottom-up procedure: pen analysis 17
      3.1.2. Top-down procedure: inter-laboratory comparison exercises 18
   3.2. EuroGO-SHIP deliverable 2.4 approach 18
      3.2.1. Monte Carlo approach: introducing random uncertainty (noise) in the 2QC crossover analysis 19
      3.2.2. Impact of weighting options in the final adjusments 27
      3.2.3. Selected approach to assess uncertainty in the 2QC crossover analysis 30
4. Case studies 31
   4.1. Rationale for each case study 31
   4.2. Northeast Atlantic RADPROF cruises 31
   4.3. Cruises in the Western Mediterranean Sea 35
5. Conclusions 38
   5.1. Summary and main findings of the deliverable 38
   5.2. Contributions to the project and the European hydrography community 40
   5.3. Limitations and outlook 40

References 40

1. Introduction

The EuroGO-SHIP project aims to develop a concept for a European Research Infrastructure (RI) for hydrography, that is, the measurement of ocean physical and biogeochemical properties from the platform of a marine vessel. Measurements of parameters like ocean salinity and velocity, dissolved oxygen and inorganic carbon, inorganic nutrients and transient tracers throughout the water column are essential for understanding the ocean's role in local and global climate and ecosystems, and thus society's need for hydrographic data is growing. Currently many observations are made on a nation-by-nation or even institution-by-institution basis, producing data of variable quality in a fragmented way. The EuroGO-SHIP RI would address gaps in facilities and best practices, enabling the European hydrographic community to increase the quality, traceability, and availability of hydrographic data.

1.1. Background and motivation of the deliverable

International hydrographic programs such GO-SHIP (Global Ocean Ship Based Hydrographic Investigations Program) and predecessor programs WOCE (World Ocean Circulation Experiment) and CLIVAR (Climate Variability and Predictability) aim to assess the strength and pattern of climate change in ocean physical and chemical properties. Discerning climate over natural variability, given the low signal to noise ratio, requires a detailed evaluation of the observational errors. This evaluation is directly related to a well-documented and quantified data quality. In fact, GO-SHIP cruises have strict data quality requirements and need to follow the GO-SHIP Hydro Manual (Hood et al., 2010, updated in 2019) with specific standard operation procedures (SOP) and quality assurance and quality control (QA/QC) for accuracy and precision (Figure 1). Accuracy is checked against certified reference materials (CRMs) where available which are considered to have a known true value, and precision, repeatability, is assessed with repeated measurements on the same sample. In this regard, information about the quality of each variable should be included in the cruise report and the corresponding metadata information aligned with the data compilation. Finally, data and metadata should be properly submitted and stored into a National Oceanographic Data Centre following FAIR (Findable Accessible Interoperable Reusable) principles (Tanhua et al., 2019).

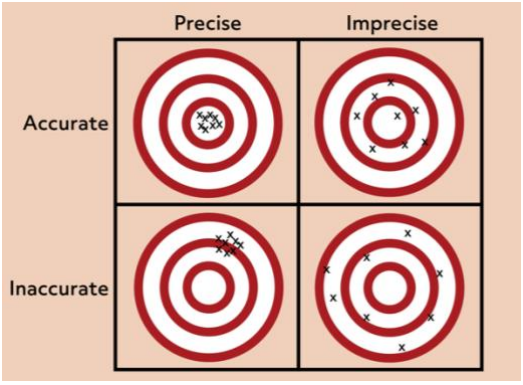

Figure 1. Schematics explaining the difference between accuracy and precision, two important concepts to assess data quality.

Data products for ship-based variables combine, format and quality control data from different cruises to make them more accessible and coherent. The individual data sets are managed by different laboratories, from different eras with evolving technology and measuring methods, and consequently different quality standards. Even on recent cruises where gold standard SOPs, such as the Hydro Manual (Hood et al., 2010) are followed, ad hoc adaptations to the procedure or peculiarities of the cruise/lab might introduce errors in the final individual data sets. EuroGO-SHIP Deliverable 3.3 (Firing et al., 2024) compared replicate salinity samples analysed in different laboratories and found mean differences of 0.001 to 0.005 psu. Biases in measurement equipment, its calibration or standardisation (e.g. Uchida et al., 2020) may also contribute.

Therefore, a data product which is the sum of individual datasets might contain a combination of random errors (mainly affecting the precision of the measurements) and systematic biases in the form of constant differences between datasets (cruises). Both random and systematic errors (Figure 2) contribute to a lack of coherence and homogeneity, i.e. increased uncertainty, in the final product.

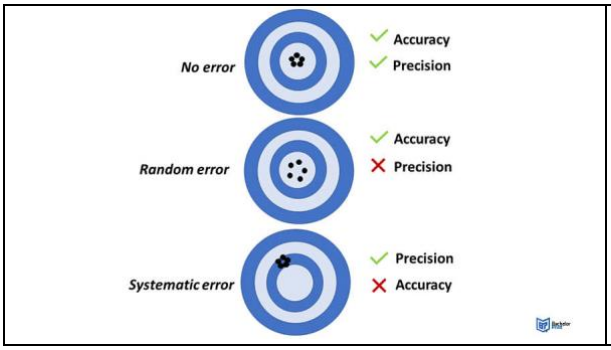

Figure 2. Schematics explaining the difference between random and systematic errors or uncertainty.

Identifying random errors is the subject of the primary quality control (1QC) procedures, which inspect the per variable and per cruise homogeneity in the data, including comparison to CRMs, standards, and alternate measures of some parameters, to finally add quality flags to each measurement. The secondary quality control procedure (2QC) evaluates the data consistency between several cruises. Typically, 2QC procedures inspect data at intersection points between cruise tracks, or crossovers, to detect systematic differences or biases in areas with low temporal variability and homogeneous characteristics, i.e., deep and bottom waters. The 2QC procedure will be further described in Section 2,

From the historical perspective, the first efforts to detect systematic biases in hydrographic physical and chemical data mainly started in the 1990s, with WOCE, an international effort to sample all ocean basins following the same procedures and quality requirements. When oceanographers combined those higher quality data with recovered historical data from previous programs with different methods, instruments, sampling density and therefore quality, the need to detect, objectively quantify and correct the “intercruise offsets” or systematic differences between cruises (Gouretski and Jancke, 1999) arose.

Currently, the quantity of hydrographic and biogeochemical measurements is rapidly increasing thanks to autonomous systems (Chai et al., 2020) that complement the traditional ship-based observations. Ship-based observations themselves comprise not only the high-

quality GO-SHIP cruises but also regional hydrographic cruises usually near coastal areas, for example cruises contributing to ICES-WGOH (International Council for the Exploration of the Seas - Working Group on Oceanic Hydrography) (González-Pola et al., 2023) or regional programs such as Med-SHIP (Mediterranean Sea Ship-based Hydrographic Investigations Program) (Schroeder et al., 2015). Merging such data sets to release comprehensive and quality assessed data products raises the need for a careful examination of the uncertainty.

EuroGO-SHIP is aware of the increasing demand to integrate multiple European hydrographic observations to assess the impact of climate change and extreme events in European waters. A perfect example is GLODAP (Global Ocean Data Analysis Project). GLODAP aims to provide high-quality and bias corrected water column bottle data from the ocean surface to bottom, in order to document the state and the evolving changes in physical and chemical ocean properties (Tanhua et al., 2021). GLODAP has been regularly updated with new data since GLODAPv1.1 (Key et al., 2004), followed by GLODAPv2 (Olsen et al., 2016), and yearly updates since 2019, with the most recent update happening in 2023 (Lauvset et al., 2024). Another example is the annually published ICES Report on Climate Change (IROC) for thermohaline properties (González-Pola et al., 2023).

1.2. Objectives and scope of the deliverable

The objective of this EuroGO-SHIP Deliverable 2.4 is to define a framework to improve the uncertainty estimate of data synthesis products merging data with different initial uncertainties, with the final aim of providing a more objective and coherent quantification of the final product uncertainty, which will increase its usability.

The EuroGO-SHIP framework for 2QC will build on the procedures used in GLODAP, refining the existing method in two ways:

(i) Introducing noise (analogous to reducing measurement precision) to both physical and chemical variables

(ii) Comparing different weighting schemes to calculate systematic biases (analogous to measurement accuracy) and final corrections, considering differences in time and space

The framework will be demonstrated using case studies in the Mediterranean Sea, the Nordic Seas, and the Northeast Atlantic Ocean.

1.3. Previous concepts and vocabulary regarding uncertainty

1.3.1. Metrology and Oceanography

In this section we define the concepts and vocabulary regarding quality assurance, quality control (QA/QC), secondary quality control (2QC) and uncertainty assessment used in this deliverable, following, where possible, metrology concepts.

Metrology is the science of measurement that establishes a universal system for measuring, in order to ensure the comparability of measurement results over time and space. Metrology formally started after the French Revolution with the introduction of the decimal metric system relying on a single definition of the unit “meter” (law of the 18 Germinal an III, 7th April

1795). Metrological consistency is now guaranteed through a close collaboration and interplay between national (National Metrology Institutes, NMIs), regional (Regional Metrology Organizations as EURAMET, European Association of National Metrology Institutes) and international (BIPM, *Bureau de poids et des mesures*) governance. Metrology requires a clear definition of the measurand (quantity to be measured), the validation of the measurement procedure, the uncertainty estimation or budget, the traceability and standardisation, to ensure the comparability of measurement results.

Applying Metrology concepts into Oceanography is a current ambition but also an urgent requirement, as monitoring the impact of climate change in the ocean requires space and time comparability of the gathered data for Essential Ocean and Climate Variables (EOVs and ECVs). Projects such as EU INFRAIA-02-2020 MINKE (Metrology for Integrated marine maNagement and Knowledge-transfer nEtwork) brings European marine science and Metrology Research Infrastructures together to identify synergies and create an innovative approach to QA/QC of some EOVs (Hartman et al., 2023). Other projects such as EMPIR SapHTies (Metrology for standardised seawater pH_T measurements in support of international and European climate strategies) deal with the metrological concept for pH in the ocean. Embedded in the EuroGO-SHIP initial concept (Table 2 within the proposal) are several important aspects related to data quality assurance (Best Practices, Quality Control and Data Management and Access) that touch on Metrology concepts (standards, traceability, uncertainty, reference materials).

1.3.2. Metrology terms used in this deliverable

Using the VIM guide ("International Vocabulary for Metrology", 2012), the table below lists the metrological concepts and terms used in this deliverable.

Metrology term and definition
1. Quantity • Property of a phenomenon, body, or substance where the property has a magnitude that can be expressed as a number and a reference
2. Measurand • Quantity intended to be measured. The specification of a measurand requires knowledge of the kind of quantity, description of the state of the phenomenon, body, or substance carrying the quantity, including any relevant component, and the chemical entities involved.
3. Measurement Accuracy • Closeness of agreement between a measured quantity value and a true quantity value of a measurand • A measurement is said to be more accurate when it offers a smaller measurement error.
4. Measurement trueness • Closeness of agreement between the average of an infinite number of replicates measured quantity values and a reference quantity value • Measurement trueness is not a quantity and thus cannot be expressed numerically

• Measurement trueness is inversely related to systematic measurement error but is not related to random measurement error.
5. Measurement Precision • Closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions. Sometimes measurement precision is erroneously used to mean measurement accuracy. • Repeatability: Precision evaluated under a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time • Reproducibility: Precision evaluated under a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects
6. Measurement error • Measured quantity value minus a reference quantity value • When there is a single reference quantity value to refer to, which occurs if a calibration is made by means of a measurement standard with a measured quantity value having a negligible measurement uncertainty or if a conventional quantity value is given, in which case the measurement error is known • Systematic measurement error or bias: estimate of a systematic measurement error, is a component of measurement error that in replicate measurements remains constant or varies in a predictable manner • Random measurement error: component of measurement error that in replicate measurements varies in an unpredictable manner
7. Uncertainty • Parameter associated with the results of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand • Uncertainty is usually expressed as a standard deviation and understood as a confidence interval
8. Traceability • property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty • the traceability chain should be linked to a primary reference measurement procedure using primary standards and linked to International System of Units (Figure 3)
9. Reference Material • material, sufficiently homogeneous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement or in examination of nominal properties • Certified Reference Material: reference material, sufficiently homogeneous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement or in examination of nominal properties; the certification

is provided according to ISO (International Organization of Standardization) Guide 33401:2024

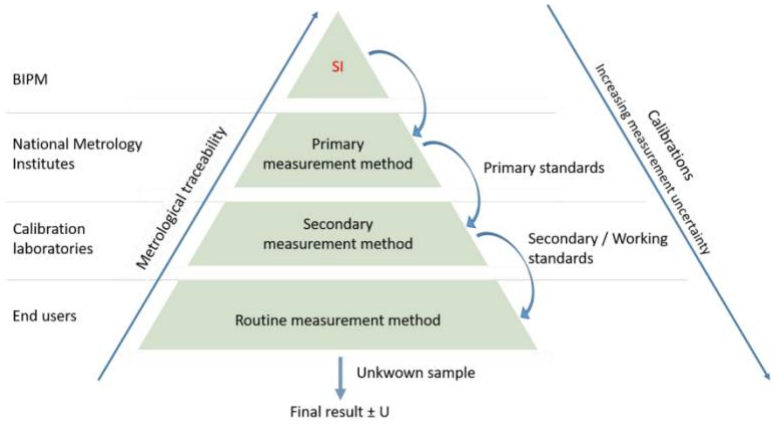

Figure 3. Metrological traceability chain showing the connection of end-users' (e.g., oceanographers) measurement results to internationally agreed reference. BIMP stands for International Bureau of Weights and Measures, SI for the International System of Units, and U for uncertainty.

Figure 4 is a schematic of most of the terms defined above.

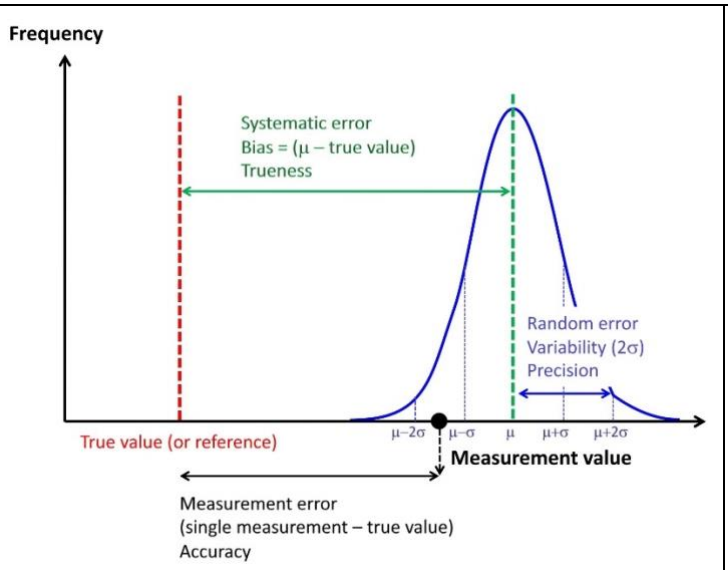

Figure 4. Schematic representation of the different types of measurement errors: random and systematic. The curve shows the frequency of the measurements with μ = mean and σ = standard deviation. The systematic error is the mean minus de true value, and as the frequency distribution is normal, the random error is expressed as ±2σ/μ, a confidence interval of 95%. A confidence interval of 68% would be ±σ/μ. From Squara et al. (2015)

In oceanography, any physical, chemical or biological variable expressed in measurement units would be a measurand. The quality of these measurements is expressed in terms of precision, within the same conditions (repeatability) and/or within different conditions (reproducibility). In oceanography precision is a measure of how reproducible a particular experimental procedure is, it can refer to the final analysis or the entire procedure including sampling, conservation, analysis. Precision is estimated by performing replicated measurements, calculating a mean and a standard deviation of the results. Accuracy is a

measure of the degree of agreement of a measurement value and the true value. Measurements are compared to the true values assigned to commercial reference materials available for some seawater variables (inorganic nutrients, practical salinity and some CO_2 variables). Certified reference materials in natural or artificial seawater are available for some chemical compounds (Ebeling et al., 2022). In house or secondary reference material with an assigned true value are those prepared by oceanographic labs to QA/QC their procedures. Accuracy is expressed as a mean difference with a standard deviation or percentage. Error and uncertainty are understood as in metrology (Figure 4).

2. Secondary quality control (2QC) procedures: crossover analysis

2.1. Introduction

When combining independent datasets, typically cruise data (See Section 1.1) systematic biases might appear. A 2QC procedure aims to detect, quantify and correct those systematic biases (constant differences) which exceed a prescribed uncertainty limit for a particular oceanographic variable within a particular dataset, usually a cruise. The procedure is designed to make the systematic bias clearly identifiable over the random error and natural variability in the data. It is important to consider that the detected systematic bias is assumed to be constant over the whole of each dataset.

The commonly used procedures to detect systematic biases in cruise data are:

(i) multilinear regression (MLR) analysis: given a set of assumptions (no biases in the input data and the same relationship, or processes, affecting the variables), biases would appear as the mean residual of the modelled MLR variable,

(ii) crossover analysis: inspects averaged differences between datasets at intersection areas where the spatial and temporal natural variability is low and may compute the offsets that would minimise these differences using inverse methods.

Those two approaches were described in the CARINA ESSD 2010 special issue by Jutterström et al. (2010) and Tanhua et al. (2010), respectively.

Crossover analysis stems from the “intercruise offsets” method proposed by Gouretski and Jancke (1999) and later refined by Gouretski and Jancke (2001) and Johnson et al. (2001). This method is the reference one used by GLODAP, it is briefly described in the next section as we aim to refine it within this deliverable.

2.2. GLODAP 2QC crossover analysis description

This subsection focuses on describing the GLODAP 2QC crossover analysis, which is the starting point for the 2QC crossover analysis with enhanced uncertainty quantification described in Section 3 and applied to several case studies in Section 4. The original method is detailed in Tanhua et al. (2010) and Lauvset and Tanhua (2015) as the *running cluster* 2QC crossover analysis.

Briefly, the crossover analysis is an objective analysis to detect systematic differences between two cruises conducted in the same area by comparing measurements in the deep part of the water column (typically>1500 m). The deep ocean is typically a low variability environment in time and space. Cruise to cruise pairs of stations to be compared are sought within a given distance, usually ≈200 km (2° arc distance), although the distance threshold could conceivably be varied depending on region. This selection of stations is called cruise-pair or crossover (Figure 5 top plot). Within a cruise-pair, each station profile from both cruises is vertically interpolated to the same set of intervals in pressure or density (sigma4) space. Cruise A and cruise B interpolated profiles (Figure 5 left) are differenced (or divided depending on the variable of interest, see below) to finally obtain the mean offset profile and the offset standard deviation profile (Figure 5 right, black and grey lines). The interpolated values in the mean offset profile are averaged with weighting by the offset standard deviation profile, to calculate the weighted mean offset (μ_{AB}) and the weighted mean offset standard deviation (σ_{AB}) (Figure 5 right, red lines).

Mean offset profiles are calculated as differences (additive or absolute biases) for variables such as salinity, pH, dissolved inorganic carbon and total alkalinity, and as ratios (multiplicative or relative biases) for dissolved oxygen and inorganic nutrients. The latter is used for ocean variables where concentrations may be very close to zero and variables where problems with standardization are the most likely source of error.

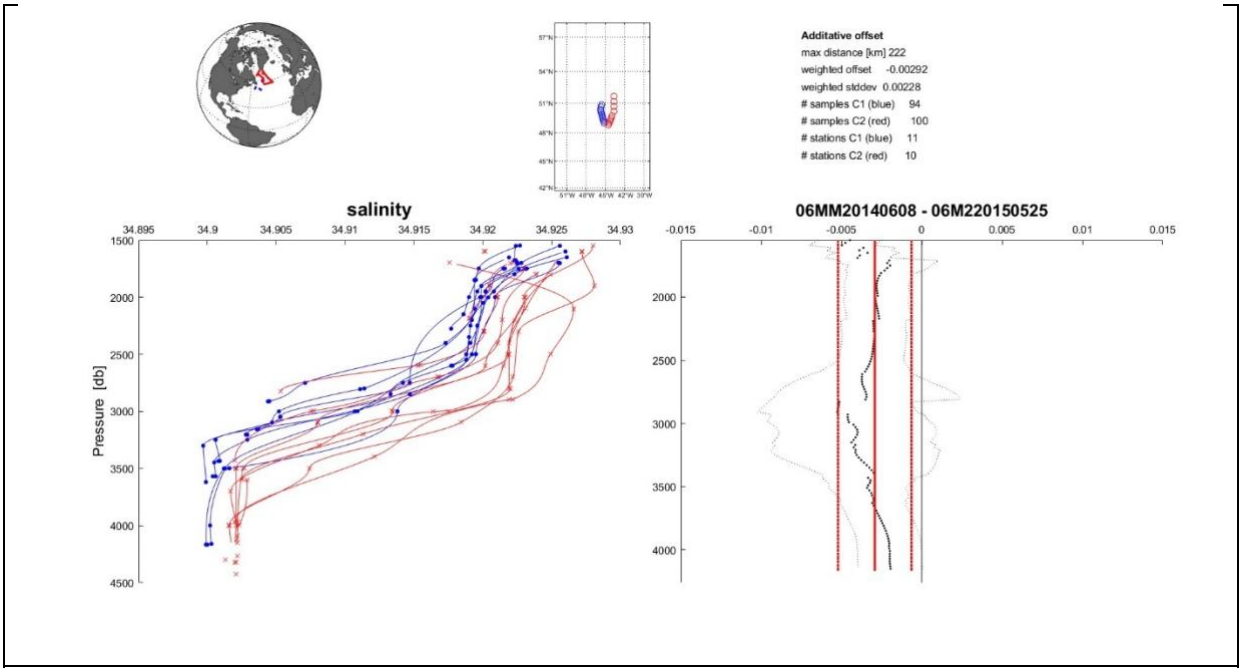

Figure 5. Example of a cruise-pair crossover result. The left plot shows the profiles for cruise A in blue and cruise B in red. The right plot shows the mean and standard deviation difference profile (black points) and in red the weighted mean offset and standard deviation.

Figure 5 shows the crossover result for salinity in one cruise-pair between cruise A (blue) and cruise B (red). Salinity profiles are interpolated (left plot) and difference (Δ) profiles are calculated for regular intervals (typically 5 dbar) between stations A and B. For each vertical

interval the mean and standard deviation of all the Δs is calculated. The result is a mean difference profile with a standard deviation (black lines/dots in Figure 5 right). The weighted mean offset and standard deviation over all vertical intervals is then calculated (red lines in Figure 5 right):

$$\mu_{AB} = \frac{\Sigma\left(\frac{\mu}{\sigma^2}\right)}{\Sigma\left(\frac{1}{\sigma^2}\right)}, \sigma_{AB} = \frac{\Sigma\left(\frac{1}{\sigma}\right)}{\Sigma\left(\frac{1}{\sigma^2}\right)} \qquad \text{Equation (1)}$$

where μ_{AB} = arithmetic mean of Δ at each vertical surface, and σ_{AB} = standard deviation of Δ at each vertical surface.

If a systematic bias between two cruises, such as detected in Figure 5 for salinity, is identified, further detail about the QA/QC for that variable in both data sets needs to be collected and evaluated to identify, if possible, the cause of the bias.

When assembling cruises covering an ocean basin or wide ocean area there would be multiple cruise-pairs or crossover results, i.e., cruise A can have several cruise-pairs with cruise B, but also with one or more other cruises. In the GLODAP 2QC procedure, the final correction or adjustment applied to each cruise is assessed using an inversion scheme where all biases between all data sets in an ocean region are calculated and then compared with each other using a least squares model (Equation 2) following the methodology described in Johnson et al. (2001). The method solves an inversion considering all cruise-pairs and looks for the solution that minimizes the bias between all cruise-pairs or crossovers, finally suggesting individual adjustments for each data set or cruise, which, when applied, produce a more internally consistent data product.

The least squares or inversion method described in Johnson et al. (2001) minimizes and solves this equation:

$$m = (G^T \times W \times G)^{-1} \times G^T \times W \times d \qquad \text{Equation (2)}$$

where G is the model matrix of size *o* x *n*, where *o* is number of crossovers and *n* number of cruises, *d* is the length-*o* crossover offsets matrix and *m* is the length-*n* corrections matrix needed. W is a weighting matrix, populated with information about 1) the precision of the original data that determines the standard deviation of the crossover offset (see Figure 5); 2) the distance (time and space) between the original profiles (i.e., a crossover weighs more heavily if the repeat of a station was performed within a short time frame) and 3) the stability of the region inspected and the degree of ocean variability in the region (i.e., crossovers in an ocean region with small variability far away from ocean fronts could weigh more heavily); 4) the expected limit of adjustment for each cruise or data set, i.e., an a priori idea about the accuracy of the data set. This information is introduced in the weighting scheme: no W would be a simple least squares, including information (1) and (2) would be a weighted least squares and including information (4) would be a weighted damped least square.

GLODAP uses the global mean *d* (basically the remaining offsets) after the *m* (corrections or adjustments) have been applied as a measure of the inter-cruise or data product consistency or reproducibility. Over decades the corrections applied to GLODAP cruises decrease in

magnitude, and the number of cruises corrected also decreases (see Figure 8 in Lauvset et al., 2022). This finding is related to the improvement of (i) precision: over time, precision tends to be improved due to improved methodology (for example, automated systems), more advanced technology and access to technical support and (ii) accuracy: this improves thanks to widely accepted and traceable to the Standard International System best practices or analytical procedures, along with the availability of (certified) reference materials for different conditions and properties encountered in the open and coastal oceans (Figure 3). A more consistent data product means that all variables are comparable and coherent. Therefore, applying a proper Metrology concept, in theory, the data product would present a much higher reproducibility (Section 1.3.2), than the individual data sets. Measurements from different labs, slightly different methods, methodological differences and research vessels contained in the data product, would be comparable, as any systematic bias has been corrected. Several activities within the EuroGO-SHIP work packages and their final deliverable reports will contribute to addressing these issues, but the requirement to evaluate the uncertainty in combined data products through 2QC will remain. Using the GLODAP 2QC procedure as baseline, within this deliverable we aim to define a more thorough assessment of the uncertainty in data products, considering different sources of errors both systematic and random, using metrological terms and approaches adapted to current practices in oceanography.

2.3. 2QC crossover analysis flowchart and software tool

The detailed flowchart of the 2QC crossover analysis used in GLODAP is presented in Figure 3 in Tanhua et al. (2010), while the software tool is described in Lauvset and Tanhua (2015). The tool can be accessed via this link: https://github.com/sivlauvset/2nd_QC_tool. In deliverable 2.4 we simplify the 2QC flowchart in Tanhua et al. (2010) as also accessible here https://gitmind.com/app/docs/f3v84ukk) to highlight specific modifications used in Section 3, particularly for the case studies in Section 4.

In brief, the 2QC tool seeks the intersections within a given distance (typically ≈200 km, 2° arc distance) between a reference cruise data set and the cruise data to be checked. The reference cruise data comprises a collection of cruises in which all variables, especially biogeochemical ones, are measured and reported according to the highest quality standards. These cruises include comprehensive QA/QC information in the cruise report and/ or metadata, ensuring the highest levels of precision and accuracy, which assures the reproducibility and traceability of the measured variables (Section 1.3). Steps 1 to 4 noted in the flowchart were explained in Section 2.2 and steps 5 and 6 will be detailed in Section 3.2.2. Within this deliverable we adapted the 2QC toolbox developed by Lauvset and Tanhua (2015) in ©Matlab. No inversion (Equation 2) was applied in any of the case studies (Section 4).

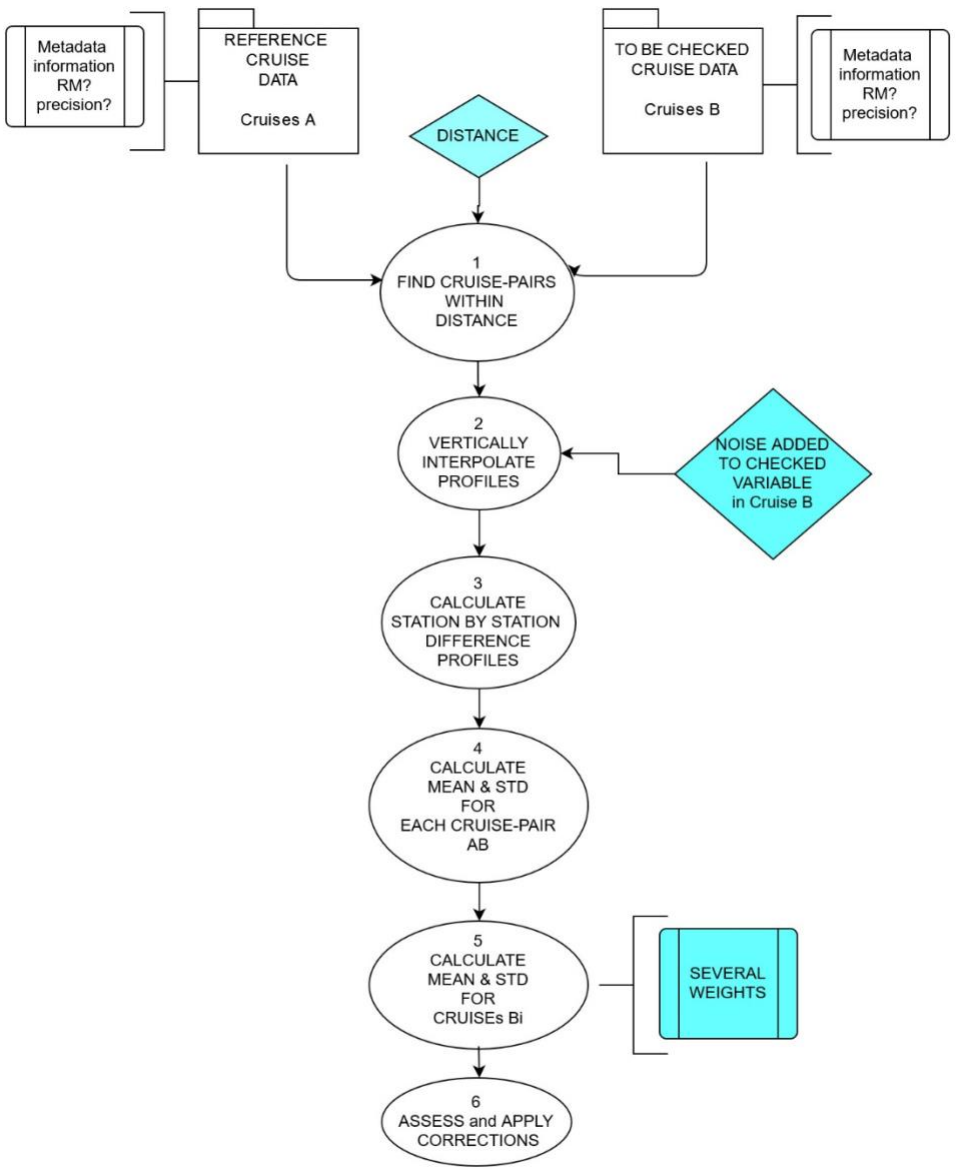

Figure 6. Simplified flowchart for the 2QC crossover analysis used in Deliverable 2.4.

3. Framework to assess the uncertainty in a 2QC crossover analysis

This deliverable aims to establish a framework to quantify the uncertainty in a 2QC crossover analysis when combining datasets, mainly ship-based cruise data, with different origins (Section 1.2). This coherence would be based on the quality of the combined original data

sets, thus related to their QA/QC procedures: physical (mostly sensor-based) and biogeochemical (mainly discrete analysis) variable contained in the cruise data set should be traceable to SI, should be precise (good repeatability) and accurate (checked against, certified or in-house, reference materials) along the cruise. Therefore, individual data sets would be metrologically reproducible (Section 1.3) in time and space

3.1. Applicability of metrological methods to assess uncertainty in a 2QC crossover analysis

Following a metrological rationale, if our aim is to assess the uncertainty of the 2QC analysis and thus of the final data product, our measurand, quantity or property to be assessed would be the reproducibility of the different variables between the different cruise data sets. According to the definition Section 1.3.2, the uncertainty in the reproducibility would be understood as a dispersion or confidence interval and expressed as a standard deviation.

The Guide for expression of uncertainty in measurement (GUM, JCGM 100:2008, 2008) GUM, JCGM 100:2008, 2008) provides detailed information about the quantification of the different sources of uncertainty involved in a measurement process and how they are combined to determine the final uncertainty budget and combined standard uncertainty of the evaluated property. The focus of the GUM are laboratory measurements. Thus, it is not directly applicable to our case. However, it is worth considering the GUM as very few oceanographic measurement procedures have been properly evaluated following the GUM (Feistel et al., 2016; Waldmann et al., 2022; Seitz et al., 2011), especially biogeochemical variables (Hartman et al., 2023); and some procedures can be adapted for the objective of our deliverable.

Generally, two methods can be applied to quantify uncertainty:

(i) the bottom-up approach relies on a detail identification and quantification of every source of uncertainty involved in quantifying the measurand

(ii) the top-down approach, which relies on experimental data to compute the uncertainty budget

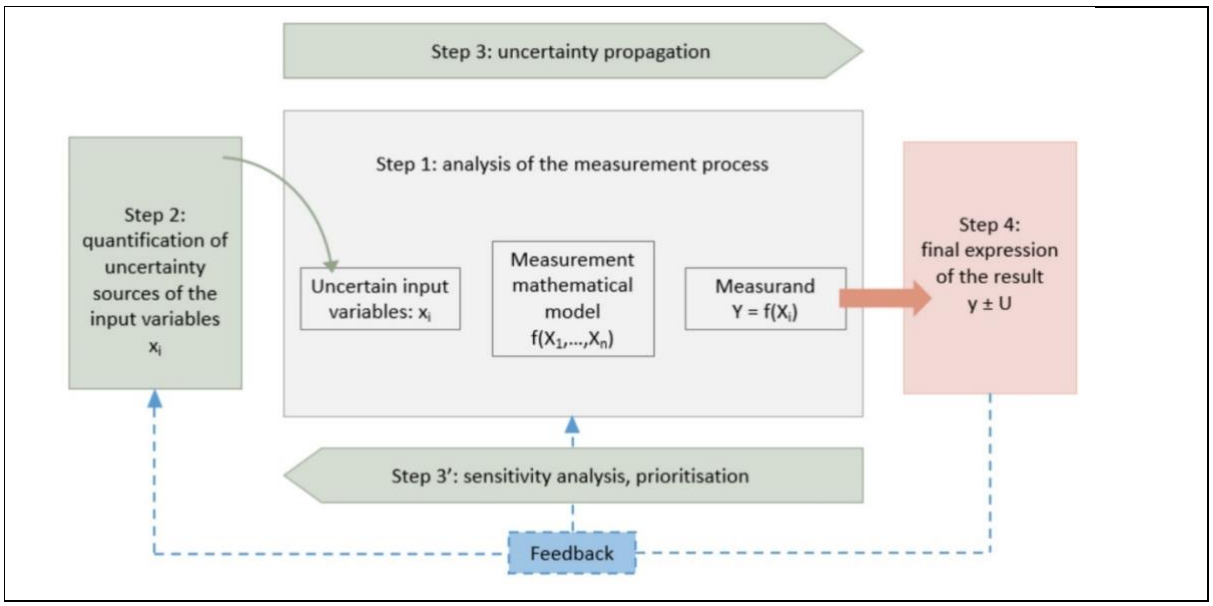

Figure 7. Schematic representation of the bottom-up process to assess the measurement uncertainty following the GUM.

3.1.1. Bottom-up procedure: pen analysis

This approach requires several steps as those depicted in Figure 7 taken from Allard and Fischer (2015). The first step is an evaluation of the measurement process, define the measurand (Y), the quantities involved in the measurement process and the influence quantities (Xi) and the mathematical model linking them with Y (Y=f(Xi)). Identifying the input variables that affect the measurand and the measurement process, and designing an Ishikawa diagram or pen analysis is important in Step 1. The pen analysis includes devices, materials, method steps, environment and user, all affecting Y. An elaborated pen analysis for the determination of seawater spectrophotometric pH is presented in Figure 8 as example.

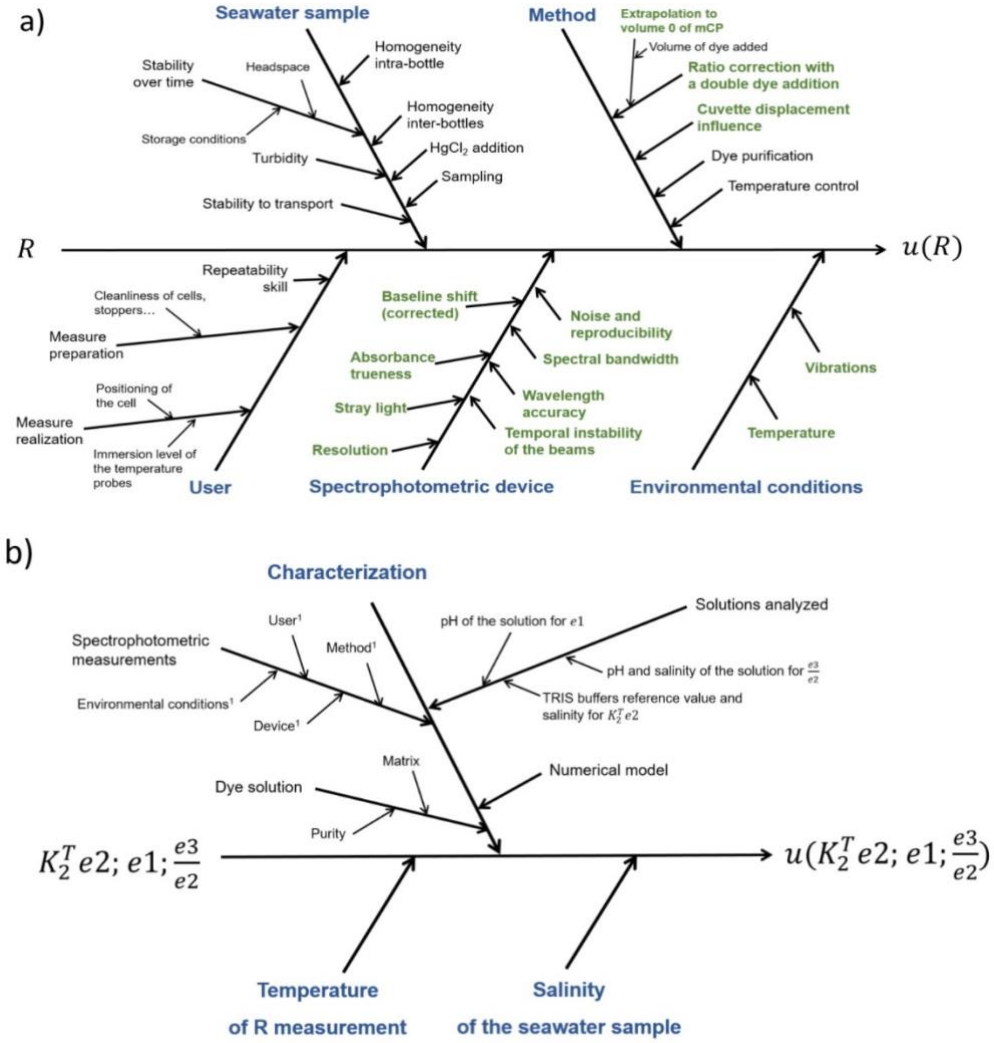

Figure 8. Ishikawa or pen diagram with the uncertainty sources for seawater discrete spectrophotometric pH measurements where a) shows the uncertainty sources affecting the absorbance Ratio, and b) those affecting the indicator dye characterization. Figures elaborated by Gaëlle Capitaine (PhD, 2024).

The second step in Figure 7 is the evaluation of each uncertainty source affecting X, accounted as standard uncertainties. Two categories of standard uncertainty can be distinguished:

(i) Type A uncertainty is obtained from experimental repeatability and
(ii) (ii) Type B uncertainty requires experience, expertise or auxiliary information to define the distribution form of the uncertainty and the range of variation, for example, the resolution of a device given by the manufacturer specification.

Step 3 propagates the uncertainty associated to the input variables into the mathematical model.

Step 4 corresponds to the uncertainty budget calculating the expanded uncertainty (U) of the measurand (Y). The expanded uncertainty is calculated multiplying the standard uncertainty by the coverage factor, which is usually 2, corresponding to a level of confidence of 95%.

The feedback step in Figure 7 indicates an optimization process to reduce U, improving one or several input variables to adjust the uncertainty budget to the expected or required uncertainty.

This bottom-up approach is typically quite cumbersome in oceanography (Bushnell et al., 2019; Le Menn et al., 2023; Waldmann et al., 2022), especially for seawater biogeochemical variables where not only the measurement procedure should be considered but also the sampling and preservation method. A good example is the complex evaluation of the measurement uncertainties in the seawater CO_2 system variables, as highlighted in Carter et al. (2023, 2024).

An alternative method to assess uncertainty in a bottom-up process is using a Monte-Carlo approach, i.e. propagating the uncertainty distribution of all input quantities in the mathematical model to obtain directly the distribution of the measurand value (Allard and Fischer, 2015; “JCGM 101:2008,” 2008).

3.1.2. Top-down procedure: inter-laboratory comparison exercises

The top-down approach to assess the uncertainty of a measurand relies on experimental data from measurement results, based on inter-laboratory comparison exercises to check a standardized method within laboratories, or for just one laboratory, using its quality control data over a long period of time. In both cases, the laboratory would use reference materials (Section 1.3.2) to detect any bias from the considered true value for the measurements. The ISO standard 21748 *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation* (2017) can be used to establish an uncertainty budget with the results on an inter-laboratory exercise, only if it follows the rules given in the ISO standard 5725-2 (2020).

3.2. EuroGO-SHIP deliverable 2.4 approach

Although none of the previously commented procedures can be easily and thoroughly applied to quantify the reproducibility of ship-based variables over time and space, they help set confidence on the proposed approach to assess the uncertainty in a 2QC crossover analysis for cruise data.

(i) The bottom-up procedure requires a deep knowledge about the measurement practices and the QA/QC for every variable and every cruise, and even so the uncertainty budget or propagation would be cumbersome, even unfeasible for initiatives as GLODAP where more than 1000 independent cruise data sets, globally distributed, are combined.

(ii) The top-down procedure, the inter-laboratory exercises, are usually adopted when trying to improve and refine a measurement technique once the method is clearly established. Proficiency test exercises, as those yearly organized by WEPAL QUASIMEME for several chemical seawater properties (now including seawater CO_2 variables and dissolved inorganic nutrients) are properly metrologically set examples. Nice examples for inter-comparison exercises are recently provided within the EU projects MINKE and SapHTies. These exercises might be a more expensive alternative, and probably not enough and feasible, alternative for 2QC analysis.

The best option would be a Monte Carlo approach, where we introduce a distribution of random uncertainty, noise, in the cruise data and assess the effect on the results for the 2QC crossover analysis, and the final adjustment or correction factors obtained (

Figure **6**). As commented in Section 2 the cruise, or set of cruises, to be quality checked are compared to a reference data set that is supposed to comply with the highest standards for hydrographic and chemical ship-based measurements (

Figure **6**), following GO-SHIP best practices (Hood et al., 2010). These reference cruises could be understood as true values (sort of reference materials in terms of metrology), so containing precise and accurate measurements for physical and chemical variables. The term noise is not usually used in the GLODAP 2QC analysis, instead, uncertainty is understood as a combination of random and systematic errors, directly related with precision and accuracy, respectively (Section 1.3.2). For a given oceanographic profile, precision might be quantified by measuring the same sample multiple times, but vertical individual samples are measured just once. Accuracy might be quantified for an instrument/measurement procedure by measuring certified reference materials. Combining these two numbers (precision and accuracy) allows calculating an uncertainty: a cloud or range of possible values around the measured/reported one, typically, this cloud is normally distributed (Figure 4).

3.2.1. Monte Carlo approach: introducing random uncertainty (noise) in the 2QC crossover analysis

In the following, we demonstrate the effect of introducing different random uncertainty (noise or random error) on the crossover offsets. This is applied to physical (salinity) and

chemical (dissolved oxygen) cruise data profiles in the Northeast Atlantic Ocean. We aim to assess the effect on both the weighted offset and the standard deviation (Equation 1) of the probability distribution for an introduced uncertainty (typical noise value) for each variable.

Following are some trials showing the effect on the crossover offsets of introducing different probability distribution functions of random uncertainty (noise or random error) in the to be checked physical (case for salinity) and chemical (case for dissolved oxygen) cruise data profiles in the Northeast Atlantic Ocean. We aim to assess which is the effect on the weighted offset and standard deviation (Equation 1) of the probability distribution for an introduced typical value of uncertainty (noise) for each variable.

Our working area would be the Northeast Atlantic Ocean covering the Iberian Abyssal Plain where depths higher than 5500 meters are reached (Figure 9). Deep and bottom waters, deeper than 3000 dbars, correspond to North Atlantic Deep Water (NADW), upper NADW is established from sigma2>37 to sigma4<45.84 and lower NADW presents sigma4>45.84 (Lherminier et al., 2007). Both layers, below 3000 dbars, are considered having a low temporal variability environment, and already used as an environmental standard for cruises in the 1980s (Saunders, 1986; Mantyla, 1994).

The reference data set are the GLODAPv2.2022 cruises in the Northeast Atlantic, while the to be checked data would be the IEO RADPROF (Finisterre Deep Section, Tel et al., 2016) sampled in 2019, expocode (i.e., unique identifier for a cruise containing the research vessel ICE code and the year month and day of cruise first day) 29RM20190818. The RADPROF hydrographic monitoring program has been running since 2003. Since 2014, carbonate chemistry variables and dissolved oxygen are measured following the GO-SHIP manual. We selected the 2019 cruise as it reaches bottom depths and the western most longitude station (-15.40 ºW, station 134). Two variables are evaluated, salinity from CTD (CTDSAL) that was calibrated with salinometer measurements and dissolved oxygen (OXYGEN) data measured following a potentiometric Winkler method (Langdon et al., 2010). The a priori uncertainty is 0.005 for CTDSAL and 1% for OXYGEN for both the reference and the RADPROF 2019 cruise. CTDSAL offsets will be evaluated as differences (additive relationships). OXYGEN offsets will be evaluated as ratios (multiplicative relationships). Crossovers stations will be evaluated within 2° arc distance (approximately 222 km) and considering data below 4000 dbars. For simplicity, pressure will be used as the reference variable to define the profiles. Offsets for the original 29RM20190818 data will be evaluated.

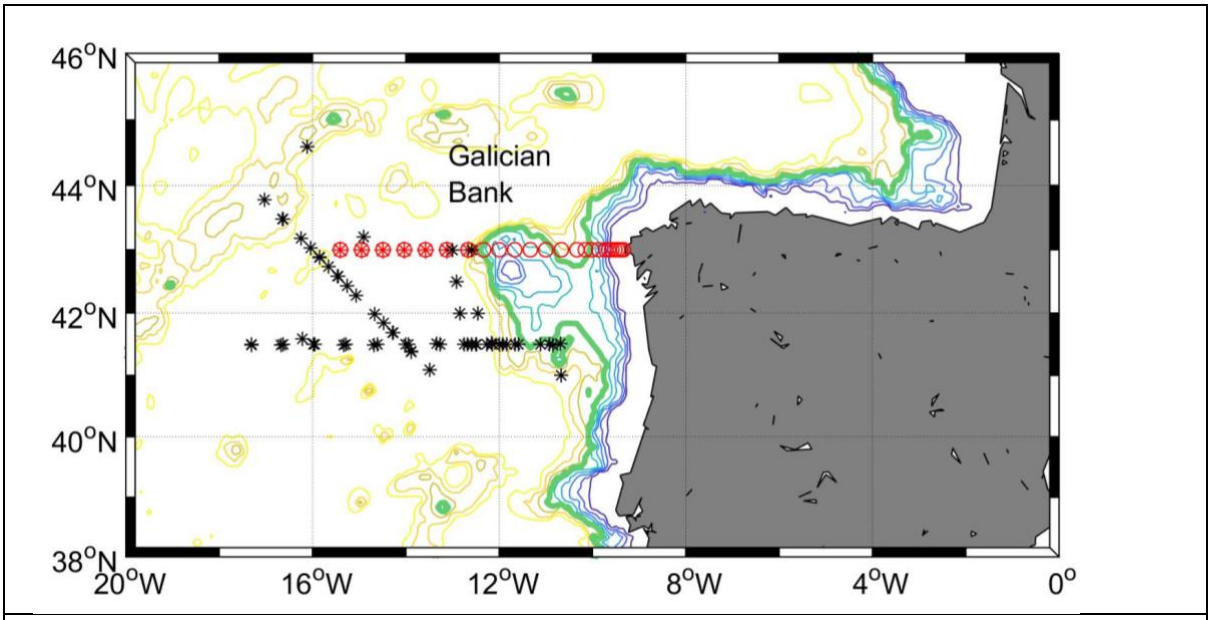

Figure 9. Map of the Northeast Atlantic Ocean showing the RADPROF 2019 cruise stations (in red) and the GLODAPv2.2022 stations from different reference cruises (in black). RADPROF stations west of the Galician Back are highlighted as they are deeper than 3000 dbars. The isobath of 3000 dbars in green is highlighted.

Random noise

Each data point from the RADPROF 2019 cruise is modified randomly by applying different probability distributions or cases. The forced random error introduced would have a magnitude called typical uncertainty or noise varying from 1 to 3 times the usual measurement uncertainty. We will explore introducing a typical noise for CTDSAL of 0.005, 0.010, 0.015 and 0.02. And in the case of OXYGEN, it would be 1% (usual uncertainty), 1.5%, 2% and 3%. The probability distribution of this random noise would be:

- Case 1. No noise, the original data is used.
- Case 2. Random uniform distribution for the introduced typical uncertainty, for example, noise from -0.01 to 0.01 for CTDSAL, or -1% to 1% for OXYGEN.
- Case 3. Random normal distribution within the introduced typical uncertainty, for example, from -0.01 to 0.01 for CTDSAL, or -1% to 1% for OXYGEN.
- Case 4. Random normal gaussian distribution with a zero mean and a standard deviation equal to the typical noise value introduced.

A weighted mean and standard deviation considering all crossover results are calculated for RADPROF 2019 and at each case option and typical noise introduced.

CTDSAL evaluation

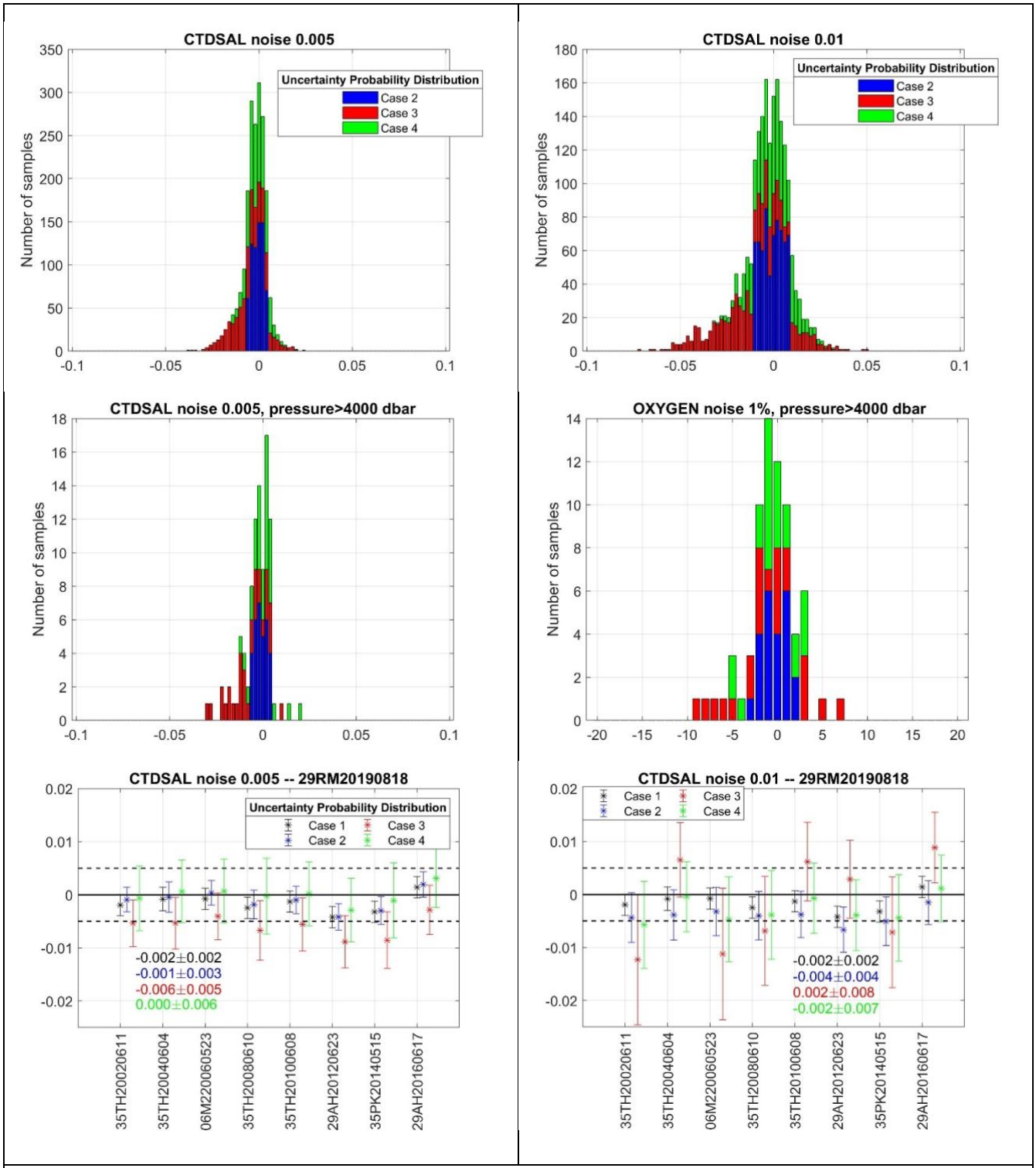

Figure 10. Upper and middle plots: histograms with the differences between the original and modified data, according to the probability distribution (Cases) of typical noise introduced in the RADPROF 2019 CTDSAL data (0.005 left plots and 0.01 in the right plots). The top panel shows the entire data set, the middle panel only data deeper than 4000 dbars which is used in the crossover analysis. The bottom plot shows the weighted mean and standard deviation of the offsets between RADPROF 2019 and the reference cruises. The bottom plots also show the overall weighted mean and standard deviation. Left: the typical noise introduced is 0.005. Right: the noise introduced is 0.01.

Figure 10 shows the impact on the original data after introducing different probability distributions of random uncertainty for several typical values of noise, the usual one 0.005 (Figure 10, left plots), and a higher one 0.01 (Figure 10, right plots) for CTDSAL. A higher value of typical noise would mean that the data set to be checked has a lower quality. Except for Case 3 where the noise is introduced randomly with a normal distribution between -0.005 and 0.005 (equally for -0.01 and 0.01), i.e., the red points in the bottom plots of Figure 10 indicate that the mean values for each crossover and the overall weighted mean remain stable. However, the standard deviation changes and increases with a higher noise level, particularly in Case 3 and 4.

Figure 11 shows the overall weighted mean and standard deviation for the RADPROF 2019 cruise considering different probability distributions (Cases 1 to 4) introducing different typical noise values in the original data, from the expected usual one (0.005) to a very high value (0.02) indicating a very bad precision in the CTDSAL data. Introducing a uniform random noise (Case 2 blue points) does not affect the overall mean, so the data is still accurate, as the mean offset keeps within the limits of correction (±0.005), but the standard deviation increases with the noise introduced, as the precision of the data, would be very poor. The random normal distribution (Case 3, red points) modifies the data, causing the mean offset to change sign significantly and introducing an artificial bias to the data. Additionally, it introduces a high random uncertainty as the standard deviation clearly increases. If noise is introduced with no bias (zero mean value) and with a gaussian distribution, i.e., the standard deviation equals the typical noise, it means that 68% of the modified data are within the ±typical noise range introduced (Case 4, green points). The final weighted mean of the offsets can change sign even surpassing the limit for the adjustment (±0.005), while the weighted standard deviation increases with the increasing typical noise introduced.

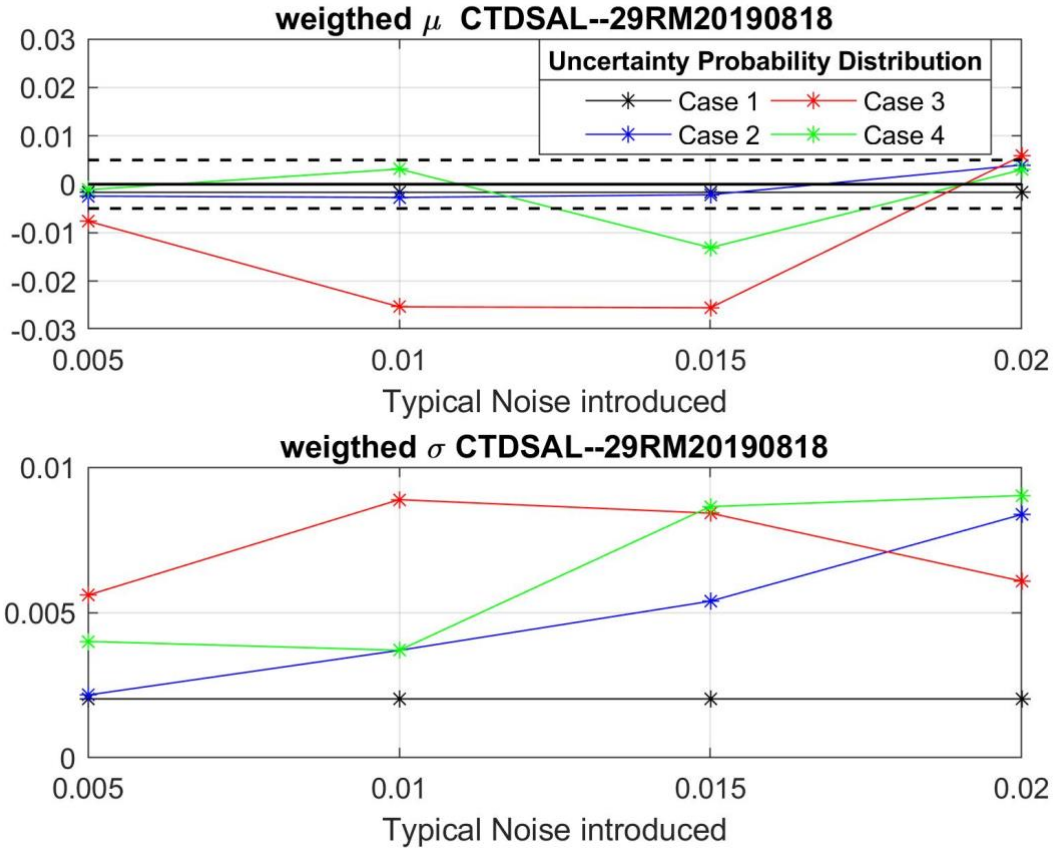

Figure 11. Weighted mean and standard deviation for the crossover results comparing the RADPROF 2019 CTDSAL with GLODAPv2.2022 cruises introducing different probability distributions of several typical noise values showed in the x-axis. Results with the original data are shown as reference values (black points and lines).

OXYGEN evaluation

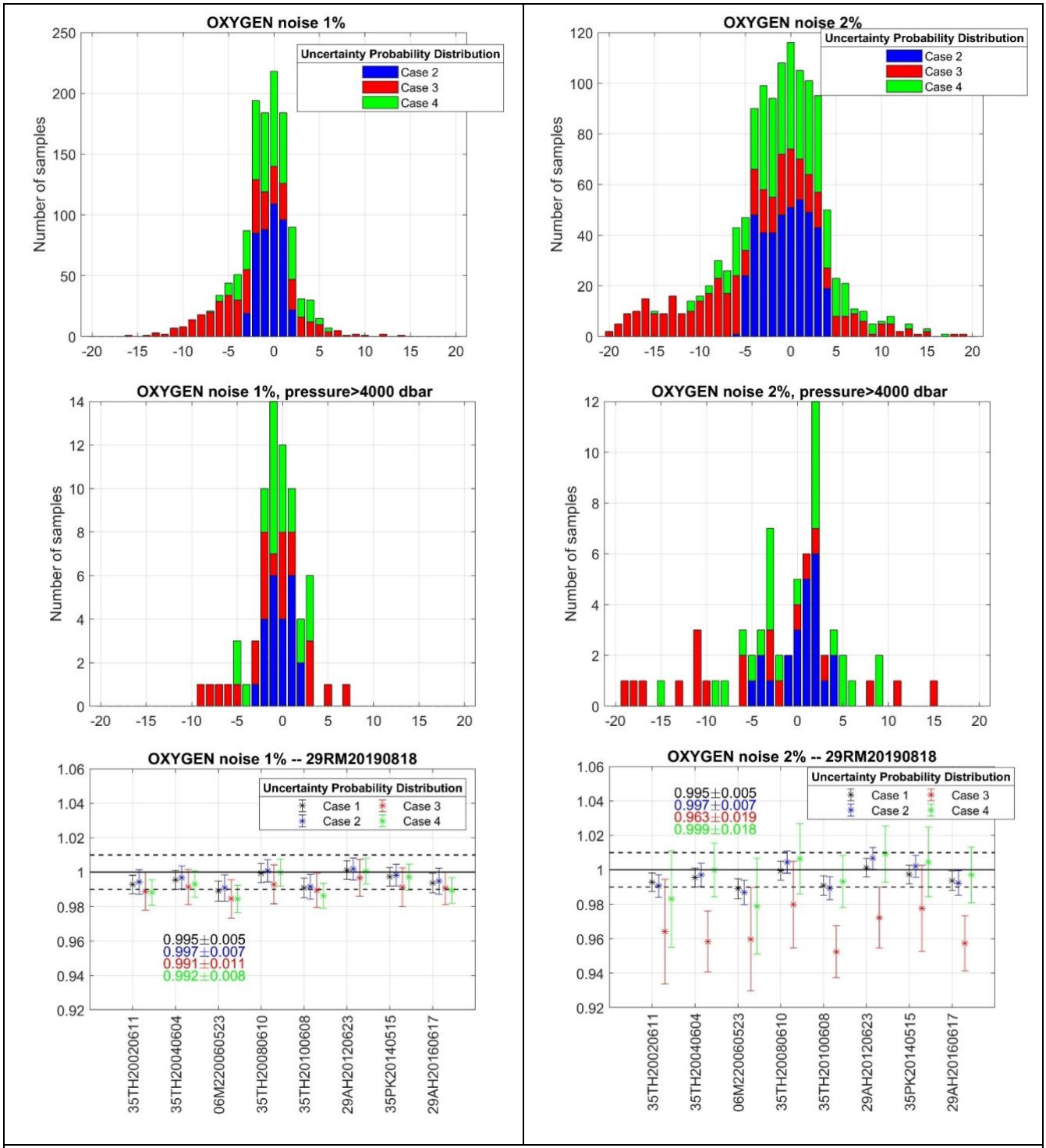

Figure 12. Upper and middle plots: histograms showing the differences between the initially measured and modified data, according to the probability distribution (Cases) of typical noise introduced in the RADPROF 2019 OXYGEN data (1% left plots and 2% in the right plots). The top panel shows the whole data set, and the middle panel only the data deeper than 4000 dbars which is used in the crossover analysis. The bottom plot shows the weighted mean and standard deviation of the offsets between RADPROF 2019 and the reference GLODAPv2 cruises. The bottom plots also show the overall weighted mean and standard deviation. Left: the typical noise introduced is 1%. Right: the noise introduced is 2%.

Figure 12 shows the impact on the original data of introducing different probability distributions of random uncertainty for several typical values of noise, 1% (Figure 12, left plots) and 2% (Figure 12, right plots) for OXYGEN. A higher value of typical noise would mean that the data set to be checked has a lower quality. Except for Case 3 where the noise is introduced randomly with a normal distribution between -1% and 1% (equally for -2% and 2%), i.e., bottom plots in Figure 12 show the mean values of each crossover (red points) and the overall weighted keep quite stable within the ±1% limit of adjustment, but changes are evident for the 2% noise. Weighted standard deviations remain comparable, except for Case 3, where for both the 1% and 2% typical noise values, the standard deviations are very large.

Figure 13 shows the overall weighted mean and standard deviation for the RADPROF 2019 cruise, considering different probability distributions (Cases 1 to 4) that introduce different typical noise values into the original data, ranging from the expected usual one value 1%, to a very high value of 2%, which indicates a very bad precision in the OXYGEN data. Weighted mean values except for Case 3 and Case 4 with 2% typical noise, remain within the limits of no correction (±1%). The weighted standard deviation values are quite similar across the different typical noise levels introduced, except for Case 3, where they are very high.

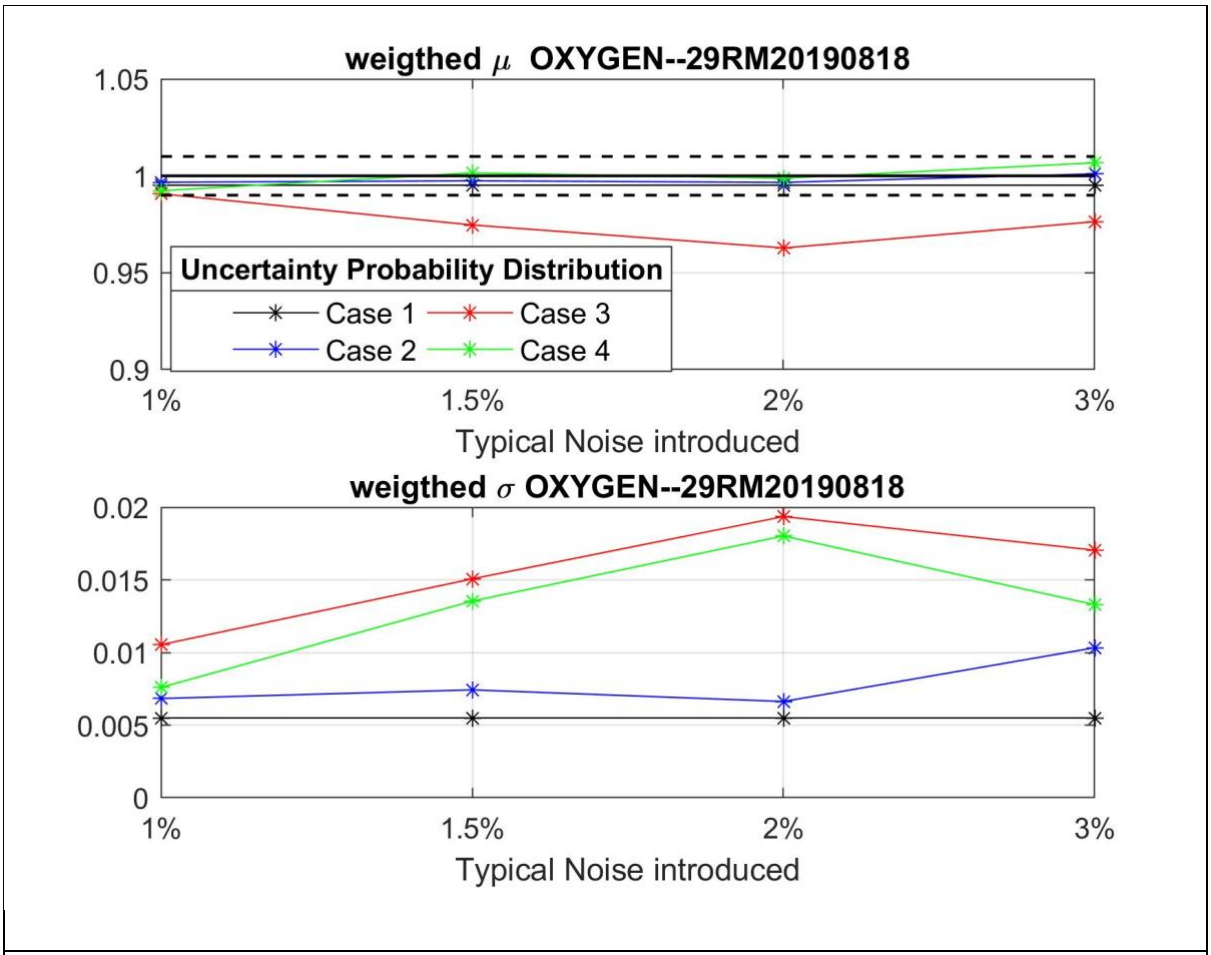

Figure 13. Weighted mean and standard deviation for the crossover results comparing the RADPROF 2019 OXYGEN with GLODAPv2.2022 cruises introducing different probability distributions of several typical noise values showed in the x-axis. Results with the original data are shown as reference values (black points and lines).

3.2.2. Impact of weighting options in the final adjusments

As described in Section 2.2, the final adjustments for a set of cruises evaluated with a 2QC crossover analysis is obtained with Equation (2). This inversion is needed to minimize biases between cruises when any of them could have systematic biases. Different weighting schemes with additional information could have an impact on the final adjustments.

In this exercise, applied to the RADPROF 2019 (cruise B) the data set is checked against reference GLODAPv2 data (cruises A) in the Northeast Atlantic, therefore the inversion cannot be applied. However, the weighting schemes can be evaluated to obtain the overall result for the 2QC analysis and calculate the mean offset and standard deviation ($\mu_B \pm \sigma_B$) for cruise B as showed in Equation (3).

$$\mu_B = \frac{\sum(W_i * \mu_i)}{\sum(W_i)}, \sigma_B = \frac{\sum(W_i * \sigma_i)}{\sum(W_i)} \quad \text{Equation (3)}$$

where the cruise-pair (i) offset mean values (μ_i) and standard deviations (σ_i) are weighted (W_i) according to different considerations:

1) The precision of the original data that determines the standard deviation of the crossover offset, ($1/\sigma^2$), so that, cruise-pair offsets with lower standard deviation weight more.

2) Additional weights as the distance (time and space) between the station profiles increases (i.e., a crossover is weighted more heavily if stations from cruise A are closer in time and space to those from cruise B). Specifically, distance in time will be expressed in years, and distance in space in degrees considering the centroids in latitude and longitude for each set of stations to be compared from cruise A (GLODAPv2 in this case) and B (RADPROF 2019). Both distances will be in absolute values. As the order of magnitude of the standard deviation, time and space distance values are greatly different, each weight is scaled and modified to combine them:

$$W\sigma = 1/\sigma^2;\ WTime = abs[1/(year\_B - year\_A)];\ \ WSpace = abs(1/distance\ B\text{-}A)$$

Each weight is divided by the corresponding standard deviation, then each weight minimum value is subtracted, zeros are substituted by the new minimum value, then each weight is divided by the corresponding mean value. At this step we get homogeneous weight values, but we want σ and time distance to be more important than space distance, and therefore, Wσ is multiplied by 10, and WTime by 5. Therefore, in addition to just $1/\sigma^2$ the following options for W are explored in Equation (3):

$W = WTime \cdot W\sigma$
$W = WSpace \cdot W\sigma$
$W = WTime \cdot WSpace \cdot W\sigma$

Table 1. shows the values for the different weighting options using the original RADPROF 2019 data to calculate $1/\sigma^2$ CTDSAL and $1/\sigma^2$ OXYGEN for each cruise-pair. If using the original weights for σ, time and space, combining them would be useless, as $1/\sigma^2$ are in order of magnitudes higher than time or space differences. The modified weights need to be used to incorporate time and space information in the 2QC crossover analysis final results.

Table 1. Cruise-pair information for the crossover analysis comparing the RADPROF 2019 cruise data with GLODAPv2 cruises. The original values for the weights[1] are several orders of magnitude different and need to be modified and scaled to be comparable and useful.

	Original Values				Scaled and Modified Weights			
EXPOCODE	CTDSAL σ	OXYGEN σ	Time years	Space degrees	CTDSAL Wσ	OXYGEN Wσ	W Time	W Space
35TH20020611	0.0020	0.0054	17.5	1.34	10.47	12.28	0.26	1.00
35TH20040604	0.0022	0.0054	15.5	1.34	9.11	11.71	0.26	1.00
06M220060523	0.0020	0.0058	13.5	1.34	10.47	5.09	0.90	1.00
35TH20080610	0.0020	0.0054	11.5	1.34	9.45	12.35	2.29	1.01
35TH20100608	0.0020	0.0056	9.5	1.34	10.47	6.22	2.94	1.00
29AH20120623	0.0020	0.0054	7.5	1.34	9.11	13.42	6.01	1.00
35PK20140515	0.0020	0.0054	5.5	1.34	10.45	13.85	7.71	0.99
29AH20160617	0.0020	0.0056	3.5	1.48	10.47	5.09	19.62	0.99

[1]As example the unscaled weights for the 35TH20020611 cruise would be 250000 (CTDSAL $1/\sigma^2$), 34168 (OXYGEN $1/\sigma^2$), 0.059 (1/Time) and 0.747 (1/Space), where σ is the standard deviation of the crossover result considering the original data.

.

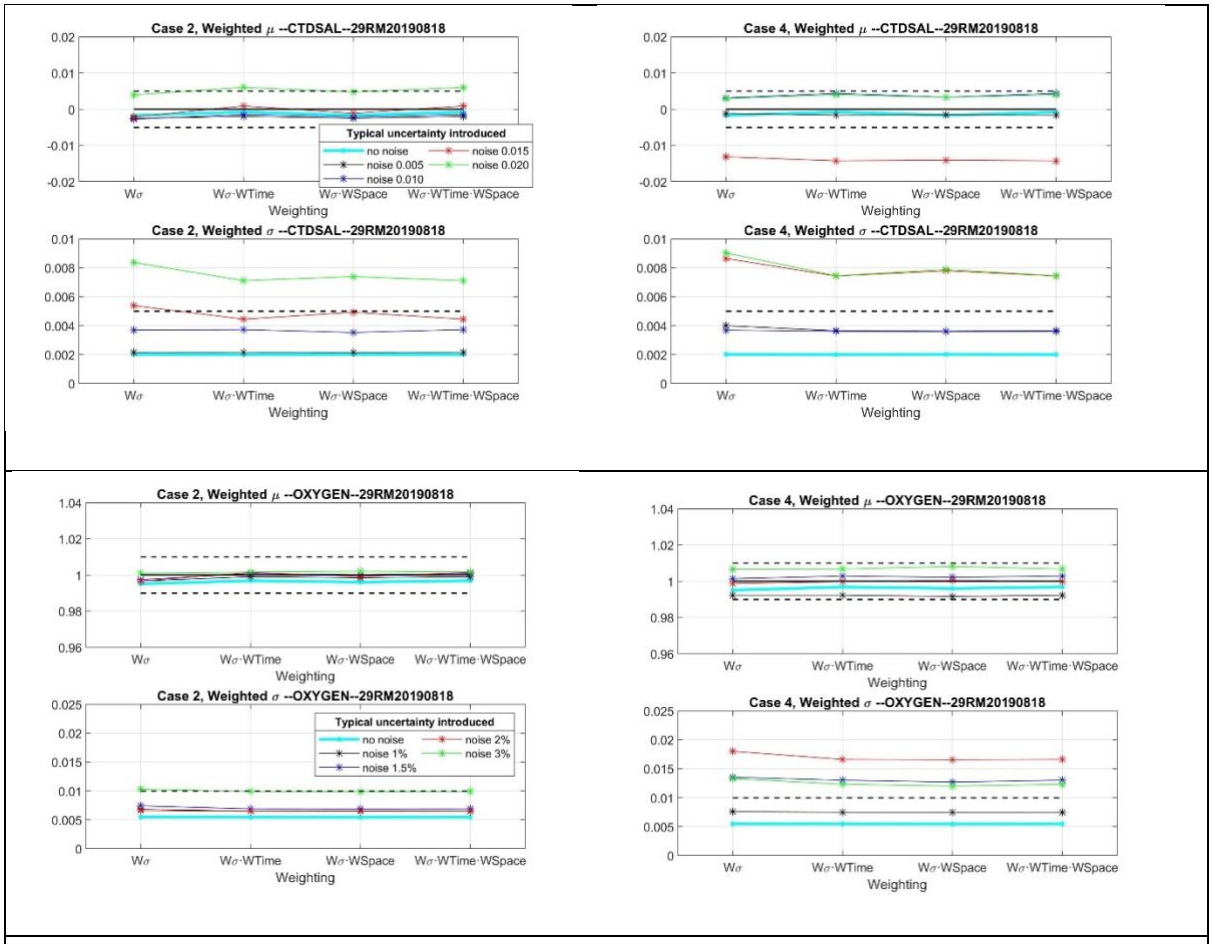

Figure 14. Impact of different weighting schemes to calculate the mean and standard deviation for the 2QC crossover analysis checking the RADPROF 2019 CTDSAL and OXYGEN data. The left column shows the results for Case 2 (random uniform noise), and the right column for Case 4 (gaussian noise) introducing different magnitudes of typical noise in the original CTDSAL (upper row) or OXYGEN (lower row) RADPROF 2019 data.

Figure 14 shows that the impact of the different weighting schemes on the final $\mu_B \pm \sigma_B$ that is small. We forced the weight containing the standard deviation of each cruise-pair (Wσ) to be the most important component (black points in Figure 15), only comparable to WTime if the reference cruise is less than three years apart as for cruise 29AH20160617 (see WTime, blue points in Figure 15).

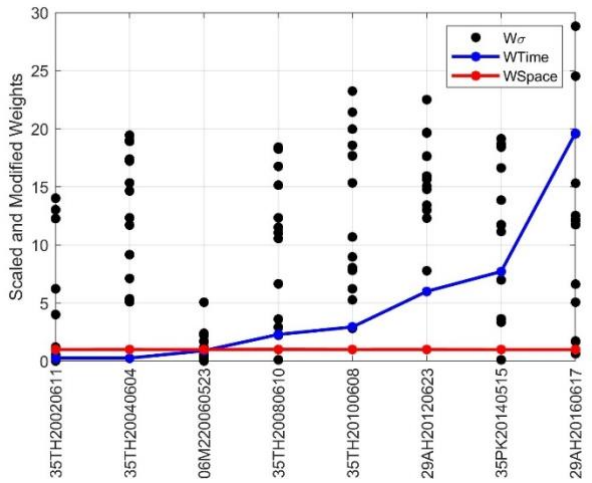

Figure 15. All possible values for the modified weights used in the OXYGEN 2QC crossover analysis between RADPROF 2019 and GLODAPv2 cruises (x-axis): considering the standard deviation of the cruise-pair crossover comparison (Wσ), the distance in time (WTime) and space (WSpace). Wσ varies according to the noise introduced with different probability distributions (Cases) and typical noise (1%, 1.5%, 2% and 3%).

3.2.3. Selected approach to assess uncertainty in the 2QC crossover analysis

We will apply the Monte Carlo approach assuming that:

- Measurements in the checked cruises are done with the same QA/QC along each batch of analysis, therefore, keeping the same precision and accuracy for the whole cruise data set.
- GLODAPv2.2022 cruises are used as reference true values, assuming they comply with the highest accuracy and precision in oceanographic measurement procedures.

In every case study for the different oceanographic areas (Section 4), we will compare the results for the 2QC crossover analysis for the original data (Case 1: no noise) with the Monte Carlo approach for Case 4 (with noise). This case introduces random noise in the original data with a probability distribution following a gaussian curve (mean zero and a prescribed standard deviation value), where the standard deviation will equal a priori selected random uncertainty typical values:

- For salinity, we will introduce a typical uncertainty value of 0.005 and 0.01.
- For oxygen, we will introduce a typical uncertainty value of 1% and 2%.

Two types of weighting will be explored, Wσ and WTime·WSpace·Wσ (Section 3.2.2)

The cruise data would be corrected if final $\mu \pm \sigma$ for each cruise (Equation 3) surpasses a prescribed limit. After correction, the improvement in data quality can be assessed by checking the overall consistency or repeatability (in metrological vocabulary) of the checked cruises. A proper approach would involve rerunning again the 2QC analysis and checking the reduction of each cruise $\mu \pm \sigma$ for the corrected data with respect to the reference data.

However, due to time and computing restrictions, an alternative method is to quantify the changes in standard deviation and skewness value of the anomaly from the overall mean value of the checked variable at the corresponding depth layer. Systematic biases in the original data would appear as high values of skewness and/or standard deviation. Once

corrected, the random uncertainty in the data would remain, both or any of the two quantities would be reduced. Assuming a gaussian distribution of the random uncertainty in the data (Figure 4), approximately 68% of the cruise data would have an uncertainty corresponding to the new standard deviation.

4. Case studies

4.1. Rationale for each case study

This section presents specific case studies for two different oceanographic areas to test the proposed framework for 2QC analysis uncertainty assessment described in Section 3.2.3.

- The first case in the Northeast Atlantic assesses a collection of repeat annual RADPROF cruises expanding 10 years from 2014 to 2023. This area is characterized by a low temporal variability in the deep and bottom water masses, where quite homogeneous profiles are found for thermohaline and biogeochemical properties. Here we test the viability of our approach to salinity and oxygen discrete data from a data set of cruises run by the same research institute, which aim to comply with the GO-SHIP cruises quality standards.
- The second case assesses a collection of 10 cruises in the Western Mediterranean Sea (WMED). This region is characterized by an interannual thermohaline variability in the deep and bottom waters associated with deep-water formation processes in the Gulf of Lion. Here, we test the viability of our approach applied to sensor-based oxygen data from a set of cruises led by different laboratories.

4.2.Northeast Atlantic RADPROF cruises

A collection of 10 repeat RADPROF cruises conducted from 2014 to 2023 was selected to assess the coherence of salinity (CTDSAL) and dissolved oxygen (OXYGEN) measurements with GLODAPv2.2022 cruises in the region (Figure 9). RADPROF is a structural monitoring program lead by the IEO (Spanish Institute of Oceanography, CSIC), that has been running from 2003. Since 2014, inorganic biogeochemical variables, including CO_2 measurements, have been collected following GO-SHIP standards (Hood et al., 2010) with the aim of recognising this annual hydrographic section as an Associated GO-SHIP line, and to include it in the GLODAPv3 data product.

The RADPROF 2014-2023 cruise data can be found both in the OCADS (NOAA) and IEO (NODC IEO) repositories. For the 2QC analysis, data below 4000 dbars are compared, as this is an area with low temporal variability and quite homogeneous property profiles, as commented for the RADPROF 2019 case (Section 3.2.1). Three crossover analysis for each RADPROF cruise are run using the adapted 2QC script as presented in Section 2 and

Figure **6** with a maximum distance of 200 km between stations. The reference data utilized is the GLODAPv2.2022 North Atlantic data. Specifically, the 2QC evaluates every cruise-pair between the 10 RADPROF cruises and 10 biannual French-Spanish OVIDE cruises conducted crossing the North Atlantic from 2002 to 2016 (Table 1.). he results of the 2QC analysis for

the original RADPROF data will be compared with those obtained after modifying the RADPROF data, as explained in Section 3.2.3:

- Salinity data (CTDSAL) is modified using Case 4 and two values for the typical uncertainty, 0.005 and 0.01.
- Oxygen data (OXYGEN) is modified using Case 4 and two values for the typical uncertainty, 1% and 2%.

The results of 2QC analysis, which introduces random noise with a gaussian distribution (Case 4) at two different magnitudes, a medium one (in magenta dots) and a high one (in blue dots) are compared with the results obtained from the original data (in cyan dots) in Figure 16. When a moderate value of random uncertainty is introduced, simulating lower precision than expected the crossover results for both salinity and oxygen remain largely unchanged: the number of cruise-pair results above or below the fixed adjustment limits shows only slight variation, and the standard deviation values remain moderately low. However, when a higher noise is introduced in the original data, the tail of results above and below the limits remarkably increases, also presenting a high standard deviation.

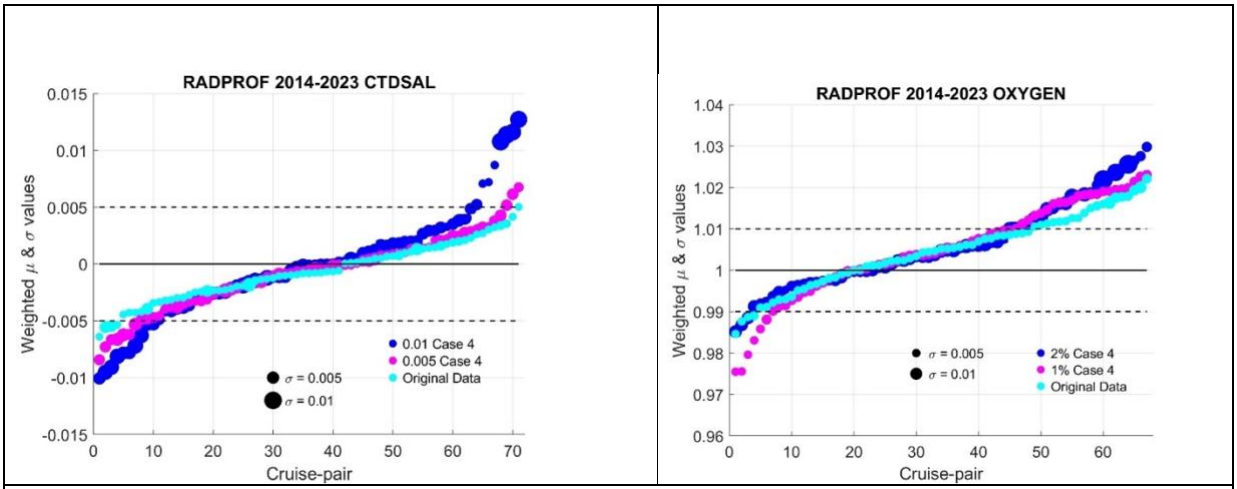

Figure 16. Summary of all cruise-pairs or crossover results sorted in ascending order comparing RADPROF (2014-2023) and OVIDE (2002-2016) cruises. Each cruise-pair has a weighted mean (μ, dot value) and standard deviation (σ, size of the dot) according to Equation (1). Values for CTDSAL are additive, positive values indicate that RADPROF data is higher than OVIDE data. Values for OXYGEN are multiplicative, higher than 1 indicate that RADPROF data is higher than OVIDE data. Different colours represent the typical noise value introduced in the original RADPROF data using Case 4 (random noise with mean zero and a standard deviation equal to the typical noise value). The reference adjustments' limits used in GLODAP are showed as dotted lines, ±0.005 for CTDSAL, and ±1% for OXYGEN.

Each RADPROF cruise weighted mean and standard deviation offset (Equation 3) considering all the crossover or cruise-pairs results is calculated with two types of weighting, Wσ and WTime·WSpace·Wσ (Section 3.2.2).

Systematic biases in the RADPROF data are evidenced when the weighted mean of all crossover results for each RADPROF cruise surpasses a predefined adjustment limit. Along

with the former conditioning, adjustments are applied if the weighted standard deviation of the offset is sufficiently low to consider the mean offset meaningful. In Figure 17 we explore the impact of the introduced uncertainty and the weighting scheme (Section 3.2.2) on the final offset results for each RADPROF cruise.

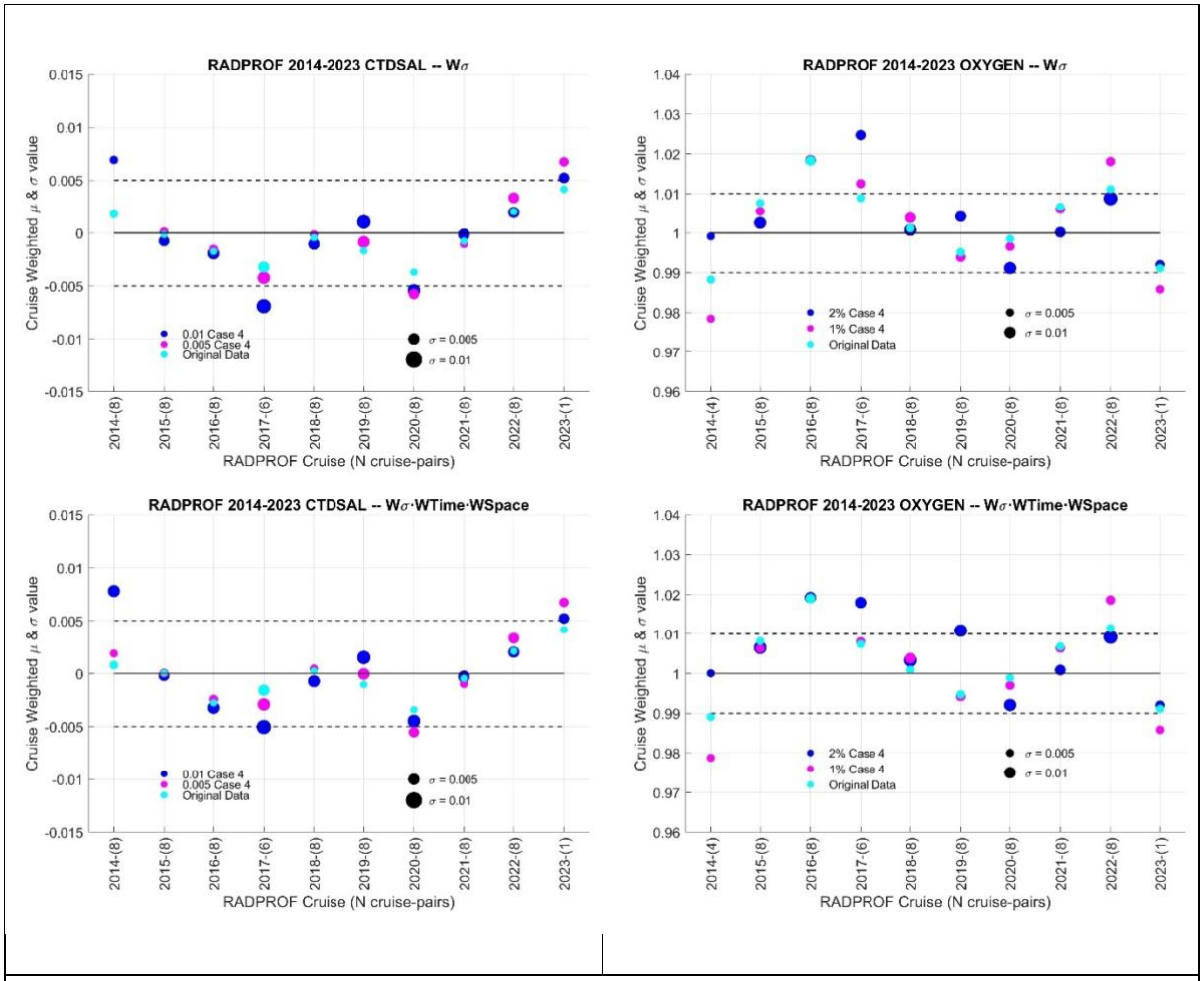

Figure 17. Overall crossover result for each RADPROF cruise showed as the weighted mean (μ, dot value) and standard deviation (σ, size of the dot) according to Equation (3), each cruise is identified with the year and the number of cruise-pairs in brackets in the x-axis. Values for CTDSAL are additive and values for OXYGEN are multiplicative. Different colours represent the typical noise value introduced in the original RADPROF data using Case 4 (random noise with mean cero and a standard deviation equal to the typical noise value). The reference adjustment limits used in GLODAP are showed as dotted lines, ±0.005 for CTDSAL, and ±1% for OXYGEN.

Regarding CTDSAL in Figure 17, most of the results using Wσ are comprised within the ±0.005 limit, and when over the limit they have a high standard deviation (e.x., cruise 2017 with 0.01 noise) or appear only in the high noise case (e.x., 2014 cruise). We think that a more robust estimation of the offsets should consider both also time and space in the weighting scheme. For example, the 2014 CTDSAL data would remain uncorrected as both the original and medium noise results are within the acceptable limits, as also obtained for the 2017 cruise. The results for 2023 are not robust as there is only one crossover point, during that cruise due

to an emergency onboard, IEO was unable to complete the deepest stations. But salinity data seems high in that cruise where the cruise report also evidences some problems with the salinity calibration. Therefore, we only recommend two corrections for CTDSAL data: an increase of 0.0055±0.0040 for RADPROF 2020 and a decrease of 0.0068±0.0034 for RADPROF 2023, based on the medium noise results that includes all weights. Considering the usual 2QC analysis, no cruise would be corrected.

Regarding OXYGEN in Figure 17, most of the results using Wσ are comprised within the ±1% limit except cruises 2014, 2016, 2022 and 2023 that present original and medium noise results slightly above the limit. We think a more robust estimation of the offsets is considered using also time and space in the weighting scheme. We propose the following corrections for OXYGEN data: RADPROF 2014 is divided by 0.9788±0.0050 and RADPROF 2016 by 1.0190 ±0.0069, and RADPROF 2022 by 1.0186±0.0066 based on the medium noise results including all weights. Considering the usual 2QC analysis only the RADPROF 2016 would be corrected.

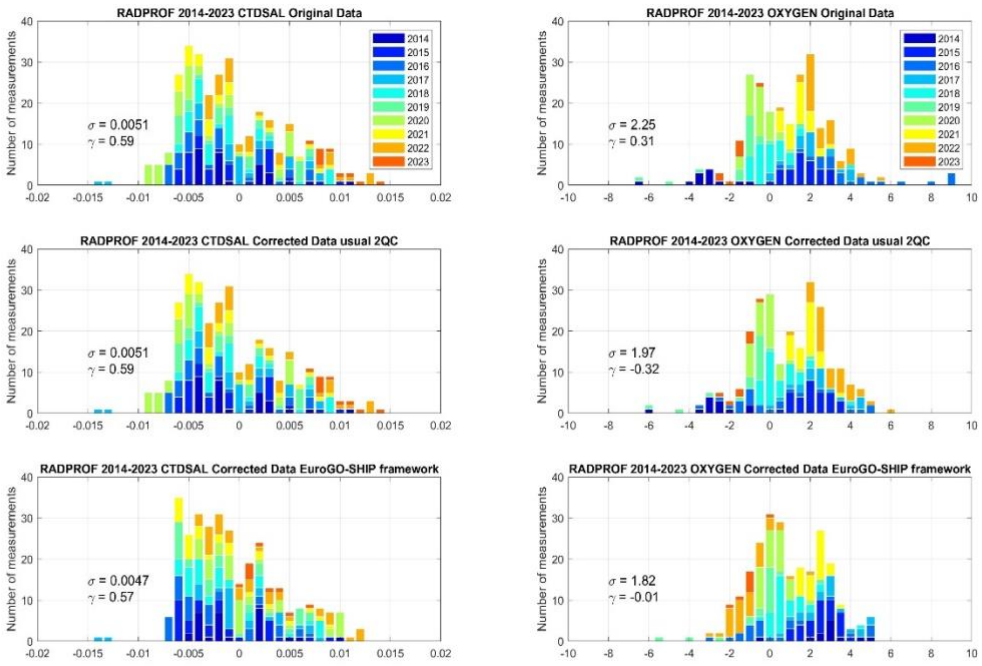

Figure 18. Histograms showing the distribution of the deep waters (>4000 dbars) anomalies from the mean value of CTDSAL (right plots) and OXYGEN (left plots) considering all RADPROF 2014-2023 cruises. Original data is showed in the upper plots, the middle plots show the results after the usual 2QC analysis corrections are applied and the lower plots the results after the proposed framework results are applied. Each plot shows the standard deviation and the skewness of the anomaly values.

By applying the 2QC analysis corrections, the coherence (repeatability) of RADPROF CTDSAL and OXYGEN is quantified using the standard deviation and the Fisher coefficient of asymmetry, calculated for the differences or anomalies from the mean deep water CTDSAL

or OXYGEN value (>4000 dbars) considering both original and corrected RADPROF cruises data. **Figure 18. shows the improvement of the coherence or repeatability of RADPROF data after applying the 2QC analysis corrections. Data can be considered coherent to better than ±0.005 for CTDSAL and better than ±2 µmol.kg^{-1} for OXYGEN.**

4.3. Cruises in the Western Mediterranean Sea

In this case study we assess the repeatability of sensor-based dissolved oxygen data (CTDOXY) calibrated against discrete Winkler dissolved oxygen measurements for a series of 10 cruises in the Western Mediterranean Sea (West MED) (Figure 19.). Those cruises contribute to the Western Mediterranean Sea oxygen data collection (O2WMED) that has been compiled, presented and quality controlled in a recently submitted manuscript by Belgacem et al. (2024). This work complements the inorganic nutrient data product for the same area compiled and quality controlled in Belgacem et al. (2020).

The collection of cruises to be checked consists of Italian cruises conducted by different laboratories, crossing the Algero-Provençal subbasin in the southern WMED. The crossover analysis excludes the Gulf of Lion region where annual deep water formation processes take place, and is confined to the geographical area between 2°E and 8.1°E longitude and 35.5°N to 41.5°N latitude.

As reference cruises, we use three cruises: 06MT20110405, 29AH20140426 and 29AJ20160818, during which discrete oxygen measurements were performed following Langdon (2010). Crossover stations were evaluated within the 2° arc distance (approximately 222 km) for data collected at depths below 1000 dbars. The 2QC tool presented in Section 2.2 & 2.3 is adapted to compare continuous CTDOXY data pressure profiles with discrete oxygen data from the reference cruises. This case study aims to assess the noise propagation (Case 4) within the to be checked dataset with 1% and 2% typical noise added. This noise would stem from a misleading sensor calibration issue, i.e, reduced precision of Winkler measurements used in the calibration, lower coverage of Winkler data, sensor hysteresis, and any other potential issues leading to systematic and random uncertainty in the original CTDOXY data, including potential natural variability affecting deep waters in the area.

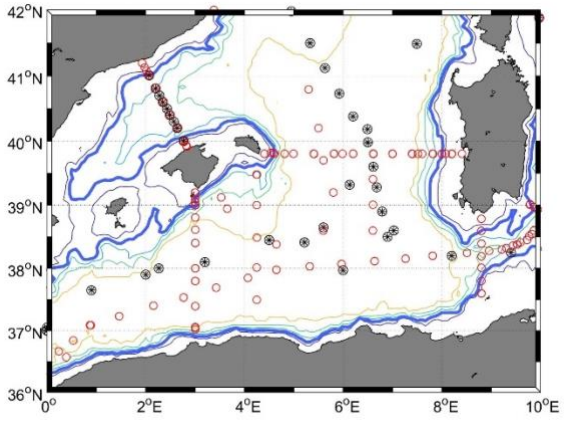

Figure 19. Map of the Western Mediterranean Sea showing the ten O2WMED Mediterranean cruise stations (in red) and the reference cruise stations (in black). The 1000 dbars isobath is highlighted.

Like the previous case study, Figure 20. presents the weighted mean and standard deviation for every cruise-pair or crossover, calculated using Equation (3). The figure clearly shows that several crossover results surpass the lower and upper adjustment limit marked at ±1%. The mean crossover values for the different levels of uncertainty remain relatively stable; but, as expected, the overall standard deviation increases progressively with the addition of noise. Specifically, standard deviation rises from 0.007 (no noise added) to 0.01 (with 1% noise added) and further to 0.017 (with 2% noise added).

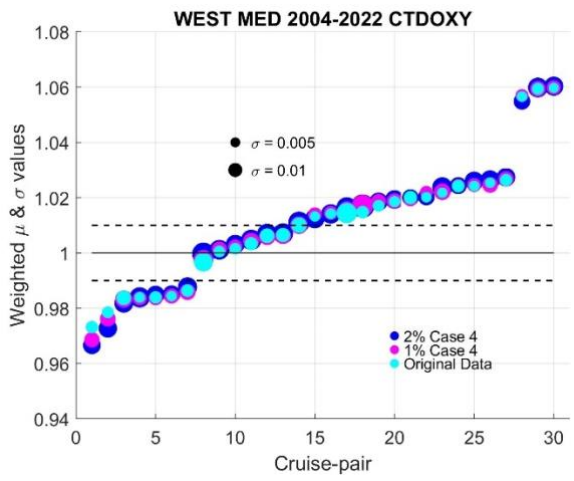

Figure 20. Summary of all cruise-pairs or crossover results sorted in ascending order comparing cruises in the WMED. Each cruise-pair has a weighted mean (μ, dot value) and standard deviation (σ, size of the dot) according to Equation (1). Values are multiplicative, higher than 1 indicate that the checked data is higher than the reference. Different colours represent the typical noise value introduced in the original data using Case 4 (random noise with mean zero and a standard deviation equal to the typical noise value). The reference adjustment limits used in GLODAP are showed as dotted lines ±1%.

Systematic biases in the WMED CTDOXY data are evidenced when the weighted mean of all crossover results for each cruise (Equation 3) surpasses a predefined adjustment limit. For example, the 2006 cruise is quite higher than the mean. Along with the former conditioning, adjustments are applied if the weighted standard deviation of the offset is low enough to consider the mean offset meaningful. Figure 21. explores the impact of the introduced uncertainty and the weighting scheme (Section 3.2.2) on the final offset results for each cruise. In this case, the weighting has minimal or no impact on the 2QC evaluation due to the limited geographical area under consideration and the predominance of Wσ over time and space. It is important to note that this analysis considers the water column below 1000 dbars, which is less homogenous compared to the water column below 4000 dbars in the Northeast Atlantic. Therefore, the standard deviation for each cruise pair σ is higher than that observed in the Northeast Atlantic (Figure 16. & Figure 20. As in the previous case study for the Northeast Atlantic, we consider the results for Case 4 with 1% uncertainty and apply the complete weighting scheme (Figure 21.). The CTDOXY data would be divided by the following offsets:

Expocode	CTDOXY Offset
48UR20041006	1.024±0.009
48UR20060608	1.059±0.012

48UR20071005	0.977±0.013
48UR20100731	1.019±0.012
48UR20131015	1.025±0.012
11BG20220517	0.984±0.010

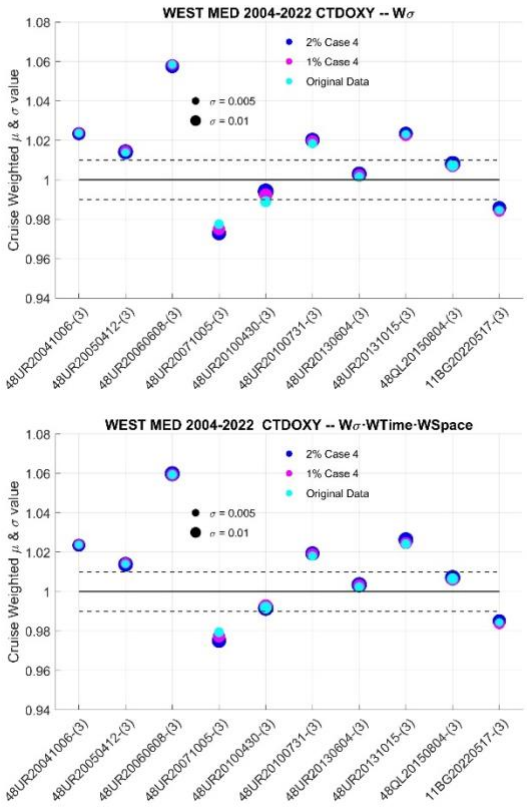

Figure 21. Overall crossover result for each West MED cruise showed as the weighted mean (μ, dot value) and standard deviation (σ, size of the dot) according to Equation (3), each cruise is identified with the expocode and the number of cruise-pairs in brackets in the x-axis. Values are multiplicative, higher than 1 indicate that the checked data is higher than the reference. Different colours represent the typical noise value introduced in the original RADPROF data using Case 4 (random noise with mean cero and a standard deviation equal to the typical noise value). The reference adjustment limits used in GLODAP are showed as dotted lines ±1%.

By applying the 2QC analysis corrections, the coherence (repeatability) of WMED CTDOXY data is quantified using the standard deviation and the Fisher coefficient of asymmetry, calculated for the differences or anomalies from the mean deep water CTDOXY value (>1000 dbars) considering both original and corrected WMED cruise data. Figure 18. **shows the improvement of the WMED data after applying the 2QC analysis results. Several systematic biases were detected and corrected, as evidenced by a reduction in the skewness value, which decreases from 0.54 to 0.03, while the standard deviation exhibited a slight decrease. In this marginal sea, where the natural variability is higher than in the Northeast Atlantic, and considering sensor-based data that originally had higher uncertainty, we can assure that systematic biases are reduced in the CTDOXY WMED data product. But the random uncertainty remains, as indicated by the relatively stable standard deviation before and after the 2QC corrections. The overall CTDOXY product consistency would be about ±9 μmol.kg^{-1}.**

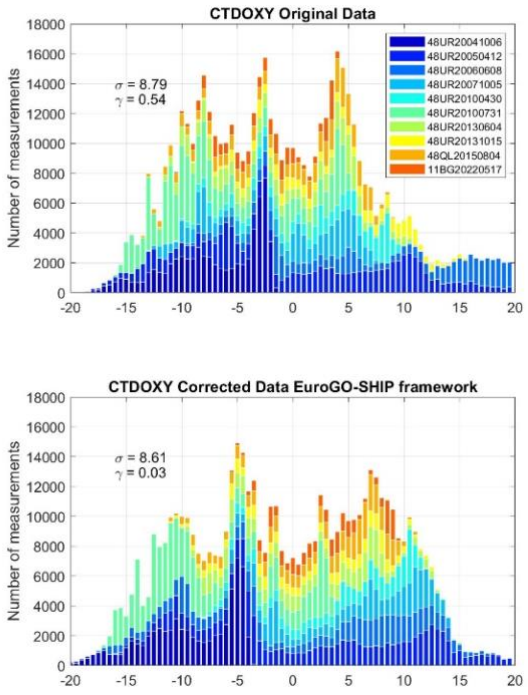

Figure 22. Histograms showing the distribution of the deep water (>1000 dbars) anomalies from the mean value of CTDOXY considering all West MED cruises. Original data is showed in the upper plots, the lower plot shows the results after the proposed framework results are applied. Each plot shows the standard deviation and the skewness of the anomaly values.

5. Conclusions

5.1. Summary and main findings of the deliverable

his deliverable underscore the difficulty in designing a framework to properly quantify the uncertainty in a data product compiling cruise data from diverse origins. Established Metrological approaches to quantify the uncertainty in measurements are unfeasible to apply, as they are designed to quantify the uncertainty propagation in each step of a measurement procedure(Figure 7 & Figure 8). Indeed, a secondary quality control procedure assesses the coherence over space and time, the repeatability in a metrological sense, of the same measurand (variable) performed under different conditions, maybe or not following the same procedure.

In this report, we explore the effect of introducing random uncertainty into the set of measurements (cruise data) on the 2QC offset results. Here, we are considering the reference cruises as true values with no a priori uncertainty. This proposed framework for assessing uncertainty in a 2QC analysis is a first approach applying a Monte Carlo method within the 2QC process.

Our findings explored different probability distributions for the random uncertainty (noise) introduced in the original data, as well as different magnitudes of this noise and different weighting schemes used in the final calculation of the weigthed mean offsets and standard deviations for each checked cruise. We explored two different variables: salinity, a physical

property where offsets are additive, and dissolved oxygen, a biogeochemical property where offsets are multiplicative.

We conclude that:

- Uniform random noise (Case 2) with moderate to high values (salinity 0.005 to 0.01, and oxygen 1% to 1.5%) does not significantly affect the final 2QC mean offset results; but it does lean to an increase in standard deviation. Therefore, data with a low precision, regardless of the presence of systematic bias, can be identified and included in a data product.

- A more likely distribution of the noise in oceanographic data would be Case 4, a gaussian distribution of the random noise (zero mean and standard deviation equal set to an uncertainty value) with different magnitudes according to the QC/QA reported in the checked data.

- Typical uncertainty values for salinity (0.005) and oxygen (1%) can be propagated through high precision cruise data and help identify suspicious offset values in the adjustment limits for discrete salinity and oxygen data in RADPROF cruises. While in the WMED cruises where sensor-based oxygen data has lower precision, confirm the original 2QC results with the findings from Case 4 at 1% noise.

- A crossover or cruise-pair result is considered meaningful according to its standard deviation, a combination of each cruise data precision. An offset result is the combination of several cruises-pair results weighted by each cruise-pair standard deviation. Uncertain offset results could be confirmed or disregarded if additional weights are considered, as the time and space distance for each cruise pair.

- A simple method to quantify the coherence or repeatability of a set of cruises involves checking the standard deviation and skewness of the residual values within the assumed low temporal variability layer (usually deep and bottom waters). Residuals would be calculated based on the overall mean values of both the original and 2QC corrected data. The standard deviation would indicate the random uncertainty of the data, while the skewness would indicate the presence of any remaining systematic biases.

- Adjustment limits are defined based on a minimum uncertainty value for high quality measurements', according to the gold standard best practices procedures. A metrological approach would consider the individual uncertainties of each data set. For example, if Cruise A reports a salinity uncertainty of 0.003 and Cruise B reports an uncertainty of 0.005, systematic biases would be considered if they exceed (0.005^2+0.003^2)^0.5 = 0.0058.

- A dialogue between the oceanographic and metrology communities is highly recommend to improve the quality assurance and quality control procedures of not only individual data, but also to establish statistical metrology methods to quantify the reproducibility of oceanographic data products.

5.2. Contributions to the project and the European hydrography community

This deliverable highlights the need for detailed information about QA/QC for every variable in a cruise data report and/or cruise metadata. The 1QC application developed within EuroGO-SHIP could also require this information to be integrated in the final formatted and quality-controlled cruise data.

Oceanographers need to be aware that precision and accuracy are not the same: the first is usually associated with the random uncertainty of the data and the second with any systematic bias from a true value. Best practices procedures need to clearly account for QA/QC details to quantify precision and accuracy. Both of them should be guaranteed for all batches of measurements during a cruise. In this sense, there is an urgent need to develop and characterize procedures for in-house and certified reference materials production. EuroGO-SHIP will contribute to this urgent need for the oceanographic community.

Clearly, if individual data sets' precision and accuracy are assured, so will the reproducibility and repeatability between different data sets and no 2QC would be needed.

5.3. Limitations and outlook

The proposed MonteCarlo approach requires computing time and memory. If properly done, each case study should have been repeated several times to check the propagation of the introduced noise and provide mean values of those repetitions.

Information on the noise or uncertainty to be propagated for each cruise and variable should be included in the cruise report and/or metadata, to have an idea about the precision and accuracy of the original data, but it is not always present or quantified.

We have proposed a quite simple approach to assess the uncertainty in a 2QC analysis for two case studies with a limited number of cruises. Further development and final implementation of the proposed framework for a global data product as GLODAP, compiling more than 1000 cruises, will be a great challenge. We envisage this challenge will require computing resources and a coherent strategy to implement the Monte Carlo approach within the inversion in Equation (2), where the time and space weighting schemes would be easy to implement. Results would be computed automatically. However, the combined expertise of oceanographers, metrologists and statisticians will be always required.

References

Allard, A., Fischer, N., (2015). Note technique n°002: Guide pour l'évaluation des incertitudes de mesure LNE internal document

Belgacem, M., Chiggiato, J., Borghini, M., Pavoni, B., Cerrati, G., Acri, F., Cozzi, S., Ribotti, A., Álvarez, M., Lauvset, S. K., and Schroeder, K. (2020). Dissolved inorganic nutrients in the western Mediterranean Sea (2004–2017), Earth Syst. Sci. Data, 12, 1985–2011, https://doi.org/10.5194/essd-12-1985-2020.

Belgacem, M., Schroeder, K., Lauvset, S. K., Álvarez, M., Chiggiato, J., Borghini, M., Cantoni, C., Ciuffardi, T., and Sparnocchia, S. (2024). A consistent regional dataset of dissolved oxygen in the Western Mediterranean Sea (2004–2023): O2WMED, Earth Syst. Sci. Data Discuss. [preprint] , https://doi.org/10.5194/essd-2024-365, in review.

Bushnell M, Waldmann C, Seitz S, Buckley E, Tamburri M, Hermes J, Henslop E and Lara-Lopez A (2019). Quality Assurance of Oceanographic Observations: Standards and Guidance Adopted by an International Partnership. *Front. Mar. Sci.* 6:706. doi: 10.3389/fmars.2019.00706

Capitaine, Gaëlle (2024). Etablir la traceabilité métrologique des mesures de l'acidification des eaux marines. These Doctorat, Aix Marseille Université.

Carter, B., Sharp, J., Dickson, A., Álvarez, M., Fong, M., García-Ibáñez, M.I., Woosley, R., Takeshita, Y., Barbero, L., Byrne, R., Cai, W.J., Chierici, M., Clegg, S., Easley, R., Fassbender, A., Fleger, K., Li, X., Martín-Mayor, M., Schockman, K., Wang, Z.A. (2024). Uncertainty sources for measurable ocean carbonate chemistry variables. Limnology and Oceanography, Review, 69, 1, 1-21, DOI: 10.1002/lno.12477.

Carter, B.R., Sharp, J.D., García-Ibáñez, M.I., Woosley, R.J., Fong, M.B., Álvarez, M., Barbero, L., Clegg, S.L., Easley, R., Fassbender, A.J., Li, X., Schockman, K.M. and Wang, Z.A. (2024). Random and systematic uncertainty in ship-based seawater carbonate chemistry observations. Limnol Oceanogr. https://doi.org/10.1002/lno.12674

Chai, F., Johnson, K.S., Claustre, H. *et al.* (2020). Monitoring ocean biogeochemistry with autonomous platforms. *Nat Rev Earth Environ* **1**, 315–326. https://doi.org/10.1038/s43017-020-0053-y

Ebeling, A., Zimmermann, T., Klein, O., Irrgeher, J. and Pröfrock, D. (2022). Analysis of Seventeen Certified Water Reference Materials for Trace and Technology-Critical Elements. Geostand Geoanal Res, 46: 351-378. https://doi.org/10.1111/ggr.12422

Fanton, J.-P., (2019). A brief history of metrology: past, present, and future. Int. J. Metrol. Qual. Eng. 10, 5. https://doi.org/10.1051/ijmqe/2019005

Feistel, R., Wielgosz, R., Bell, S. A., Camões, M. F., Cooper, J. R., Dexter, P., Dickson, A. G., Fisicaro, P., Harvey, A. H., Heinonen, M., Hellmuth, O., Kretzschmar, H. J., Lovell-Smith, J. W., McDougall, T. J., Pawlowicz, R., Ridout, P., Seitz, S., Spitzer, P., Stoica, D. and Wolf, H. (2015) Metrological challenges for measurements of key climatological observables: Oceanic salinity and pH, and atmospheric humidity. Part 1: Overview. Metrologia, 53(1),pp. R1-R11. DOI:10.1088/0026-1394/53/1/R1

Firing, Y., Álvarez, M., Balan, S., Karstensen, J., Musat, T., O''Donnell, G., Sanchez-Franks, A., Schroeder, K. Vasiliu, D. (2024) EuroGO-SHIP D3.3: Report on updating salinity best practice.

González-Pola, C., Larsen, K. M. H., Fratantoni, P., and Beszczynska-Möller, A. (Eds.). (2023). ICES Report on ocean climate. ICES Cooperative Research Reports Vol. 358. 123 pp. https://doi.org/10.17895/ices.pub.24755574

Gouretski, V. and Jancke, K. (1999). A description and quality assessment of the historical hydrographic data for the South Pacific Ocean, Journal of Atmospheric and Oceanic technology, 16 , pp. 1791-1815 .

Gouretski, V. V. and Jancke, K. (2001). Systematic errors as the cause for an apparent deep water property variability: global analysis of the WOCE and historical hydrographic data, Prog. Oceanogr., 48, 337–402.

Hartman SE, Gates AR, Lopez-Garcia P, Bozzano R, Delory E, Favali P, Lefevre D, Chirurgien L, Pensieri S, Petihakis G, Nair R, Neves S, Dañobeitia JJ, Salvetat F, Le Menn M, Seppälä J, Schroeder K and Piera J (2023) Proposed synergies between oceanography and metrology. *Front. Mar. Sci.* 10:1192030. doi: 10.3389/fmars.2023.1192030

Hood, E.M., C.L. Sabine, and B.M. Sloyan, eds. (2010). The GO-SHIP Repeat Hydrography Manual: A Collection of Expert Reports and Guidelines. IOCCP Report Number 14, ICPO Publication Series Number 134. Available online at http://www.go-ship.org/HydroMan.html.

ISO 21748 (2017). Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation.

ISO 5725-2 (2020). Accuracy (trueness and precision) of measurement methods and results-Part 2: basic method for the determination of repeatability and reproducibility of a standard measurement method

JCGM 100:2008 (2008). Evaluation of Measurement Data-Guide to the Expression of Uncertainty in Measurement.

JCGM 101:2008 (2008). Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method.

JCGM 200:2012 (2012). International Vocabulary of Metrology (VIM) – basic and general concepts and associated terms, 3rd ed., https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf

Johnson, G. C., Robbins, P. E., and Hufford, G. E. (2001). Systematic adjustments of hydrographic sections for internal consistency, J. Atmos. Ocean. Technol., 18, 1234–1244.

Jutterström, S., Anderson, L. G., Bates, N. R., Bellerby, R., Johannessen, T., Jones, E. P., Key, R. M., Lin, X., Olsen, A., and Omar, A. M. (2010). Arctic Ocean data in CARINA, Earth Syst. Sci. Data, 2, 71–78, https://doi.org/10.5194/essd-2-71-2010.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanninkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T. H. (2004). A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cy., 18, GB4031, https://doi.org/10.1029/2004GB002247.

Langdon, C. (2010). "Determination of dissolved oxygen in seaweater by Winkler titration using Amperometric Technique," in *The GO-SHIP Repeat Hydrography Manual: A Collection of Expert Reports and Guidelines. Version 1 (IOCCP Report Number 14; ICPO Publication Series Number 134)*, eds E. M. Hood, C. L. Sabine and B. M. Sloyan (Lyon: ICPO), 18. doi: 10.25607/OBP-1350

Lauvset, S. K., Lange, N., Tanhua, T., Bittig, H. C., Olsen, A., Kozyr, A., Alin, S., Álvarez, M., Azetsu-Scott, K., Barbero, L., Becker, S., Brown, P. J., Carter, B. R., da Cunha, L. C., Feely, R. A., Hoppema, M., Humphreys, M. P., Ishii, M., Jeansson, E., Jiang, L.-Q., Jones, S. D., Lo Monaco, C., Murata, A., Müller, J. D., Pérez, F. F., Pfeil, B., Schirnick, C., Steinfeldt, R.,

Suzuki, T., Tilbrook, B., Ulfsbo, A., Velo, A., Woosley, R. J., and Key, R. M. (2022). GLODAPv2.2022: the latest version of the global interior ocean biogeochemical data product, Earth Syst. Sci. Data, 14, 5543–5572, https://doi.org/10.5194/essd-14-5543-2022.

Lauvset, S. K., Lange, N., Tanhua, T., Bittig, H. C., Olsen, A., Kozyr, A., Álvarez, M., Azetsu-Scott, K., Brown, P. J., Carter, B. R., Cotrim da Cunha, L., Hoppema, M., Humphreys, M. P., Ishii, M., Jeansson, E., Murata, A., Müller, J. D., Pérez, F. F., Schirnick, C., Steinfeldt, R., Suzuki, T., Ulfsbo, A., Velo, A., Woosley, R. J., and Key, R. M. (2024). The annual update GLODAPv2.2023: the global interior ocean biogeochemical data product, Earth Syst. Sci. Data, 16, 2047–2072, https://doi.org/10.5194/essd-16-2047-2024.

Le Menn M., Seitz S., Nair R., Ntoumas M. (2023). White paper on advances in absolute salinity measurements. Metrology for Integrated Marine Management and Knowledge-Transfer Network, MINKE. doi: 10.5281/zenodo.7599993

Levitus, S., Burgett, R., & Boyer, T. P. (1994). World Ocean Atlas (1994). Volume 1: Salinity. *NOAA Atlas Series*, Washington D.C

Lherminier, P. et al. (2007). Transports across the 2002 Greenland-Portugal Ovide section and comparison with 1997. *J. Geophys. Res. Ocean.* **112**, 1–20.

Mantyla, A. (1994). The treatment of inconsistencies in Atlantic deep water salinity data. *Deep-Sea Research*, *41*, 1387–1405.

Olsen, A., Key, R. M., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishii, M., Pérez, F. F., and Suzuki, T. (2016). The Global Ocean Data Analysis Project version 2 (GLODAPv2) – an internally consistent data product for the world ocean, Earth Syst. Sci. Data, 8, 297–323, https://doi.org/10.5194/essd-8-297-2016.

Saunders, P. M. (1986). The accuracy of measurements of salinity, oxygen and temperature in the deep ocean. *Journal of Physical Oceanography*, *16*, 189–195.

Schroeder, K., T. Tanhua, H.L. Bryden, M. Álvarez, J. Chiggiato, and S. Aracri (2015). Mediterranean Sea Ship-based Hydrographic Investigations Program (Med-SHIP). *Oceanography* 28(3):12–15, https://doi.org/10.5670/oceanog.2015.71.

Seitz, S., Feistel, R., Wright, D. G., Weinreben, S., Spitzer, P., and De Bièvre, P. (2011). Metrological traceability of oceanographic salinity measurement results, Ocean Sci., 7, 45–62, https://doi.org/10.5194/os-7-45-2011.

Squara, P, Imhoff MD, Michael MD, Cecconi M, (2015). Metrology in Medicine: From Measurements to Decision, with Specific Reference to Anesthesia and Intensive Care. Anesthesia & Analgesia 120, 66-75.DOI: 10.1213/ANE.0000000000000477

Tanhua T, Pouliquen S, Hausman J, O'Brien K, Bricher P, de Bruin T, Buck JJH, Burger EF, Carval T, Casey KS, Diggs S, Giorgetti A, Glaves H, Harscoat V, Kinkade D, Muelbert JH, Novellino A, Pfeil B, Pulsifer PL, Van de Putte A, Robinson E, Schaap D, Smirnov A, Smith N, Snowden D, Spears T, Stall S, Tacoma M, Thijsse P, Tronstad S, Vandenberghe T, Wengren M, Wyborn L and Zhao Z (2019). Ocean FAIR Data Services. *Front. Mar. Sci.* 6:440. doi: 10.3389/fmars.2019.00440

Tanhua, T., Lauvset, S.K., Lange, N., Olsen, A., Álvarez, M., et al. A vision for FAIR ocean data products. Commun Earth Environ 2, 136 (2021). https://doi.org/10.1038/s43247-021-00209-4.

Tanhua, T., van Heuven, S., Key, R. M., Velo, A., Olsen, A., and Schirnick, C. (2010). Quality control procedures and methods of the CARINA database, Earth Syst. Sci. Data, 2, 35–49, https://doi.org/10.5194/essd-2-35-2010.

Tel, E., Balbin, R., Cabanas, J.-M., Garcia, M.-J., Garcia-Martinez, M. C., Gonzalez-Pola, C., Lavin, A., Lopez-Jurado, J.-L., Rodriguez, C., Ruiz-Villarreal, M., Sánchez-Leal, R. F., Vargas-Yáñez, M., and Vélez-Belchí, P. (2016). IEOOS: the Spanish Institute of Oceanography Observing System, Ocean Sci., 12, 345–353, https://doi.org/10.5194/os-12-345-2016.

Uchida, H., T. Kawano, T. Nakano, M. Wakita, T. Tanaka, and S. Tanihara (2020). An expanded batch-to-batch correction for IAPSO standard seawater. *J. Atmos. Oceanic Technol.,* 1507-1520. 10.1175/JTECH-D-19-0184.1

Waldmann, C., Fischer, P., Seitz, S., Köllner, M., Fischer, J-G., Bergenthal, M., Brix ,H,. Weinreben, S., and Huber, R. (2022). A methodology to uncertainty quantification of essential ocean variables. Frontiers in Marine Science, 9:1002153, 16pp. DOI: https://doi.org/10.3389/fmars.2022.1002153