DATA QUALITY CONTROL IN THE U.S. IOOS

Matthias Lankhorst[1], Fred Bahr[2], Emmanuel Boss[3], Patrick Caldwell[4], Orest Kawka[5], Michael F. Vardaro[6]

1. *Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA.*
2. *Monterey Bay Aquarium Research Institute, Moss Landing, CA.*
3. *School of Marine Sciences, University of Maine, Orono, ME.*
4. *NOAA National Oceanographic Data Center, Honolulu, HI.*
5. *School of Oceanography, University of Washington, Seattle, WA.*
6. *College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR.*

Contact email: mlankhorst@ucsd.edu

Abstract

The end product of an observing system is data. Data quality is therefore of utmost importance to the system, as it is synonymous with quality of the end product. Every step beyond data generation, such as analysis and interpretation of the data or use of observational data in computer models of the ocean and the atmosphere, relies on the observational data to be fit for the particular purpose: if the input data are not accurate enough, the outcome of the analysis or the computer simulation becomes invalid. Yet, observing systems historically struggle with data quality control, for reasons of both technical and budgetary nature–it may be difficult or expensive to do. This document calls for an end-to-end approach towards design and implementation of data quality control in the IOOS, to be a core element of the system implementation and the budget. While touching upon design aspects of the field instrumentation, the focus is on the data management and data processing techniques, bearing two goals in mind: to properly assess whether a given data point is "good" or "bad" when considered against a target uncertainty, and to maximize the number of data points that are "good" within this uncertainty.

Key words: data quality control, qc, calibration, validation, verification, algorithm, bio-fouling

1. Introduction

a. Requirements and Summit Proceedings

The IOOS High-Level Functional Requirements document (version 1.5, January 2009) calls out the following requirements for data quality control in the Data Management and Communications (DMAC) subsystem:

> Requirement ID 4.2.1: The DMAC shall provide a mechanism for ensuring that data are of known and documented quality. Quality control operations are a partnership among data observation/collection components, processors, analysts, other users, and the DMAC.
>
> Requirement ID 4.2.2: Regional data centers shall apply quality control measures to data and derive specialized products.

The purpose of this document is then to elaborate on these requirements and work towards an implementation thereof, which is quantifiable in terms of time and effort involved, as well as objectivity in analysis. This will broadly map to the following sections of the 2012 IOOS Summit Proceedings:

- Section 2 (User Requirements: Revisiting / Updating)
- Section 5 (Vision for Next 10 Years).

b. Example

Figure 1 depicts a fictitious data set, which has been artificially degraded by phenomena commonly seen with in-situ data. Data quality control then is supposed to detect these degradations, some of which may be correctable, and help determine how accurate the data really are. The figure shows large and small outliers, such as those caused by telemetry failures or temporary contamination of the sensor element, as well as a small drift, such as sensor aging or gradual bio-fouling. In this context, the words "large" and "small" mean that the anomaly is large or small when compared to the natural variability in the surrounding data points. The wording "outlier" means that the anomaly occurs in isolation or a short succession of a few data points. In contrast, "drift" indicates a relatively slow and monotonic change in sensor behaviour that affects many successive, or even all, data points. These distinctions drive different approaches to detecting their presence. Obviously, large errors are easier to spot than small ones. Outliers can potentially be found by comparing against earlier or later values from the same instrument, which is less likely to work for drift since these other data points are erroneous as well.

Figure 1: Example of a data set illustrating the failure modes that data quality control ought to address. The data are fictitious, but can be thought of as salinity measurements from a mooring as a function of time. The black arrows denote a (fictitious) target accuracy, data outside of which would be considered bad. Left panel: data as reported by the instrument. Right panel: data as reported by the instrument (red, same as left panel), and in addition the true data (green). The red data have been derived from the green data by artificial modifications introducing three large outliers (marked purple), a temporary episode with a small constant offset (marked blue), and a linear trend increasing with time (marked orange at the end). The challenge is to identify these cases, when only the "red points" are known.

2. Observing System Design

The vision is for the IOOS to be built such that for each observation made, there is a clear objective, such as a scientific question or a monitoring function, which the observation is supposed to address. Each objective constitutes a high-level functional requirement. Derived from these requirements, there are values for the permissible data uncertainties, and the observing platforms as well as the instruments and the operational cycles are designed to achieve these uncertainties. The resulting data undergo routine quality control, the outcome of which is an assessment of the actual uncertainty of the data. This assessment is available to the user along with the data, as is the documentation about the observational objective and the particular platform design. Availability of this information to the user constitutes the "end-to-end" aspect of data quality from inception at the requirement-level to the final assessment of data quality at the end-product level. Figure 2 shows the hierarchy by which these requirements flow down to the observations, with data quality control as an embedded step.

3. Available Technology

a. In-Situ Instrumentation

As shown in figure 2, the design of the instruments and platforms is an integral part of the process flow between science goals and the resulting data. Therefore, this design process is interwoven with data quality control, as the latter depends on the former. In addition to standard engineering aspects in the design, protection against bio-fouling is a particular concern for marine measurements. The following instrument design options are listed as examples that can improve data quality by reducing the effects of bio-fouling (see also Delauney et al., 2010, and Manov et al., 2004):

- Use of shutters or wipers on optical sensors
- Use of toxins near sensors
- Enclosing sensors in opaque plumbing circuitry together with pumps and toxins
- Use of copper guards or plumbing

Figure 2: Concept of the design logic of the observing system. The left column shows how the implementation proceeds from a science goal to the final data, and the right column provides some examples for each step along the way.

Similarly, the following platform design features can improve data quality by reducing bio-fouling:

- Use of anti-fouling paint on all structural underwater surfaces, particularly near sensors, to avoid growth affecting sensors directly and to avoid creating a micro-environment in which the data are not representative of the surrounding environment
- Fencing of surface buoys to protect against seals and other animals
- Deterrents on surface buoys (spikes, wire fences) to protect against bird excrements
- Use of profiling platforms (gliders, floats, moored profilers) to pressure-cycle the platforms and to reduce time spent in the euphotic zone

The following are operational considerations of the instrument life cycle which address issues such as bio-fouling:

- Instrument turn-around cycles should be frequent enough to minimize the effects of bio-fouling and sensor drift.
- Collection of calibration/verification data both at deployment and recovery of instruments, to assess drift in-between.
    - At recovery, collect these data while the instrument is still in its fouled state (before cleaning and servicing).

    - Additional verification points during the mission can improve drift correction to something more appropriate than a simple two-point linear interpolation.
- Proper documentation of each calibration step is paramount, e.g. which computations an instrument already makes internally vs. which ones are adjusted externally. This is to protect against applying the same adjustments twice, and to make re-processing possible e.g. if updated calibration data become available. Preserving and making the raw data available is part of this consideration.

c. Post-Processing

Most of what is usually called “data quality control” occurs during post-processing, i.e. after the data have been collected, and includes all activities in “software space”. One of these steps is to determine possible adjustment factors to the data, based on the calibration and verification data available. This step usually requires human intervention: a data analyst reviews the data, determines and applies the adjustments, and then publishes the updated data product as well as the documentation of the processing steps.

Another component of the post-processing quality control is the detection of outliers and drifts, as shown in figure 1. There are computer algorithms that can automatically detect some of these anomalies, whereas others rely on human inspection of the data. Applicability of one versus the other heavily depends on the particular data product and how stringent the uncertainty requirements are, but as a general rule of thumb, the following apply:

- **Platform health monitoring** by algorithms that monitor engineering parameters can provide a general indication when data might be compromised due to engineering issues such as low battery voltages, or buoys broken loose and drifting away from their nominal positions. Such monitoring is also routinely used for surface drifters, which have a sensor that detects whether the drogue has broken off the float element. If platform engineering data are available, it is usually simple and straight-forward to implement such checks.
- **Self-contained computer algorithms** examining the data can reliably detect large outliers, but are usually not able to detect small outliers or signal drifts. However, they are essential for real-time data delivery and can reduce the workload for the data quality analyst performing the visual inspection procedures. The real-time data QC algorithms of the Argo project are a prototype example of these kinds of algorithms (see Argo quality control manual). Typical algorithms are range checks that compare data against climatological values, and other filters that search for specific failure modes such as spikes. Another class of algorithms inspects not the data values themselves, but rather, related or ancillary data, in order to assess the quality of the actual data. An example is given by ADCP instruments (acoustic Doppler current profilers). This instrument class measures currents acoustically and also inherently reports correlation values between different acoustic paths as well as echo intensity of the acoustic signals, all of which can be used in fully automated data QC algorithms. There are specific cases where availability of such ancillary information, and analysis thereof, are sufficient to make human inspection obsolete even in cases that demand high data quality. However, this is rarely the case if the data streams from the instrument do not contain multiple and independent parameters.
- **Visual Inspection** of data by a subject-matter expert can often detect small signal drifts and outliers if sufficient ancillary data (e.g. calibration and verification data, engineering data) are available. Depending on the data type and quantity, and the available personnel and financial resources, data thus examined are of the highest quality available. Some software exists to aid the operators by providing aggregate statistics and specialized plots tailored to a particular sensor behaviour. The commonly used T-S diagrams are an example, in that they reveal deviations from historical temperature and salinity relationships with much higher sensitivity than a climatology of each parameter plotted as a function of depth. The delayed-mode QC procedures of the Argo project are a prototype example of procedures combining visual inspection by a human expert with purpose-built software tools to increase efficiency (see Argo quality control manual). In conventional research grants issued to a single investigator, that investigator (or his team of postdoctoral scholars and students) may spend a significant amount of time inspecting and adjusting data before using it in publications. In contrast, in many operational systems such as weather stations there is no human processing at all, perhaps because the uncertainty requirements are not so stringent as to make this a necessity, because the sensors have been sufficiently studied and documented to provide data within the required accuracy over the time they are deployed for, or because the need for real-time data outweighs other needs. As a consequence, observation systems that blend research and operational systems, such as IOOS, typically struggle to find the right balance between research-quality data that requires a certain amount of human processing, and complete automation. There remains a need to develop and adopt semi-automated procedures like those successfully developed by Argo, but for other parameters than temperature and salinity. Known issues with human-operated quality control are inter-operator discrepancy (two operators might come to different conclusions about identical data) and time delay (human processing takes time; a delay of months to years is common).

4. Recommendations

The following are recommendations for IOOS development over the next decade:

- Each observing platform should have a clear and documented design approach to mitigate bio-fouling and other data-degrading processes.
- The DMAC and data formats should support storing and distributing such design artefacts along with the data streams, e.g. as links to persistent documents or as embedded metadata.
- Operational cycles and instrument life cycles of IOOS assets should be managed such that bio-fouling and sensor degradation are minimized.
- The scope of requirement ID 4.2.1 should be expanded to apply also to time and location of the measurement, i.e. requiring that time and place are also of known and documented quality.
- Elaborating on requirement ID 4.2.1, the DMAC capability should include metadata fields for data uncertainty, and ancillary QC flags that describe the outcome of specific QC tests, which may either be computer- or human-generated.
- Elaborating on requirement ID 4.2.1, the DMAC capability should include metadata fields that allow a full back-tracing of all calibration steps from the raw data to the final data product, including versioning control of data sets as needed, and should always store the raw data.
- Elaborating on requirement ID 4.2.2, each data provider should have a defined approach to test, and successively flag or correct, for each of the following, balancing the effort against the desired data uncertainty (which should be quantified to the best ability of the operator):
    - gross outliers in the data (suggested implementation: automated algorithms)
    - signal drifts in the data (suggested implementation: pre-deployment and post-recovery calibrations, and additional inter-sensor verification during mission as applicable)
    - small outliers (suggested implementation: computer-aided visual inspection by subject-matter expert)
- Elaborating on requirement ID 4.2.2, IOOS should actively participate in development of data QC techniques to improve our capacity to detect small outliers and signal drifts in data, e.g. by advancing computer algorithms to increase efficiency versus the visual inspection processes by an operator, and to reduce inter-operator subjectivity of these inspections.

References

Argo quality control manual. Version 2.7. Argo data management, 3 January 2012. Available online at: http://www.argodatamgt.org/content/download/341/2650/file/argo-quality-control-manual-V2.7.pdf

Integrated Ocean Observing System. **High-Level Functional Requirements**. Version 1.5, January 2009. Available online at: http://www.ioos.gov/library/noaa_hlrd_v1_5_01_13_09.pdf

Delauney, L., Compère, C., and Lehaitre, M.: **Biofouling protection for marine environmental sensors**, Ocean Sci., 6, 503-511, doi:10.5194/os-6-503-2010, 2010.

Manov, D. V., Chang, G. C., and Dickey, T. D.: **Methods for reducing biofouling of moored optical sensors**, J. Atmos. Ocean. Tech., 21, 6, 958–968, 2004.