

EGU2020-17071

<https://doi.org/10.5194/egusphere-egu2020-17071>

EGU General Assembly 2020

© Author(s) 2022. This work is distributed under the Creative Commons Attribution 4.0 License.



Exploiting SeaDataCloud Temperature and Salinity time series data collections and comparing with Copernicus - a novel approach with SOURCE tool

Paolo Oliveri¹, Simona Simoncelli¹, Pierluigi Di Pietro¹, and Sara Durante²

¹Istituto Nazionale di Geofisica e Vulcanologia, Ambiente, Bologna, Italy

²Istituto di Scienze Marine (ISMAR) - CNR, Sede Secondaria Napoli, Italy

One of the main challenges for the present and future in ocean observations is to find best practices for data management: infrastructures like Copernicus and SeaDataCloud already take responsibility for assembly, archive, update and publish data. Here we present the strengths and weaknesses in a SeaDataCloud Temperature and Salinity time series data collections, in particular a tool able to recognize the different devices and platforms and to merge them with processed Copernicus platforms.

While Copernicus has the main target to quickly acquire and publish data, SeaDataNet aims to publish data with the best quality available. This two data repository should be considered together, since the originator can ingest the data in both the infrastructures or only in one, or partially in both. This results sometimes in data partially available in Copernicus or SeaDataCloud, with great impact for the researcher who wants to access as much data as possible. The data reprocessing should not be loaded on researchers' shoulders, since only skilled users in all data management plan know how merge the data.

The SeaDataCloud time series data collections is a Global Ocean soon-to-be-published dataset that will represent a reference for ocean researchers, released in binary, user friendly Ocean Data View format. The database management plan was originally for profiles, but had been adapted for time series, resolving several issues like the uniqueness of the identifiers (ID).

Here we present an extension of the SOURCE (Sea Observations Utility for Reprocessing, Calibration and Evaluation) Python package, able to enhance the data quality with redundant sophisticated methods and simplify their usage.

SOURCE increases quality control (Q/C) performances on observations using statistical quality check procedures that follows the ocean best practices guidelines, exploiting the following issues:

- Find and aggregate all broken time series using likeness in ID parameter strings;
- Find and organize in a dictionary all different metadata variables;
- Correct time series time to match simpler measure units;
- Filter devices that are outside of a selected horizontal rectangle;

- Give some information on original Q/C scheme by SeaDataCloud infrastructure;
- Give information tables on platforms and on the merged ID string duplicates together with an errors log file (missing time, depth, data, wrong Q/C variables, etc.).

In particular, the duplicates table and the log file may be helpful to SeaDataCloud partners in order to update the data collection and make it finally available for the users.

The reconstructed SeaDataCloud time series data, divided by parameter and stored in a more flexible dataset, give the possibility to ingest it in the main part of the software, allowing to compare it with Copernicus time series, find the same platform using horizontal and vertical surroundings (without looking to ID) find and cleanup duplicated data, merge the two databases to extend the data coverage.

This allow researchers to have the most wide and the best quality possible data for the final users release and to use these data to calibrate and validate models, in order to reach an idea of a whole area sea conditions.