DAM DEUTSCHE ALLIANZ MEERESFORSCHUNG

AI-Application for Scientific Sensor Data collected onboard German Research Vessels

Michael Schlundt (GEOMAR), Julia Oelker (UOL/ICBM), Robert Kopte (CAU/KMS), Gauvain Wiemer (DAM)

PERMANENTLY MOUNTED SCIENTIFIC SENSORS

DAM

FS Polarstern
- ADCP
- CTD
- Ferry Box
- Multibeam
- TSG

FS Meteor
- ADCP
- CTD
- Multibeam
- TSG

FS Maria S. Merian
- ADCP
- CTD
- Multibeam
- TSG
- Fluorometer

RSWS

FS Sonne
- ADCP
- CTD
- Multibeam
- TSG
- Fluorometer

SMB

blog.ankerherz.de


Foto: LDF


Foto: N.Verch


Foto: Hartig

DATAMANAGEMENT
SUPPORT FOR SCIENCE

DAM

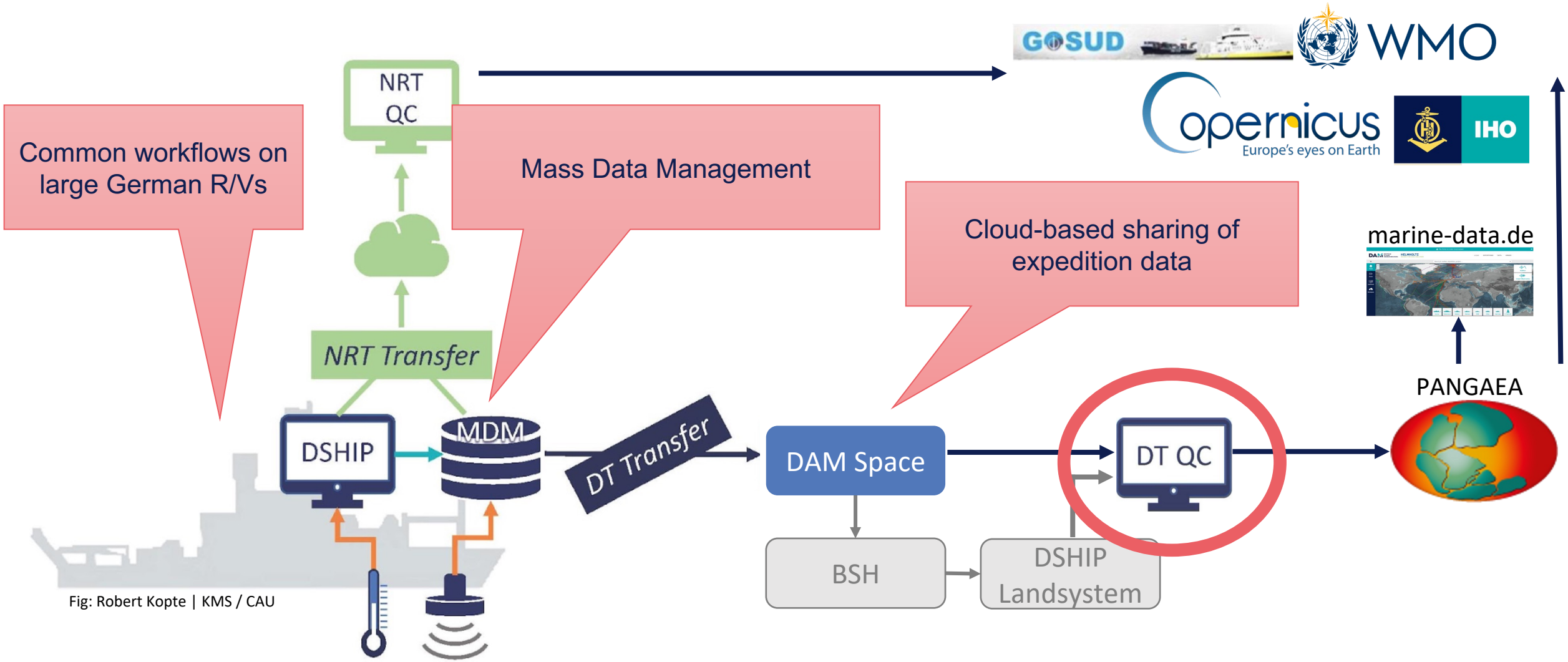


Fig: Robert Kopte | KMS / CAU

Can we improve the QC of underway data with AI approaches?

- Several QC steps are conducted to get a final quality-controlled dataset
- For **TSG/fluorometer**-timeseries data those are mostly: skip dummies and non-usable data (in port etc.), outlier detection, time-averaging, correction against independent data, flagging of data values
- For **ADCP**-profile data: binary raw-data conversion, screening for bottom and acoustic interference, ship-speed calc. and averaging, water-track calibr., flagging
- Is this possible by AI as well?
  - **Faster processing?**
  - **Unknown features?**
  - **...?**

Helmholtz AI-consulting

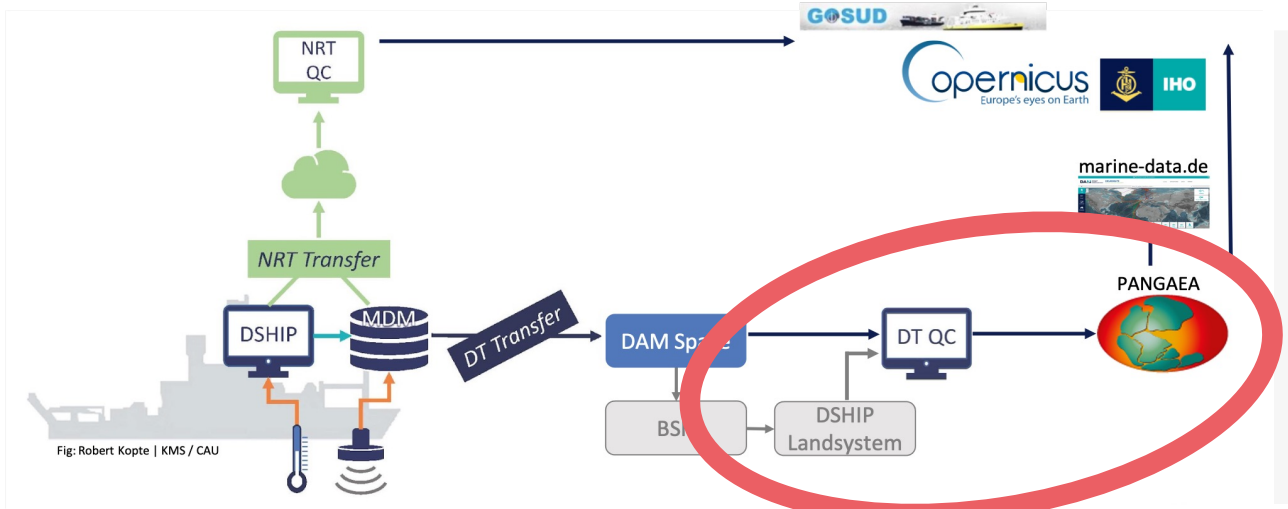
DAM

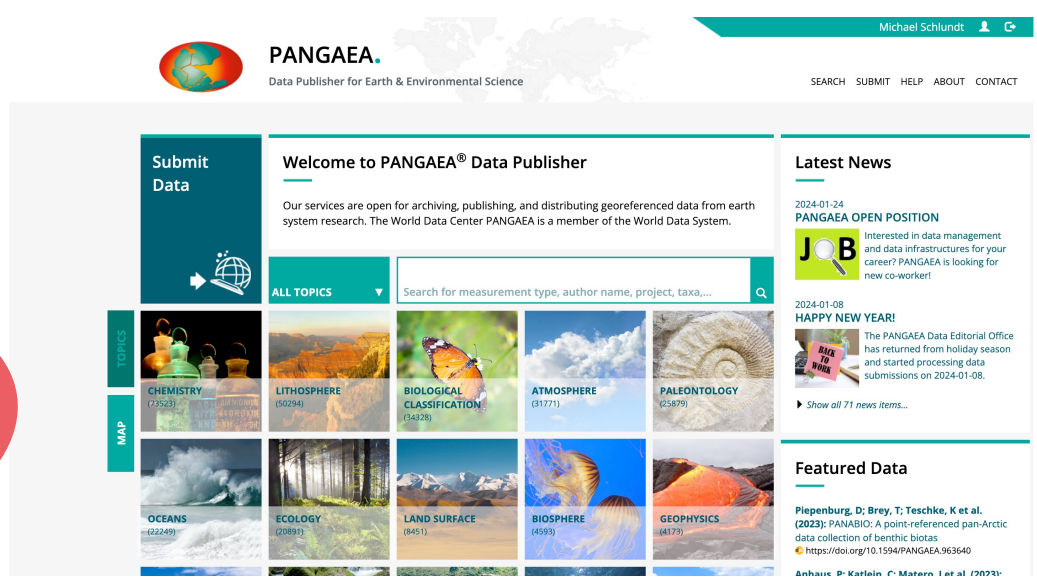
- **Helmholtz AI-voucher**: Helmholtz AI Consulting helps with finding right approaches and/or methods
- **Exploration voucher**: Initial exploration of idea and related background, assessment of feasibility, first results
- **Realisation voucher**: follow-up of exploration, deeper digging into the topic
- We applied for a voucher and got assigned to a consultant team at DLR

DLR

Data Selection for AI-consulting

- **RV Maria S. Merian**: all „raw“ data obtained from BSH DSHIP Landsystem from **2021** (**48** parameters (system, RSWS, weather station)) in 1-sec resolution >> **raw-files**
- **TSG**: as a first step TSG data were used (temperature and salinity of seawater in 6.5m depth)
- **PANGAEA**: archived QC TSG-datasets (MSM98-MSMX14) in 1-min resolution >> **qc-files**



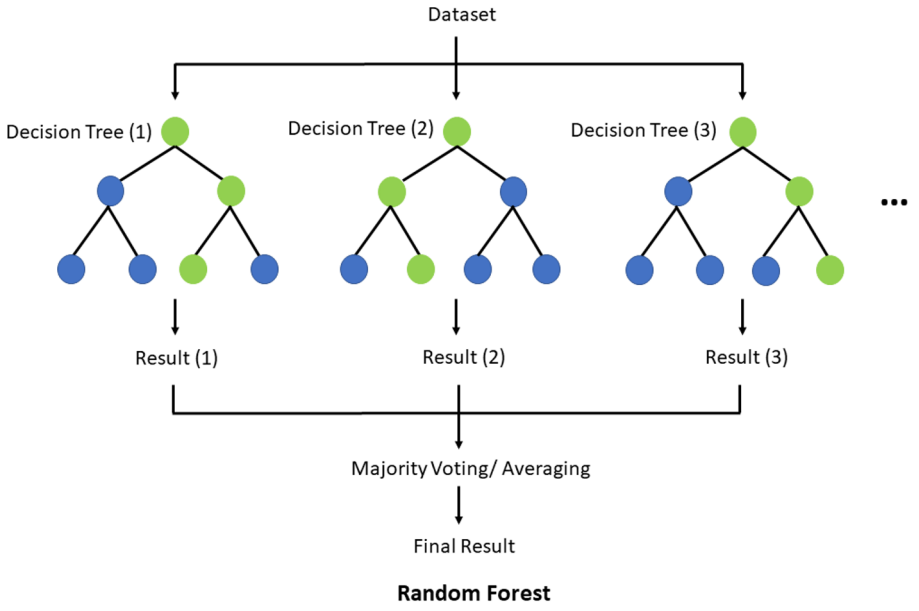


Data Preprocessing

- **Temporal Aggregation**: computing 1-minute averages from 1-second resolution samples
- **Alignment with Ground Truth**: Discarded rows in the **raw-files** with timestamps not present in the **qc-files**, ensuring alignment between data sources.
- **Data Synchronization**: Ordered **raw** and **qc-files** to establish a one-to-one correspondence between their rows, facilitating accurate comparison.
- **Handling Missing Values**: Replaced missing (or invalid) values with NaN
- **Data Consolidation**: Preprocessed all ship cruises data to create a dataset comprising approximately 0.5 million samples. Columns of the **raw-files** file served as training data, while **qc-files** columns *Temperature* and *Salinity* acted as ground truth references.

First approach: Random Forest Regression

- **Random Forest** is an ensemble **Machine Learning algorithm** that uses many decision tree models.
- Can be used for **classification** or **regression** (like in this case).
- For regression problems, the value predicted by the random forest is the mean of the values predicted by the individual trees.
- One main advantage is the opportunity to «reveal» which **parameters** are most important for output



Random Forest – How does it work?

- **Ensemble of Decision Trees:** A Random Forest is a collection of decision trees.
- **Random Feature Selection**: Each tree in the forest considers only a random subset of the available features. This allows to avoid «overtraining»
- **Bootstrap Aggregating (Bagging)**: To train each tree, we randomly sample the original data with replacement. This allows to decrease the variance of the model without increasing the bias.
- **Fully Grown Trees**: In a random forest, each tree is allowed to grow fully without any pruning.
- **Voting for Prediction**: The final prediction is determined by majority vote among all trees. Or in the case of regression problems the average prediction is taken.
- **Out-of-Bag Error Estimation**: The performance of the random forest model is estimated using the out-of-bag (OOB) samples, which are the data points not included in the bootstrap sample for each tree. This provides an unbiased estimate of the model's generalization error.

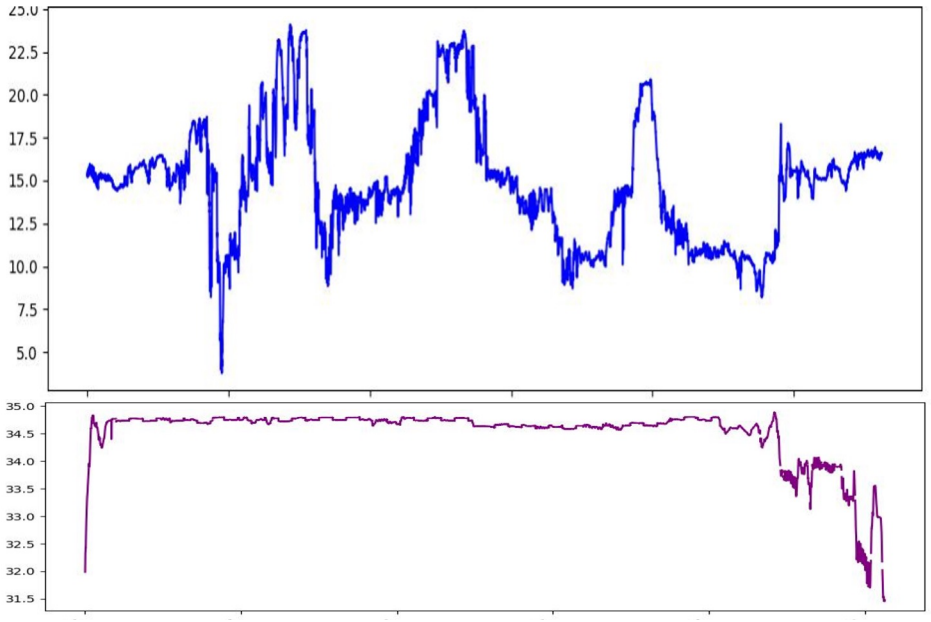
Experimental Set-Up for TSG data

- **Training Data**: The entire **raw** dataset was used for training the model, ensuring that the model learns from the full range of available information.
- **Cross Validation**: To assess the model's performance, we employed 5-fold cross validation. This technique divides the data into five subsets, using four subsets for training and one for testing in rotation.
- **R^2 Score**: The R^2 score measures what percentage of the variance in the dependent variable is explained by the model. Values close to 1 indicate that the model predicts the output very well. Value close to 0 indicate that the model predicts no better than simple average. Values can be arbitrarily low since the prediction can be arbitrarily bad.
- **Decision Tree Depth**: Depth of the decision trees was reached when R^2 score on the test set began to decrease. Number of trees in the random forest was set to 100, constrained the maximum depth of each tree to 20. These parameters help control the complexity of the model and prevent overfitting.

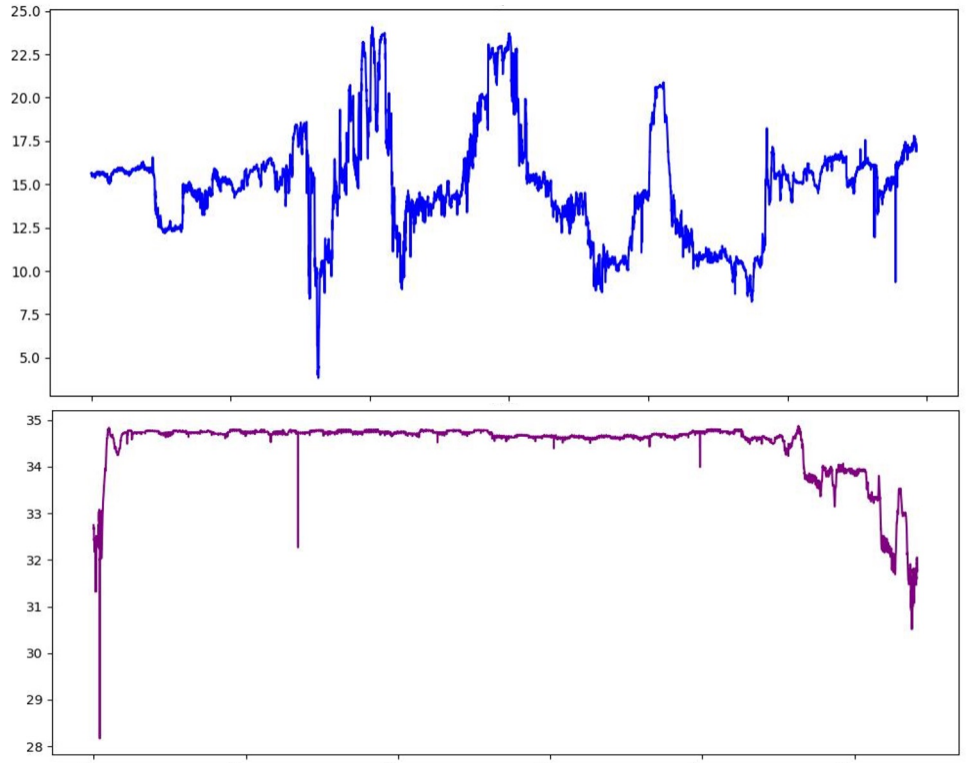
First Results

- The R2 score (0.99) and visual analysis of the plots indicate that our Random Forest regressor demonstrates highly promising alignment with the reference data.

Reference (**qc data**)



Prediction (from **raw data**)



Temperature

Salinity

Further Work

- **Realisation phase**
  - Capture the missing steps in QC-processing of timeseries data
  - Analysis of raw data archive in terms of important **parameters** (obtained from **Random Forest** algorithm)
  - Explore the use of other methods, such as Deep Learning methods.
  - Apply method(s) on other timeseries (**Fluorometer** chl-a and turbidity) and profile data (**ADCP** velocities)

DAM DEUTSCHE ALLIANZ MEERESFORSCHUNG

DLR

Carl von Ossietzky Universität Oldenburg

ICBM

THANKS FOR YOUR ATTENTION

GEOMAR

KIEL MARINE SCIENCE CONNECTED RESEARCH

C | A | U Christian-Albrechts-Universität zu Kiel