Enriched metadata for hybrid data compilations with applications to cryosphere research

Anna Simson (simson@mbd.rwth-aachen.de), Marc S. Boxberg and Julia Kowalski

Methods for Model-based Development in Computational Engineering, RWTH Aachen University, Aachen, Germany

Chair of Methods for Model-based Development in Computational Engineering

RWTH AACHEN UNIVERSITY

www.mbd.rwth-aachen.de

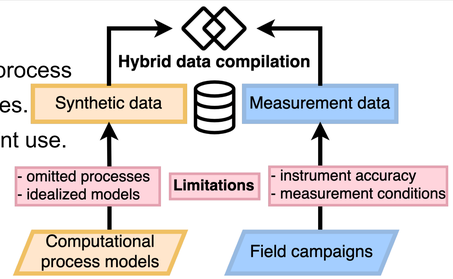
Hybrid data compilations for cryosphere science

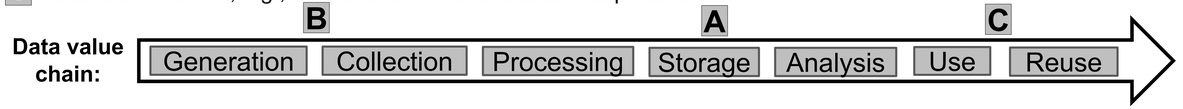
Data platforms provide a large variety of **measurement data**, and computational process models generate **synthetic data** of physical, geochemical, and biological processes.

It is the **GOAL** to combine complementing measurement and synthetic data for joint use.

General **CHALLENGES**, here exemplified by **sea ice** data are:

- **A** **FAIR**ness of existing data repositories
- **B** Hybridizing measurements and simulations for knowledge generation
- **C** **Data readiness** for, e.g., simulations or model-based interpretation



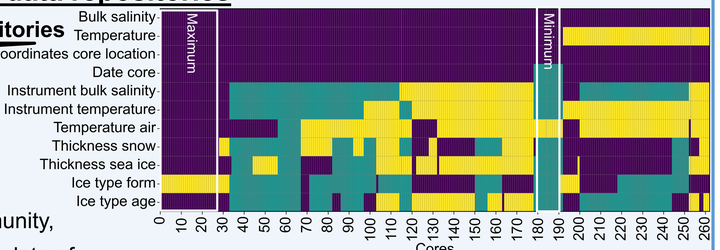


A FAIRness of sea ice core measurement data repositories

Review of data from **262 sea ice cores** from **105 repositories**

\+ Data is **F**indable and **A**ccessible via, e.g., Pangaea

– but mostly not (easily) **I**nteroperable and **R**eusable.



Data and metadata often lack **cross-repository consistency** and **completeness** due to

- missing established **keyword** standards in the community,
- varying **repository content** that ranges from all core data of one campaign to data of one property of a single core, and
- scope of **included (meta)data** as illustrated in the figure.

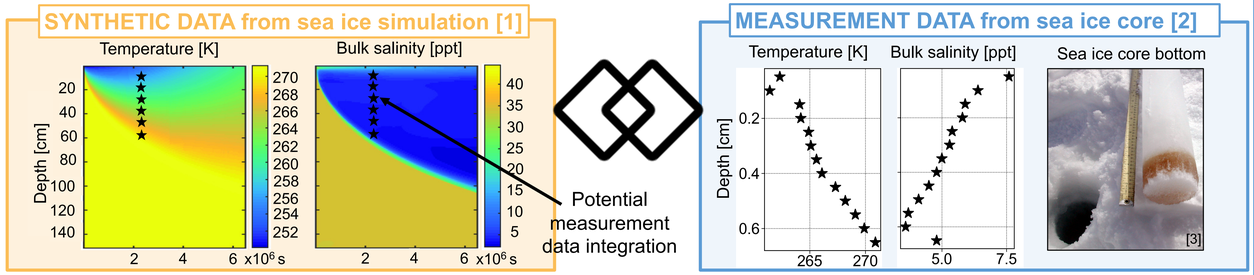
Critical knowledge necessary for hybridization such as measurement uncertainty is often **missing** in data repositories and requires a **data fill** for (meta)data enrichment.

Data fill strategy:

1) Search for metadata in further repositories/papers/reports

2) Use existing definitions for, e.g., ice type from [4]

B Simulation and measurement (meta)data to enable hybridization

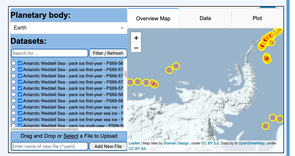


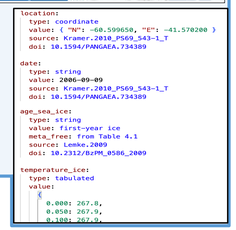
State variables (SV)	Selection of **model**/**measurement** information that should be provided as (meta)data for successful hybridization	Measurables (M)
Model uncertainty		Measurement uncertainty
Computational domain		Observational domain
Initial and boundary condition of SV		M at observation start and boundaries
Simulated time		Observed time
Temporal and spatial discretization		Temporal and spatial resolution

C Data readiness with Ice Data Hub [5] as flexible data management tool

The Ice Data Hub is a **python module** that contains

- a web browser-based **GUI** to view, manipulate, add, plot and map data,
- an interface to directly load data into python simulation or data workflows,
- a database consisting of human readable **YAML files** each representing one location that allows for (meta)data enrichment (i.e., adding information on instruments, ice type and spatial resolution), and
- an interpolation method to directly adjust measurement resolution to match with required spatial discretization and vice versa.





Summary and conclusions

- **A** Critical data for hybridization is missing, and data repositories are **FA**IR due to lack of cross-consistency and completeness.
- **B** Hybridization requires rich metadata from data generation.
- **C** Ice Data Hub allows later metadata enrichment and increases data readiness for hybridization.

References:

[1] Buffo et al. (2018), doi: 10.1002/2017JC013345.

[2] Pučko et al. (2011), doi: 10.1594/PANGAEA.818647.

[3] Krembs and Deming (2011), ice_core_algae_band.

[4] WMO (2014), https://library.wmo.int/doc_num.php?explnum_id=4651 .

[5] Boxberg et al., doi: 10.5194/egusphere-egu21-13052.

This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) and received funding from the Helmholtz Association of German Research Centres. The project is supported by the Federal Ministry of Economics and Technology, Germany, based on the decision by the German Bundestag (FKZ: 50NA1908, 50NA2009)

HDS-LEE HELMHOLTZ SCHOOL FOR DATA SCIENCE IN LIFE | EARTH | ENERGY

DLR

TRIPLE

Gefördert durch: Bundesministerium für Wirtschaft und Energie aufgrund eines Beschlusses des Deutschen Bundestages