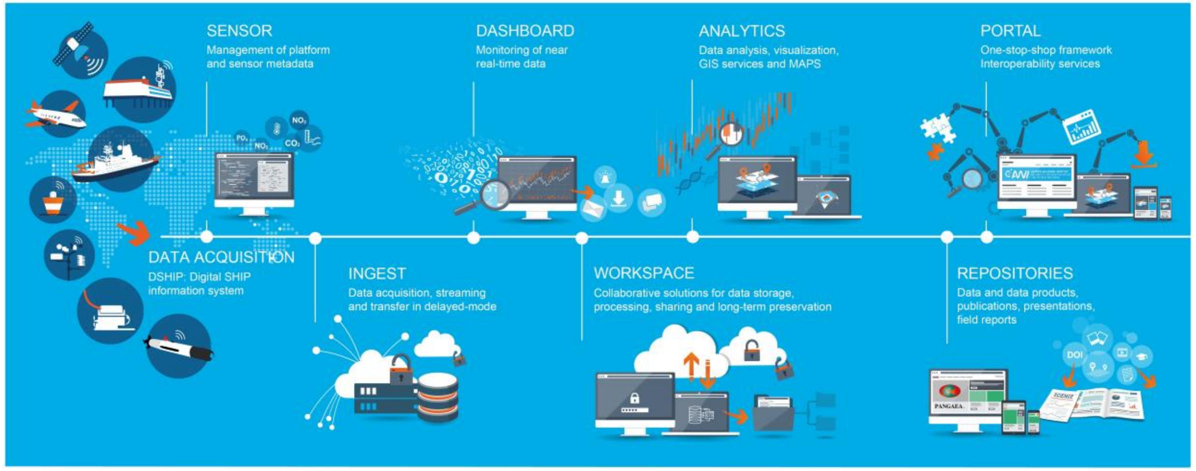
Automatic quality control and quality control schema in the Observation to Archive

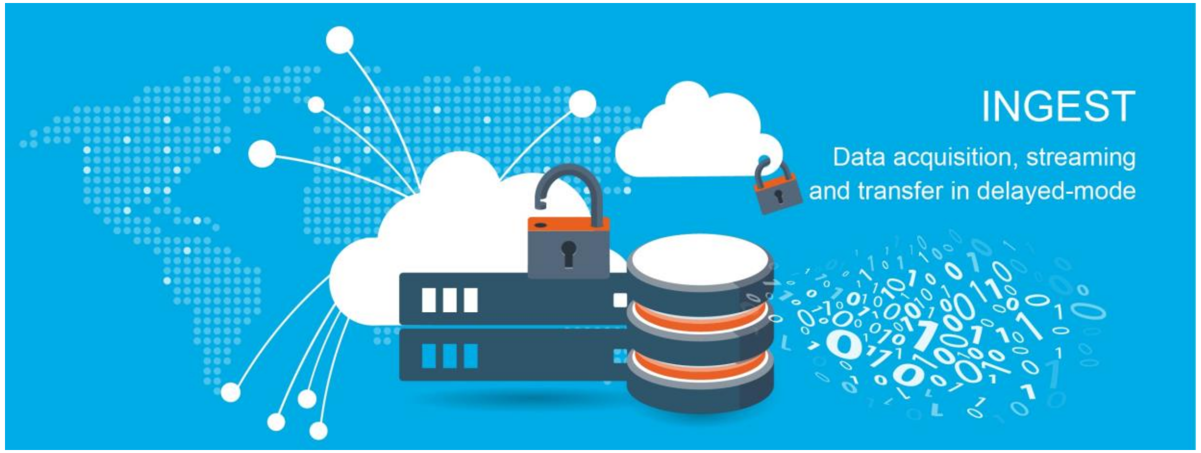
Brenner Silva[1]**,** Najmeh Kaffashzadeh[2], Erik Nixdorf[3], Sebastian Immoor[1], Philipp Fischer[1], Norbert Anselm[1], Peter Gerchow[1], Angela Schäfer[1], Roland Koppe[1]

[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

EGU2020-15961, presented on May 6, 2020.
https://doi.org/10.5194/egusphere-egu2020-15961
EGU General Assembly 2020

The Observation to Archives and Analysis (O2A)


SENSOR
Management of platform and sensor metadata
DASHBOARD
Monitoring of near real-time data
ANALYTICS
Data analysis, visualization, GIS services and MAPS
PORTAL
One-stop-shop framework Interoperability services
DATA ACQUISITION
DSHIP: Digital SHIP information system
INGEST
Data acquisition, streaming and transfer in delayed-mode
WORKSPACE
Collaborative solutions for data storage, processing, sharing and long-term preservation
REPOSITORIES
Data and data products, publications, presentations, field reports
DOI
PANGAEA



The O2A is a data-flow framework that is operational and under developed at the Alfred-Wegener-Institute, Bremerhaven, Germany.

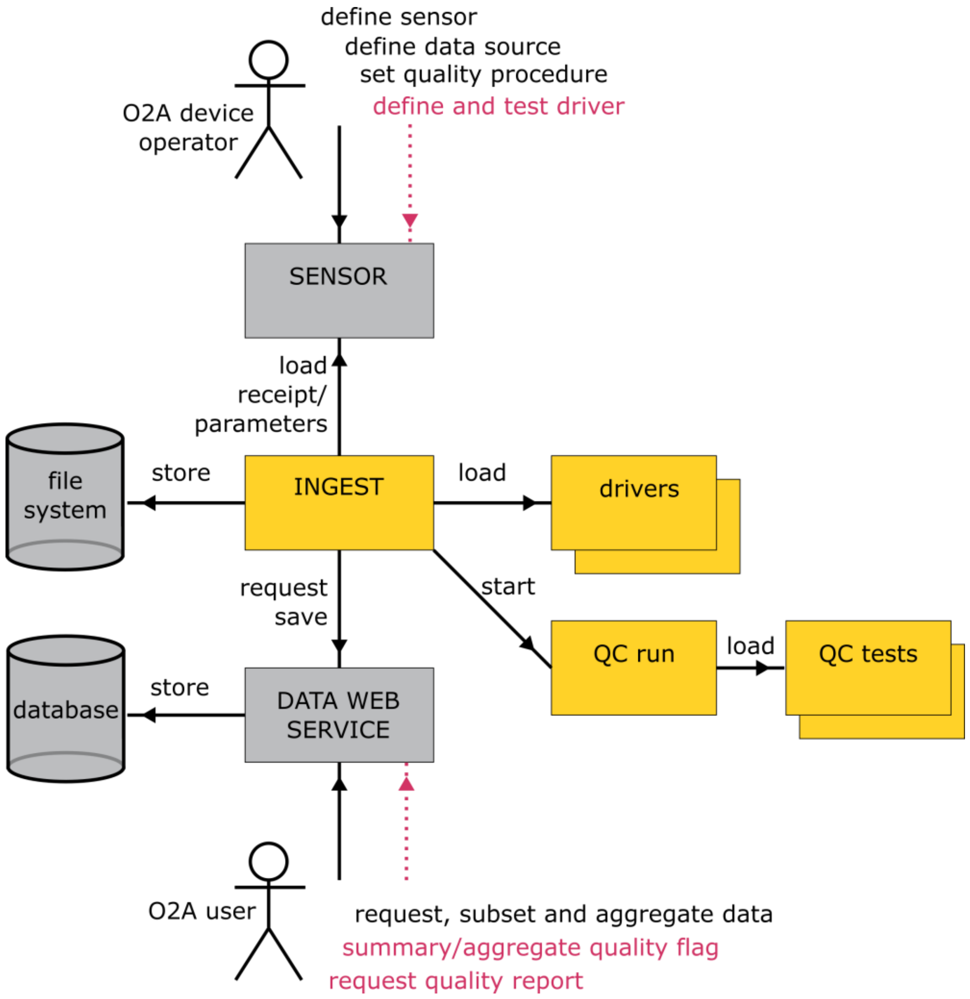
The O2A supports the sensor management of heterogeneous sources, by gathering together metadata, near-real-time ingestion, quality control, monitoring, analytics, and a data-driven knowledge base of scientific and technical documentation.

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

This work (EGU2020-15961) focus on time series data and the ingestion part of the framework, where the automatic quality control takes place.

O2A-INGEST

Near Real Time System



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

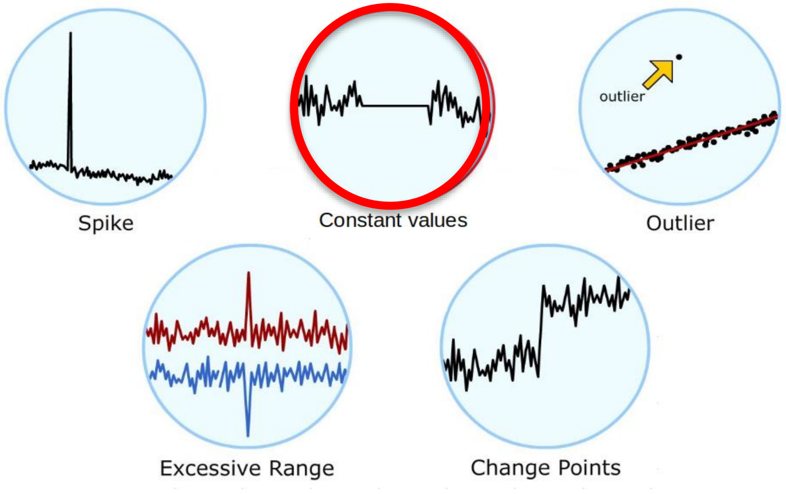
Test name	Description	Property required	Ancillary data required	Status
Operation temperature range	Test for temperature conditions (air and/or surface temperature) under which the instrumentation is deployed	Operation Temperature	Temperature observation	operational
Manufacturer range	Test if value is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

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

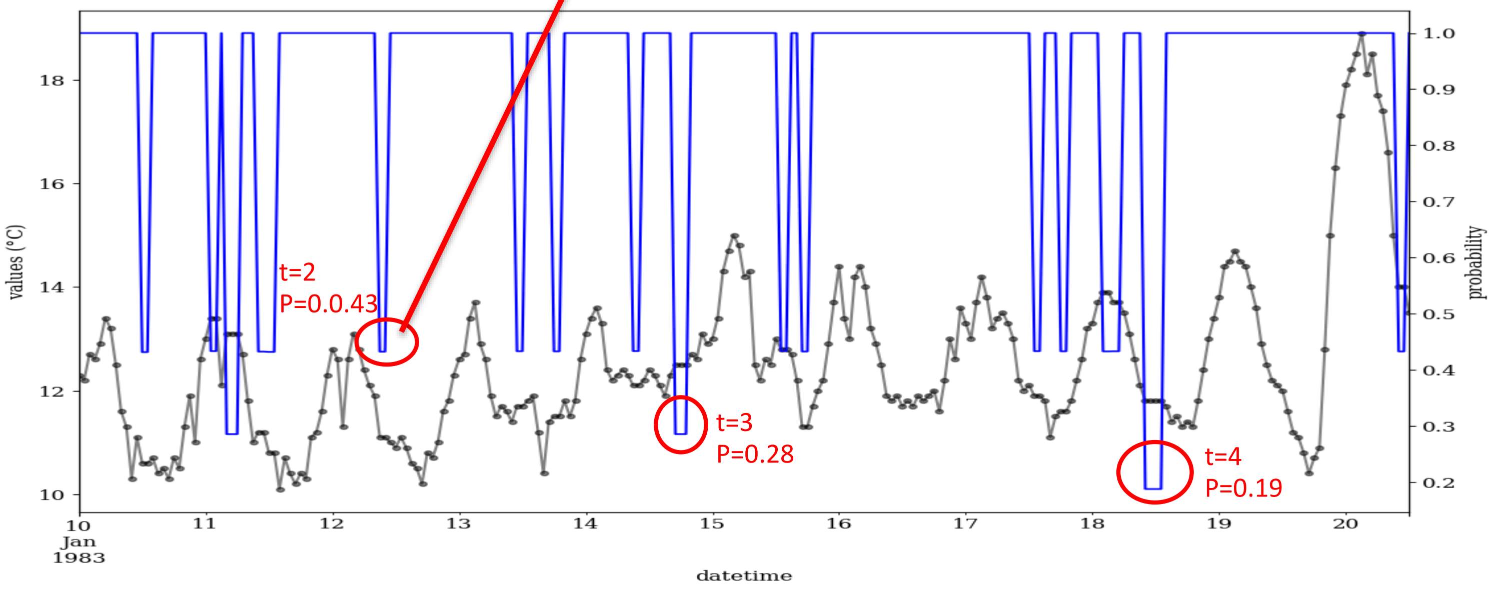
Probabalistic QC

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.

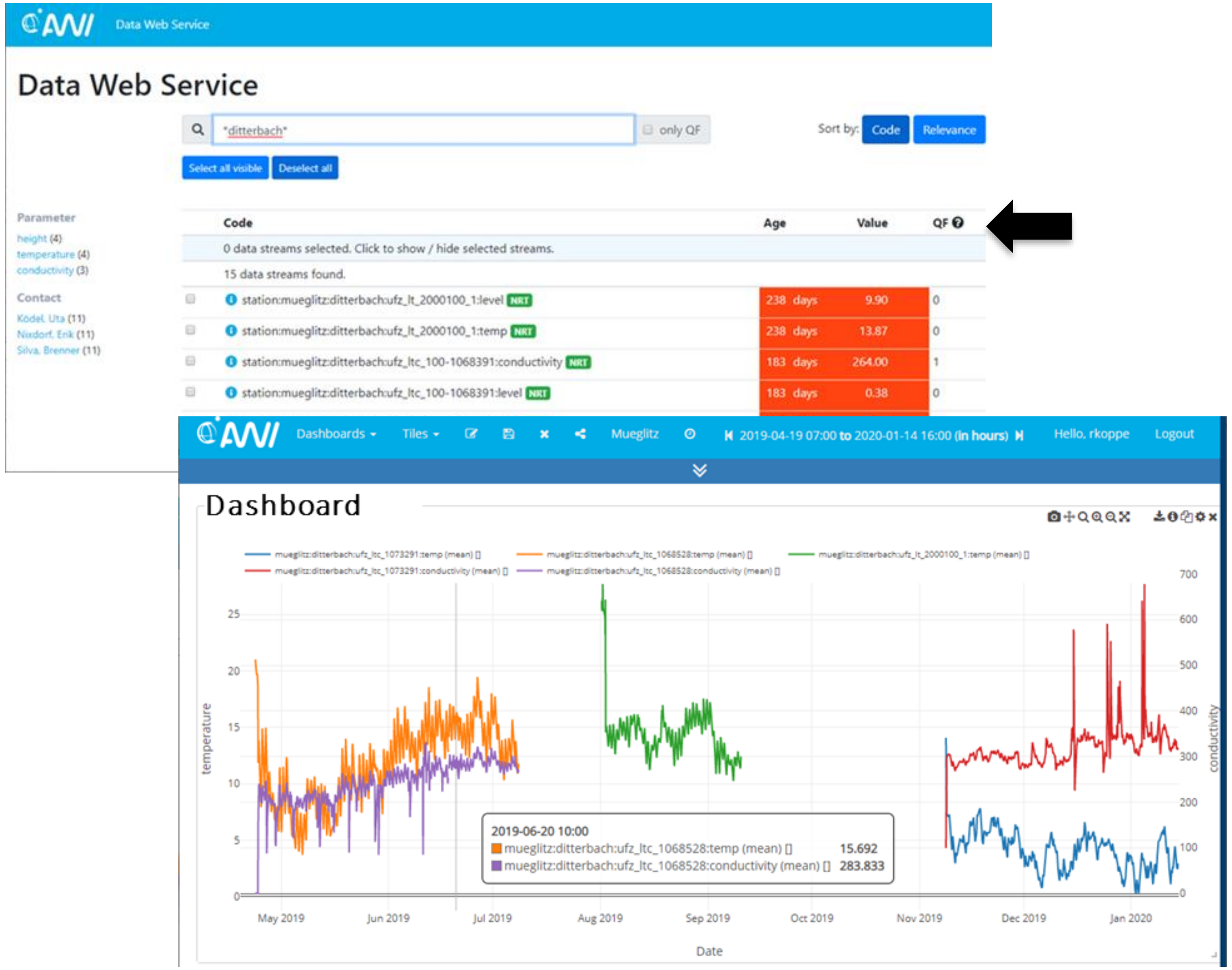
The occurrence of successive constant values episode (CVE) can not be always interpreted as an indicative of sensor (system) failures or other measurement errors. There is 0.43 likelihood that this episode consists of valid data values.



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.

Data Web Service, Dashboard and the Quality Flag

Overview and monitoring of time-series in NRT



Data Web Service

Data Web Service

ditterbach □ only QF Sort by: Code Relevance

Select all visible Deselect all

Parameter
height (4)
temperature (4)
conductivity (3)

Contact
Ködel, Uta (11)
Nixdorf, Erik (11)
Silva, Brenner (11)

	Code	Age	Value	QF
	0 data streams selected. Click to show / hide selected streams.			
	15 data streams found.			
□	station:mueglitz:ditterbach:ufz_lt_2000100_1:level NRT	238 days	9.90	0
□	station:mueglitz:ditterbach:ufz_lt_2000100_1:temp NRT	238 days	13.87	0
□	station:mueglitz:ditterbach:ufz_ltc_100-1068391:conductivity NRT	183 days	264.00	1
□	station:mueglitz:ditterbach:ufz_ltc_100-1068391:level NRT	183 days	0.38	0

Dashboards Tiles Mueglitz 2019-04-19 07:00 to 2020-01-14 16:00 (in hours) Hello, rkoppe Logout

Dashboard

2019-06-20 10:00
mueglitz:ditterbach:ufz_ltc_1068528:temp (mean) 15.692
mueglitz:ditterbach:ufz_ltc_1068528:conductivity (mean) 283.833

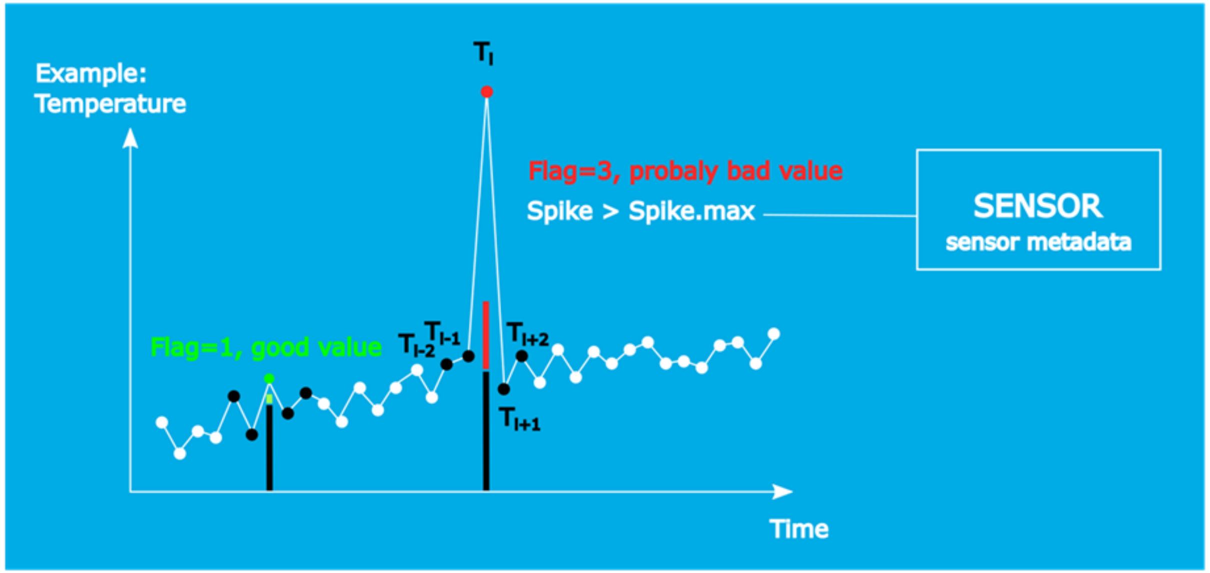
mueglitz:ditterbach:ufz_ltc_1073291:temp (mean)
mueglitz:ditterbach:ufz_ltc_1068528:temp (mean)
mueglitz:ditterbach:ufz_lt_2000100_1:temp (mean)
mueglitz:ditterbach:ufz_ltc_1073291:conductivity (mean)
mueglitz:ditterbach:ufz_ltc_1068528:conductivity (mean)
temperature
0
5
10
15
20
25
conductivity
0
100
200
300
400
500
600
700
May 2019
Jun 2019
Jul 2019
Aug 2019
Sep 2019
Oct 2019
Nov 2019
Dec 2019
Jan 2020
Date

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.

Test Sensitivity - Example

Revision and implementation



Manufacturer range: $T_i \in [a_m, b_m]$

- Temperature $\in [-2.5, 35.]$°C
- Pressure $\in [-5, 3140^*]$ dbar
- Salinity $\in [2.,41.]$ PSU
- Conductivity $\in [0.0,65.]$ mS/cm

Operation range: = $T_i \in [a, b]$

- Temperature $\in [-1.9, 10.]$°C
- Pressure $\in [2, 314^*]$ dbar
- Salinity $\in [30., 37.]$ PSU
- Conductivity $\in [25.,45.]$ mS/cm

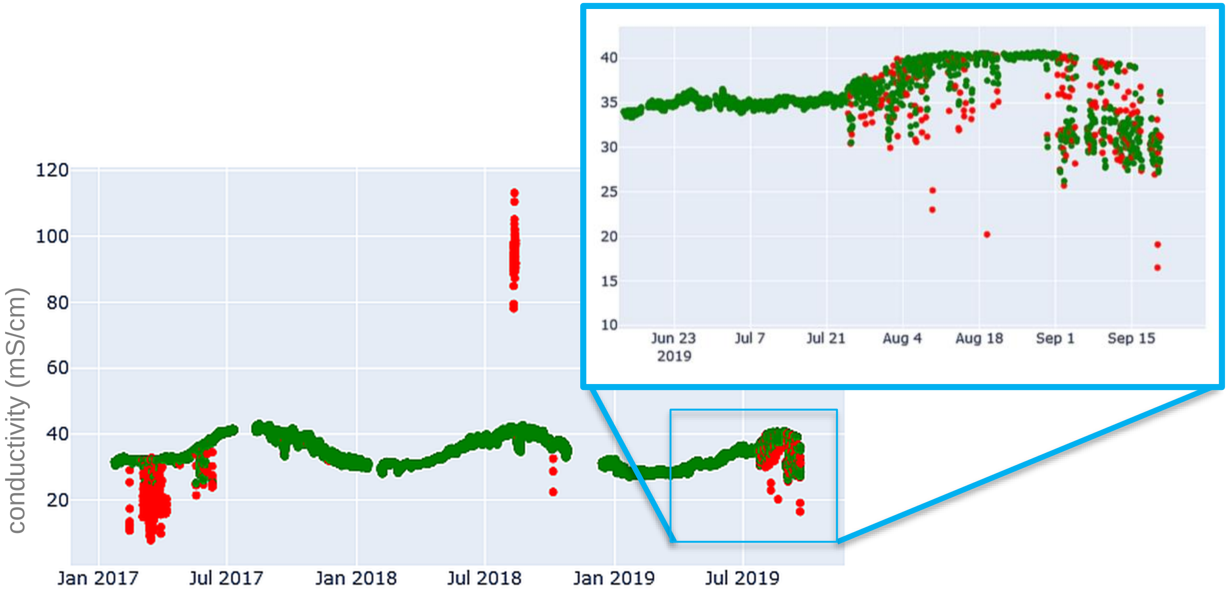
Spike: $|\text{median}[T_{i-2},.., T_{i+2}]| + |\text{SD}[T_{i-2},.., T_{i+2}]| < \text{Spike}_{max}$

- Temperature, Spike_{max} = 1°C
- Pressure, Spike_{max} = 1 dbar
- Salinity, Spike_{max} = 0.9 PSU
- Conductivity, Spike_{max} = 0.75 mS/cm

Gradient: $| T-(T_{i+1} + T_{i-1})/2 | < \text{Gradient}_{max}$

- Temperature, Gradient_{max} = 1.5°C
- Pressure, Gradient_{max} = 1.5 dbar
- Salinity, Gradient_{max} = 1.25 PSU
- Conductivity, Gradient_{max} = 1.0 mS/cm

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 .