Marine Data Quality Control

Simona Simoncelli

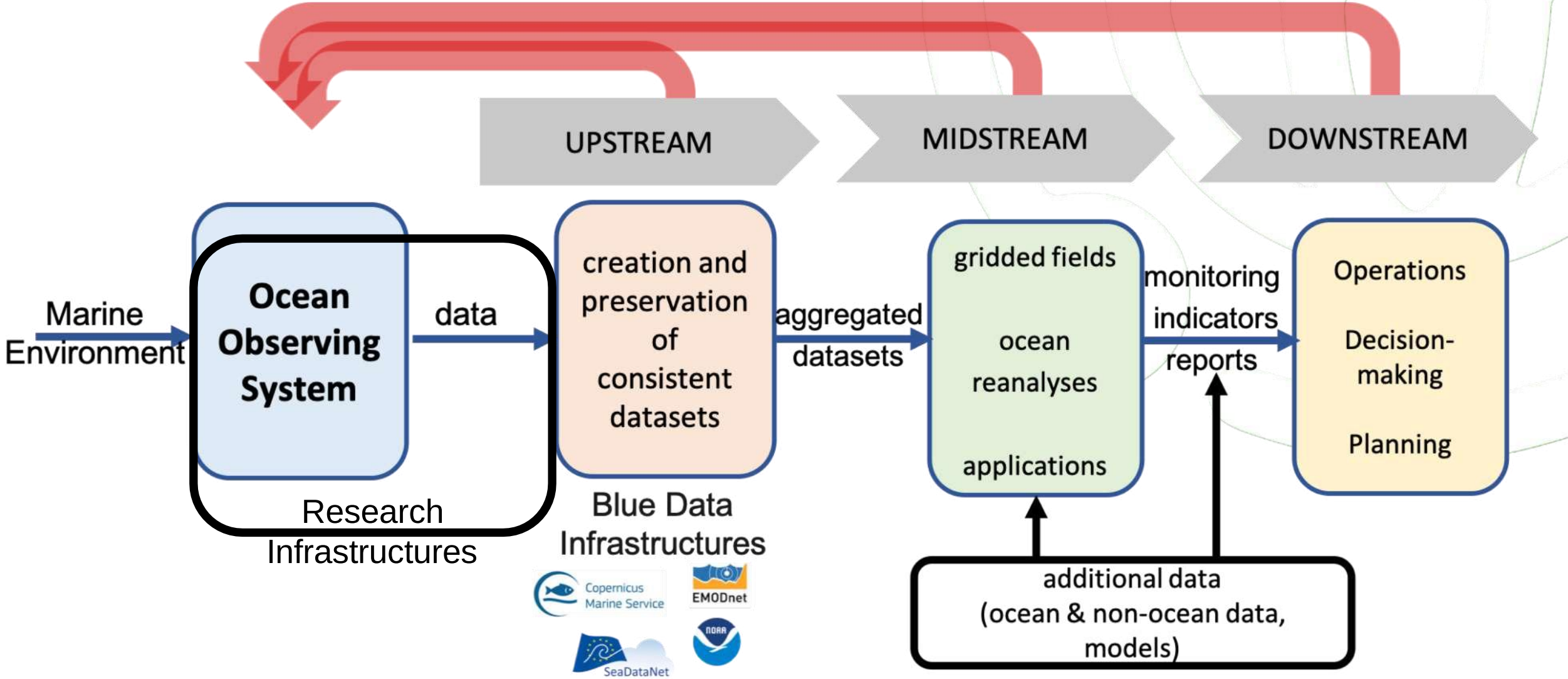
IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Outline

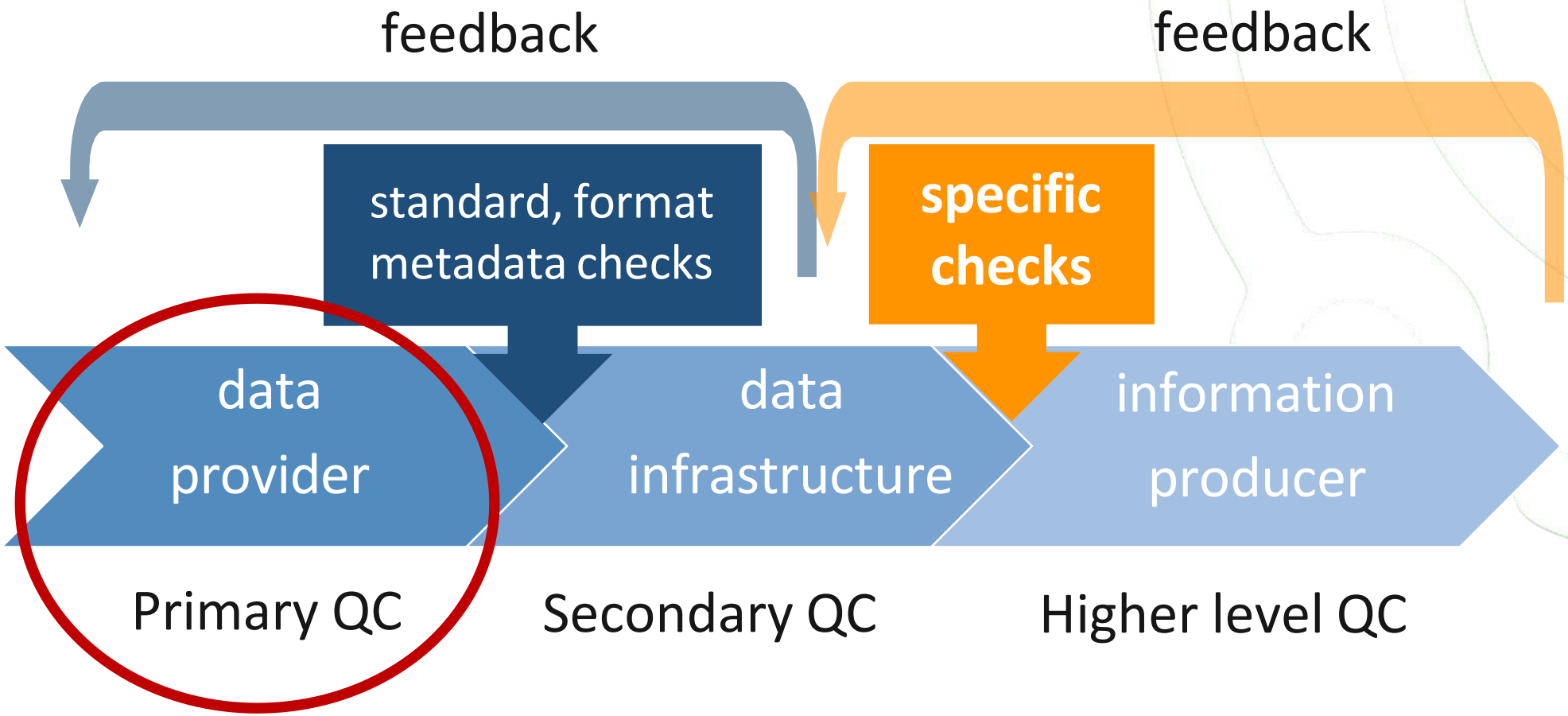
- Marine data value chain and Quality Control levels
- Data life cycle, Quality Assessment and Quality Control
- Best Practices (BPs)
- Data processing levels, latency and QC
- Data quality dimensions
- Data Processing steps with some examples
- automatic vs manual Quality Control
- Data flagging schemes
- Some data products and QC examples

Marine Data Value Chain



Simoncelli et al. 2022 https://doi.org/10.1016/B978-0-12-823427-3.00001-3

Quality along the data value chain



- **integrity**
- **consistency**
- **completeness**

→ several actors apply QC procedures along with the data value chain

→ **provenance** and **lineage information** are key elements to preserve

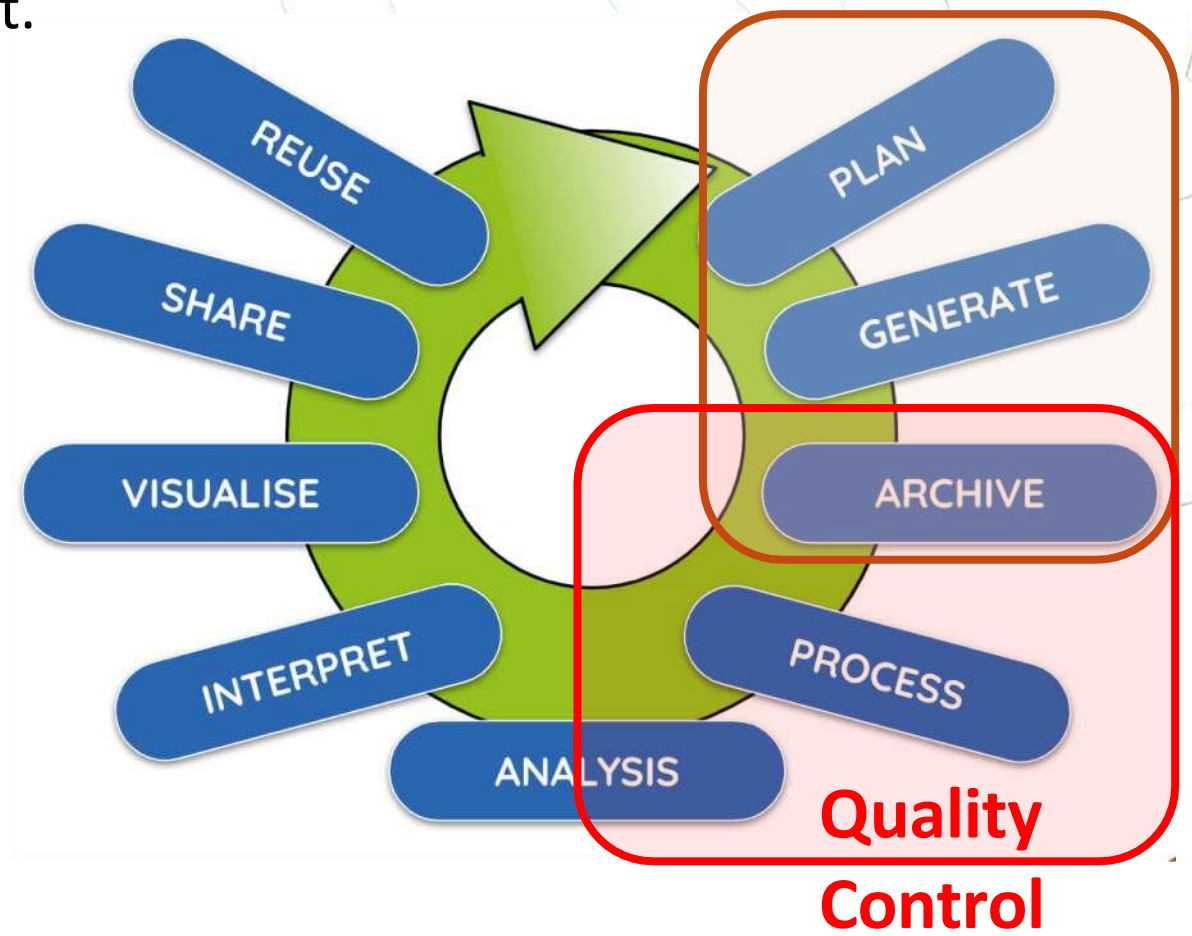
Simoncelli et al. (2022) https://doi.org/10.1016/B978-0-12-823427-3.00001-3

Research Data Life Cycle

It is described as a cycle because the lessons learned and insights gleaned from one data project typically inform the next and the final step of the process feeds back into the first.

Open research data are both raw and processed, and the accompanying metadata, …, protocols, analysis code and workflows that can be openly used, reused, retained and redistributed by anyone, subject to acknowledgement

Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship (FAIR principles) supported by regular curation and maintenance



Data Quality

Scientists who observe and distribute oceanographic data require a process to ensure high-quality data. This process includes QA, QC, quality assessment, standards and best practices

Quality Assurance

QA: processes that are employed to support the generation of high-quality data (instrument selection, preparation, deployment, maintenance and data acquisition)

Quality Control

QC: the follow-on steps that support the delivery of high-quality data, requiring both automation and human intervention

Data Quality Assessment

DQA: the process of empirically evaluating data to determine whether they meet the necessary quality requirements (e.g., accuracy, precision, spatial, and temporal resolution, etc.).

Monitoring activity

↓

data product L0-L4

↓

downstream product

Data Quality

International Organization for Standardization (ISO) (ISO9000:2105) states “While **quality assurance** relates to how a process is performed or how a product is made, **quality control** is more the inspection aspect of quality management”

QA and QC procedures may be specific to a sensor technology or even to a particular manufacturer’s model, so the establishment of a standard that is applicable to every sensor is difficult.

«Consider a post-deployment calibration. Under certain circumstances, it may be used to establish sensor drift and determine a correction to be applied to the data – a QC action. The same post calibration might then also serve as a pre-deployment calibration for the next deployment – a QA action. These semantic conundrums can be clarified through documentation when operators (data providers) describe their data acquisition methods. Metadata should clarify and document the efforts undertaken»

Bushnell et al. (2019) https://doi.org/10.3389/fmars.2019.00706

Quality Assurance (QA)

- **Prior planning** is mandatory to document through metadata and standards data creation, content and context but also to fulfill data quality requirements.
- **Data quality requires predefined QA strategies** based on the selection of internationally validated methodologies (best practices) for sampling and analysis and the use of reference materials

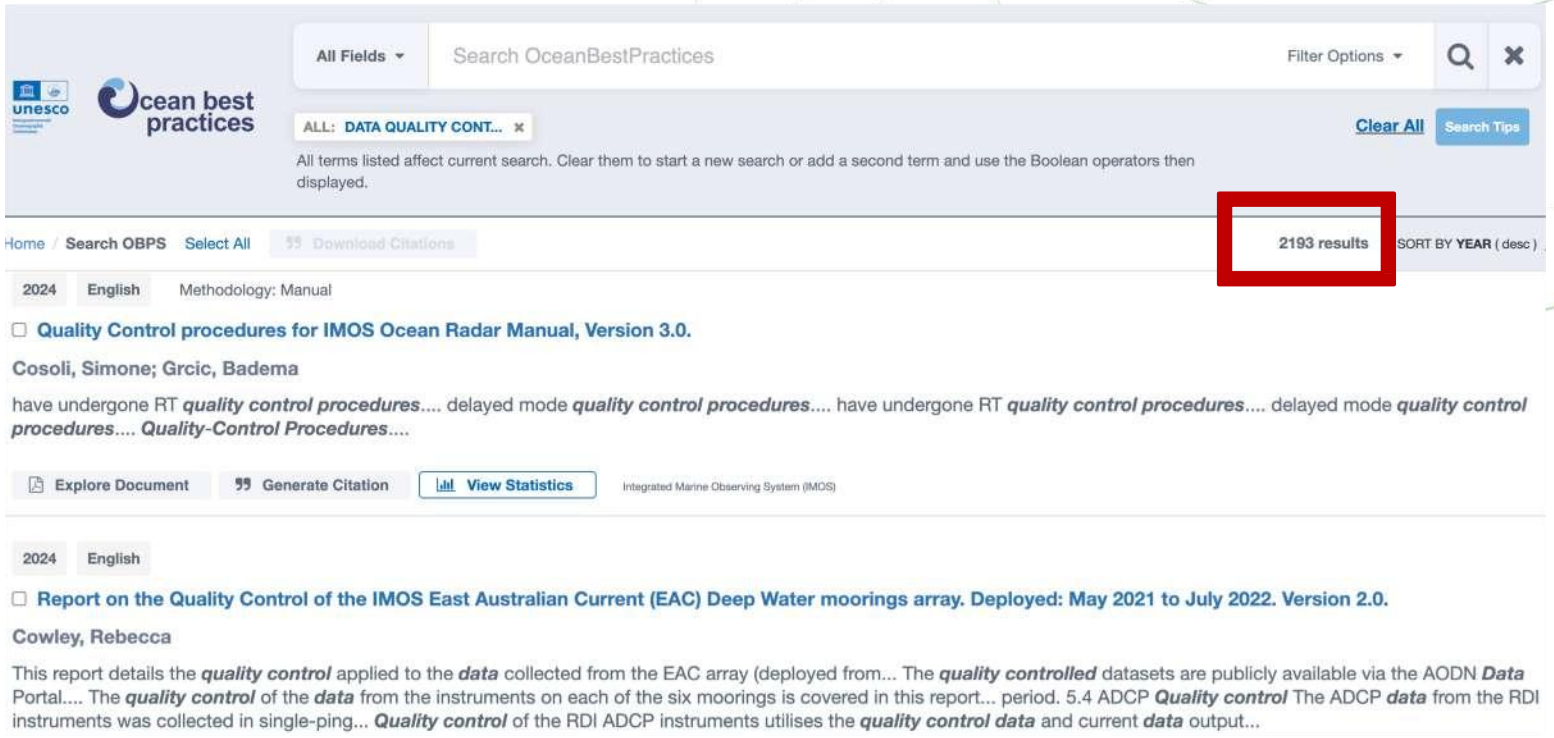
The **data providers**

- are responsible of the adequacy of the sampling strategy to the scope
- must follow specific QA procedures and protocols applied before and during the dataset creation → to maximize the quality of gathered data
- have responsibilities to document through metadata the sensors and the methodologies adopted (platform, instrument type, sensor's accuracy, calibration info)

QA/QC protocols and Standard Operating Procedures

Best Practice (BP): a methodology that has repeatedly produced superior results relative to other methodologies with the same objective and has been adopted and used by multiple organizations or communities (Simpson et al., 2017)

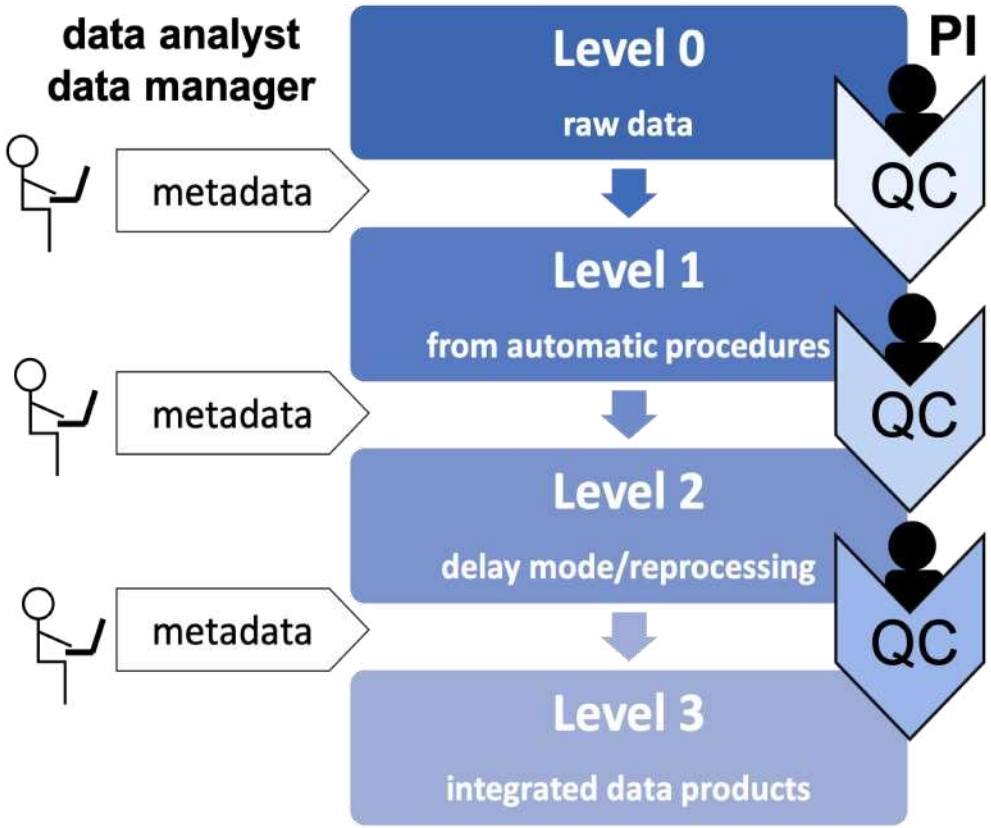
→ sustained QA and QC practices ensure credibility and value to both producers and users



Quality Control (QC)

- Many decisions today are increasingly data-driven and the importance of high quality data cannot be overstated → no one wants to make decisions based upon incorrect or misleading data
- QC is generally presumed to improve data quality, but many chices that influence data quality must be made before data arrive (i.e. adopting a flagging scheme)
- QC is vital for data reuse: without QC data from different sources cannot be ingested in data infrastructures and combined to gain value
- Scientific, analytical and statistical evaluations must determine if data present adequate quality to support the **intended data usage**, resulting in labeling each numerical value with a Quality Flag (QF) and **avoid modifying the original data record**
- QFs ensure that the quality of the data is apparent to the user, who holds suffcient information to decide the suitability for a specific task applying the proper data filtering

Data Processing Levels (DPL)



Latency

Real Time

Near Real Time

Delay Mode

Reprocessing

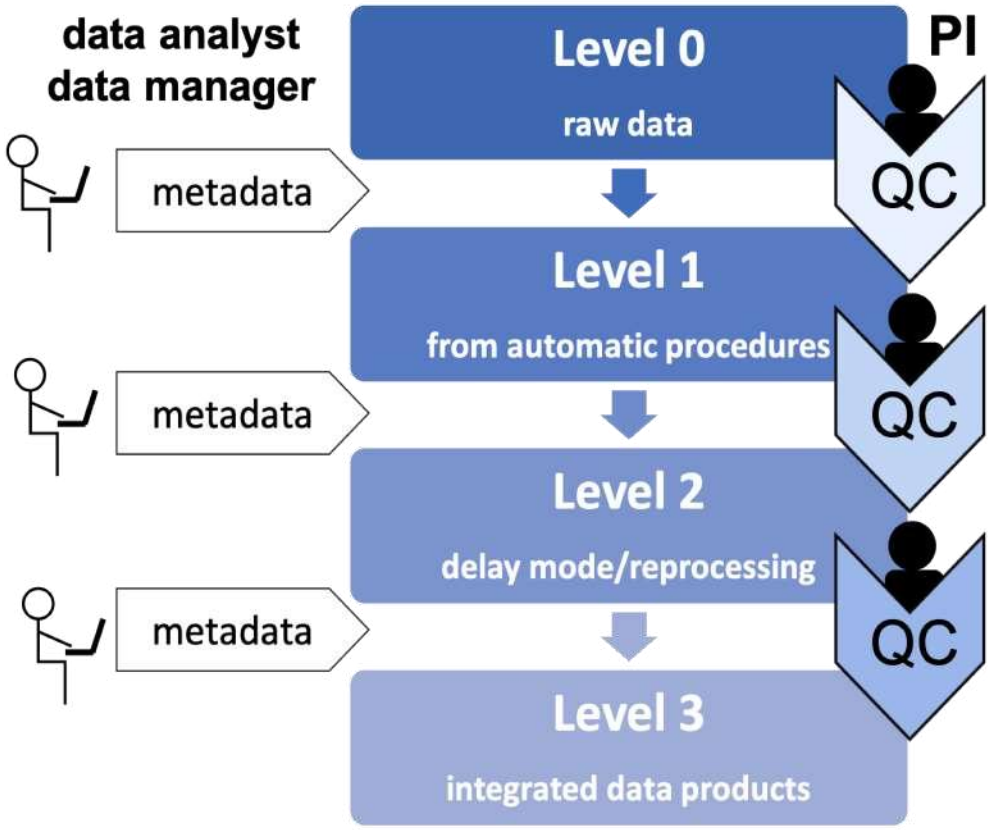
Purpose

- ocean prediction
- early warning

- ocean state assessment
- climate studies

different data versions and duplicates could generate issues in data infrastructures if not well documented

Data Processing Levels (DPL) and QC



Latency

Real Time

Near Real Time

Delay Mode

Reprocessing

QC of the data generated from an observing system can occur on several time scales:

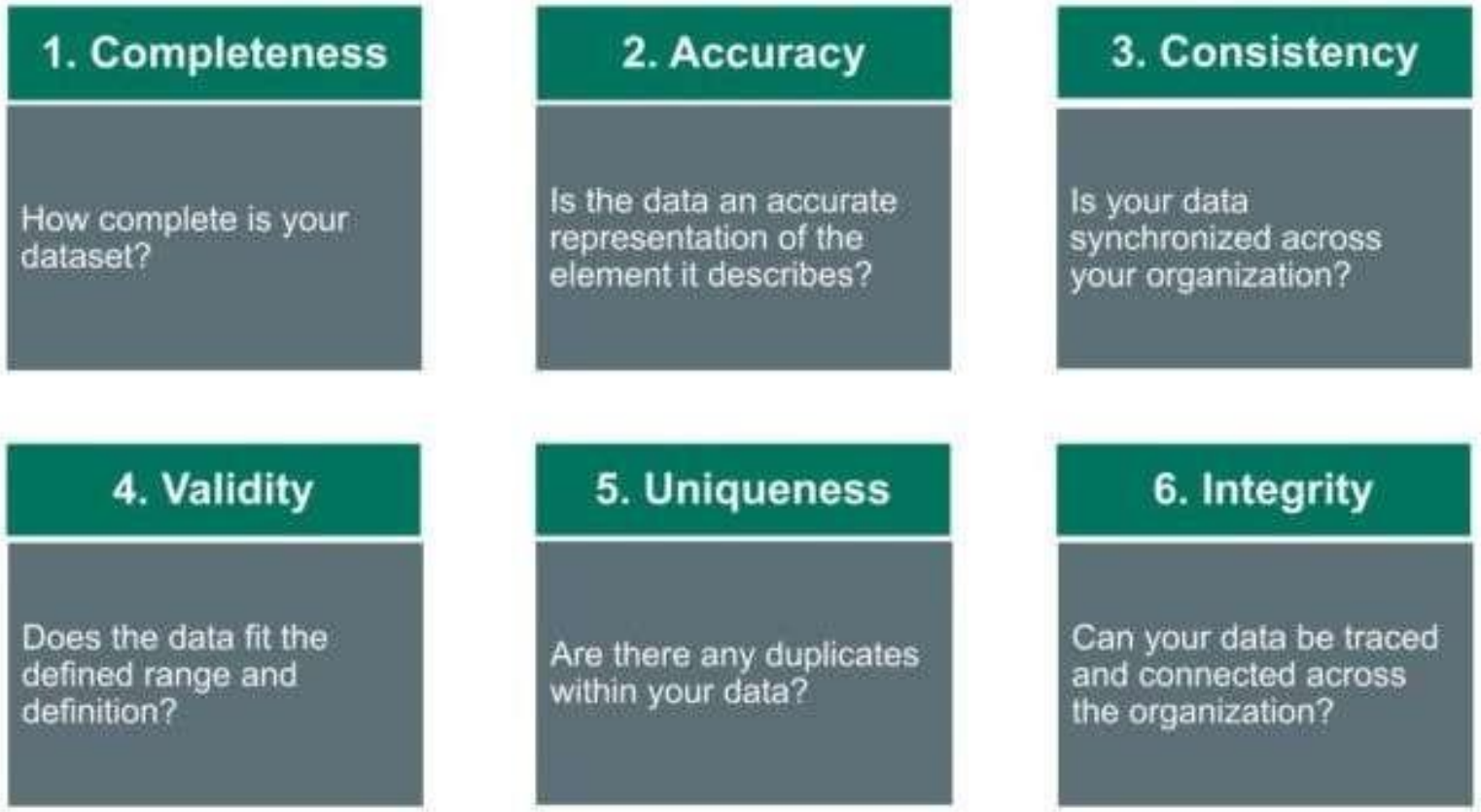
QC operating on the most recent data points, depending on the tests employed (i.e. QARTOD)

QC evaluating recent data points by using subsequent values in the time series

Best QC performed to improve data archives

It may be conducted at infrequent intervals. It can use multiple variables and relationships without regard to timely delivery of data

Dimensions of data quality



Data quality checks determine metrics that address both quality and integrity

https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality

Data processing steps

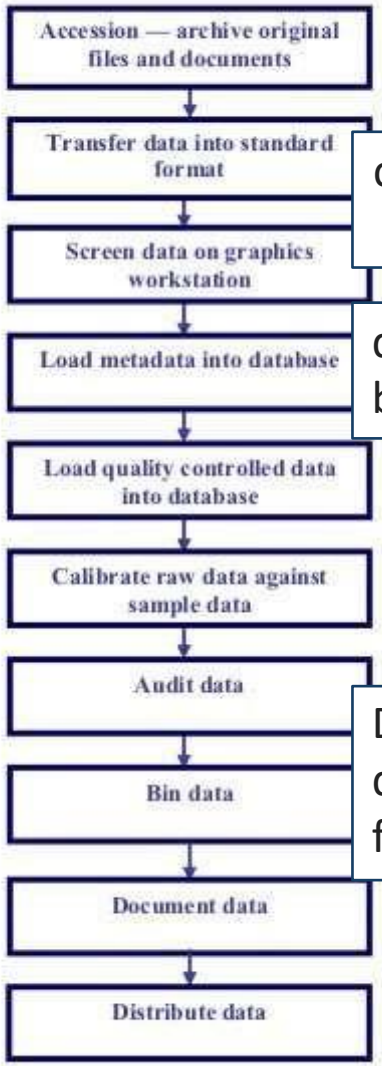
- data must be first securely archived in their original form along with any associated documentation (log sheets, cruise reports) before being processed
- QC practices include:
  - **data integrity checks** → data has been correctly transmitted and stored (correct data types, formats, mandatory fields, null values, missing values, duplicates, syntax as required by each field)
  - **data validity checks** → value range checks, spikes and outliers checks, neighbor checks, climatology checks)
- QC highly depend on the data thematic (physics, chemistry, biology), sensor type and the available amount of time for the analysis
- the processing procedure varies depending on whether the data are
  - **discrete samples** (e.g. CTD bottle, zooplankton net, optics rig deployment)
  - **profiles** (e.g. profiling floats, XBT, CTD)
  - **time series** (moorings, fixed platforms)
  - underway data (underway instruments take continuous measurements of sea surface data e.g. salinity, temperature, transmissivity and fluorescence)

Data processing steps

discrete samples

info about measurement, sample processing or analytical technique, mean values standard units

CTD



calibrations applied

date and time, position, CTD type, bottle information and rig geometry

Data samples are usually averaged to one decibar for shallow casts or to two decibars for deep casts

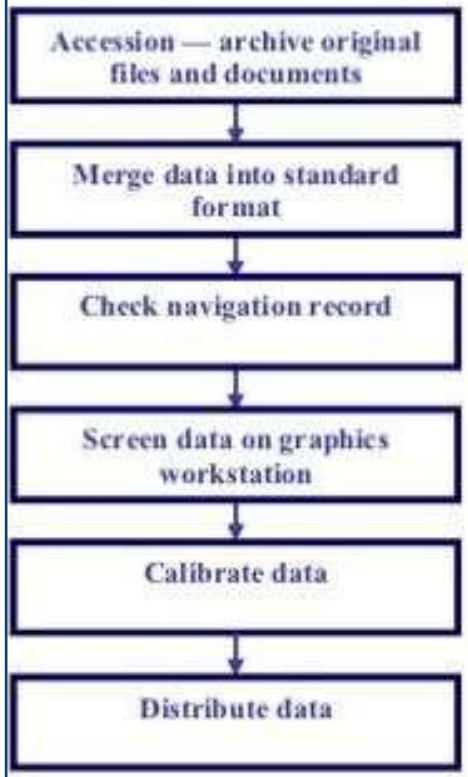
https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/data_processing_steps/

Data processing steps

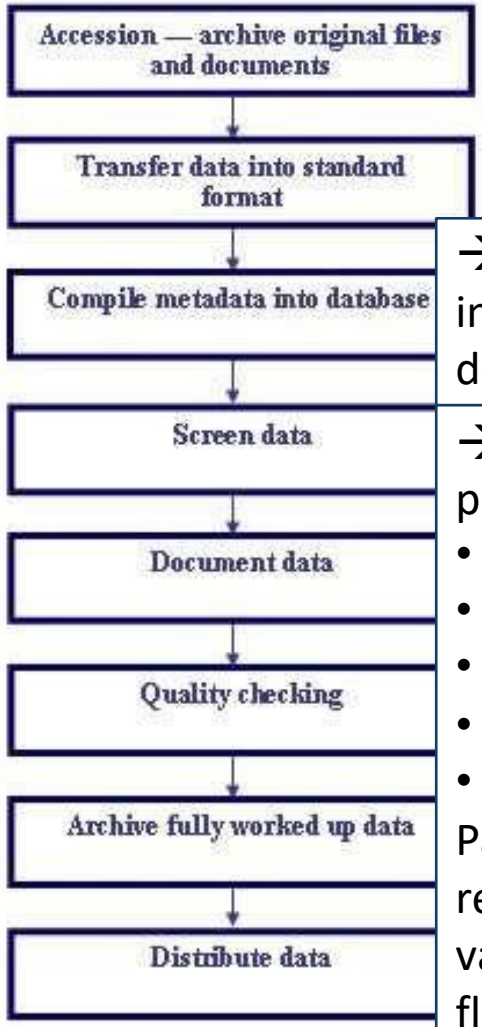
underway data

continuous underway data set includes:

- navigation from the ship
- data from the meteorological package
- data from the thermosalinograph
- data from other instruments attached to the pumped water supply (e.g. fluorometer and transmissometer)



moored instrument data



→ collection date and times, mooring position, instrument type, instrument depth and sea floor depth

→ show a map of mooring positions

plot time series:

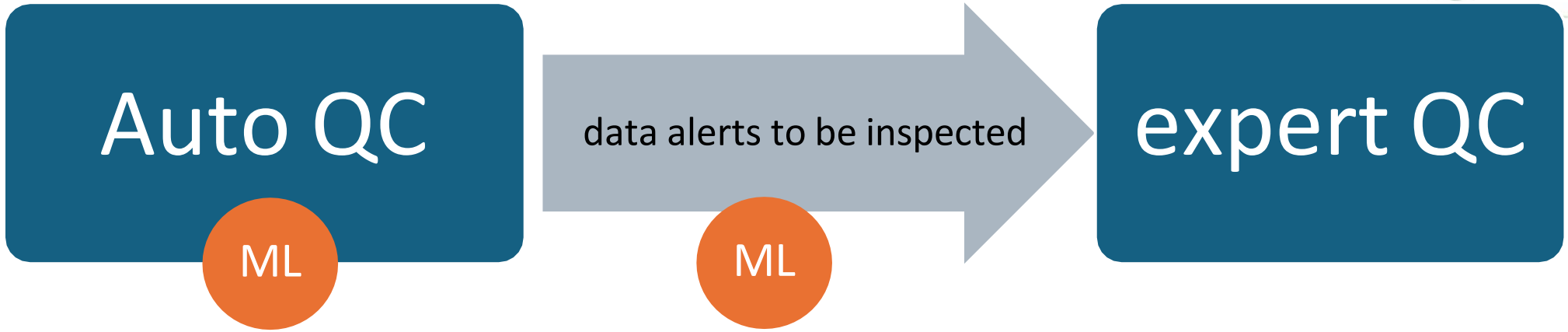
- current meter data
- water level recorder data
- transmissometers
- nutrient analysers
- pressure recorders

Parameters can be plotted concurrently and records from different instruments compared. Data values are NOT changed or removed, but may be flagged if they appear suspect.

https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/data_processing_steps/

automatic vs manual Quality Control

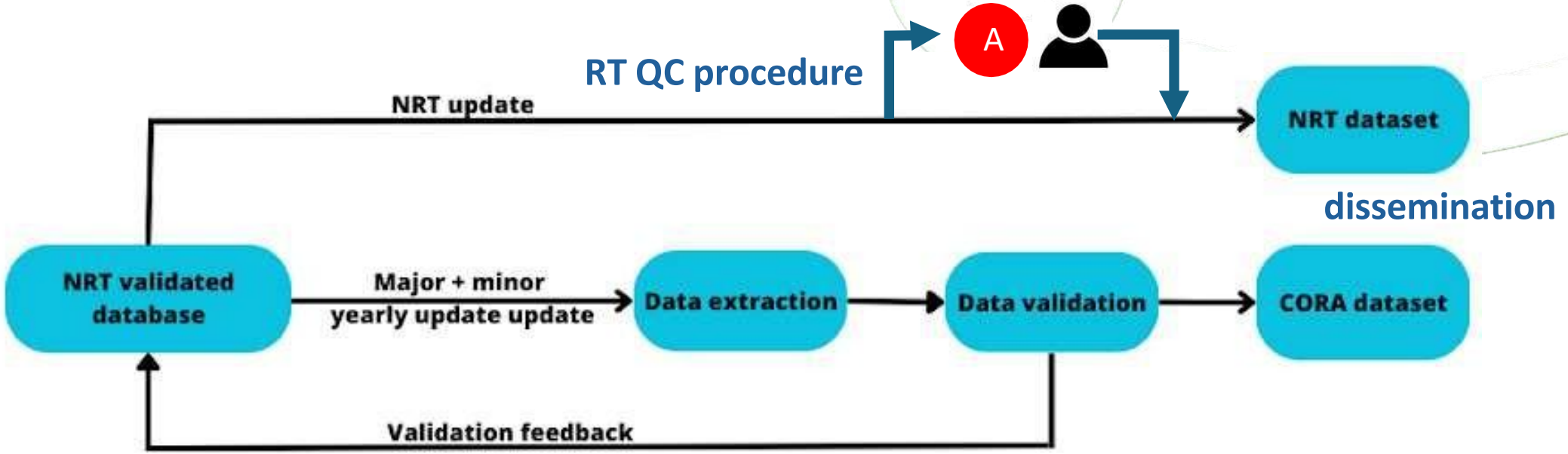
- QC might require both automation and human intervention
  - Visual screening → small datasets
  - Automatic QC (RT, NRT) → algorithms
  - Automatic + manual/visual/expert QC (DM) → tools with graphical interfaces (i.e. OceanDataView)
  - Machin Learning can assist auto QC procedures or can help in reducing the work load of experts during visual inspection of alerts



CMS In situ TAC QC procedures

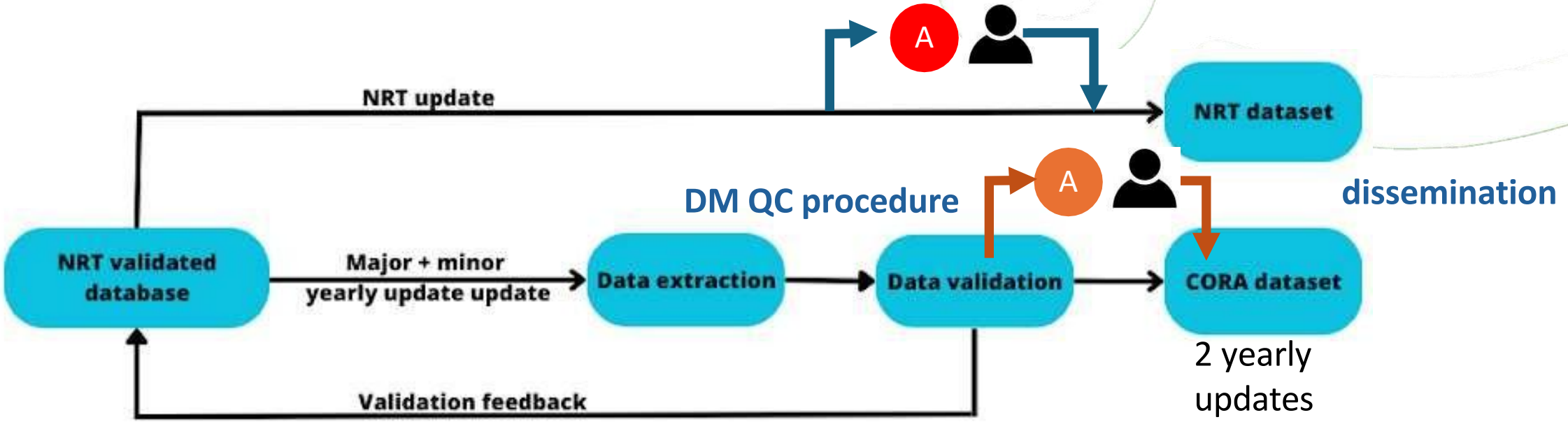
- **NRT product** gathers data from various types of platforms. Depending on the observation and transmission networks, the level of QC is highly variable, e.g. quite low for data arriving through the Global Transmission System (GTS)
- **RT QC procedures** in the NRT production to detect as much erroneous data as possible (bad position, frozen value, spikes, unrealistic vertical gradients, density inversion,…)
- MinMax test to check if the measurement lies inside a range of values obtained from historical databases → **global range and regional range** tests

- data outside the validity interval raise an alert
- the alert (A) is submitted to the NRT QC operator for further inspection and decision



CMS In situ TAC QC procedures

- DM product undergo a more rigorous validation process compared to the methods used in NRT data processing.
- Range check, stuck value test, ascending immersion check, duplicate levels test, **DM MinMax test**, spike test, depth vs press field check, duplicates, XBT correction.



Data Flagging Schemes

- Data are evaluated by visual inspection or using QC tests, and the results are indicated using flags in the data files
- QFs ensure that the quality of the data is apparent to the user, who holds sufficient information to decide the suitability for a specific task applying the proper data filtering
- Multiple flagging standards exist and there is flexibility within them → mappings of QFs from one scheme to another (ODV tool)
- Operators may incorporate additional flags into metadata records to provide more detailed information to assist with troubleshooting

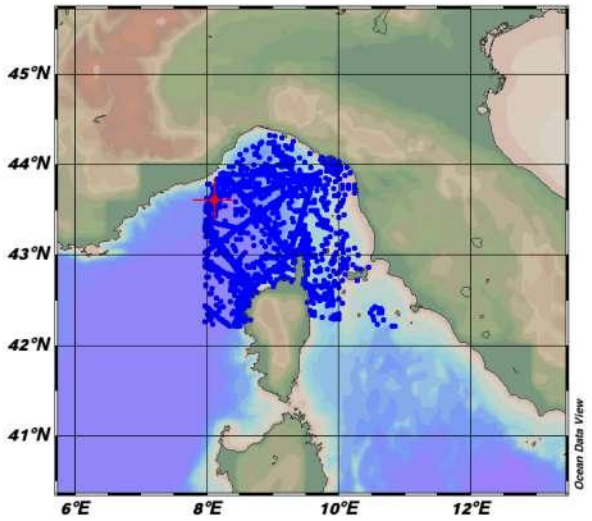
Quality Flags: SeaDataNet scheme

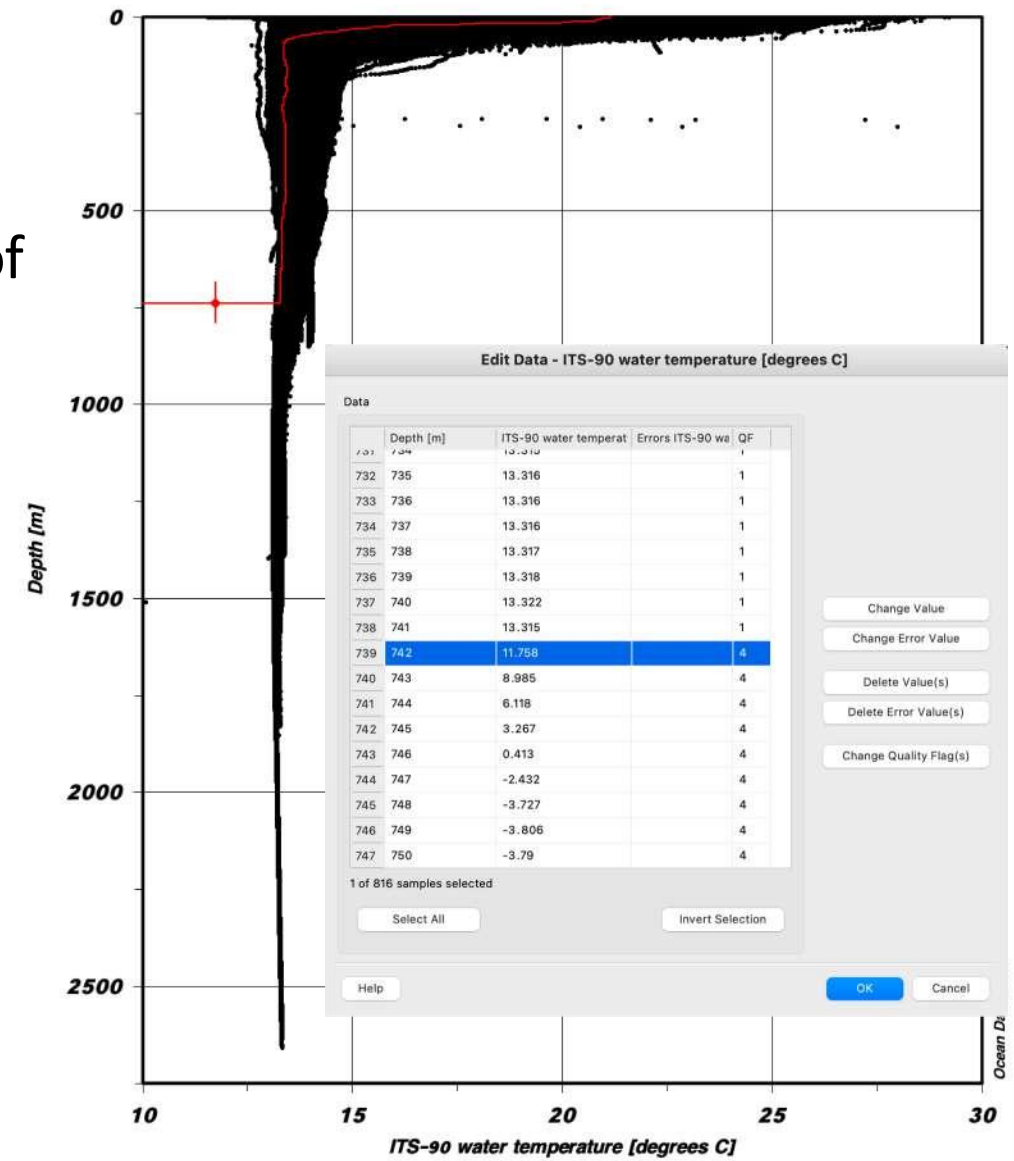
- QFs are labels associated to each measurement indicating its quality level that the user can use to filter data according to the needs
- QC flag scale is available in the SeaDataNet Common Vocabularies as list L20

→ adopted by Argo, CMS In situ TAC, EMODnet

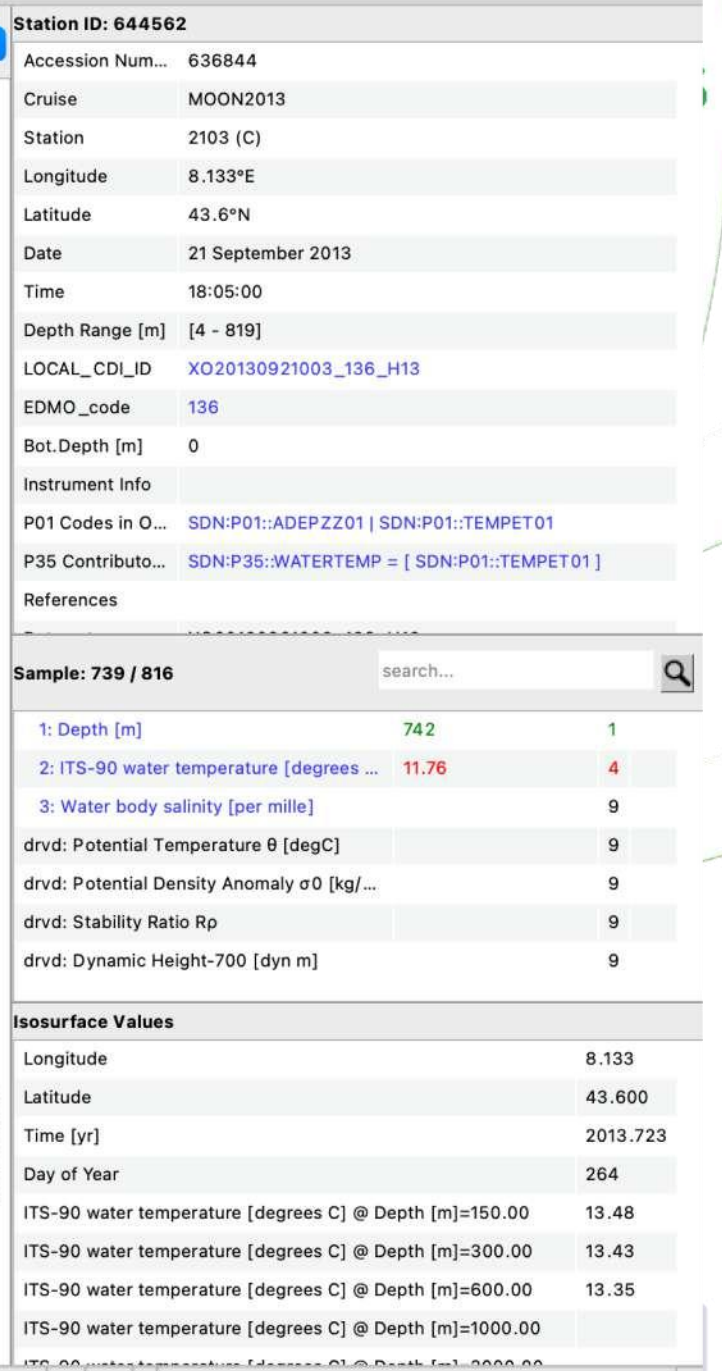
Key	Entry Term	Abbreviated term	Term definition
0	no quality control	none	No quality control procedures have been applied to the data value. This is the initial status for all data values entering the working archive.
1	good value	good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	probably good value	probably_good	Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	probably bad value	probably_bad	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	bad value	bad	An obviously erroneous data value.
5	changed value	changed	Data value adjusted during quality control. Best practice strongly recommends that the value before the change be preserved in the data or its accompanying metadata.
6	value below detection	BD	The level of the measured phenomenon was too small to be quantified by the technique employed to measure it. The accompanying value is the detection limit for the technique or zero if that value is unknown.
7	value in excess	excess	The level of the measured phenomenon was too large to be quantified by the technique employed to measure it. The accompanying value is the measurement limit for the technique.
8	interpolated value	interpolated	This value has been derived by interpolation from other values in the data object.
9	missing value	missing	The data value is missing. Any accompanying value will be a magic number representing absent data.
A	value phenomenon uncertain	ID_uncertain	There is uncertainty in the description of the measured phenomenon associated with the value such as chemical species or biological entity.

Ocean Data View (ODV) software package for the interactive exploration, analysis and visualization of oceanographic and other geo-referenced profile, time-series, trajectory or sequence data





https://odv.awi.de/

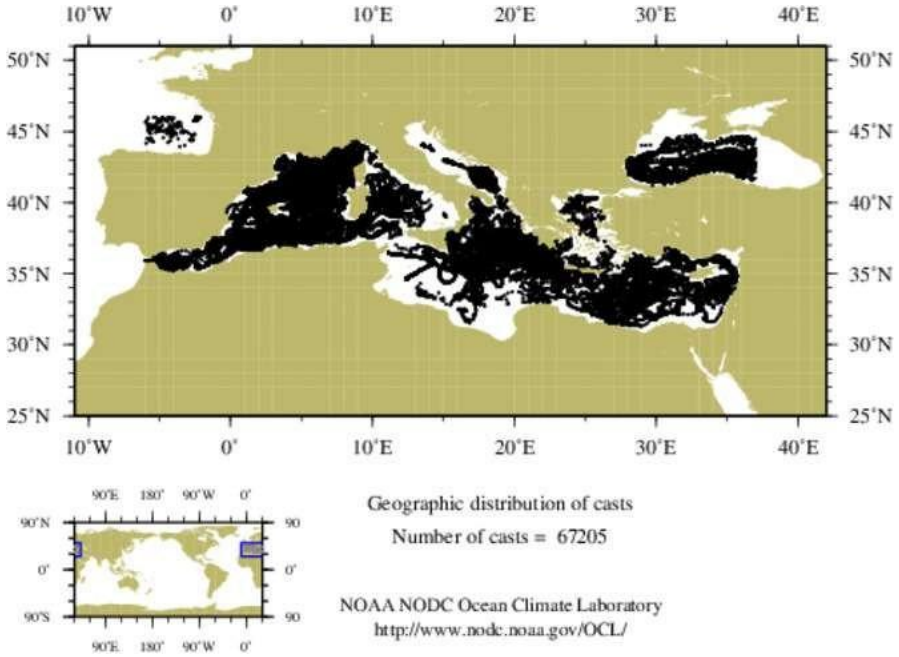


World Ocean Database and World Ocean Atlas

- WOD is a collection of uniformly formatted, quality controlled, publicly available ocean profile data → result of more than 20 years of coordinated efforts to incorporate data from institutions, agencies, individual researchers, and data recovery initiatives into a single database
- each major WOD release is associated with a concurrent release of the World Ocean Atlas (WOA) and contains final quality control flags used in the WOA, which includes manual as well as automated steps
- WOA is a collection of objectively analyzed, quality controlled temperature, salinity, oxygen, phosphate, silicate, and nitrate means based on profile data from the WOD
- WOA objectively analyzed and statistical data at 102 standard depth levels of the World Ocean.
  - Temperature and salinity data are available on 1 deg and 0.25 degree lat-lon grids
  - time periods: 1955-1964, 1965-1974, 1975-1984, 1985-1994, 1995-2004, 2005-2014, 2015-2022 and
  - three 30-year 'climate normal' periods: 1971-2000, 1981-2010, and 1991-2020

https://www.ncei.noaa.gov/products/world-ocean-database
https://www.ncei.noaa.gov/products/world-ocean-atlas

WOD and WOA data products



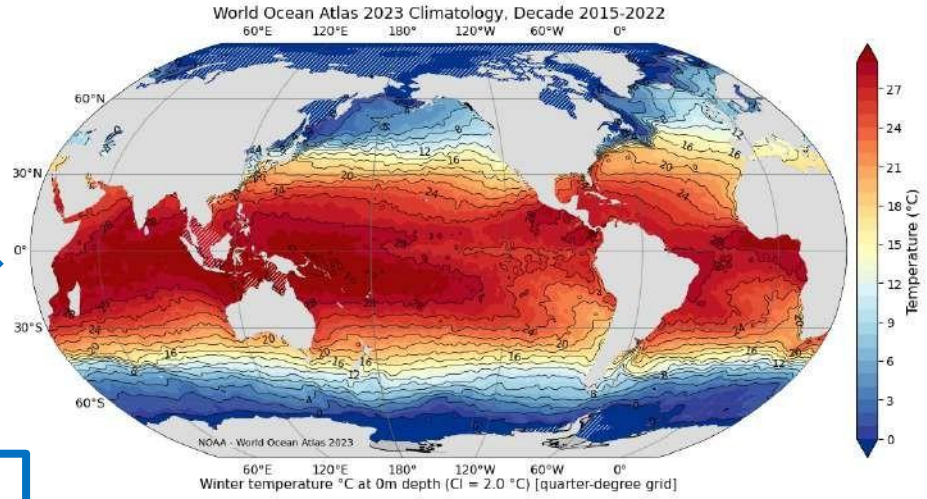
COPY OF YOUR DATABASE SEARCH CRITERIA:

OBSERVATION DATES:	Year from 2020 to 2023; Month from 1 to 12; Day from 1 to 31
GEOGRAPHIC COORDINATES:	*Longitude* from -6.0000 to 37.0000; *Latitude* from 46.0000 to 30.0000
DATASET:	OSD,CTD,XBT,MBT,PFL,GLD
MEASURED VARIABLES (extract):	Temperature

QC

WOD QC procedure is targeted to the generation of WOA

https://www.ncei.noaa.gov/access/world-ocean-atlas-2023f/



World Ocean Atlas 2023: Temperature

Grids	○ 1° ◉ 1/4°
Time Span	2015-2022 years
Field Name	Objectively analyzed mean
Time Period	Winter
Depth	Surface

https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html

Quality Flags: World Ocean Database scheme

(1) FLAGS FOR ENTIRE CAST (AS A FUNCTION OF VARIABLE)	
0	accepted cast
1	failed annual standard deviation check
2	two or more density inversions (Levitus, 1982 criteria)
3	flagged cruise
4	failed seasonal standard deviation check
5	failed monthly standard deviation check
6	failed annual and seasonal standard deviation check
7	bullseye from standard level data or failed annual and monthly standard deviation check
8	failed seasonal and monthly standard deviation check
9	failed annual, seasonal and monthly standard deviation check

(2) FLAGS ON INDIVIDUAL OBSERVATIONS	
(a) Depth Flags	
0	accepted value
1	duplicates or inversions in recorded depth (same or less than previous depth)
2	density inversion

(b) Observed Level Flags	
0	accepted value
1	range outlier (outside of broad range check)
2	failed inversion check
3	failed gradient check
4	observed level "bullseye" flag and zero gradient check
5	combined gradient and inversion checks
6	failed range and inversion checks
7	failed range and gradient checks
8	failed range and questionable data checks
9	failed range and combined gradient and inversion checks

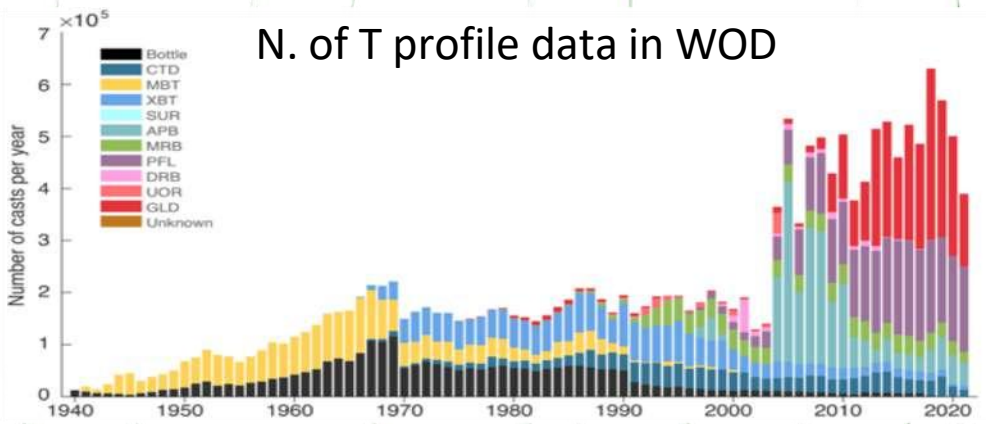
(c) Standard Level Flags	
0	accepted value
1	bullseye marker
2	density inversion
3	failed annual standard deviation check
4	failed seasonal standard deviation check
5	failed monthly standard deviation check
6	failed annual and seasonal standard deviation check
7	failed annual and monthly standard deviation check
8	failed seasonal and monthly standard deviation check
9	failed annual, seasonal and monthly standard deviation check

(d) Biological data flags (applied only to Comparable Biological Value - CBV Taxa code 27)	
0	accepted value
1	range outlier (outside of broad range check)
2	questionable value ("bullseye flag")
3	group was not reviewed
4	failed annual standard deviation check

The WOD contains the data on the originally measured depth levels (observed) and also interpolated to standard depth levels (WOA) to present a more uniform set of iso-surfaces for oceanographic and climate studies

https://www.ncei.noaa.gov/access/world-ocean-database/CODES/definition-quality-flags.html

WOD Quality Control example



QC of T data is a major task: difficulty related to

lack of data and metadata (for some areas) upon which to base statistical checks

- certain empirical criteria were applied , and
- subjective judgment was used as part of the last processing step

Individual T data, and in some cases entire profiles or all profiles for individual cruises, have been flagged and not used further because these data produced features **non representative** or questionable.

The flags mark either individual measurements or entire profiles which were not used in the next step of the procedure - either interpolation to standard depth levels for observed level data or calculation of statistical means in the case of standard depth level data.

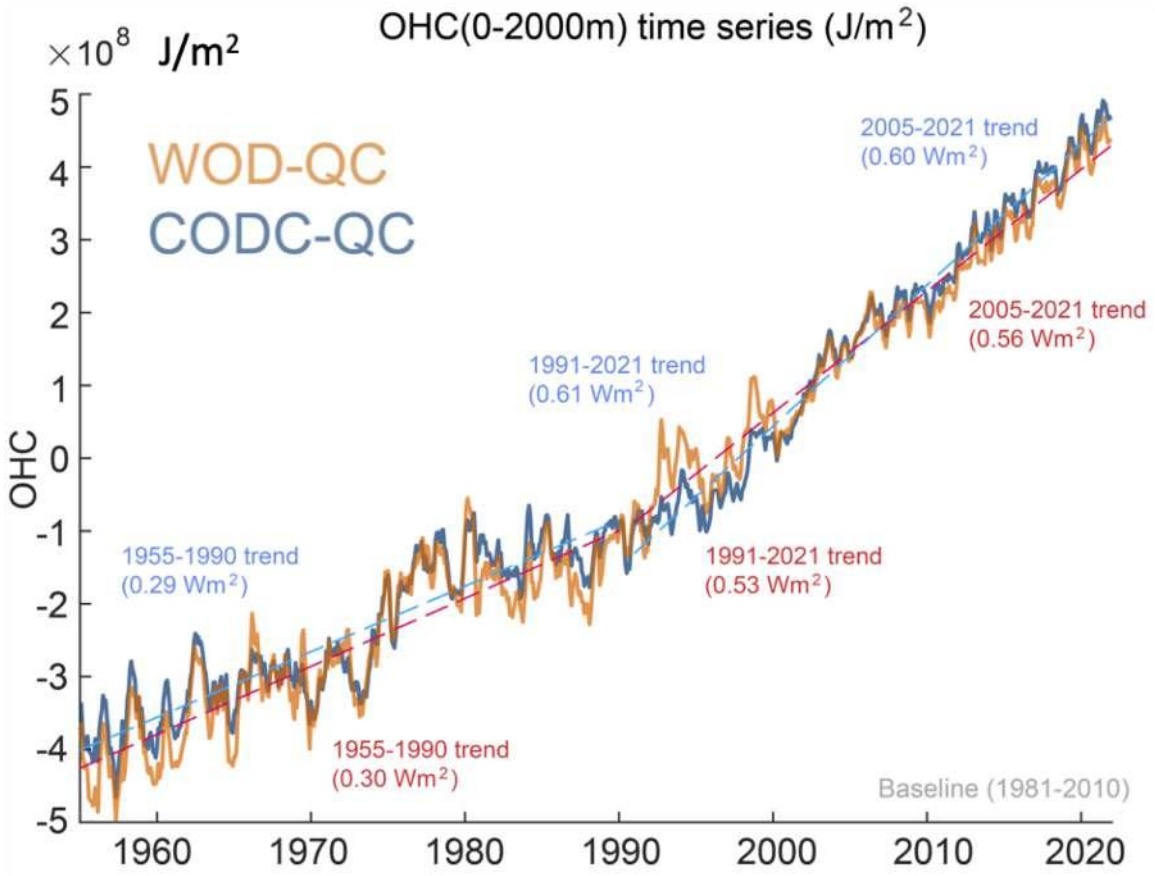
file:///Users/simonasimoncelli/ITINERIS/training/data_management/noaa_60599_WOA_T.pdf

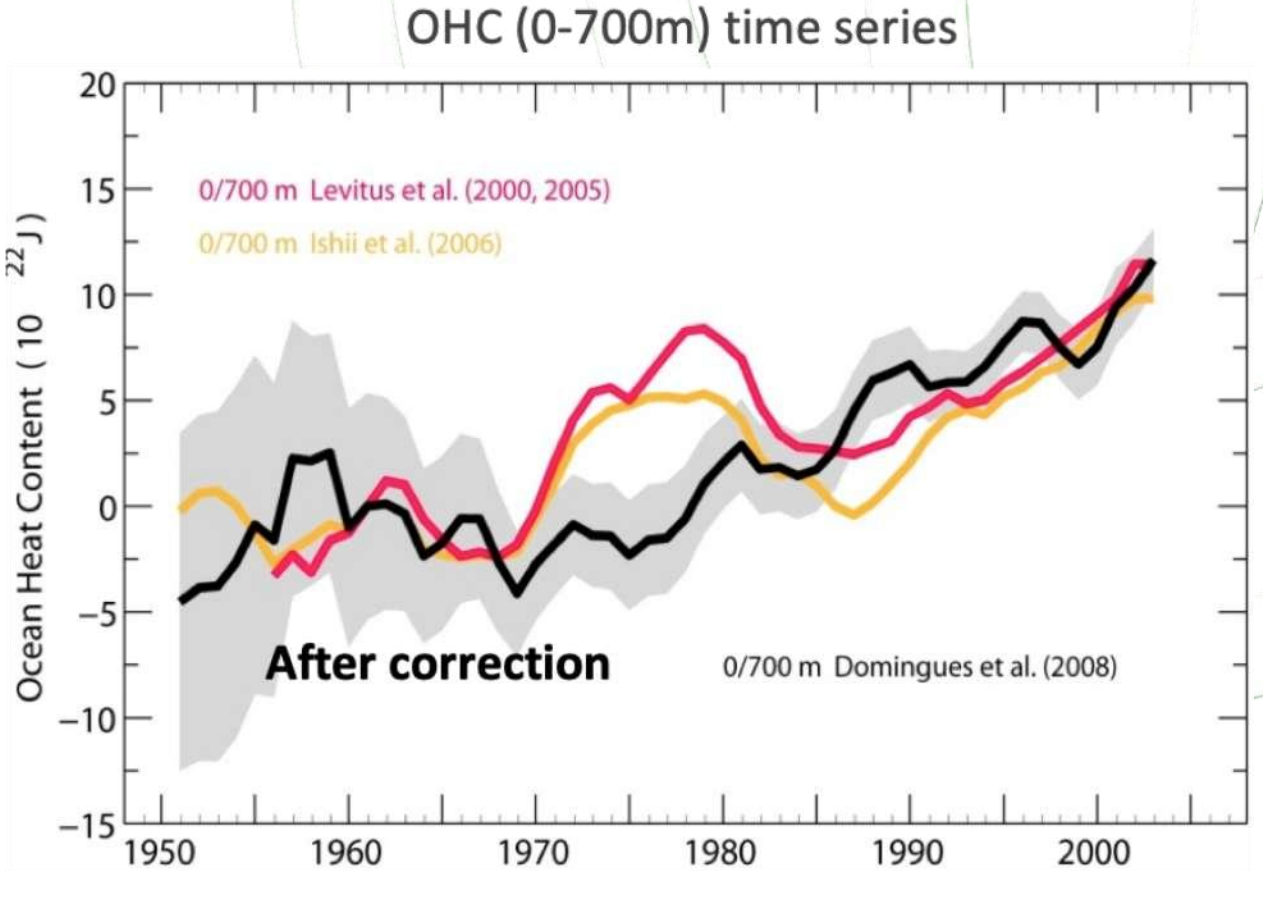
WOD Quality Control example

- **Duplicate elimination** - data are received from many sources, sometimes the same data set is received more than once but with slightly different time and/or position and/or data values, and hence are not easily identified as duplicate stations
- **Range check** - checking whether a value is within preset min and max values as a function of depth and major ocean basin
- **Gradient check** - excessive vertical gradients occur in the data both in terms of positive and negative gradients
- **Statistical checks** - data at each standard depth level, were averaged within 5 deg lat-lon squares to produce a record of the n. of obs, *mean*, and *standard deviation* in each square computed for the annual, seasonal, and monthly compositing periods → validity intervals estimated based on *mean +/- N*std*
- **Static stability check**
- **Subjective flagging -** bull's-eyes or spatial gradients features indicate that some of them were due to profiles

file:///Users/simonasimoncelli/ITINERIS/training/data_management/noaa_60599_WOA_T.pdf

‘Climate quality’ ocean database





Impact of QC on OHC 0-2000 m:

~8% trend difference from 2005-2021

Impact of instrumental bias on OHC 0-700 m:

~50% trend difference from 1970-2000

AUTOMATED QC: IQuOD example

Good et al. (2023) developed a methodology to assess the performance of AutoQC tests and define fit for purpose combinations of them:

- **60 AutoQC checks** and a WOD data reader (wodpy) have been coded in Python
- open code repository (MIT license) → it can be used by anyone
- It has been used to benchmark the AutoQC checks and make recommendations for which to use to QC historical data
- performance has been benchmarked against three reference datasets of certified quality with the final aim to recommend an optimal set of tests
- AutoQC checks are being applied to WOD data and will be used in a future release of the IQuOD dataset

International Quality-controlled Ocean Database

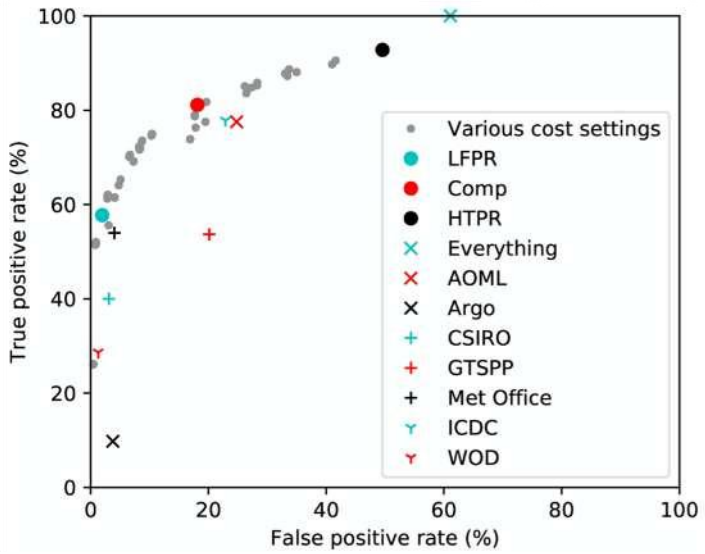
https://github.com/IQuOD

open-source collaborative software infrastructure

Popular repositories

AutoQC — Public
A testing suite for automatic quality control checks of subsurface ocean temperature observations
Python 27 15

wodpy — Public
A package to consume WOD format data.
Python 13 8



Evaluation of different QC systems

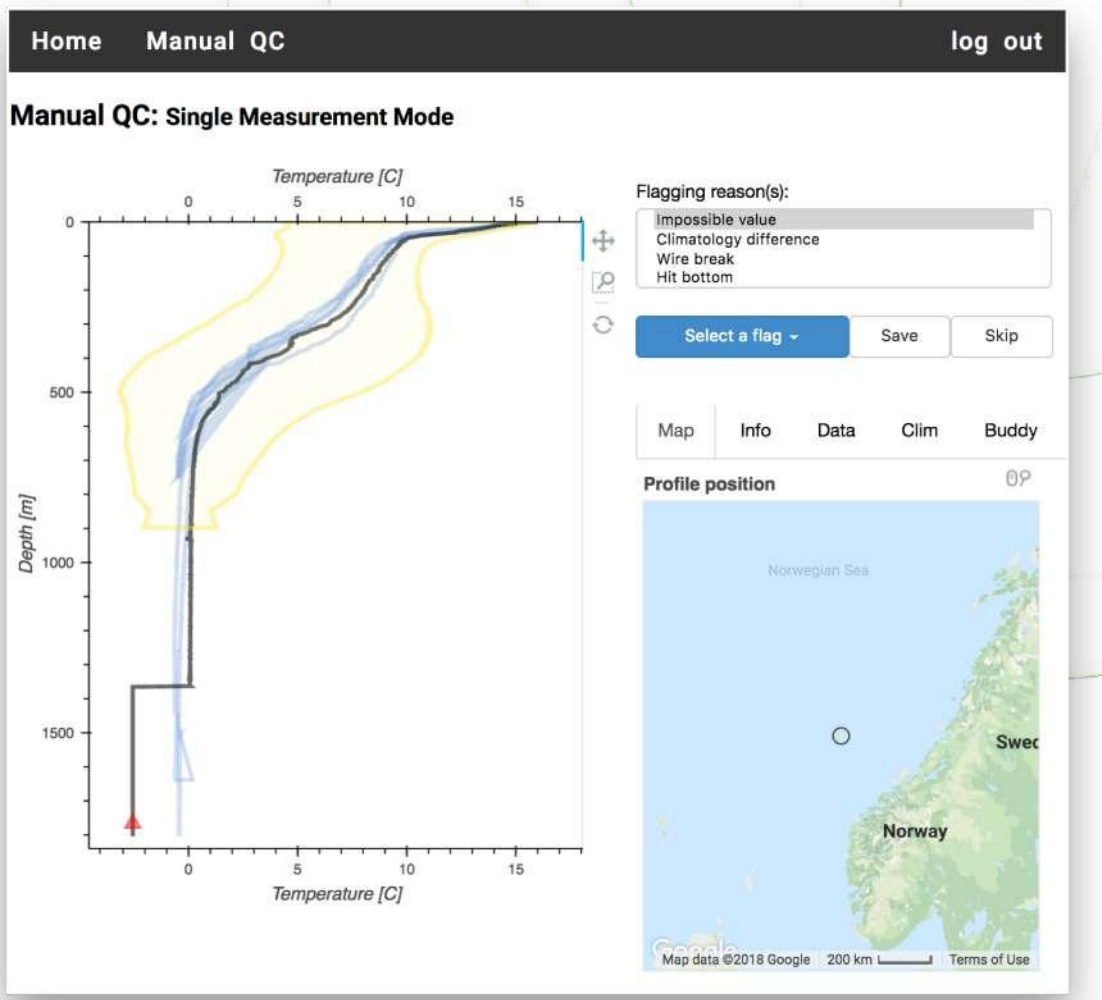
Benchmarking metrics

- True Positive Rate % **TPR**
- False Positive Rate% **FPR**

general aim: to maximize the TPR and minimize the FPR, but different applications might have different requirements

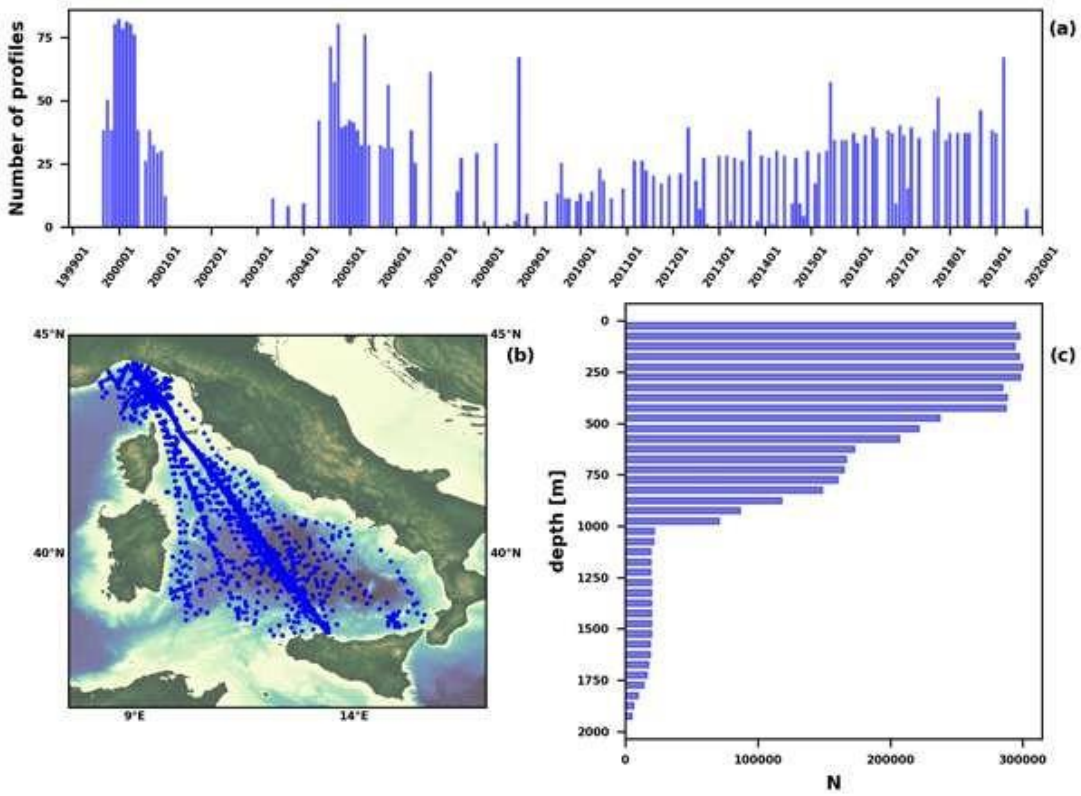
Expert QC & Machine Learning

- Machine Learning Approach to QC Oceanographic data
- Web App to integrate experts around the world https://expertqc.castelao.net
- To improve efficiency of the manual QC, the experts are paired with an interactive learning schema of Machine Learning to combine the high skill of the human with the speed of the machine
- Twofold return to the community: Expert QC flags on the WOD and public access to the calibrated open source CoTeDe https://github.com/castelao/CoTeDe



Castelão (2021) https://doi.org/10.1016/j.cageo.2021.104803

XBT data reprocessing



- full metadata upgrade
- calibration, correction and uncertainty information
- data dissemination through ERRDAP server

Simoncelli et al. (2024) https://doi.org/10.5194/essd-16-5531-2024

Take home messages

- Data might have varying quality due to differing monitoring program requirements in terms of accuracies, capabilities and resources;
- A QA/QC strategy must be pre-defined from the data provider consulting the available community best practices (OBPs repository);
- QC processing depends on the data type, sensor type, thematic, sea region and the delivery time/mode (RT, DM), can be automatic, manual or a combination of both;
- Secondary QC analysis can be conducted for products generation (climatology, Ocean Heat Content estimation) or application (assimilation in numerical model) and different methodologies can give different results;
- It is the responsibility of the users to understand and appropriately utilize data of varying quality, and operators must provide guidance by documenting and publishing their QA/QC procedures;
- The sharing/publication of QA/QC procedures in GitHub or zenodo are becoming a common practice, in line with the Open Science principles.

References

Bushnell et al. (2019). Quality Assurance of Oceanographic Observations: Standards and Guidance Adopted by an International Partnership. Front. Mar. Sci. 6:706. https://doi.org/10.3389/fmars.2019.00706

Cummings, J.A. (2011). Ocean Data Quality Control. In: Schiller, A., Brassington, G. (eds) Operational Oceanography in the 21st Century. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0332-2_4

Good et al. (2023) https://doi.org/10.3389/fmars.2022.1075510

Locarnini et al. (2024). World Ocean Atlas 2023, Volume 1: Temperature. A. Mishonov, Technical Editor. NOAA Atlas NESDIS 89, 52 pp. https://doi.org/10.25923/54bh-1613

Simoncelli et al. 2022. A collaborative framework among data producers, managers, and users. https://doi.org/10.1016/B978-0-12-823427-3.00001-3

Tan et a. (2021). Quality control for ocean observations: From present to future. Science China Earth Sciences, 64, https://doi.org/10.1007/s11430-021-9846-7

Simoncelli et al. (2024) Reprocessing of eXpendable BathyThermograph (XBT) profiles from the Ligurian and Tyrrhenian seas over the time period 1999–2019 with a full metadata upgrade, Earth Syst. Sci. Data, 16, 5531–5561, https://doi.org/10.5194/essd-16-5531-2024

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani