Novel automated quality control procedures for BGC data developed by the Copernicus Marine Service In-Situ TAC Team

Vidar S. Lien, Institute or Marine Research (Norway), vidar.lien@hi.no
Jan Even Øie Nilsen, Institute of Marine Research (Norway), jan.even.oeie.nilsen@hi.no
Håvard Vindenes, Institute of Marine Research (Norway), haavard.vindenes@hi.no
Leonidas Perivoliotis, HCMR (Greece), lperiv@hcmr.gr
Seppo Kaitala, SYKE (Finland), seppo.kaitala@ymparisto.fi
Virginie Racapé, Ifremer (France), virginie.racape@ifremer.fr

Abstract

The Copernicus Marine Service is a "one-stop-shop" providing freely available operational data on the state of the marine environment for use by marine managers, advisors, and scientists, as well as intermediate and end users in marine businesses and operations. The Copernicus Marine service offers operationally updated and state-of-the-art products that are well documented and transparent. The European Commission's long-term commitment to the Copernicus program offers long-term visibility and stability of the Copernicus Marine products. Furthermore, Copernicus Marine offers a dedicated service desk, in addition to training sessions and workshops.

Here, we present the in situ biogeochemical (BGC) data products distributed by the Copernicus Marine Service since 2018. It offers available data of chlorophyll-*a*, oxygen, and nutrients collected across the globe. These products integrate observation aggregated from the Regional EuroGOOS consortium and Black Sea GOOS as well as from SeaDataNet2 National Data Centers (NODCs), JCOMM global systems, including bio-argo, and the Global telecommunication system (GTS) used by the Met Offices.

Through the evolution of its delayed-mode, re-processed offering of BGC data, the Copernicus Marine In-Situ Thematic Assembly Centre (INS TAC) has developed novel automated quality control procedures applied for BGC data to identify questionable data for visual inspection, thereby greatly enhancing the team's delayed-mode quality control capabilities. Moreover, there is a potential for utilizing these procedures to improve also the near-real time quality control. The in-situ re-processed BGC product is updated two times per year and made freely available through the Copernicus Marine website. The product is delivered on NetCDF4 format compliant with the CF1.7 standard.

Introduction

The Copernicus Marine Service is the European Commission's infrastructure for providing marine Earth Observation data and products to research, management, and businesses, among other. As a part of the Copernicus Marine Service, the INS TAC provides in-situ ocean data through two main channels: near-real time (NRT) products where observation data are provided within 24 hours post measurement, and re-processed (REP) products that consist of delayed-mode quality controlled data updated twice a year. Here, we present novel, automated quality control procedures for delayed-mode quality control of BGC data implemented for the INS TAC delayed-mode BGC REP product.

Data and products

In the production of the BGC REP product, we re-process available BGC data from 1993 and up to present date and divide the data into three main categories: chlorophyll-*a*, oxygen, and nutrients. Currently, the nutrients dataset consists of measurements of nitrate, silicate, and phosphate. The measurements include both discrete samples from bottle data, as well as sensor data. Moreover, measurements come from a wide variety of observation platforms, including CTD, bio-argo, gliders and ferryboxes, among other. Furthermore, the data originates from a large host of data providers.

Quality control procedures

Most of the data have been subject to delayed-mode quality control by the different data providers. However, the quality control is inconsistent between providers and the procedures applied may be unknown to the INS TAC partners that channel the data to Copernicus Marine. The INS TAC quality control procedures need to take these matters into account and make sure that data of high or low quality are flagged accordingly. Therefore, we apply consistent, well-documented quality control procedures for all data going into the BGC REP product, including visual inspection of data of questionable quality. To reduce the need for visual inspection and thereby increase the efficiency of the quality control process, we have developed automated quality control based on statistical testing. Moreover, where possible, we have in addition applied tests that check the data against physical constraints. Data that pass the tests are flagged as good data, while data that fail any of the tests are visually inspected before being flagged as good or bad data.

For chlorophyll-*a* a purely statistical approach has been chosen. The world ocean is divided into coastal and pelagic regions, and each region is divided into upper and deeper ocean. Then, the 99th percentile for chl-*a* concentration is computed for each region individually and used as upper boundary for accepting data, i.e., any data point outside the 99th percentile of any given region is visually checked before being flagged as either '1 – good' or '4 – bad'. For oxygen, on the other hand, the concentrations are compared with the calculated maximum saturation of oxygen in seawater (allowing for over-saturation in the upper ocean), in addition to a regional range test based on a statistical approach. Any data points that are above 100% saturation (but allowing for over-saturation in the surface layer) are rejected and flagged as '4 – bad' data. Moreover, any data points that falls outside the regional fixed ranges are visually inspected before having its final quality flag decided. For nutrients, a profile test is applied in addition to the statistical range test. The profile test identifies all profiles where the concentration in the surface layer exceed the concentration at intermediate depths (Figure 1), and all profiles that are marked are visually inspected before the final quality flag is applied. Note that the profile test is only applied in pelagic (non-coastal) regions to avoid potential impact of runoff from land, yet, as Figure 1 shows, impact from coastal waters may still occur. We find that for some variables in some regions, e.g., nutrients in the Mediterranean Sea, a substantial part of the data are flagged during the regional range test and to a lesser extent during the profile test. Moreover, generally more measurements of nitrate and silicate obtained by CTD are flagged compared to measurements from bottle data. This result was found in both the range test and the profile test.

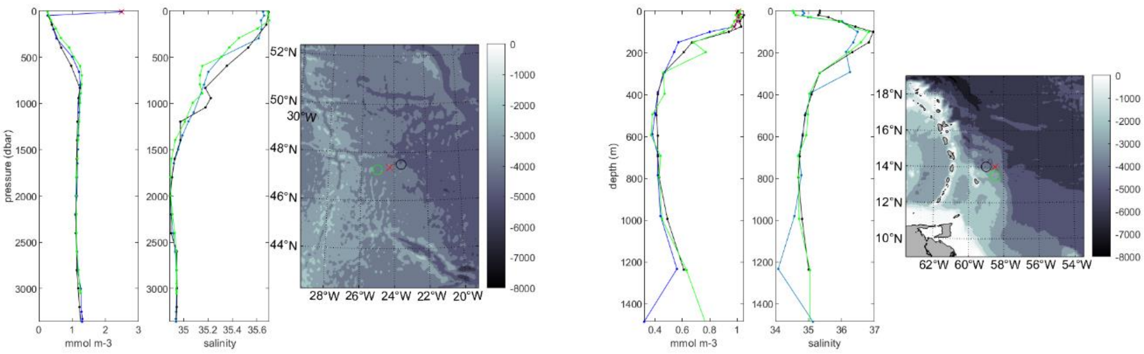

Figure 1: Examples of the profile test for nutrients (phosphate). Left: surface value marked (far left) and flagged as bad data (surface value larger than subsurface values; salinity profile to the right). Right: surface values are marked (left profile) but flagged as good data after visual inspection (influence of coastal water; salinity profile to the right).

Summary & Conclusion

We present automated quality control procedures for identifying BGC data that require visual inspection for delayed-mode quality control. The proposed automation procedures greatly reduce the required resources for delayed-mode quality control, and also provide opportunity for further improvements of near-real time quality control of BGC data.