A Framework to Quality Control Oceanographic Data

Guilherme P. Castelao[1]

1 Scripps Institution of Oceanography

Summary

The ocean is an intrinsically challenging environment to collect data, which makes spurious measurements inevitable. Thus, the quality of oceanographic datasets is highly dependent on the ability to identify and remove bad samples. Quality control (QC) of oceanographic data has mostly relied on manual QC by experts, which, despite resulting in the best data quality, is not scalable and becomes impractical to handle large datasets or real-time data streams. To address this issue, automatic QC procedures have been proposed and widely used for decades (e.g., IOC/IODE, 1993; DATA–MEQ working group, 2010; *GTSPP real-time quality control manual*, 2010; Morello et al., 2014; QARTOD group, 2016; Wong, Keeley, Carval, & Argo Data Management Team, 2015); however, these procedures are seldom organized and distributed as packages, so it is still common for new users to have to implement them from scratch. Additionally, different applications of the same dataset may require different QC procedures. For example, a particular user faced with the QC of a small dataset might be willing to apply a less conservative QC in order to preserve a larger number of data points, paying the price of having some false positives. CoTeDe is an Open Source Python package that provides a flexible way to automatic QC oceanographic data by combining multiple QC standards while allowing the users to fully control and tune the parameters according to their own needs.

Automatic QC traditionally consists of a sequence of tests that identify spurious measurements based on different criteria; thus, it is common to aggregate multiple tests to cover a more extensive possibility of problems. With that in mind, the tests of the most common QC procedures were implemented in CoTeDe in a modular fashion to facilitate expanding with new tests and to permit alternative arrangements. The user can choose from one of the built-in standard procedures, for example, the Argo recommendations (Wong et al., 2015), or compose a custom arrangement of tests. This freedom allows a better data assessment by fine-tuning the methods – A Spray underwater glider operating on the California coast might benefit from slightly different thresholds than used to QC the same Spray in the Mediterranean. The custom QC procedures can be saved as JSON descriptors to be re-used or shared so that experts can develop a diverse collection of alternative procedures without necessarily requiring more coding. The implementation of such a versatile system was simplified by adopting a generic data model object that standardizes the data and metadata access, resulting in a platform-agnostic system that could equally evaluate data from an XBT or a Spray underwater glider with the appropriate tests. For instance, the concept of the World Ocean Atlas climatology comparison test (*GTSPP real-time quality control manual*, 2010) is the same for a profile from an XBT or a CTD, being only necessary to inform the observed temperature, geolocation, and time. Finally, CoTeDe dependencies were intentionally minimized to simplify its adoption by other packages.

Modern oceanography relies on an unprecedented abundance of measurements that keeps increasing every day (see Boyer et al., 2019). In the past couple of decades, traditional observational systems have been complemented by autonomous platforms like Argo, Spray underwater gliders, and Saildrones, which can provide nearly continuous data streams. Thus, QC

procedures for oceanographic data must evolve to keep pace with the ever-growing dataflow. In this context, CoTeDe was developed to reduce the burden on manual expert QC with automation, which enables the expert to focus their effort on the most challenging cases. In addition, the flexibility of CoTeDe provides a framework to implement novel QC approaches such as fuzzy logic and machine learning. CoTeDe is the result of several generations of quality control systems that have evolved since 2006, and it was designed to attend not only individual scientists but also real-time operations on large data centers.

Acknowledgements

References

Boyer, T. P., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Locarnini, R. A., Mishonov, A. V., et al. (2019). World ocean database 2018. In T. E. A. V. Mishonov (Ed.), NOAA atlas nesdis 87.

DATA–MEQ working group. (2010). *Recommendations for* in–situ *data near real time quality control* (EG10.19 ed.). European Global Ocean Observing System.

GTSPP real-time quality control manual. (2010). (First Revised Edition.). United Nations Educational, Scientific; Cultural Organization 7, Place de Fontenoy, 75352, Paris 07 SP: UNESCO–IOC.

IOC/IODE. (1993). *Manual of quality control procedures for validation of oceanographic data*.

Morello, E. B., Galibert, G., Smith, D., Ridgway, K. R., Howell, B., Slawinski, D., Timms, G. P., et al. (2014). Quality Control (QC) procedures for Australia's National Reference Station's sensor data—Comparing semi-autonomous systems to an expert oceanographer. *Methods Oceanogr.*, *9*(September), 17–33. doi:10.1016/j.mio.2014.09.001

QARTOD group. (2016). *Manual for real–time quality control of in-situ temperature and salinity data* (Version 2.0.). IOOS.

Wong, A., Keeley, R., Carval, T., & Argo Data Management Team. (2015). *Argo quality control manual for CTD and trajectory data*. doi:http://dx.doi.org/10.13155/33951