

1 R1 - Hernan Garcia Review

Seth and co-authors describe the BCG-Argo+ dataset and code used. The addition of the BCG-Argo+ dataset is welcome for Argo and BGC delayed-mode data users. This paper describes their effort to provide the ocean community with a dataset subjected to additional primary level QC using an Argo delayed-mode data set downloaded January 24, 2025. The code helps ensure independent repeatability of the QC.

Thank you for your comments and for taking the time to review. We have worked to address all of your comments and concerns, with detailed responses provided below after each comment.

I am not understanding in what context the authors use the term "secondary quality control" as stated in the title and in the text. The identification and flagging of large outliers and vertical gradients is what is generally part of a primary level QC. A secondary QC approach would generally aim to goals such as reducing data uncertainty, improving internal data quality consistency between newer and old instruments and methods, reducing known data biases, new data re-calibrations or data adjustments (e.g., precision, target accuracy, or some other metric). If the authors mean that this is a second QC iteration following the delayed mode level, then this might cause confusion with each future BCG-Argo+ QC iteration. To help minimize confusion, please clearly define what the authors mean by primary and secondary level QC.

Your point makes it clear that we did not do a good job of defining what we mean by secondary quality control. We have now explicitly defined what we mean and why we use that term. To us, this work does not equal primary quality control because it is being done after the data has passed through the data centers and has been served to the public. We will label different versions w/ a version number. For this version we will call it BGC-Argo+ 1.0. Future improvements on these float data will be designated by changing the version decimal, while future changes to time period considered / the number of floats included will change the version integer. We have changed the title and dataset description throughout to reflect this.

It would be beneficial if the authors include a version number or label if they consider potential future QC iteration of the BCG-Argo+ (e.g., in the metadata, BCG-Argo+ ver number and/or date).

See comment above.

I suggest that the methods section should include or define metrics for what the authors consider large outliers, hooks, spikes, and large vertical gradients; perhaps in a table form for simplicity. I also would encourage the authors to point to the python code in the table label and/or in the text.

We have worked to clarify how we define outliers in the methods section. All criteria for the automated outliers are described and we have added a table to the Appendix that states which steps are included in which Python files. For the manual detection we strive to be as clear about decision making as possible. However, we acknowledge that the manual detection can be inherently subjective. If we could write an exact definition for the manual outliers then we would have been able to write an algorithm to do that detection manually. That is one eventual outcome of this work, that all manual outlier detection is eventually replaced by automated detection. Unfortunately, we are not there yet.

Below I provide some additional comments and suggestions for the authors consideration that I hope prove useful. Please use them as you see fit. The numbers before each comment refer to line numbers in the manuscript.

10 Simply as a historical note, I note that there were open ocean dissolved O₂ measurements collected even before the Winkler method was introduced (Winkler, 1888). For example, about 10 years earlier, O₂ data were collected in 1878 at the surface and at 157 m depth during the 1876-1878 Norwegian North-Atlantic Expedition (Tornøe H. The Norwegian North-Atlantic Expedition, 1876-1878. Chemistry. Christiania: Grrendal and Son bogtrykkeri, 1880).

Thank you, added

12 "Nisken"? should be Niskin. I'd probably also include the use of the Nansen bottles (e.g., pre-1970s).

Changed, thanks.

44 Add a reference to SOCAT here?

Added

41-47. I find the meaning of this paragraph confusing. In what way the authors use the GLODAP and SOCAT QC protocols in this paper? When the authors state "In this work we follow the lead of two ship-based quality control and data-inter-comparability programs to ...". Consider merging here the paragraphs starting in line 55 and/or lines 60-67.

We have changed the first sentence of the paragraph to better explain why we introduced these two datasets here and to no longer make it seem like we exactly followed protocols from GLODAP or SOCAT. The new paragraph starts (line 40):

"Examples of this type of work can be found in the community-led efforts to provide a second layer of data assessment and quality control of shipboard datasets, which have greatly enhanced their usability. The Global Ocean Data Analysis Project..."

48-49. It seems to me that the measurement quality of O₂, pH, Nitrate shipboard and sensor chemical observations is also quite different. Modern shipboard data are of higher quality (e.g., uncertainty, target accuracy, precision) than those from current chemical sensors mounted on floats. This is an important distinction to make in the text because it gives the reader the erroneous impression that their quality are similar and invariant over time.

Good point and not something we meant to gloss over. We have restated this paragraph and added more about sensor stability and accuracy (starting at line 48):

"BGC-Argo data are fundamentally different than shipboard bottle measurements in terms of the quantity of data collected, the time over which measurements are returned, and the quality and stability of the observations made. When an individual researcher on a research cruise has responsibility over a limited number of measurements and runs all samples themselves, there is a high initial level of oversight on the generation of data. BGC-Argo data undergo evaluation by individual researchers and automated routines at the Data Assembly Centers (DACs) that process Argo data, bringing a depth of knowledge on float and sensor types and broad consistency, but data are collected data over years and individual floats may not get the same degree of consistent attention as shipboard measurements. Additionally, the biogeochemical sensors on Argo floats have known issues with accuracy and stability that have long been the source of on-going engineering and calibration efforts (Johnson and Coletti, 2002; Johnson et al., 2017; Maurer et al., 2021; D'Asaro and McNeil, 2013; Martz et al., 2010; Johnson et al., 2016; Demas et al., 1999; Bushinsky et al., 2016, 2025)."

55-57: It seems to me that the identification and removal of outliers is part of any primary level QC approach. Why would this be considered a secondary level QC approach? Please define/explain what you mean by primary and secondary QC in the text (e.g., methods).

We now explicitly define what we mean by primary and secondary quality control at the end of the introduction where we describe what we do in the manuscript (starting at line 58):

"In this manuscript we describe a secondary quality control approach for BGC-Argo data designed to remove and identify outliers from the global BGC-Argo dataset, combining new routine data quality checks and a systematic outlier detection approach specifically adapted to Argo data. We refer to all of the work done by the individual float PIs and the DACs toward quality controlling and correcting float data as the "primary" quality control step. Everything that we do to the dataset happens after it has been published in an Argo snapshot; we therefore refer to our work as a "secondary" quality control step. Our focus is on the oxygen, nitrate, and pH sensor data collected by BGC Argo floats. By analyzing all available data we are able to identify issues common..."

60: What do you mean by "improving" data quality? How is this quantified? I suggest rewriting this sentence as, for example, (1) improve the internal data quality consistency by minimizing the number of large vertical gradients and outliers. Then define in the text what are "large" vertical gradient and outliers.

We have tried to clarify what we mean by "improving" data quality. Essentially we are acknowledging that we cannot have a perfect dataset, but we can continue to make the data better. This is, and will be, an

on-going process and we view this work as a step in that direction.

71 Please define "Sprof" and "meta" and/or add a reference where these terms are defined.

Definition added along with reference.

76 For consistency, if the chemical measurements (e.g., O₂, nitrate) are reported using units on a per mass basis (e.g., umol/kg; Molal) then the recommended term to use is "content" (e.g., Nitrate content). On the other hand, reporting such data in units on a per volume basis (e.g., umol/l; Molar), then the term to use is "concentration" (e.g., Nitrate concentration).

While we appreciate the distinction that you are making, we would like to point out that the use of "concentration" for units per mass (umol/kg) is common throughout the literature and chemical oceanography text books (see for example Emerson and Hamme, Chemical Oceanography: Element Fluxes in the Sea. Cambridge University Press. 2022). Furthermore, content often refers to the volume or mass integrated amount of a given tracer and therefore, in our opinion, would add more confusion if it were used in place of concentration.

90-97. These steps are basic primary QC (see comment lines 55-57).

See our note above and the new text at the end of the Data section.

96 Please change "variables" to "parameters" in the text to differentiate between measured and derived quantities.

In cases where we wrote "variables" but were referring to derived quantities we have replaced "variables" with "parameters".

99-107. Please cite or refer to a manual or a reference that describes the Argo QC flags and metadata.

Added a reference to Bittig et al. 2022:

Bittig, H., Wong, A., & Plant, J. (2022). BGC-Argo synthetic profile file processing and format on Coriolis GDAC (Version 1.3). Ifremer. <https://doi.org/10.13155/55637>

102 Please clarify here that the primary Argo delayed-mode data used in this effort already has received QC by the Argo data management team. I would also note that the delayed-mode data may have been adjusted as part of a secondary QC step by the Argo data management teams.

Clarified. See new text above at the end of the Data section.

109. Suggest changing "highest quality data is present the new dataset" to "data with higher level of QC in the new dataset" or similar wording. What do you mean by data of "highest quality"? How do you know that the BGC-Argo+ data are of the highest quality possible?

We have re-written this sentence to better reflect our intended meaning. The start of this paragraph now reads (line 117):

"For oxygen, nitrate, and pH data we only retained the Delayed Mode data, setting profiles in the other modes to NaN. Since one of the goals of this dataset is to make it easier for non-experts to work with this data, we decided to limit the dataset to data that had gone through the full DAC QC process."

110. "salvaged". What does this term mean?

We removed this sentence as we agree that it was not clear and unnecessary.

130-131 "sharply deviated from clear mixed layer mean values" Please define what you mean by a "sharp deviation". I did not understand what the authors mean by "clear mean values"; suggest using "mixed layer mean values" instead?

Re-written with a more clear explanation (line 135+):

"As we examined the BGC-Argo files for outliers, we identified several categories of common errors for

which we developed automated detection algorithms. Our first automated step was for any above surface and near-surface measurements to be removed. In-air measurements are used for in-situ calibration of oxygen sensors and should not be present in the main float profiles. However, a significant number of floats contained data at the top of profiles that sharply deviated (differences greater than $1 \mu\text{mol kg}^{-1} [\text{O}_2]$) from mixed layer (ML) mean values (Fig. A3), often with associated pressures that were ≤ 0 db (example shown in Figure 4). These anomalous measurements may be due to wave action that brings the top of the float up above the surface briefly. This problem has been noted in the Argo O_2 quality control manual for PROVOR/ARVOR floats (one of the three main types of Argo floats) due to the optode being suspected of measuring a mix of air/water (Thierry et al., 2025b). The quality control manual recommends that data with a pressure ≤ 1 db for bin-averaged data and ≤ 0.5 db for spot sampled data be flagged as "probably bad data" (QC flag = 3). Based on our experience from the first several hundred floats that we examined, we determined that these outliers were not all removed and were still common in some floats, necessitating an automated approach to handle these outliers. A 2 db threshold was set, with any data at a lower pressure set to NaN. "

155 If the authors set certain observations to "NaN" or remove values, then there should be an independent way to replicate your QC data analysis from the same delayed-mode version used. Why not just flag the data following the Argo QC flag scheme? As the authors indicate in lines 291-292 "maintain the original data format to make it easier to compare between different versions" Why not do the same here? See comments for line 218.

We have found that one of the barriers to using float data for new users are the number of pre-processing steps needed before the data can be confidently used. We tried to create a dataset that was usable "as is", while retaining the ability for existing users to only change a few variable names in order to use these data. Despite that goal, we did consider just flagging the data using the Argo QC scheme. However, one complication is that if we were to simply update the existing QC flags, it would be unclear whether a flag was put in place by the DAC or during our processing. If we created a new QC flag, then someone would have to apply two separate QC flags before using the data, making it harder, not easier to use. For these reasons we decided that flags of "bad" or "probably bad" should be set to NaN, along with all outliers identified here. The original, unchanged "ADJUSTED" fields still exist, along with the QC flags, so users can undo individual processing steps if they so choose.

I note that the O_2 data "hooks" in Figure 5 for profiles 48 and 49 near the bottom are approximately less than $0.5 \mu\text{mol/kg}$. Is the precision of the O_2 sensors capable of consistently resolving or measure such O_2 content differences over sensor deployment time? Please check and provide additional context. I also note that profiles 47, 49, and 49 also show what appears to be constant or stuck values over some depth ranges (e.g., near bottom values in profile 47). Are these a product of slow sensor response? In comparison, I note the temperature profile show a vertical gradient over the same depths/pressure.

True, these are small differences shown in profiles 48 and 49. However, small differences in a consistent direction could falsely create an implied change in oxygen when comparing ship to float data, especially since most of these hooks occur at similar depths. We decided not to focus on the magnitude of the hook, but whether it was present and exceeded the variability in the profile above. For the temperature gradient in those pressures, it is fairly constantly getting cooler with depth, whereas the oxygen hook shows the opposite direction of change than the rest of the profile over those depths. We do not think that represents a real change in oxygen. We are not certain of the cause of the oxygen hooks and it may differ between different sensor types / float configurations. For now we are just trying to identify and remove them.

172. Instead of "at each point", suggest using "at each depth or pressure level in the profile".

Changed.

188 Please spell acronyms the first time used (e.g., DIC, TALK, pCO_2) or cite a reference. Some readers might not necessarily be familiar with the nomenclature.

Definitions have now been added where they were missing, thanks.

218. See comments for line 155.

See comments above.

227 It should be "hook" not "book". Perhaps run a spell check on the rest of text.

Fixed, thanks for catching that.

247-248. "goo covered". Jargon. Please cite a reference where this term is described.

We have changed the description of this form of outlier to a slow ML response.

Table 3. Change "Variable" to "Parameter" for consistency with the text and common community use for derived parameters and measured variables.

Changed.

265-266. Please spell out IF, IN, AO, HZ, and BO and/or point to a reference.

Definitions of acronyms are now in the figure caption, including institution/country for each data center.

Figure A2. Please label axes and their units. Consider adding a-d to the panels and refer to them in the figure caption as needed.

Axes are now labeled, have units, and panels are labeled by letter, thanks.

Figure A4: PSAL is not a unit. Salinity is unit less in the practical salinity scale. Suggest changing $+3.466\text{e}1$ to $+34.66$ (or is it 34.660 for consistency with the salinity significant figures for differences in the graph?).

Changed. The Argo netcdf file has "psu" listed as the unit for the salinity sensor. We decided to follow that, rather than change the unit to "unitless" in order to maintain consistency with the DACs.