27-29 May 2024

imdis

International conference on **M**arine **D**ata and **I**nformation **S**ystems

SeaDataNet AISBL

Havforskningsinstituttet

Ifremer

MARIS

MEDIN

National Oceanography Centre

INGV

iMagine

IODE

eosc Blue-Cloud2026

International Conference on Marine Data and Information Systems
IMDIS 2024 - Bergen (Norway), 27-29 May 2024

The International Quality-controlled Ocean Database (IQuOD)

Simona Simoncelli (INGV, Italy)*
R. Cowley (CSIRO, Australia), **Z. Tan** (CAS, China), **R. Killick** (Met Office, UK),
G. Castelão (Scripps, USA), **L. Cheng** (CAS, China), **S. Good** (Met Office UK),
T. Boyer (NCEI, USA), **W. Mills** (University of Colorado USA),
U. Bhaskar (INCOIS, India), **R. Locarnini** (NCEI, USA)

*on behalf of the IQuOD team

simona.simoncelli@ingv.it ORCID: 0000-0003-1283-2798

IQuOD in a nutshell

IQuOD

www.iquod.org

GOAL: to **maximize the quality, consistency and completeness** of the long-term global subsurface ocean temperature database (EOV & ECV) **by developing and implementing an internationally-agreed framework**

co-chairs

Lijing Cheng — **Gui Castelao**

steering team

Task Teams (TT)

- **UNCERTAINTIES**
- **AUTOMATED QUALITY CONTROL**
- **INTELLIGENT METADATA**
- **DUPLICATES**
- **EXPERT QUALITY CONTROL**
- **G**LOBAL **D**ATA **AC**QUISITION CENTER
- **EDUCATION & OUTREACH**

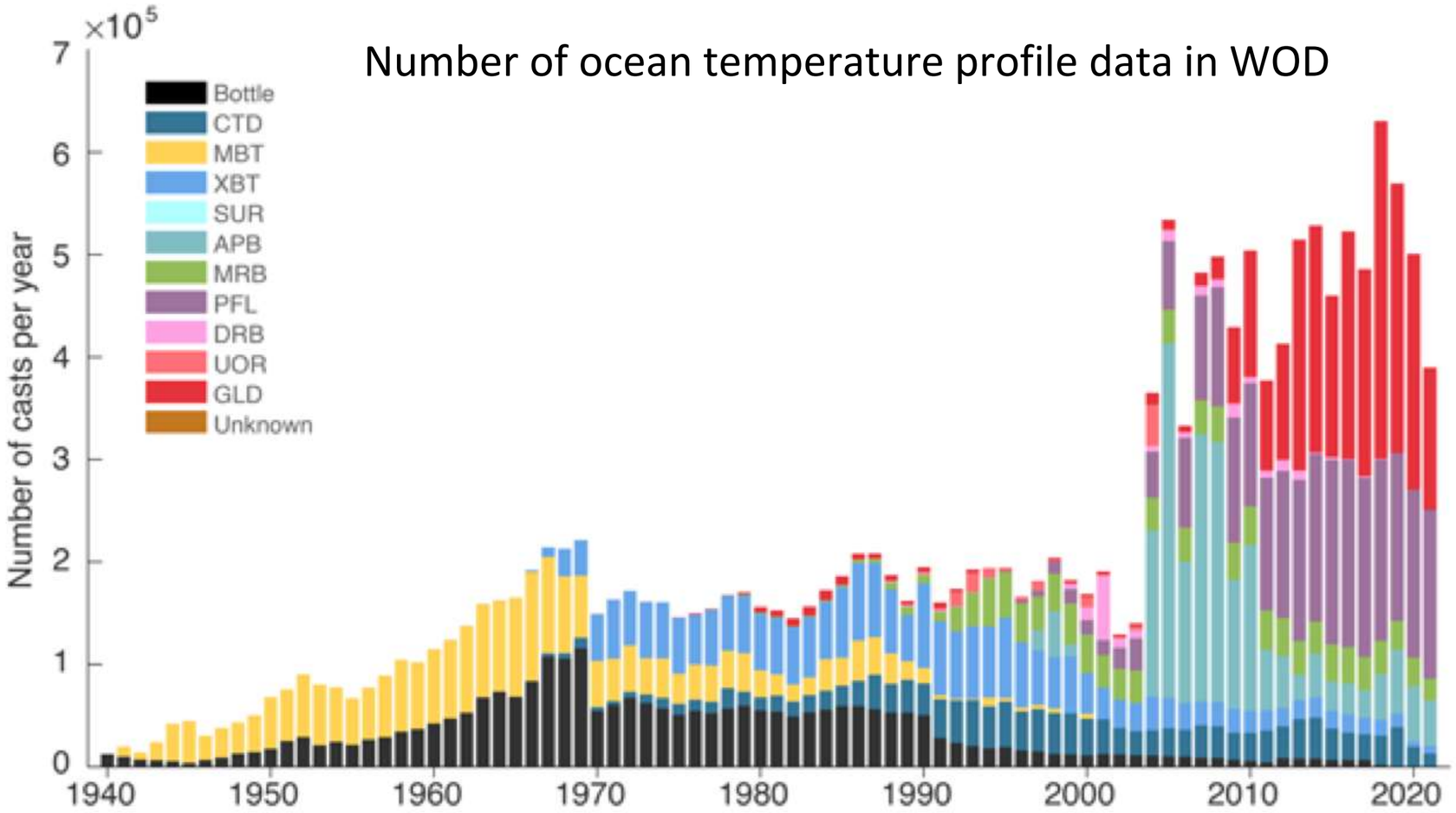
IQuOD's operational structure comprises 17 nation members and a dynamic workforce of 30-50 international members, organised into specialised TTs

CLIVAR SCOR IODE

Close collaboration between experts and users (data quality and management, climate modelers and the broader climate-related community) with support from: CLIVAR, SCOR, IODE

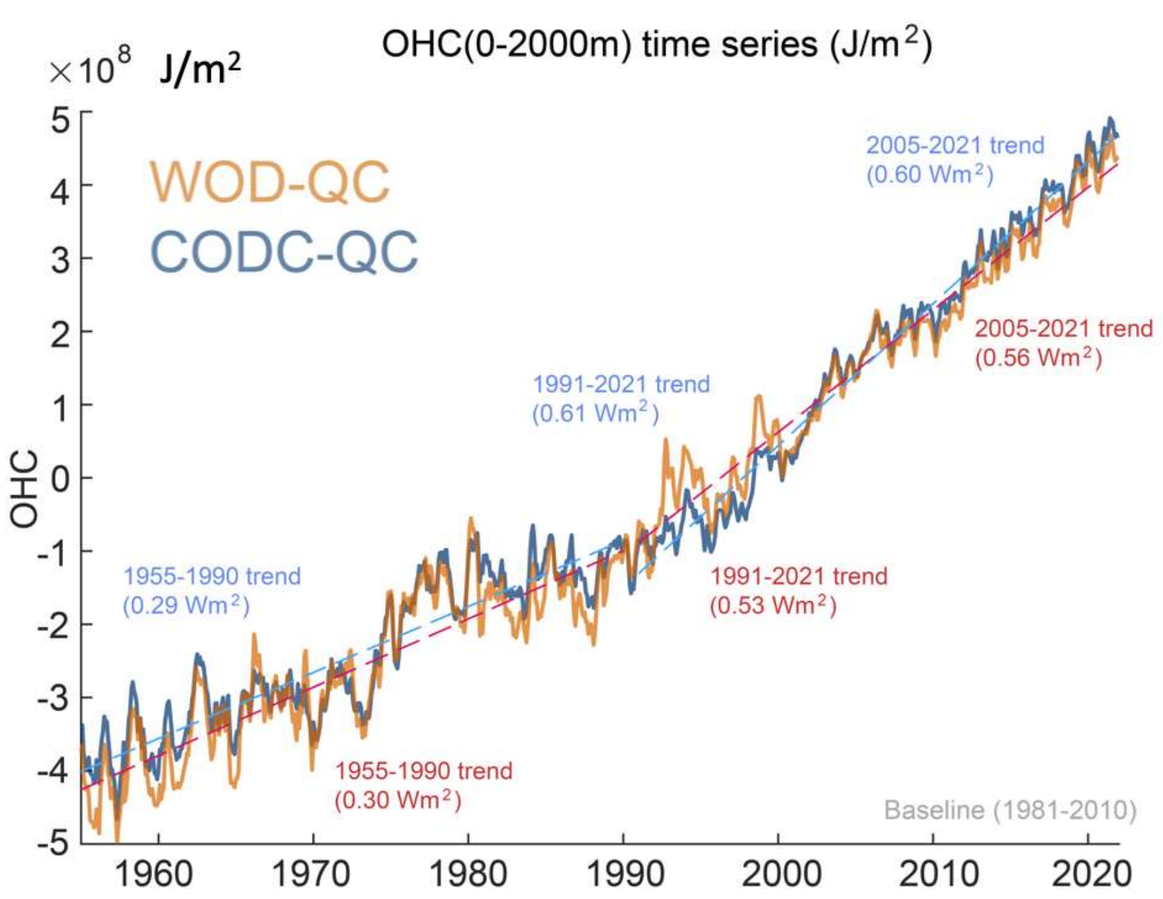
‘Climate quality’ ocean database

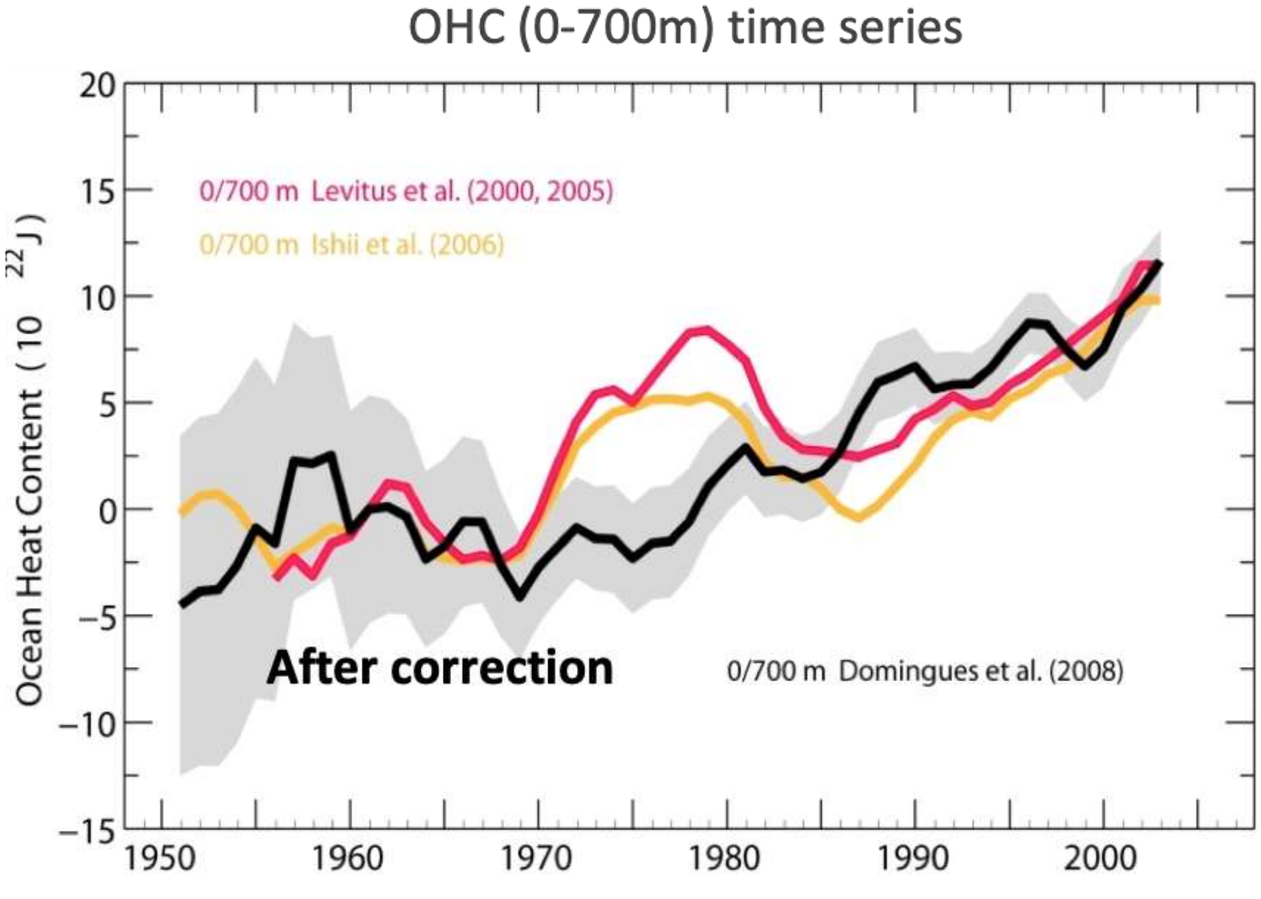
Number of ocean temperature profile data in WOD



Tan et al. (2023) https://doi.org/10.1016/j.dsr.2022.103961

‘Climate quality’ ocean database





Impact of QC on OHC 0-2000m:

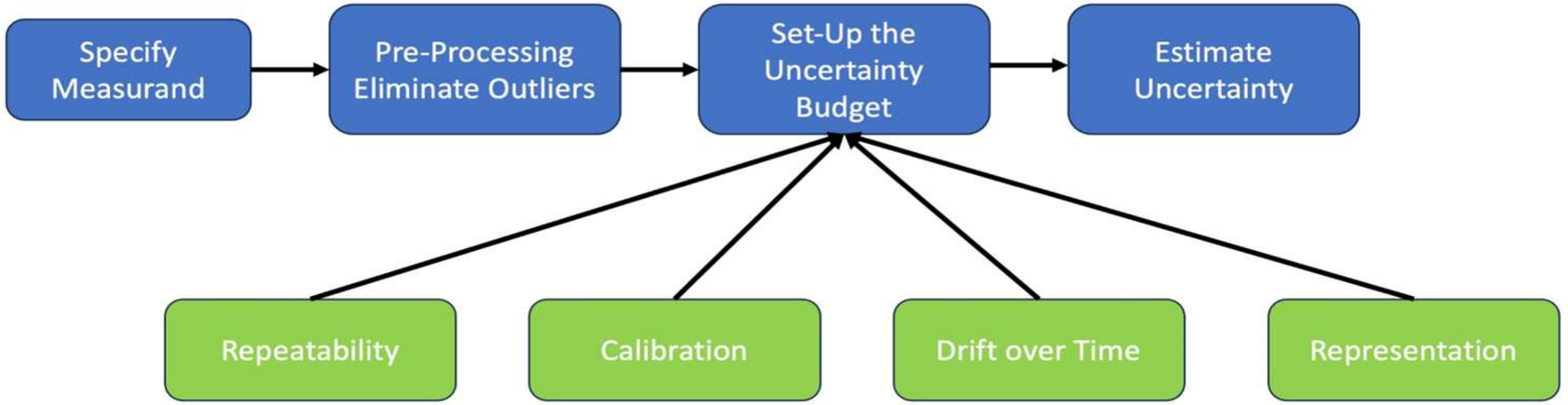
~8% trend difference from 2005-2021

Impact of instrumental bias on OHC 0-700m:

~50% trend difference from 1970-2000

UNCERTAINTIES

Uncertainty quantification flow chart



Measurement uncertainty **changes over time** depending on the main components of the observing system, measurement method, instrument and platform used

Initial uncertainty estimates:

IQuOD v0.1 (2018) release contains 'Type B' measurement uncertainties determined from manufacturer specifications and other publications (*Cowley et al, 2021* https://doi.org/10.3389/fmars.2021.689695)

Representativeness Errors:

When considering incorporating measurements into model applications (eg, reanalysis, ocean heat content mapping), representativeness describes the uncertainty of using a single measurement to represent the gridded averages for a certain spatial and temporal resolution

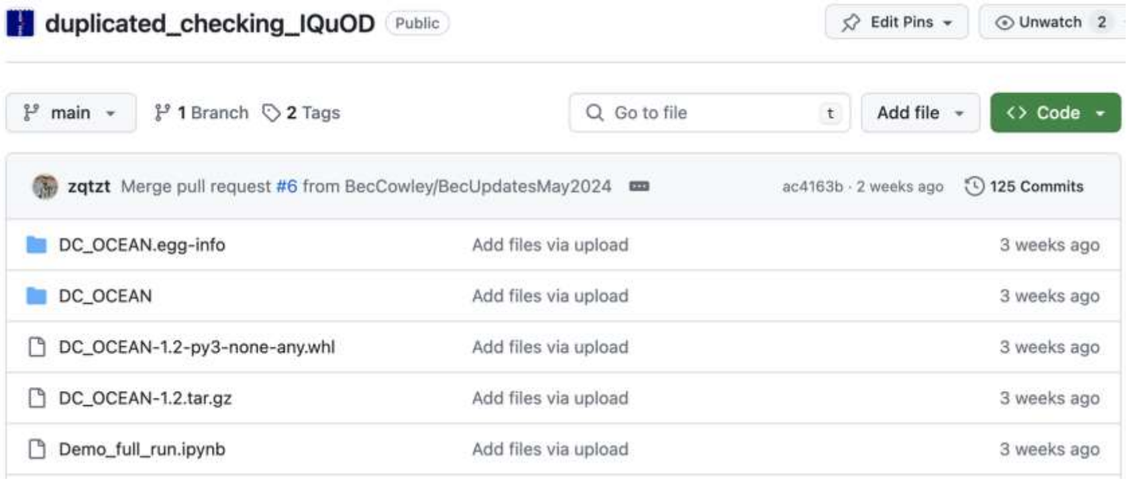
Plans for IQuOD:

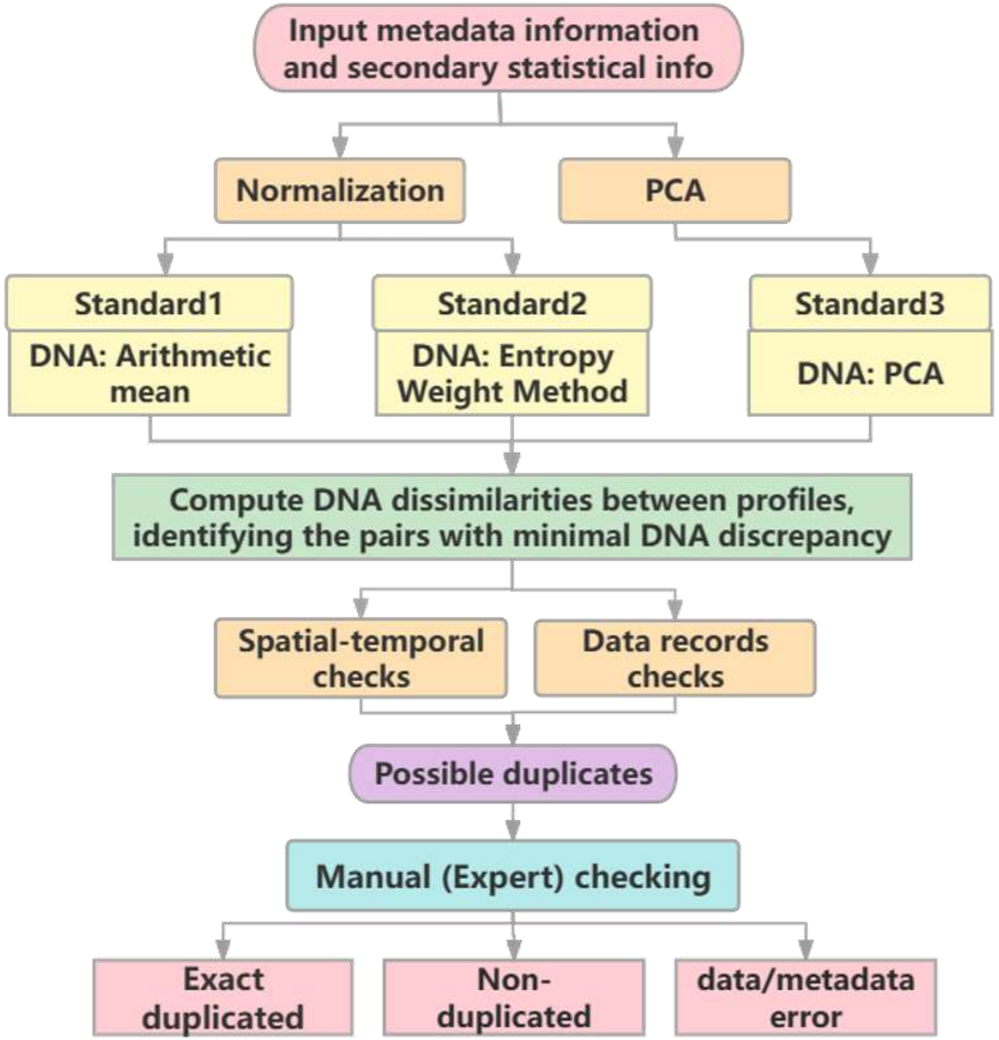
- Supplying gridded uncertainties for typical applications, including publication of a set of algorithms for different use cases to calculate representativeness errors
- IQuOD will also provide monthly estimates for the upper ocean

DUPLICATE CHECKING

Integrating data from different data infrastructures needs a duplicate detection

- **An semi-automatic system to detect duplicates**
  - Automatic check: crude screening and target screening
  - Manual (expert) check
- **Definition of duplicates**
  - Exact duplicates
  - Possible duplicates
  - No duplicates
- **Open-source Python Packages**
  - DC_OCEAN
  - https://github.com/IQuOD/duplicated_checking_IQuOD
  - Version 1.2





X. Song, Z. Tan et al. (2024) An open-source algorithm for identification of duplicates in ocean database (under review)

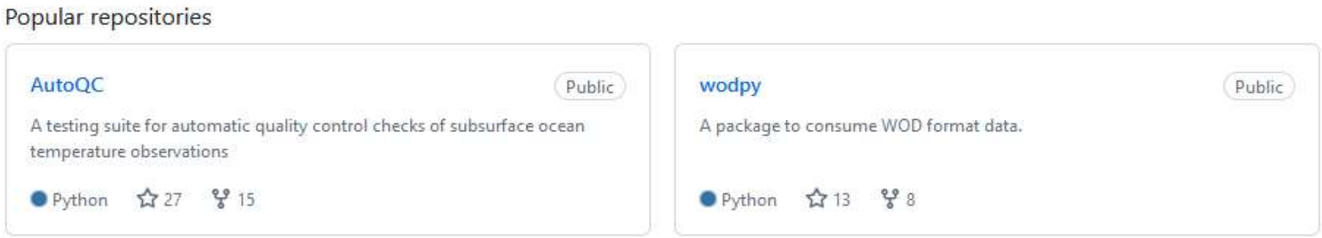
AUTOMATED QUALITY CONTROL

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

- AutoQC checks (60) and a WOD data reader (wodpy) have been coded in Python
- code repositories are open (MIT license) so the code 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

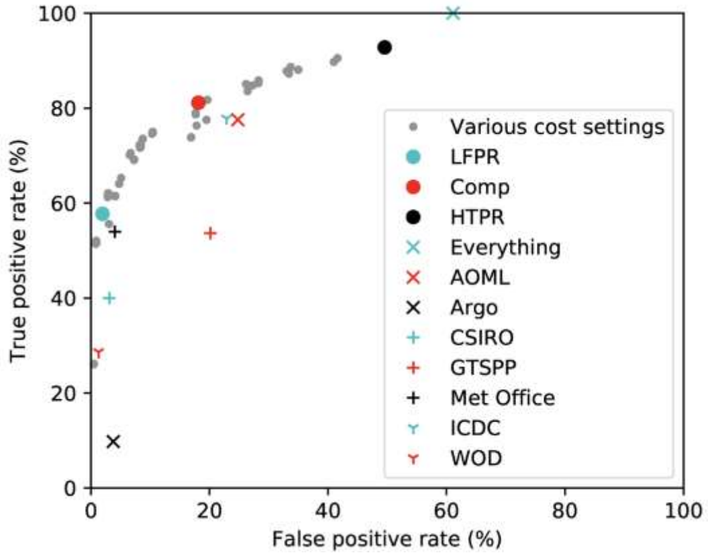
International Quality-controlled Ocean Database

https://github.com/IQuOD



open-source collaborative software infrastructure

Evaluation of different QC systems



Benchmarking metrics

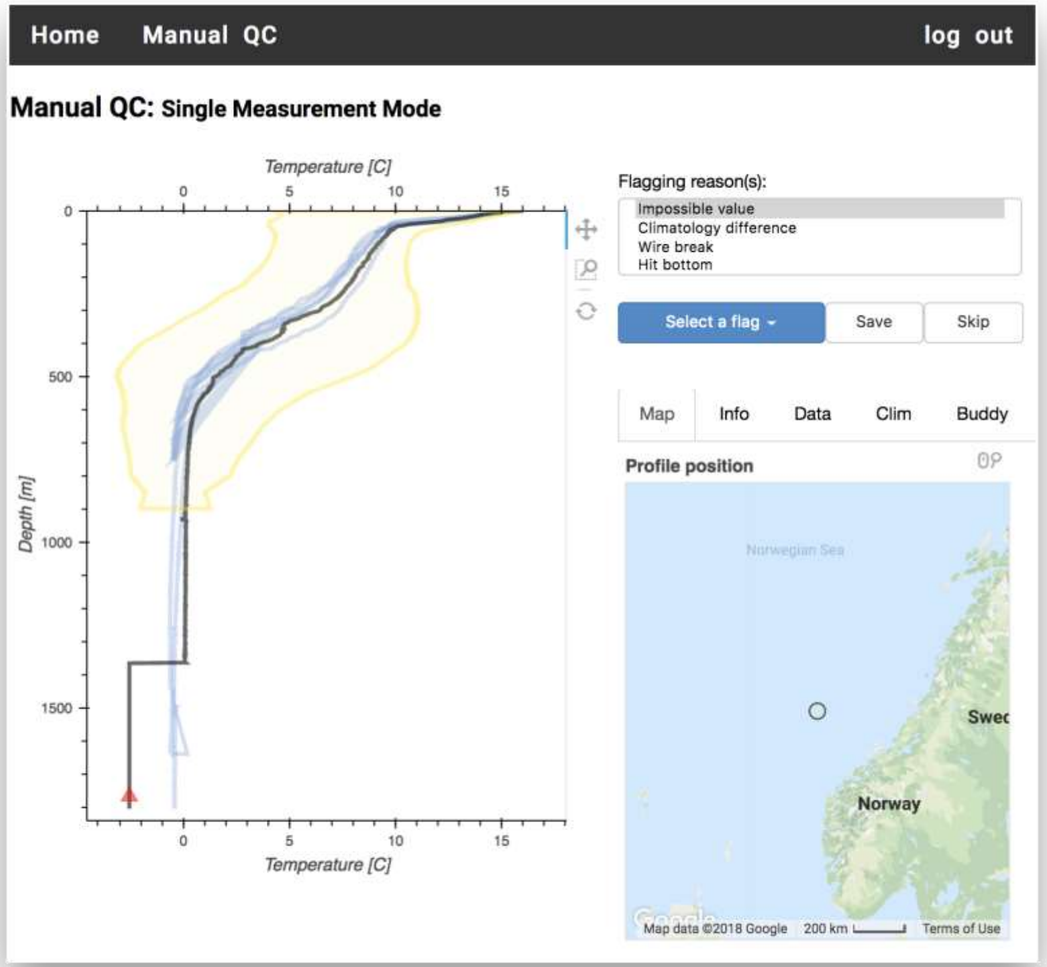
- True Positive Rate % **TPR**
- False Positive Rate% **FPR**

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

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

Expert QC & Machine Learning

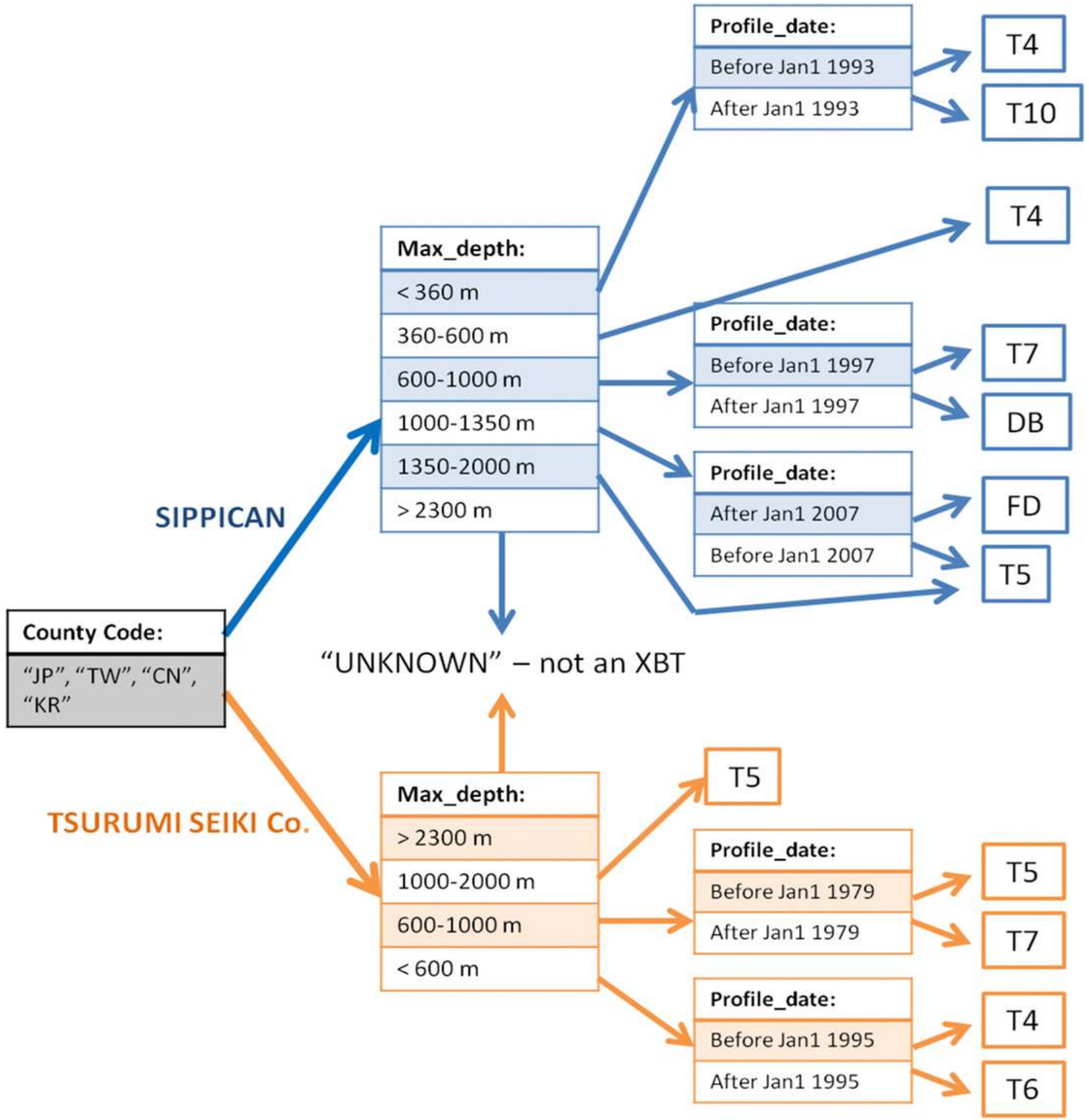
- A Machine Learning Approach to QC Oceanographic data
- Web App to integrate experts around the world https://expertqc.castelao.net
- To improve efficiency of the manual QC, the experts are paired with an interactive learning schema of Machine Learning to combine the high skill of the human with the 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

INTELLIGENT METADATA

- **What?** Using machine learning to infer metadata for eXpendable BathyThermographs (XBTs)
- **Why?** Approximately 50% of XBTs don't have manufacturer and probe type recorded in the World Ocean Database
- **How?** Increasingly sophisticated machine learning methods:
  - *Palmer et al. (2018)* - a simple human-designed decision tree
  - *Leahy et al. (2018)* - neural networks
  - Haddad et al. (2022) - multiple methods including a decision tree with an ensemble output allowing uncertainty representation
- **What next?** To incorporate the work of *Haddad et al. (2022)* into IQuOD



First decision tree idea from *Palmer et al. (2018)*

CHALLENGES

IQuOD v0.1 (2018)

intelligent metadata and uncertainty specification

→ available through NOAA/NCEI service

→ updated quarterly along with the WOD (WODselect retrieval system)

IQuOD v.1

autoQC

→ in progress 2024

- **it takes time to run** the autoQC: 1.5 months even when running each data set in parallel
- **cloud-based solutions** have been implemented to optimize the computation and production
- need to check that the final result on the global dataset are consistent with the results from Good et al. (2023)
- duplicate checking and the correction of identified metadata errors in the WOD are progressing, but slowly due to the limited capacity to confirm and make the proper fix decision

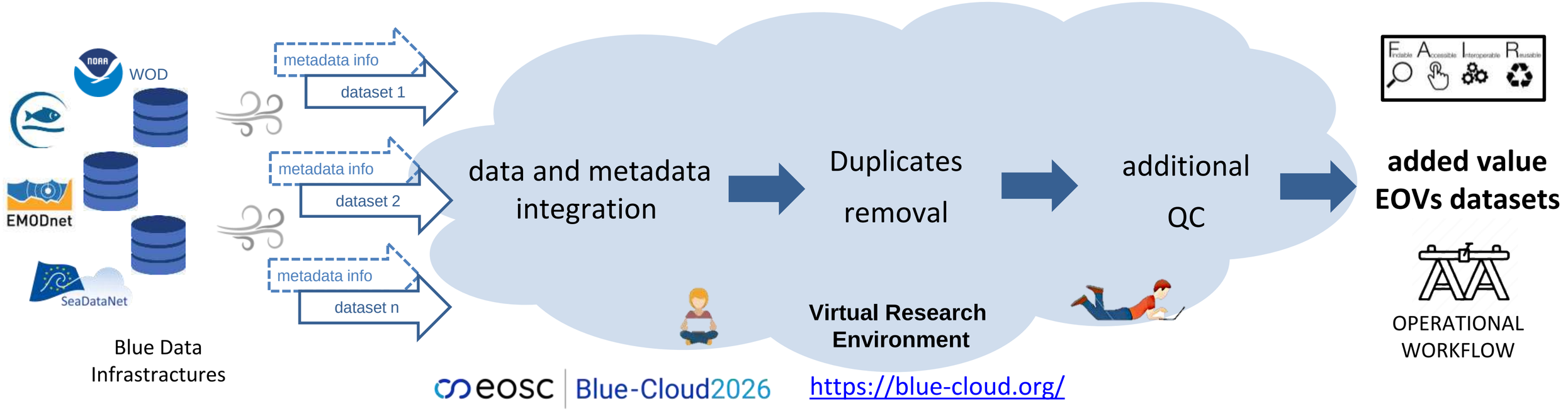
SOLUTION: to realize **World Ocean Database Cloud** which would allow for any IQuOD member to execute changes to the WOD in a cross-community version of the WOD

FAIR data

FAIR data TT → all tools and data products generated to be available and reusable to all, seeking feedback from users

- Synergy with the Blue-Cloud2026 project that established a cyber platform providing access to multi-disciplinary datasets, analytical services and computing facilities for open web-based science

- Some of the IQUOD tools (duplicate checking and QC tests) will be tested and adapted within Blue-Cloud2026 analytical workbenches → mutual feedback on data, metadata, tools and services



Conclusions

ACHIEVEMENTS under this globally-coordinated effort:

- continual improvement of QC processes, enhancing metadata and uncertainty information, duplication detection, global data assembly and **data rescue** to support a more profound understanding of our changing climate
- **30 peer-reviewed papers**
- Development and integration **of ocean best practices** into public repositories
- A curated collection of QCed data products that serve as benchmarks for the community
- Development of a **cloud-based supervised machine learning trained by experts** worldwide (pilot project)

OUTLOOK

- Outreach activities based on IODE teaching academy
- focus on Salinity
- dialog with stakeholders (ocean and climate modeling/science communities, Digital Twin of the Ocean initiatives)
- consolidate the open science strategy fully adopting the FAIR principles
- strengthen the collaborative approach through cloud-based solution to face the big data challenge
- work in synergy with ongoing initiatives like Blue Cloud 2026

Thanks!

Website: www.iquod.org

Bibliography:

https://scholar.google.com/citations?user=qYD_0r8AAAAJ&hl

Also have a publication collection at Ocean Best Practices:

http://repository.oceanbestpractices.org/handle/11329/1590

Github: https://github.com/IQuOD

IMDIS 2024 - Bergen (Norway), 27-29 May 2024

International Conference on Marine Data and Information Systems