Automatic quality control and quality control schema in the Observation to Archive

Brenner Silva¹, Najmeh Kaffashzadeh², Erik Nixdorf³, Sebastian Immoor¹, Philipp Fischer¹, Norbert Anselm¹, Peter Gerchow¹, Angela Schäfer¹, Roland Koppe¹

¹ Alfred-Wegener-Institute, Computing and Data Centre, Bremerhaven, Germany
² Forschungszentrum Jülich GmbH, Jülich Supercomputing Center, Jülich, Germany
³ Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany

https://doi.org/10.5194/egusphere-egu2020-15961
EGU General Assembly 2020
© Author(s) 2020. This work is distributed under the Creative Commons Attribution 4.0 License.
The **O2A** is a data-flow framework that is operational and under developed at the Alfred-Wegener-Institute, Bremerhaven, Germany. The O2A supports the sensor management of heterogeneous sources, by gathering together metadata, near-real-time ingestion, quality control, monitoring, analytics, and a data-driven knowledge base of scientific and technical documentation.

Please refer to [EGU2020-19631](http://data.awi.de/o2a-doc) for a complete overview and to [http://data.awi.de/o2a-doc](http://data.awi.de/o2a-doc) for a continuously improved documentation.

This work ([EGU2020-15961](http://data.awi.de/o2a-doc)) focus on time series data and the ingestion part of the framework, where the automatic quality control takes place.
The **O2A-INGEST** performs automatic quality control to deliver quality-flagged data in the **O2A-DASHBOARD**. We use a modular approach for tests and for drivers. Drivers are used to access data at specific formats (e.g. regarding data loggers of different instruments like weather stations, buoys, ferry boxes, CTD).

The quality control (QC) requests observation properties from the **O2A-SENSOR REST-API** for each corresponding sensor and each quality control test. The input data is in **NRT format**, where each column of observations is under a unique sensor-URN. At ingest, the quality control algorithm builds a table of devices and parameters to assess the input data for correctness and validity of observations.
## Quality Control Tests

**Flagging scheme and currently operational tests**

<table>
<thead>
<tr>
<th>Test name</th>
<th>Description</th>
<th>Property required</th>
<th>Ancillary data required</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operation temperature range</td>
<td>Test for temperature conditions (air and/or surface temperature) under which the instrumentation is deployed</td>
<td>Operation Temperature</td>
<td>Temperature observation</td>
<td>operational</td>
</tr>
<tr>
<td>Manufacturer range</td>
<td>Test if value is within the limits of the instrumentation (e.g. due to construction, material or filter) as given by manufacturer</td>
<td>Manufacturer range</td>
<td>None</td>
<td>operational</td>
</tr>
<tr>
<td>Operation range</td>
<td>Test if value is within a specific range valid for the location where the sensor is deployed</td>
<td>Operation range</td>
<td>None</td>
<td>operational</td>
</tr>
<tr>
<td>Gradient test</td>
<td>Test for gradient, i.e. absolute distance from the median value of neighboring (n=5) observations</td>
<td>Gradient Threshold</td>
<td>None</td>
<td>operational</td>
</tr>
<tr>
<td>Spike test</td>
<td>Test for spikes, i.e. distance from the median value subtracted by the standard deviation of neighboring (n=5) observations</td>
<td>Spike Threshold</td>
<td>None</td>
<td>operational</td>
</tr>
<tr>
<td>Range function</td>
<td>Test for physical relationships, or interdependency, among observations</td>
<td>Thresholds array</td>
<td>None</td>
<td>development</td>
</tr>
<tr>
<td>Geo location</td>
<td>Test for valid geographic location of moving and stationary sensors</td>
<td>Latitude, longitude and altitude ranges</td>
<td>Location of observations or Sensor Event</td>
<td>development</td>
</tr>
</tbody>
</table>

The current flagging scheme is of primary level ([UNESCO 2013](#)), i.e., quality flags are sequential and qualitative to describe a level of quality in the data. The starting point for the currently implemented quality control tests is the ARGO real-time quality control ([Wong et al. 2019](#)). After sending a request to O2A-SENSOR for observation properties, the O2A-INGEST performs the quality control tests. Please refer to the left table or go to [Quality Flagging](#) for current Status.
Quality Control Tests

Probabilistic QC

A score attribute has been developed, that can be either used to assess the quality flag or as an indicator for estimating plausibility of each individual data value (EGU2020-13357).

A schematic of regular data errors in a data time series. A red circle shows the focus of the developed test, i.e. constant value test (CVT). Source (EGU2020-13357) presentation.

The occurrence of successive constant values episode (CVE) can not be always interpreted as an indicative of sensor (system) failures or other measurement errors. There is 0.43 likelihood that this episode consists of valid data values.

The results of performing the CVT on temperature time series at the Cape Grim station. The black and blue lines show the time series and its associated probability. The red circles highlight several constant value episodes (CVEs) with a different length (t) and probability (P). Source (EGU2020-13357) presentation.

Presented by Brenner Silva on May 6, 2020

EGU2020-15961 - Automatic quality control and quality control schema in the Observation to Archive
Data Web Service, Dashboard and the Quality Flag

Overview and monitoring of time-series in NRT

The Data Web Service gives an overview of the data homogenized in near-real-time and the current quality flag (QF). The example on the left shows data from the Müglitz River collected by the UFZ (Nixdorf E. and Ködel U. 2019).

A secondary level of the QF has been developed using a cumulative approach to represent a combination of test results. In addition, a second quality score attribute of ratio type is under development. The quality score can be used either to assess the QF or as in the approach of the FZJ (Kaffashzadeh et al., 2019), as an indicator for estimating plausibility of each individual data value.
A set of control tests including range, spike, and gradient are currently operational. Their implementation occurs after the revision of formulations found in the literature. For instance, the spike test (as shown in the left figure) allows for the detection of the inconsistencies (in red) at the end of the time-series bellow (blue frame).
Your Data in the O2A

Four steps to get your data working with O2A

You have your sensor running and would like to work with O2A in near-real-time, you may follow these four steps:

1. Request an account for sensor.awi.de via o2a-support@awi.de.

2. Enter your metadata considering the best practices and add observation properties for sensors and quality tests.

3. Configure the data transfer, via sensor.awi.de or o2a-support@awi.de.

4. Monitor your data via overview, or create your dashboard.