Project Information	
Project full title	EuroSea: Improving and Integrating European Ocean Observing and Forecasting Systems for Sustainable use of the Oceans
Project acronym	EuroSea
Grant agreement number	862626
Project start date and duration	1 November 2019, 50 months
Project website	https://www.eurosea.eu

Deliverable information	
Deliverable number	D4.4
Deliverable title	**Quality-control procedures for ship-board biogeochemical time series data**
Description	Framework of the envisioned time-series synthesis product used to indicate the consistency of biogeochemical time-series data (in-)between different ship-based sites.
Work Package number	4
Work Package title	Data integration, Assimilation, and Forecasting
Lead beneficiary	Universitetet i Bergen (UiB)
Lead authors	Nico Lange, Benjamin Pfeil, Björn Fiedler
Contributors	
Due date	30.04.2022
Submission date	31.05.2022
Resubmission date	18.08.2023 (revised version)
Comments	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 862626.

https://doi.org/10.3289/eurosea_d4.4

Table of contents

Executive summary

This framework will be incorporated into a time-series (TS) data synthesis product. The framework will be used to indicate the consistency of biogeochemical (BGC) time-series data between different ship-based time-series sites. It differentiates between three different “consistency categories”: 1) Metadata Availability, 2) Measurement and Analyzing Techniques and 3) Applied Quality-Control (QC). For each of these categories, a flagging scheme will be implemented based upon pre-defined “consistency criteria”. All data consistency flags combined provide a comprehensive and easy to understand indication of the degree of consistency of the incorporated time-series data.

A special emphasis is put upon the third consistency category, “Applied Quality Control”, as - despite of the potential to increase the precision and accuracy of the measured data - only very few QC procedures are established within the BGC time-series community. The heterogenic nature of the time-series sites does not permit a “One-fit-all, Best-Case” QC routine, as one routine only cannot meet the needs of all time-series sites. To accommodate for this, an overarching QC guideline based upon a decision tree model has been developed, which leads to the most appropriate QC routine available. The suggested QC routines in turn are further categorized from “Best” to “Minimum”, depending on their potential to identify bad samples and/or to detect systematic biases. The guideline is applicable for all BGC ship-based time series sites and the recently developed “Regular Outlier Test” (ROT, EuroSea milestone MS13) has been incorporated and categorized as “Best”. Evaluation results of the ROT QC routines are also included.

1. Background

Biogeochemical (BGC) time-series data is of great importance for a multitude of applications ranging from the identification of temporal and spatial patterns to the validation of autonomous networks and model data. Moreover, it "[…] represent(s) the only means to distinguish between natural and anthropogenic forcings." (Benway et al., 2019). In the OceanObs'19 ocean time-series review article of Benway et al. the authors called amongst others for improved BGC time-series data integration. Following this call and embracing EuroSea's vision of "[…] a user-focused, truly interdisciplinary, and responsive European ocean observing and forecasting system […]", we are working on a BGC time-series pilot synthesis product. Its mission is to bring the ocean BGC time-series community together to jointly develop a sustainable and consistent synthesis data product and agree on Best-Practices[1], striving for OCG[2] network status and delivering timely and high impact BGC TS-data. In particular, EuroSea's objectives 3) *Improving and enhancing the readiness and integration of ocean observing networks* and 4) *Enabling FAIR[3] data, supporting integration of ocean data into Copernicus Marine Service[4], EMODnet[5] and SeaDataNet[6] portfolios* are addressed.

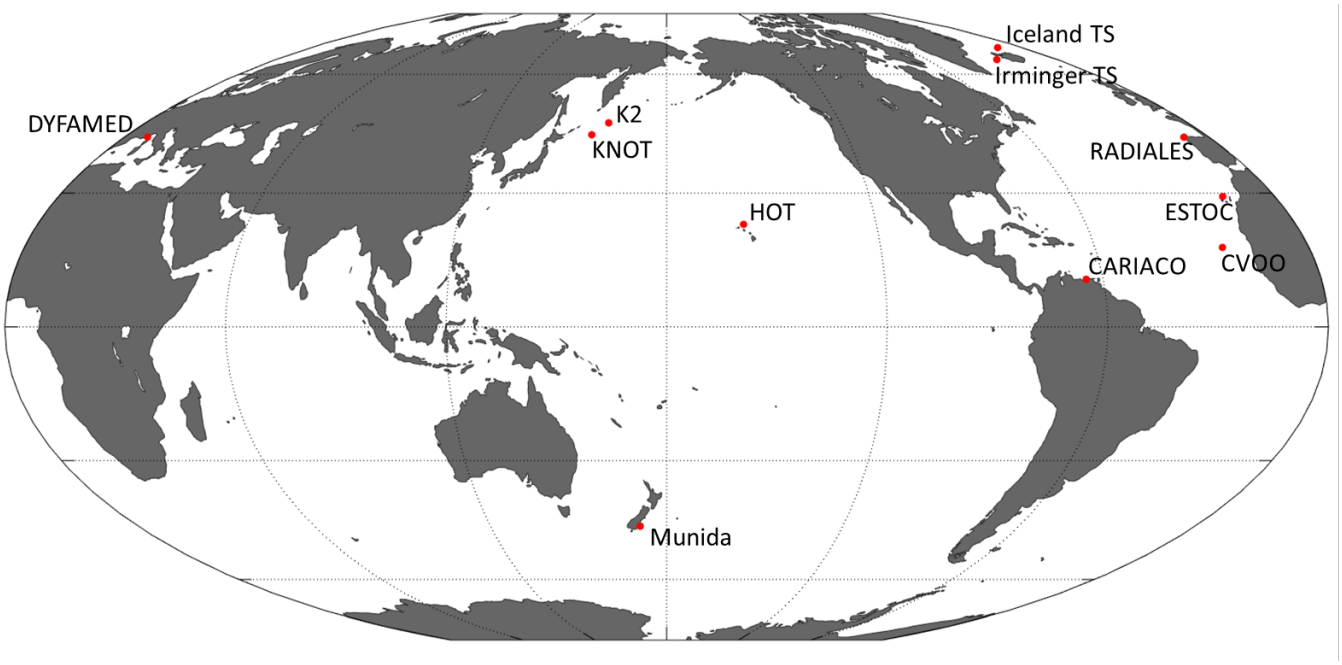

Figure 1. Station map of time-series product pilot

[1] https://www.oceanbestpractices.org/
[2] https://www.goosocean.org/index.php?option=com_oe&task=viewGroupRecord&groupID=103
[3] FAIR = findable, accessible, interoperable, and reusable
[4] https://marine.copernicus.eu/
[5] https://emodnet.ec.europa.eu/en
[6] https://www.seadatanet.org/

Table 1. Participating time-series stations with main characteristics and regions indicated.

	Location	Time Range	Frequency	Depth
CVOO	17.6°N, 24.3°W	2006 -	Seasonal	3600m
ESTOC	29.2°N, 15.5°W	1994 -	4 cruises pa	3600m
Radiales	43.4°N, 8.4°E	1989 -	Monthly	80m
ALOHA	22.8°N, 158.0°W	1988 -	Monthly	4800m
KNOT	44.0°N, 155.0°E	1997 -	1 - 3 cruises pa	6000m
K2	47.0°N, 160.0°E	2001 -	1 - 3 cruises pa	6000m
Munida	45.77° - 45.84°S 170.22° - 171.54°E	1998 -	6 cruises pa	1000m
Irminger	64.33°N, 28.0°W	1983 -	4 cruises pa	1000m
Iceland	68.0°N, 12.67°W	1983 -	4 cruises pa	1850m
CARIACO	10.5°N, 64.7°W	1995 - 2017	Monthly	1310m
DYFAMED	42.25°N, 7.52°W	1991-	Monthly	2400m

Atlantic
Pacific
Nordic Seas
Marginal Seas

The pilot product includes data from in total eleven time-series sites. The selection has been based upon the overarching goal to represent the entire spectrum of time-series sites globally present, see Table 1 and Figure 1. The pilot focuses on all BGC Essential Ocean Variables (EOVs, GOOS[7]) measured by at least one of the participating ship-based time series sites, see Figure 2. Possibly, this can be extended to additional EOVs in the future once the dataflow is established and proven to be successful.

To facilitate the development of the pilot product and to ensure community-agreed upon practices, the time-series station PIs have been consulted on a regular basis with the initiation being the international workshop on BGC time series data (November 2020). During the workshop, four different working groups on the topics of 1) Coordination, 2) Commonality of Methods, 3) QC- and Data Handling and 4) Data Policy have been formed to focus on specific tasks. Working groups 2) and 3) being in particular relevant for this deliverable. Furthermore, a general concept note has been generated to reach and consolidate consensus among all participants setting the basis for the work to come.

[7] https://www.goosocean.org/index.php?option=com_content&view=article&id=283&Itemid=441

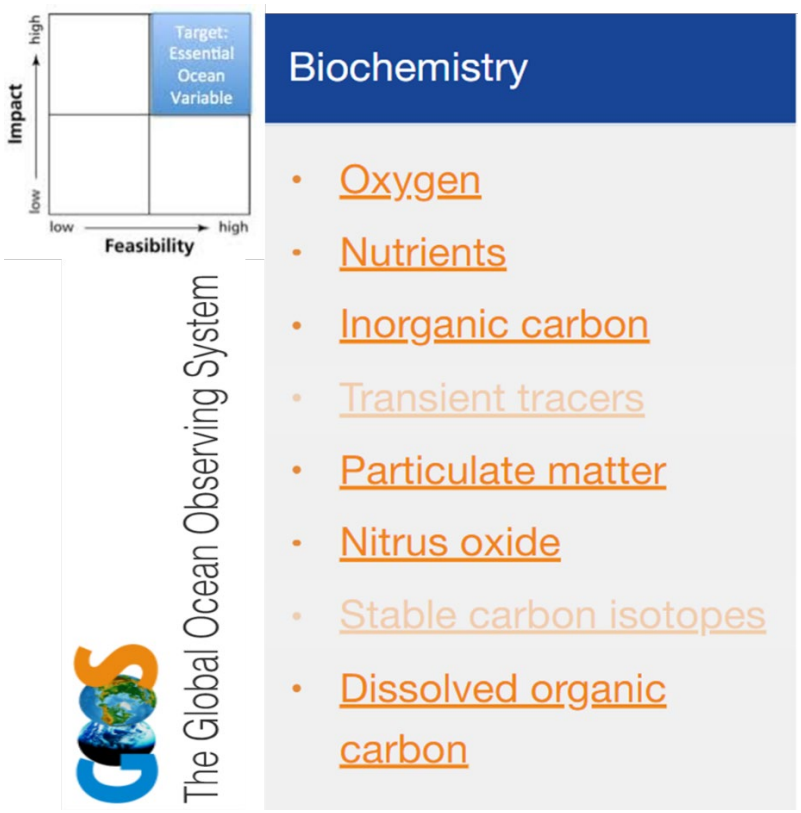

Figure 2. Focused variables of pilot product, defined as Essential Ocean Variables (EOVs) by GOOS[7]. Variables measured by at least one of the participating sites highlighted in dark orange.

2. Data Consistency

At first glance, data synthesis and the related data provision appear to be relatively straight forward tasks. However, the success and readiness of a synthesis product strongly depends on its data consistency, i.e. on the interoperability and comparability of the integrated data. Note that these characteristics are closely related to each other but are not synonyms; comparability can be understood as the goal of consistency. Obtaining truly consistent data in a heterogenic ocean observing system, even if focused on - seemingly - homogenous ship-based time series data only, is a very difficult task.

Besides the diversity of used ontologies, different applied methodologies, from the actual measurement to the analysis of the sample, are the main root of the inconsistencies of time-series data. Clearly, scientists are aware of these inconsistencies and numerous studies have been undertaken to assess their implications (Aoyama et al., 2016; Bockmon and Dickson, 2015; Álvarez et al., 2015). And moreover, nowadays strict Quality-Assurance (QA) and Best-Practices (BP) guidelines, including the application of (certified) reference material, are more commonly established and distributed among the community. Eventually, the wide-spread use of these, as well as more common inter- and intra-laboratory comparison experiments (e.g. Aoyama and Gallia, 2018), will erase most – if not all – inconsistencies. However, presently and especially in historical data, multiple inconsistencies are present. In combination with limited QC routines for marine BGC time-series data (see Quality Control Guideline), generating a synthesis product with fully comparable (historical) data might be unreachable for now. Nonetheless, additional assessments of the provided data and metadata can yield an indication for the degree of consistency of the different time-series data and thus increase data interoperability and comparability.

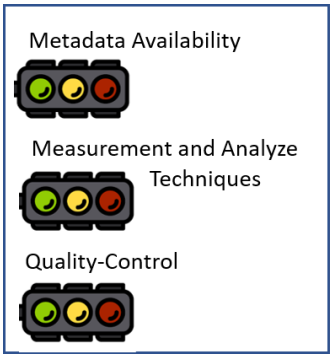

Figure 3. Illustration of data consistency flags for the different categories

To this end, the time-series product QC- and data handling working group has defined three main "consistency" categories, see Figure 3. In the final pilot product, flags are assigned to each category. These are based on the extent to which a parameter of a particular station visit meets predefined "consistency" criteria. All data consistency flags combined then provide a comprehensive way to indicate the degree of consistency of the time-series data.

Consistency categories:

1) Metadata Availability

As the name suggests, this category checks for the availability of metadata. The time-series product metadata working group has developed a metadata template that covers – in addition to CTD data - all BGC EOVs measured by at least one of the participating time-series sites, see Figure 1. Through this

coherent metadata collection, the completeness, findability and open access of the information are guaranteed. And further, the Bermuda workshop recommendations (Lorenzoni and Benway, 2013) for each parameter are more visible and "mapping" between the different ontologies out there is achieved. The template has been developed on the basis of the (NCEI[8]) SDG14.3[9] metadata template for inorganic carbon variables and based upon results from the SCOR Working Group "Towards comparability of global oceanic nutrient data (COMPONUT)"[10], e.g. the "GO-SHIP Repeat Hydrography Nutrient Manual" (Becker et al., 2020). Since the applied methodology varies with time, the metadata had to be provided for each station visit separately. As this is very labor intensive and often requires mining 20-year-old cruise reports, not all sites were able to provide all the queried metadata. The time series product differentiates between "Not present" (flag = 0); "Existing but incoherent or possibly incomplete" (flag = 1); "Provided via the community agreed BGC metadata template" (flag = 2).

2) Measurement and Analyzing Techniques

In this "consistency category" the existing metadata information are analyzed and checked for the compliance of Measurement and Analyze Techniques in respect to known Best-Practices (or established guidelines if no official Best-Practices published), see Table 2. For BGC time-series data these are mainly provided through the Bermuda Workshop Recommendations (Lorenzoni and Benway, 2013). Presently and especially in historical data, multiple inconsistencies in the applied methodology are present. The analysis of the provided metadata from the 11 participating sites revealed inconsistencies which range from the usage of different CTD-sensors, analyzing instruments, probe volumes and reagents, to more parameter specific nuances, such as whether and how nutrients are filtered and stored (analyzed within 24 hours, kept in a dark cool container or frozen to various temperatures), what pH-scale and dye (correction) has been used and whether alkalinity has been determined following spectrophotometric or potentiometric procedures. Usually, double or triple replicates are used to determine the precision of the data. However, fully traceable accuracy estimates are rarely given.

Eventually, the time series product classifies the methods as one of the following: "Do not follow known BP" (flag = 0; red light in Figure 3); "Follow known BP" (flag = 1; yellow light in Figure 3); "Follow known BP, provide accuracy- and precision estimates and participate in inter- and intra-laboratory comparison exercises" (flag = 2; green light in Figure 3).

3) Applied Quality Control

This category indicates the degree of the applied analysis checks (1st and 2nd QC), which aim at increasing the precision and/or accuracy of the data. During the QC, data that has already been analyzed and measured, is checked against neighboring samples in space and time. Note that most of the bad samples actually don't make it into the actual QC, as local experts automatically screen the data during the sample analysis. However, typically the screening and flagging process of local experts is poorly documented and often remains subjective and with time and varying local experts data inconsistencies are prone to develop (except for flags resulting from "sanity checks"). More objective and statistically based QC methods try to minimize these inconsistencies and the most advanced methods (e.g. crossover routines

[8] https://www.ncei.noaa.gov
[9] https://sdgs.un.org/goals/goal14
[10] https://scor-int.org/group/147/

(Tanhua et al., 2010)) actually try to establish baselines making data comparable solely through QC methods. However, of course QC methods have some limitations. Through the analysis of the collected metadata, we have been able to identify quite a large gap when it comes to these additional checks, especially in regard to accuracy checks (i.e. little evidence / documentation on performed 2nd QC). In the best cases, documented QC is limited to the "sigma-criteria", i.e. semi-automatic flagging of samples beyond two or three standard deviations of the total historical mean. To improve this status quo, the QC and data handling working group has developed a clear QC guideline and routine, see Figure 4.

In the pilot product the Quality Control flags either indicate "No QC checks" (flag = 0); "Individual" (flag = 1); "Follows proposed QC routine of data product" (flag = 2).

Table 2. Selected analysis recommendations for the BGC EOVs of the pilot product.

Parameter (discrete)	**Analysis Recommendations (selected examples only)**
Oxygen	Winkler titration Record draw temperature immediately after sampling Potassium iodate from OSIL or CSK
Total Dissolved Inorganic Carbon	Coulometry Follow Dickson et al., 2007 Use CRMs Report in µmol/kg
Total Alkalinity	Open cell potentiometric titration Follow Dickson et al., 2007 Use CRMs Report in µmol/kg
pH	Spectrophotometric Use a purified indicator dye Document scale, temperature, standards and dye-indicator
Nutrients	Autoanalyzer (low nutrient seawater as carrier solution) If stored silicate should be refrigerated others frozen (-20°C) Use (C)RMs Document standards and filters
Particulate Organic Matter	High Temperature Combustion Run total C (Elemental Analyzer) and PIC (Coulometrically) Assess POC by difference Ash hydrolysis for POP Reporting filtration volume (dry for 24 hours at 60°C)
Dissolved Organic Carbon	High Temperature Combustion Glass vials as containers Use combusted GF/F filters housed in polycarbonate in-line filters

3. Quality Control Guideline

The development of the community-agreed quality control techniques for discrete BGC time series data represents an essential part of the pilot product. The proposed routine follows a clearly outlined decision tree, and the applied checks make in particular use of comparisons with historical time-series data. The main goal of this routine is to support scientists in flagging of single samples as well as detecting systematic biases of particular station visits in time. I.e. the here proposed QC scheme, which combines 1st and 2nd QC (see blue panel below), only aims at indicating the (assumed) quality (accuracy and precision) of a single sample and/or station visit. We do not want to adjust for possible biases as the existing methods do not allow for such rigorous data interventions. The main result of the QC is to (re-)assess the flags of all samples. As already many different flagging schemes exist and since we do not want to introduce yet another flagging scheme, the TS pilot data product will apply the consolidated World Ocean Circulation Experiment (WOCE) water sample flagging scheme (Jiang et al., 2022), see Table 3.

Precision *vs.* Accuracy
The precision of data is a measure of the statistical variability of the data, i.e. the closer samples from repeated measurements (e.g. duplicates) are to each other, the more precise the dataset is. It is important to understand that precise data do not need to be accurate, i.e. precise data can deviate from a so called "true value". Eventually, it is the aim of 1st QC techniques to improve the precision of the dataset by identifying single bad outliers of a particular cruise/cast. The 2nd QC in turn mainly checks against historical data as a proxy for the true value. It, thus, aims at increasing the accuracy by identifying systematic biases.

We also want to stress that the developed QC methods by no means aim at replacing existing individual quality assurance (QA) routines nor the need for the application of (certified) reference materials. After all, following known Best-Practices during data acquisition in the field all the way to the analysis in the lab (e.g. using appropriate sensors, well-established methods, sampling duplicates etc.) as well as following strict QA guidelines (e.g. using appropriate storing facilities, making use of calibration curves etc.) and participating in inter- and intra-laboratory comparisons, are the strongest and most important fundaments for high quality data.

Table 3. Consolidated WOCE Flagging Scheme (following Jiang et al., 2022)

Flag	Description
0	Interpolated data
1	Not evaluated / quality unknown
2	Acceptable
3	Questionable
4	Known bad
6	Mean of replicates
9	Missing value

3.1. Decision Tree

Each time-series site has unique characteristics. The sites strongly depend on the location (depth and seasonal influence), funding opportunities (duration and frequency of visits) and scientific goals (parameters measured). All these differences have to be taken into account. Hence, a "One-fit-all" QC method for marine BGC time-series stations simply doesn't do justice to the great variability of the time-series data. To still provide a consistent routine for all time-series sites, the QC- and data working group has based the developed QC on a decision tree, see Figure 4. Using conditions which describe the characteristics of a time-series site, the tree eventually points to the best fitting methodology. Clearly, some of the suggested methods are more advanced and throughout than others, which is reflected in the color scheme of the end-nodes, i.e. suggested method.

In the "regular tree", there are in total seven different end nodes, i.e. QC methods, which can be applied to the data. These are (from least to most thorough, for details see the following sections):

1) Property-Property Plots (PPT)
2) Seasonal Sigma Test (So – Test)
3) Seasonal Outlier Test (SOT)
4) Semi Regular Outlier Test (Semi ROT)
5) Regular Outlier Test (ROT)
6) Semi Detrended Outlier Test (Semi DOT)
7) Detrended Outlier Test (DOT)

In addition, there is a "short-cut" to a very throughout check of the accuracy of an entire station visit. If the site has a known layer of no or very little long-term variability and seasonality, i.e. a "Reference Layer", then the so-called "Crossover Analysis" (CRS, Tanhua et al., 2010) can be performed. It is suggested that whenever possible this extra test should be carried out.

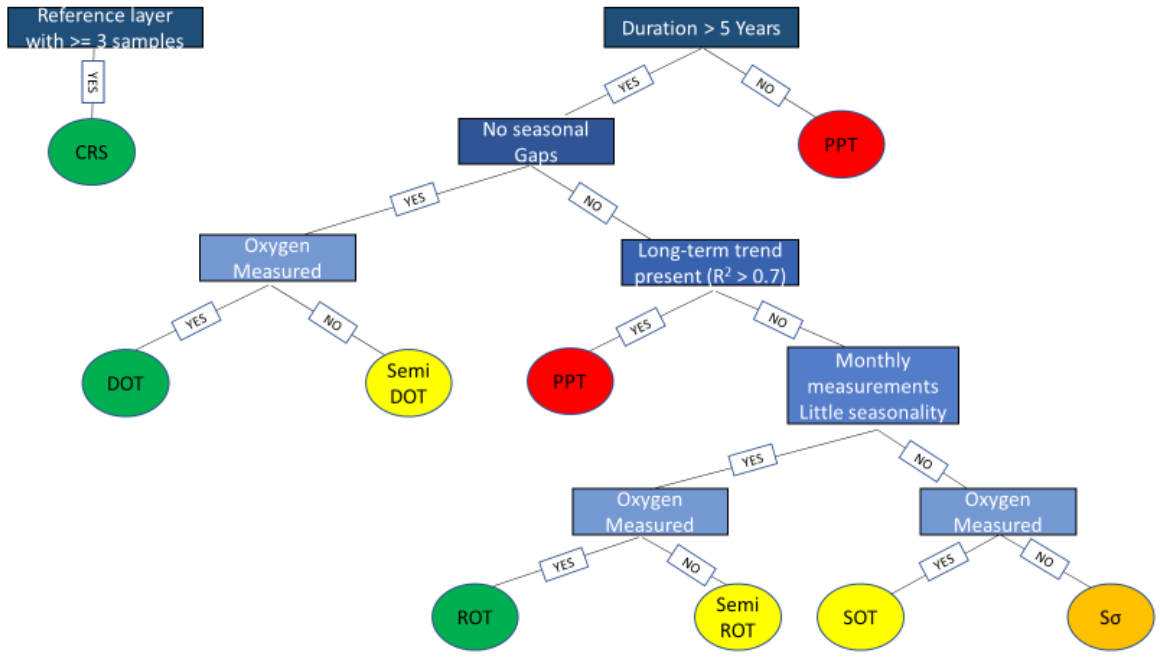

Figure 4. Time Series Niskin Bottle Quality Control Decision Tree; Colors: Green = BEST; Yellow = GOOD; Orange = OK; RED = Minimum; Abbreviations: CRS = Crossover Analysis; DOT = Detrended Outlier Test; SOT = Seasonal Outlier Test; So = Seasonal Sigma

3.2. Property-Property Plots

Time-series sites with a very short history (< 5 years) have very limited QC possibilities. The same holds for sites in areas which show signs of long-term variability (inter-annual or decadal) and infrequent measurements, i.e. with seasonal gaps. The former can be analyzed using different models, e.g. linear regression, in the form of the R^2 statistic. In both cases the only feasible QC option is to make use of property-property plots.

To support scientists, this rather visual inspection can be done using the AtlantOS QC tool (Velo et al., 2021), which in turn has been exclusively developed for the 1st QC of cruise section data. For its application the entire dataset must be on the same scale/unit and should be reported using the WOCE ontology. The data integration process of the TS data product adapts the data formats if needed automatically. The AtlantOS software in turn enables the user, i.e. the QC scientist, to graphically illustrate, analyze and manage the data by providing an interactive user-face, see Figure 5. Single profiles, i.e. measurements from a particular station visit in time, can be inspected using multiple property-property plots. By doing so and making use of known relations (e.g. Redfield ratio, Hoppema and Goeyens 1999) outliers against all other existing profiles can be detected more easily, see Figure 5. A rule of thumb is to report a suspicious sample to the PI, if it is an outlier in at least three different property-property plots. However, only very strong outliers should be flagged this way, as this method ignores any seasonality or long-term trend effects. In addition, the AtlantOS QC (Velo et al., 2021) software enhances the transparency of the QC process as evidence and version-control to the resultant flag changes is given automatically.

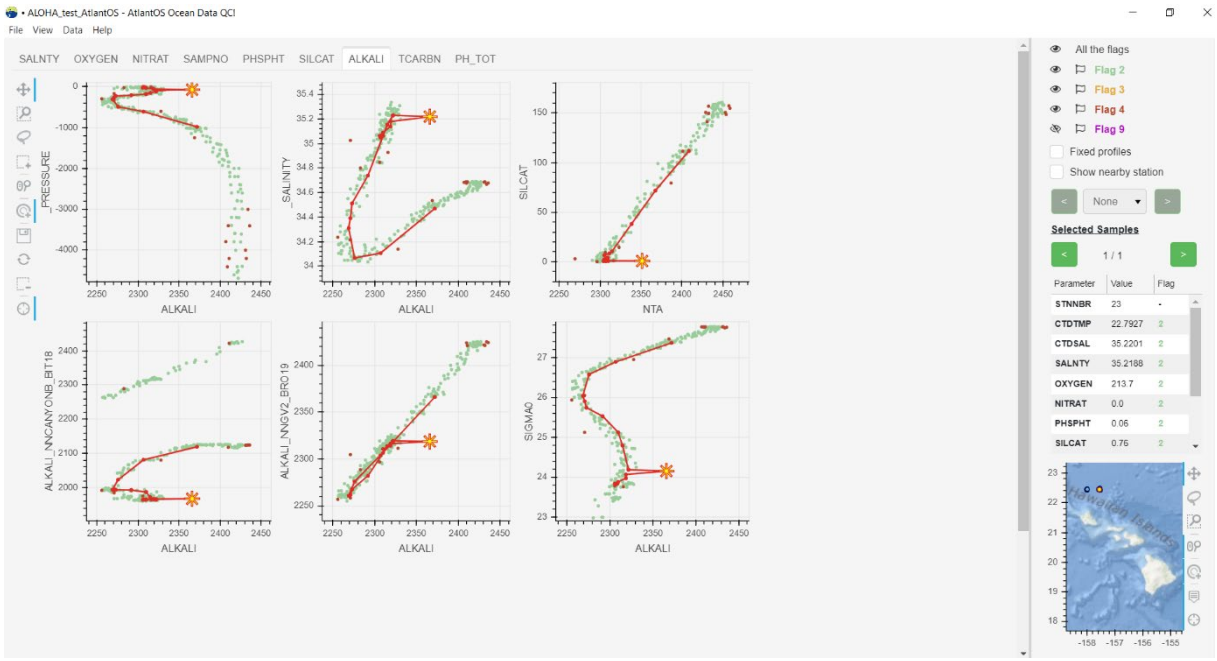

Figure 5. AtlantOS QC used for the time-series station HOT (see Figure 1); Tab of Alkalinity (ALKALI) is shown with 6 different property-property plots. The spark highlights an outlier of the profile shown in red.

3.3. Seasonal Sigma Test

The sigma test in general is the easiest statistically based method to identify outliers. Depending on the expected variability of the variable in question the scientist can choose between a 2σ or a 3σ test (z-score). The former flags samples outside of two standard deviations of the historical mean value, i.e. it flags samples smaller or larger than 95% of the data (symmetrically). The 3σ test is more significant and robust but might not catch many outliers as it only flags samples outside 99% of the mean. We only advertise to restrict the QC to sigma-tests in case of infrequent data with seasonal gaps and no or little long-term variability. Further, if the site has measured oxygen we suggest to perform the SOT instead. For the sigma test with irregular data we further propose to only compare seasonal data with each other, i.e. only winter samples with winter samples, spring samples with spring samples etc. Most effective is this method if analyzing layer by layer, usually these layers are easy to identify as most time-series sites have common depth on which samples are drawn from. Depending on the location however, it might be beneficial to analyze the data not on pressure levels but on density surfaces (sigma or gamma[11]). The "typical" density surfaces can be calculated using the typical sample depths (pressure) and the variable values should be interpolated to these levels using a quasi-Hermetian piecewise polynomial without extrapolation. Lastly, note that the sigma test works best for normally distributed data. If that is not the case (Kolmogorov-Smirnov goodness-of-fit hypothesis test), interquartile ranges around the median might be better suited to check for outliers.

3.4. Seasonal Outlier Test

As mentioned above if a time-series site meets the same condition as the ones that lead to the sigma test and if the site additionally measures oxygen, the Seasonal Outlier Test should be performed for nutrients and the inorganic carbonate system. This method also includes the Seasonal Sigma Test but further makes use of CANYON_B[12] (Bittig et al., 2018) comparisons. By doing so it increases the robustness of the statistically assigned flags. All other suggestions, such as to only compare seasonal data with each other and to perform the tests on a layer by layer bases still hold. The test needs to follow a strict order:

1. Check all variables using the Sσ-Test
2. Calculate CANYON_B values (nutrients and inorganic carbon) using oxygen, salinity and temperature
3. Calculate difference between measured value and calculated value for each layer
4. Normalize the difference to be centered around 0 (minimizing effects of seasonal biases)
5. Check for outliers using sigma test (z-score) if normally distributed otherwise use interquartile range
6. If both So-Test and the additional CANYON_B comparison indicate the same outlier, flag accordingly (assign WOCE flag = 3)
7. Local expert review

[11] Neutral density surface

[12] CANYON_B (CArbonate system And Nutrients concentration from hYdrological properties and Oxygen using a Neural-network version B) is a bayesian neural network that estimates nutrients and seawater CO_2 chemistry variables from latitude, longitude, pressure, temperature, salinity, and oxygen. The neural network was trained and validated using the GLODAPv2 data product. It mainly serves as an alternative to (spatial) climatological interpolation, and the resultant "dynamic climatology" shows a much better representation of smaller scales (40–120 days, 500–1,500 km) compared to in situ data.

3.5. (Semi-) Regular Outlier Test

If the data has highly frequent measurements, i.e. monthly, and additionally doesn't show a strong seasonal signal, the ROT method should be applied. This method additionally compares neighboring data samples for large "jumps" in time and thus consists of in total three independent tests:

1) Sigma-Test:
    a. 2σ (or 3σ) test using z-score: if time-series is normally distributed (Kolmogorov-Smirnov goodness-of-fit hypothesis test)
    b. 1.5 Interquartile range test: if time-series is not normally distributed (Kolmogorov-Smirnov goodness-of-fit hypothesis test)
2) CANYON_B normalized difference test (value - value_calculated):
    a. 2σ (or 3σ) test using z-score: if time-series is normally distributed (Kolmogorov-Smirnov goodness-of-fit hypothesis test)
    b. 1.5 Interquartile range test: if time-series is not normally distributed
3) Spike test:
   Checks for "distance", i.e. difference, between neighboring points in time. The distance is compared to a pre-set constant, i.e. meta-parameter. This meta-parameter has to be set individually for each time series site, level and parameter.

If at least two of the three tests indicate the same outlier, the sample is flagged. Again, all tests should be performed on a layer by layer bases on either pressure or density surfaces. However, the restriction of comparing data from the same season only is redundant. Figure 6 shows an example outcome of this routine for alkalinity measurements of the time-series station HOT, performed on a "typical" density level (gamma).

Subsequently, a further in-depth investigation must be performed to minimize the likelihood of flagging a "good" sample as questionable. To do so the sample in question should be qualitatively examined by evaluating it in respect to the entire profile(s) of that particular station visit, see Figure 7. If one of the following is true, the sample should not be flagged as questionable:

- Salinity, Temperature, Pressure (and/or Oxygen) are "off"
  → Density surface and/or CANYON_B calculation error
- Too large distance between the "typical" density levels
  → Interpolation error
- Eddy or similar feature present during station visit
  → Natural variability

Otherwise, the suspicious sample should be flagged accordingly (assign WOCE flag = 3) and be reviewed from a local expert. In the Semi ROT, the CANYON_B comparison cannot be performed as oxygen measurements are a mandatory condition for these. The rest of the method is identical.

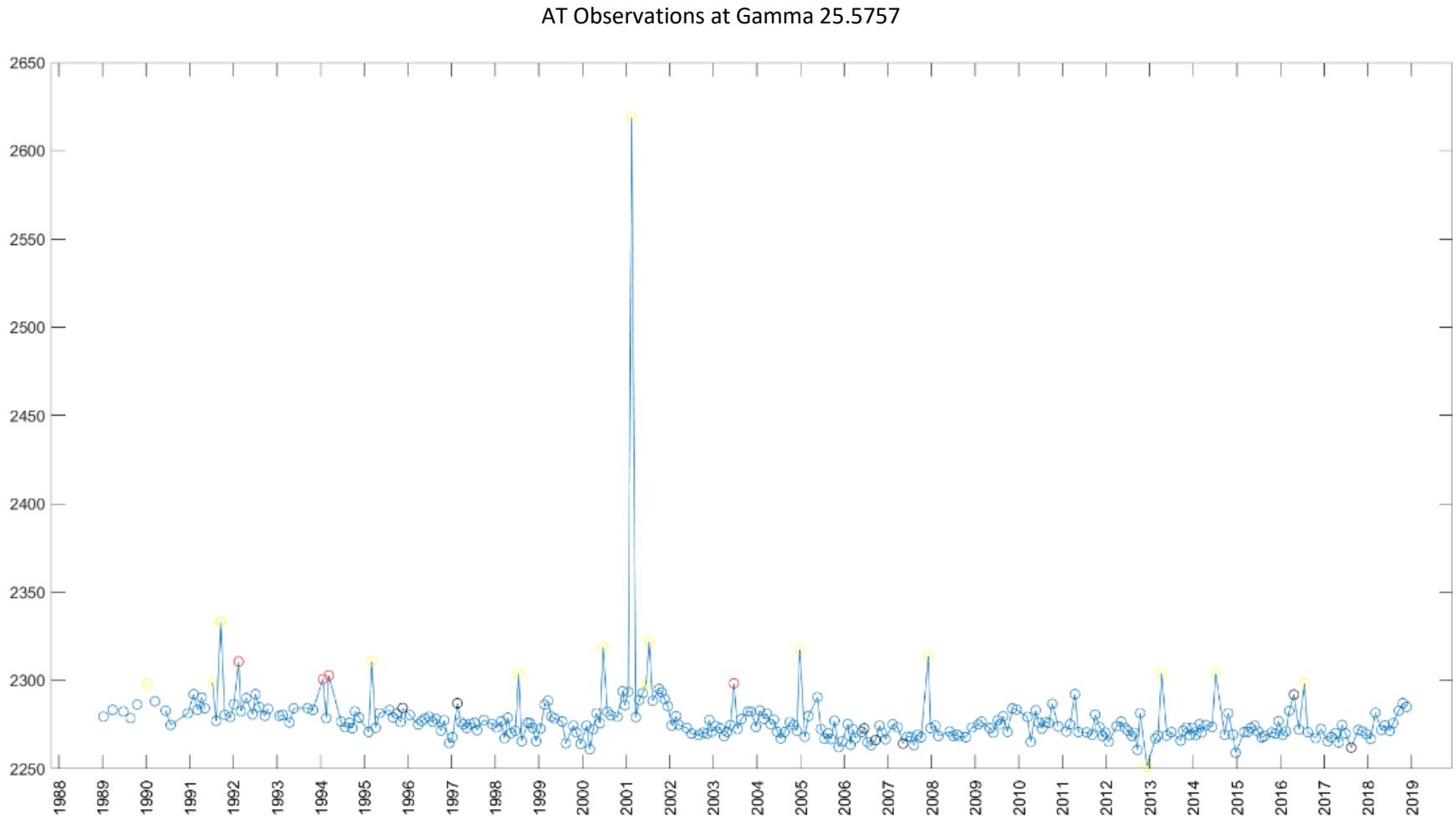

Figure 6. Station HOT gamma-interpolated ALKALI time series; Yellow dots indicate suspicious data already flagged by HOT; red dots indicate suspicious data not flagged by HOT; black dots indicate non-suspicious data flagged by HOT.

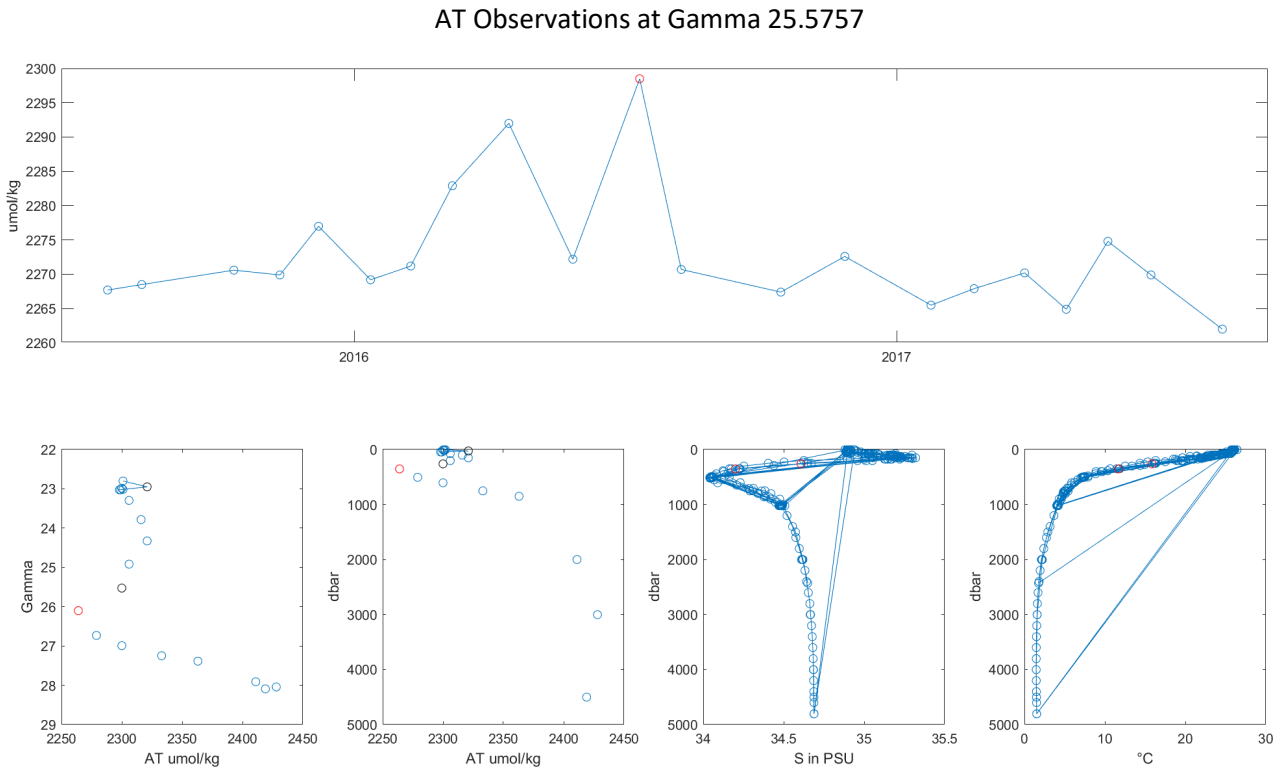

Figure 7. Final QC-scientist interface for suspicious ALKALI sample of HOT time-series. Top: Zoomed-In time-series at gamma-level 25.5757 kg/m3 (yellow dot in Figure 6); All bottom plots show all casts of particular HOT (285) cruise visit for (left to right): Gamma; Pressure; Salinity; Temperature. Red dots indicate samples used for gamma interpolation and black dots (overwriting red dots) indicate flagged data by HOT.

3.6. (Semi-) Detrended Outlier Test

Time series data with regular monthly measurements or at least seasonal measurements can be QC'd independent of long-term trends and/or variability present by detrending the dataset. Thus, the detrending not only enables comparisons of data from different seasons but also from different years. The detrending method proposed is a very simple technique to minimize detrending associated errors and requirements. A strict order must be followed:

1) Calculate seasonal mean for each year
2) Use the seasonal mean to calculate yearly means
3) Use yearly means to calculate total mean
4) Calculate yearly anomalies (difference between total- and yearly means)
5) Subtract yearly anomalies for each year

The (semi-) ROT method can then be applied to the detrended dataset in an identical manner.

However, we want to note that this method is extremely restrictive as most time-series have at least one seasonal gap during their measurement period and hence can almost be seen as an idealized case only.

3.7. Crossover Analysis

The crossover analysis (Tanhua et al., 2010) is heavily used in the well-established Global Ocean Data Analysis Project (GLODAP, Olsen et al., 2020, Lauvset et al., 2021) as the main 2nd QC method. To be applicable for the QC of time-series site, the site must have a very stable "reference layer", usually depth below 1500m, and the cruise should have measured at least three nearby (within 2°) different profiles with at least three samples within that reference layer. If that condition is met, the accuracy of the entire cruise can be assessed using this crossover routine. Eventually, this routine compares the profiles within the reference layer measured from the cruise in question against profiles from other (consistent) cruises nearby to detect systematic biases. Of course, the non-trivial problem with this method is to identify which of the cruises is biased. Anyway, as this method is already described in detail in existing literature we are not going in further detail here.

3.8. Evaluation of QC results

To evaluate the suggested "BEST" QC procedure, we compared the flagging results of the ROT method with the already QC'd and flagged data of the HOT station (for location see Figure 1; Dave, 2018). The original flagging of HOT samples is well established and often traces back to real-time observational error logging and/or instrumental errors, e.g. bottle leaking. This makes the HOT data a good data set for the evaluation of the developed QC workflow. It is important to understand that the here described QC procedure cannot detect all of the HOT flagged samples as it purely relies on statistical analysis. Clearly, methods applied by the HOT scientists go beyond QC analysis and outlier detection.

The results of the comparison between HOT flags and the QC flags for total alkalinity are shown in Figure 6 and Figure 7. In the former, yellow marked samples indicate samples, which have been flagged by HOT and by the QC procedure. The other marked samples indicate flagged samples by HOT (black) or by the QC (red) only. In the shown example for alkalinity at a gamma-level of 25.5757kg/m3 four samples are false positive (red), seven samples true negative (black) and 15 true positive (yellow). Given that some true negative samples cannot be detected by QC and that some false positive samples could be ruled out in further investigation, this example shows that the QC can support scientists in improving the precision and accuracy of the data. However, we are aware of this being one example only. By evaluating the QC at other levels and other parameters we observed especially that false positive errors rise when parameters are quality-controlled at very low concentrations, e.g. nitrate near the surface at HOT (Figure 8). Further, for locations with very little training data for CANYON_B (e.g. Southern Ocean, see GLODAP cruise map, Olsen et al., 2020, Lauvset et al., 2021) results must be treated with caution, due to the dependency of the QC on reliable CANYNON-B data.

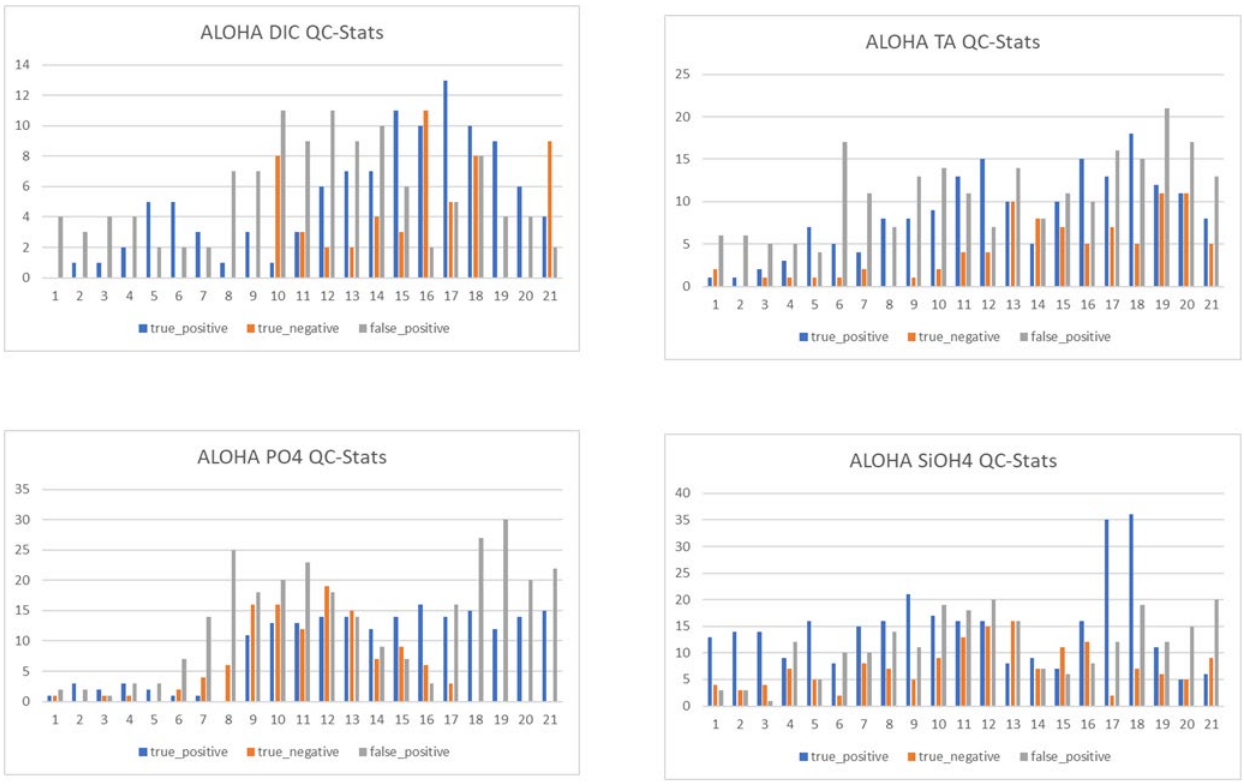

Figure 8. Evaluation of ROT QC using gamma levels for Dissolved Inorganic Carbon (DIC), Total Alkalinity (TA), Phosphate (PO4) and Silicate (SiOH4) of the HOT time series data. False negative, i.e. rightfully non-flagged samples, not shown as these would blow up the scale

Conclusions

Following the call for improved BGC time-series data integration (Benway et al., 2019) and embracing EuroSea's vision of "[…] a user-focused, truly interdisciplinary, and responsive European ocean observing and forecasting system […]", the presented framework represents an important step forward in the overarching goal to obtain interoperable and comparable BGC ship-based time-series data. The framework sets a clear focus on high consistency and is developed with the entire spectrum of BGC time-series stations in mind. The focus on consistency is also reflected in the flexible QC scheme. This scheme in turn enables consistent and more traceable as well as more statistically profound QC decision making, eventually resulting in more comparable and higher quality BGC time-series data. The framework further fosters the generation and common usage of Best-Practices, especially in terms of data handling. With the next step, the upcoming release of the BGC time-series synthesis pilot product, we will implement this framework and demonstrate the benefits of consistent BGC time-series data to the ocean observing community. On a larger scale, we hope that through this framework and the corresponding data synthesis product we can increase the impact of the BGC ship-based time-series stations and help to establish a ship-based BGC time-series network in GOOS.

References

Álvarez M., Fajar N.M., Carter B.R., Guallart E.F., Pérez F.F., Woosley R.J. and Murata A. (2015). Global ocean spectrophotometric pH assessment: consistent inconsistencies. DOI:10.1021/acs.est.9b06932

Aoyama M., Woodward E.M., Bakker K. Becker S., Björkman K., Daniel A., Mahaffey C., Murata A., Naik H., Tanhua T., Rho T., Roman R. and Sloyan B. (2016). Comparability of oceanic nutrient data. CLIVAR Open Science Conference.

Aoyama M. and Gallia R. (2018). IOCCP-JAMSTEC 2018 Inter-laboratory Calibration Exercise of a Certified Reference Material for Nutrients in Seawater. DOI:10.25607/OBP-429

Becker S., Aoyama M., Woodward E. M. S., Bakker K., Coverly S., Mahaffey C. and Tanhua T. (2020). GO-SHIP Repeat Hydrography Nutrient Manual: The Precise and Accurate Determination of Dissolved Inorganic Nutrients in Seawater, Using Continuous Flow Analysis Methods. DOI:10.3389/fmars.2020.581790

Benway H.M., Lorenzoni L., White A. E., Fiedler Björn, Levine Naomi M., Nicholson David P., DeGrandpre Michael D., Sosik Heidi M., Church Matthew J., O'Brien Todd D., Leinen Margaret, Weller Robert A., Karl David M., Henson Stephanie A. and Letelier Ricardo M. (2019). Ocean Time Series Observations of Changing Marine Ecosystems: An Era of Integration, Synthesis, and Societal Applications. DOI:10.3389/fmars.2019.00393

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

Bockmon E.E. and Dickson A.G. (2016). An inter-laboratory comparison assessing the quality of seawater carbon dioxide measurements. DOI:10.1016/j.marchem.2015.02.002

Dave K. (2018). Niskin bottle water samples and CTD measurements from the Hawaii Ocean Time-Series cruises from 1988-2016 (HOT project). Biological and Chemical Oceanography Data Management Office (BCO-DMO). DOI:10.1575/1912/bco-dmo.3773.1

Dickson A.G., Sabine C.L. and Christian J.R. (2007). Guide to Best Practices for Ocean CO2 Measurements, vol. 3. North Pacific Marine Science Organization, PICES Special Publication, Sidney, B.C., Canada. DOI:10.25607/OBP-1342

Hoppema M. and Goeyens L. (1999): Redfield behavior of carbon, nitrogen, and phosphorus depletions in Antarctic surface water. Limnol. Oceanogr., 44(1), 1999, 220–224.

Jiang L., Pierrot D., Wanninkhof R., Feely R.D., Tilbrook B., Alin S., Leticia B., Byrne R.H., Carter B.R., Dickson A.G., Gattuso J.P., Greeley D., Hoppema M., Humphreys M.P., Karstensen J., Lange N., Lauvset S.K., Lewis E.R., Olsen A., Pérez F.F., Sabine C., Sharp J.D., Tanhua T., Trull T.W., Velo A., Allegra A.J., Barker P., Burger E., Cai W.J., Chen C.T.A., Cross J., Garcia H., Hernandez-Ayon J.M., Hu X., Kozyr A., Langdon C., Lee K., Salisbury J., Wang Z.A. and Xue L. (2022). Best Practice Data Standards for Discrete Chemical Oceanographic Observations. Frontiers in Marine Science 8. DOI:10.3389/fmars.2021.705638

Lauvset S.K., Lange N., Tanhua T., Bittig H.C., Olsen A., Kozyr A., Álvarez M., Becker S., Brown P.J., Carter B.R., Cotrim Da Cunha L., Feely R.A., Van Heuven S., Hoppema M., Ishii M., Jeansson E., Jutterström S., Jones S.D., Karlsen M.K., Lo Monaco C., Michaelis P., Murata A., Pérez F.F., Pfeil B., Schirnick C., Steinfeldt R., Suzuki T., Tilbrook B., Velo A., Wanninkhof R., Woosley R.J. and Key R.M. (2021). An updated version of the global interior ocean biogeochemical data product, GLODAPv2.2021. Earth Syst. Sci. Data 13, 5565-558. DOI:10.5194/essd-13-5565-2021

Lorenzoni L. and Benway H.M. (2013). Report of Global intercomparability in a changing ocean: An international time-series methods workshop (November28-30, 2012, Ocean Carbon and Biogeochemistry (OCB) Program and International Ocean Carbon Coordination Project (IOCCP))

Olsen A., Lange N., Key R. M., Tanhua T., Bittig H. C., Kozyr A., Àlvarez M., Azetsu-Scott K., Becker S., Brown P. J., Carter B. R., Cotrim da Cunha L., Feely R. A., van Heuven S., Hoppema M., Ishii M., Jeansson E., Jutterström S., Landa C. S., Lauvset S. K., Michaelis P., Murata A., Pérez F. F., Pfeil B., Schirnick C., Steinfeldt R., Suzuki T., Tilbrook B., Velo A., Wanninkhof R. and Woosley R. J. (2020). GLODAPv2.2020 – the second update of GLODAPv2. DOI:10.5194/essd-2020-165

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. DOI:10.5194/essd-2-35-2010

Velo A., Cacabelos J., Pérez F.F., Tanhua T. and Lange N. (2021). AtlantOS Ocean Data QC: Software packages and best practice manuals and knowledge transfer for sustained quality control of hydrographic sections. Zenodo. DOI:10.5281/zenodo.2603121