9th Data Science Symposium

Thursday 4 April 2024 - Friday 5 April 2024

Haus der Wissenschaft, Bremen

Book of Abstracts

Contents

Baby steps in conducting complex simulations across HPC infrastructures 1

Local climate services for all, courtesy of large language models 1

Stakeholder involvement in CRITTERBASE - Data interoperability for marine conservation and sustainable ecosystem management 1

Best of both worlds: unlocking the potential of data science for biological research . . . 2

Embracing AI as a Tool for Science, highlights from the Helmholtz AI team for Earth & Environment 2

The NFDI4Earth distributed user support network 3

Enriching EarthServer With AI-Enabled Datacubes - Demo 3

Polar ice core micro CT super resolution segmentation with deep learning 4

How to support scientists in research data management –experiences from the DAM research missions 4

Quality controlling autonomously collected bio-optical data 5

A Convolutional Neural Network to detect bowhead whale vocalizations in passive acoustic data from the Arctic Ocean 5

Harnessing consumer grade GPU hardware for the automation of annotation processes in hydrographic data –examples from the ValidITy project 6

The Helmholtz Metadata Collaboration - Lessons from the Data Infrastructure Survey . . 6

Data conversion and aggregation for the M-VRE webODV 7

The FAIR SAMPLES template for IGSN sample registration 8

IT project management in a scientific environment to transfer knowledge into a public usable software –ValidITy project as an example 8

A.I-based characterization of seafloor habitats and megafaunal species composition: Examples using optical images from the Pacific and tropical Northeast Atlantic 9

Data quality control - an essential prerequisite of science 9

Strategies for good and comprehensive metadata in field-based permafrost research . . . 10

The ValidITy project –Transferring knowledge about bathymetric feature detection into a polished software project - Demo 10

DataPLANT - (Meta)data Annotation with Swate and Ontologies - Demo 11

Building Digital Twins of the Ocean 12

Keynote: Earth Observation and Environmental Data Science on modern HPC systems . 12

Implementing an Analysis Framework in the Marine Data Portal of the German Marine Research Alliance (DAM) 12

Using psyplot for visualizing unstructured data and vertical transects - Demo 13

Facilitating the heterogeneous scientific data sharing with the THREDDS Control Center - Demo 14

The infancy of MANIDE: Machine learning driven Assessment of polymetallic Nodule mining Impacts on Deep-sea Ecosystems 14

Enabling high performance DS capacity at the AWI's tier-3 HPC 15

Customised data services for different target groups - experiences from Meereisportal/ SEA ICE PORTAL 16

User-Customizable Map Visualizations in the Earth & Marine Data Portal - Demo 17

iLOVE: integrating Long-term Observation & Virtual Evaluation 17

autoQC: An AI based online app for ocean data quality control 18

HARMONise –Enhancing interoperability of marine biomolecular (meta)data across Helmholtz Centres 19

HMC Earth and Environment - Developing a Robust Framework for Seamless Semantic Interoperability in Earth and Environmental Research 20

NFDI4Earth Academy - Your training network to bridge Earth System and Data Sciences 20

Data science for understanding physics –modelling ship wake detectability using machine learning 21

About the the collaboration between research institutions and authorities within the German Marine Research Alliance 21

AI-Application for Scientific Sensor Data collected onboard German Research Vessels . . 22

Keynote: AI 22

Keynote: Marine authority data flow ~ Linked data - from local data nodes to harmonized services 22

DataNord: Empowering Bremen's Research Community in Data Literacy 23

Exploring the soundscapes of the world's oceans –A demonstration of OPUS, the Open Portal to Underwater Soundscapes 23

Posters, Demos and Refreshments / 3

Baby steps in conducting complex simulations across HPC infrastructures

Author: Timm Schultz[1]

Co-authors: Angelika Humbert [2]; Jonas Eberle [3]; Stephan Frickenhaus [1]; Stephan Hachinger [4]; Philipp Sommer [5]; Hannes Thiemann [6]; Noah Trumpik ; Bernadett Fritzsch [1]

[1] *AWI*

[2] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung*

[3] *DLR*

[4] *LRZ München*

[5] *Hereon*

[6] *DKRZ*

Corresponding Authors: stephan.hachinger@lrz.de, bernadette.fritzsch@awi.de, stephan.frickenhaus@awi.de, timm.schultz@awi.de, angelika.humbert@awi.de, philipp.sommer@hereon.de, noah.trumpik@tu-dresden.de, thiemann@dkrz.de, jonas.eberle@dlr.de

We present the concept of conducting complex simulations across infrastructures, an ice sheet model run on cluster A, ingesting data from cluster B and data from cluster C, and pushing results to cluster C - NEFDI4Earth Pilot CAPICE

Session 3: Governmental Data and Transfer / 4

Local climate services for all, courtesy of large language models

Authors: Antonia Anna Jost[1]; Nikolay Koldunov[None]; Thomas Jung[2]

[1] *AWI*

[2] *Alfred Wegener Institut*

Corresponding Authors: nikolay.koldunov@awi.de, antonia.jost@awi.de, thomas.jung@awi.de

Large language models can summarize, aggregate, and convey localized climate-related data. We have built a simple, proof-of-concept prototype and argue that the approach holds the potential to truly democratize climate information.

Session 3: Governmental Data and Transfer / 6

Stakeholder involvement in CRITTERBASE - Data interoperability for marine conservation and sustainable ecosystem management

Author: Katharina Teschke[None]

Co-authors: Paul Kloss ; Jennifer Dannheim ; Janine Felden ; Marianne Rehage ; Roland Koppe

Corresponding Authors: jennifer.dannheim@awi.de, mrehage@marum.de, katharina.teschke@awi.de, janine.felden@awi.de, paul.kloss@awi.de, roland.koppe@awi.de

Biogeographical research plays a pivotal role in understanding large-scale patterns of biodiversity, particularly concerning environmental change, human impacts, and sustainable ecosystem management. To conduct such research effectively, reliance on FAIR data is imperative. However, many data are not open-access and exist only in spreadsheets and local databases, which hinders the scientific analysis and data reuse. In addition, databases often employ disparate data models and (meta-)data schemas. Consequently, merging data from different sources requires time-consuming data synchronisation and often acts as a bottleneck for initiating analyses at large spatial and temporal scales. In our presentation, we elucidate our approaches to addressing data interoperability between CRITTERBASE and other data systems utilised by various decision-makers and stakeholders. Using illustrative examples, we demonstrate: (1) how seamless data exchange across different systems forms a fundamental basis for evaluating the suitability of areas for the expansion of offshore wind farms in the German Bight; (2) we delve into the implementation of (meta-)data interoperability for a federated search tool focused on Arctic and Antarctic data, with the expectation of enhancing the discoverability of CRITTERBASE data for the polar regions; and finally, (3) we present an ongoing initiative for data interoperability aimed at facilitating the smooth integration of CRITTERBASE data into leading databases such as PANGAEA for data archiving, publication, and distribution. It is highlighted how efforts related to data interoperability can contribute to capacity building, i.e., improving data handling and management for individual users and larger organisational structures, ultimately leading to biogeographical research and decision-making in the context of marine conservation and sustainable ecosystem management.

Session 1: Capacity Building and AI / 7

Best of both worlds: unlocking the potential of data science for biological research

Author: Iva Veseli[1]

[1] *Alfred Wegner Institute*

Corresponding Author: iva.veseli@hifmb.de

As the life sciences are becoming increasingly reliant on big data and an ever-growing number of computational strategies, leveraging a data science approach can substantially enhance the efficacy and robustness of biological data analyses. Yet, for these data manipulation and processing methods to reach their full potential for biological applications, it is critical for biologists and computer scientists to work together and combine biological knowledge of the research system with technical expertise to craft appropriate solutions. I will present two stories from the ‘omics field that demonstrate the power of data science when applied with biological insights to otherwise difficult problems. First, I will describe a simple technical heuristic based upon our knowledge of the extent of diversity in protein families to reduce the number of false negatives in homologous gene annotation, which in some cases enables the annotation of up to 16% more functions in a given microbial genome. Second, I will describe how adding a novel normalization step could explain the drivers of the loss of microbial diversity in individuals with inflammatory bowel disease (IBD), revealing a potential ecological explanation for something that puzzled microbial ecologists for almost two decades. In both cases, finding these solutions required a combination of technical and biological expertise, highlighting that the path forward for enhancing the scope and utility of computational work within the biological sciences will depend upon effective communication between scientists in these two fields.

Session 1: Capacity Building and AI / 8

Embracing AI as a Tool for Science, highlights from the Helmholtz AI team for Earth & Environment

Author: Danu Caus[1]

Co-authors: Adeniyi Mosaku ; Caroline Arnold [1]; Harsh Grover ; Paul Keil [1]; Tobias Weigel [1]

[1] *DKRZ*

Corresponding Authors: mosaku@dkrz.de, weigel@dkrz.de, arnold@dkrz.de, keil@dkrz.de, caus@dkrz.de, grover@dkrz.de

Artificial Intelligence is the new electricity! It is a valuable tool, that already shows its potential in numerous everyday applications. We are the Earth and Environment team at Helmholtz AI and our aim is to contribute in making this tool available within the science community, as a means towards scientific advancement. We help researchers embrace AI in their research endeavors, and work together with it to foster new ideas and scientific progress.

One of the important factors towards progress lies in bridging the gap between different fields. Within the scientific community, it is currently challenging to keep up with the latest advancements in AI, given the very rapid pace in development of machine learning technologies. Hence, we strive to occupy the role of AI experts, offering insight about the entire machine learning pipeline: starting from data pre-processing, model architecture design and implementation, result post-processing and deployment. We also give guidance regarding the usage of computational resource infrastructure that is available through the Helmholtz institution.

In this presentation we will give examples and talk about the various types of projects we worked on recently, ranging from computer vision, sequence analysis and explainable AI topics, as well as showcase relevant deployment aspects. This will convey insight about the kind of things that are currently possible, and inspire other novel ideas that can be pursued in the future. You may talk to us personally at the symposium or contact us via: https://www.helmholtz.ai/themenmenue/you-helmholtz-ai/ai-consulting/index.html

Session 4: Data Initiatives / 9

The NFDI4Earth distributed user support network

Author: Hela Mehrtens[None]

Co-authors: Klaus Getzlaff [1]; Sören Lorenz

[1] *GEOMAR*

Corresponding Authors: slorenz@geomar.de, kgetzlaff@geomar.de, hmehrtens@geomar.de

The distributed, cross-institutional User Support Network (USN) for NFDI4Earth is based on the existing and well embedded user support structures of the participating institutions. The USN serves as a single point of contact for user requests that could not be handled via OneStop4All and require individual consulting. By combining the distributed RDM knowledge of experts in the USN in conjunction with the Knowledge Hub, the NFDI4Earth team will convey the notion (knowledge) of a best practice for dealing with data and how data can be made FAIRer and open.

We will present the current status and the conncetion to D.A.M and DataHUB.

Posters, Demos and Refreshments / 10

Enriching EarthServer With AI-Enabled Datacubes - Demo

Author: Peter Baumann[1]

[1] *Constructor U*

Corresponding Author: pbaumann@constructor.university

EarthServer is a global federation of research institutions and commercial providers offering location-transparent datacube access, extraction, aggregation, analytics, and fusion. AWI is member in EarthServer since several years now.

In our live demos we will present the EarthServer principle of a single global datacube space. Further, we will present recent technical progress on AI integration, and further new features. At the same time, based on our active standardization work in OGC and ISO we link into current standardization trends and showcase OGC work on OAPI-Coverages, GeoDataCubes, and other relevant activities.

Posters, Demos and Refreshments / 11

Polar ice core micro CT super resolution segmentation with deep learning

Author: Faramarz Bagherzadeh[1]

Co-authors: Johannes Freitag [1]; Frank Wilhelms [1]; Udo Frese [2]

[1] *AWI*

[2] *University Bremen*

Corresponding Authors: ufrese@uni-bremen.de, frank.wilhelms@awi.de, faramarz.bagherzadeh@awi.de, johannes.freitag@awi.de

Precise segmentation of 3D micro-CT scans is a crucial step in analyzing the microstructure of porous materials. The polar ice core (firn column) microstructure is of great importance in polar research. Detecting environmental effects on the firn column in polar ice core studies relies on accurately digitizing the microstructure. To study the evolving microstructure of the firn column, 150 meters of core sample should be scanned with a microCT machine. With current technology in hand, it is practically impossible to scan the whole column with high resolution (e.g. 30 µm), thus, it is only possible to scan the entire column with lower resolutions such as 120 µm. Consequently, the smaller bubbles and pore structures are missed or represented vaguely in low-resolution scans. To tackle this problem, a unique pipeline for generating a data set consisting of low-resolution images and their corresponding registered high-resolution images is developed. With this pipeline, the high-resolution data were registered to the low-resolution data with the rigid image registration method. Then the patch-wised low-resolution data were fed to deep neural networks having their corresponding patch-wised high-resolution data as the ground truth. Due to the 3D nature of the project utilizing HPC is necessary for performing image registration and training the deep learning models. Finally, different deep learning models were compared on pixel-wise metrics and microstructure parameters. The trained models will be used to enhance the resolution of archived ice cores in AWI.

Session 4: Data Initiatives / 12

How to support scientists in research data management –experiences from the DAM research missions

Authors: Flavia Höring[1]; Gauvain Wiemer[2]; Tim Boxhammer[None]; Susanne Feistel[None]; Janine Felden[None]; Anneke Heins[None]; Kai Hoppe[None]; Marcus Krüger[None]; Hela Mehrtens[None]; Manja Placke[None]

[1] *MARUM / PANGAEA*

[2] *Deutsche Allianz Meeresforschung (DAM)*

Corresponding Authors: janine.felden@awi.de, hoering@uni-bremen.de, hmehrtens@geomar.de, wiemer.dam@gmail.com

Data managers are essential to support scientists for successful research data management along the whole data life cycle. At the same time, scientists as well as data managers face various challenges such as the reduction of funding and therefore missing personnel and a lack of time for data management. As part of the DAM project "Underway"Research Data, a working group has formed consisting of leading data managers of the DAM research missions, members of the PANGAEA "Data Publisher for Earth & Environmental Science"editorial team, as well as the DAM core area "Data management and digitalization". As this working group, we aim to support scientists involved in the DAM research missions to make their research data FAIR - Findable, Accessible, Interoperable and Reusable. Regularly, we discuss the various challenges in research data management, develop and offer solutions and formulate recommendations. We offer regular information and training events for scientists and consortia data managers of the missions e.g. on the general principles of FAIR data handling, DAM-related activities such as the visualisation of data in the Marine Data Portal, and data archiving in PANGAEA. During our work in the DAM research missions, we have learned that engaged data managers, good communication and training, as well as practical technical solutions for research data management are the most important factors to support scientists to publish their data FAIRly.

Posters, Demos and Refreshments / 13

Quality controlling autonomously collected bio-optical data

Author: Julia Oelker[1]

Co-authors: Daniela Voß [1]; Jochen Wollschläger [1]

[1] *Institute of Chemistry and Biology of the Marine Environment, University of Oldenburg*

Corresponding Authors: jochen.wollschlaeger@uni-oldenburg.de, daniela.voss@uni-oldenburg.de, julia.oelker@uni-oldenburg.de

German research vessels collect vast amounts of valuable data of the marine environment during each expedition. Especially underway data from autonomous sensor systems that are not the main focus of the scientific program are often not published or published with unknown quality and non-standardized and incomplete meta data. The "Underway Research Data"project of the German Marine Research Alliance (DAM) aims to increase the vessel's efficiency to create quality-assured research data by establishing well-documented workflows from the ship's data acquisition to their open publication under FAIR principles (Findable, Accessible, Interoperable, Reusable). Here, we present the current workflow for bio-optical underway data, specifically chlorophyll-a fluorescence. In general, the workflow is structurally the same for different underway oceanographic parameters. But in detail, each parameter needs special attention, in case of bio-optical data, due to its high natural variability and the proxy-nature of the underlying measurement technique. We present challenges to robustly quality control the data and make suggestion on how to overcome these by a community effort, envisioning a future standardized database for calibration coefficients. As a future perspective, all autonomous platforms with the same sensor type, such as BGC-Argo floats, can profit from such a database.

Posters, Demos and Refreshments / 14

A Convolutional Neural Network to detect bowhead whale vocalizations in passive acoustic data from the Arctic Ocean

Author: Marlene Meister[None]

Co-authors: Paul Keil [1]; Karolin Thomisch

[1] *DKRZ*

Corresponding Authors: karolin.thomisch@awi.de, marlene.meister@awi.de, keil@dkrz.de

Climate change is causing significant environmental shifts in the Arctic Ocean, affecting the habitat suitability for marine mammal species inhabiting Arctic waters seasonally or year-round. Habitat degradation or habitat loss will particularly affect Arctic endemic species, such as bowhead whales (Balaena mysticetus). Bowhead whales possess a complex and temporally variable acoustic behavior that is utilized in reproductive and social contexts. They produce single calls, usually frequency modulated vocalizations between 50 and 500 Hz, as well as highly variable songs, referring to structured series of vocalizations. Passive acoustic monitoring (PAM) represents a non-invasive tool to collect crucial year-round and multi-year information on the occurrence of bowhead whales. Since manual detection of bowhead whale vocalizations in continuous PAM data is a challenging and time-consuming task, the Ocean Acoustics group of AWI teamed up with the Helmholtz Artificial Intelligence Cooperation Unit to develop an AI-based algorithm for bowhead whale detection. To this end, we train a Convolutional Neural Network (CNN) to recognize vocalization signatures of bowhead whales in spectrograms generated from PAM data. The algorithm divides data into short-duration snippets, indicating the presence or absence of bowhead whale signals for each snippet. This approach has the potential to significantly streamline the analysis process, while enhancing objectivity of call identification. The network will be applied for the analysis of an extensive acoustic dataset (spanning 2104 recording days) collected by AWI in Fram Strait between 2012 and 2021. For training we use more than 4000 humanly labeled individual whale calls over several days. In the future, we aim to provide easy operational inference from the trained network for new data. Analyzing this acoustic data will further our understanding of trends in bowhead whale occurrence, contributing to the development of effective conservation strategies.

Session 2: Data Quality and HPC for Data Science / 16

Harnessing consumer grade GPU hardware for the automation of annotation processes in hydrographic data –examples from the ValidITy project

Authors: Flemming Staebler[None]; Valentin Buck[1]

Co-authors: Anne Hennke [2]; Jens Greinert [2]; Josephine Brauer [3]; Stephan Meyer [4]; Torsten Frey [2]

[1] *GEOMAR Helmholtz Centre for Ocean Research*

[2] *Geomar Helmholtz Centre for Ocean Research*

[3] *GEOMAR - Helmholtz-Zentrum für Ozeanforschung Kiel*

[4] *GEOMAR*

Corresponding Authors: jbrauer@geomar.de, smeyer@geomar.de, vbuck@geomar.de, ahennke@geomar.de, jgreinert@geomar.de, fstaebler@geomar.de, tfrey@geomar.de

While GPU computing has been widely used in science through the Tensorflow and Torch frameworks, and in specialized HPC applications, software that runs on end-user-devices often does not yet use these technologies.
In this presentation, we show how we used OpenGL compute shaders to accelerate key features of the software developed in the ValidITy project (https://validity-project.eu) to implement a user-friendly workflow for feature detection in gridded bathymetric data. We will explain how geomorphometric derivatives can be computed in near-real-time and show that implementing a neural network from scratch does not need to be a daunting task - even if the will need to be executed on low-powered laptop devices.

Posters, Demos and Refreshments / 18

The Helmholtz Metadata Collaboration - Lessons from the Data Infrastructure Survey

Author: Emanuel Söding[1]

[1] *GEOMAR*

Corresponding Author: esoeding@geomar.de

Emanuel Soeding, Stanislav Malinovschii (GEMAR), Andrea Poersch (GFZ Potsdam), Yousef Razeghi (UFZ Leipzig), Dorothee Kottmeier (AWI Bremerhaven)

The interconnectivity of existing data infrastructures (DIS) across national and international initiatives (e.g. NFDI, EOSC and others) is an important goal to create a common interoperable data space. To achieve this, it is critical to harmonize the existing methods and concepts of research data collection among the DIS and along the FAIR principles.

Within the Helmholtz Association we maintain more than 50 active data infrastructures in the field of Earth and Environment. Procedures of data handling, documentation and storage are hardly coordinated within Helmholtz, even less so within the larger community. To find out about the state of our infrastructures, the different approaches in data management procedures, technical capabilities, and concepts, we conducted a survey among all Helmholtz DIS. The questions asked were related to their roles in the community, self-perception, quality control, curation, technology interfaces, data re-use and demands.

Here we shed some some light on lessons and consequences for HMC strategies and what kind of recommendations may improve the state of our data infrastructures.

Posters, Demos and Refreshments / 19

Data conversion and aggregation for the M-VRE webODV

Author: Ingrid Linck Rosenhaim[1]

Co-author: Sebastian Mieruch-Schnuelle [2]

[1] *Alfred Wegener Institute*

[2] *AWI*

Corresponding Authors: sebastian.mieruch@awi.de, ingrid.linckrosenhaim@awi.de

The MOSAiC Virtual Research Environment (M-VRE) project offers three different tools for MOSAiC data exploration. One of these tools is webODV, the online version of Ocean Data View (ODV) software for analysis and visualization of oceanographic and georeferenced data.
During the MOSAiC Expedition 2019-2020, a great amount of measurements in different disciplines were collected and published in the long-term archive PANGAEA. The M-VRE project aims to increase the visibility and usage of these data sets by presenting an online and user-friendly environment where the MOSAiC data is uploaded and kept up-to-date. In webODV, the MOSAiC data are aggregated into collections and displayed in different scientific disciplines. Exploration, visualization, and analysis of MOSAiC data with webODV, is only possible after the data are converted to an ODV readable format.
In our project, we developed a semi-automatic process that queries, filters, and downloads the data set and respective metadata from the PANGAEA archive onto our server. These data sets are gathered into collections before being aggregated and converted to the ODV format. The metadata is harmonized, ensuring that all data collections have a common set of meta variables. During the conversion process, no data is altered, only the format is changed. Following the concept of FAIR data (Findable, Accessible, Interoperable, and Reusable), all data presented in webODV contain the harmonized metadata and references to the original files in the PANGAEA archive, ensuring the traceability of authors and data sources, as well as transparency.
The MOSAiC data is very diverse in its scientific disciplines and different measurements. However, these data sets are also very different in format and structure, and many require an extra step before being converted into the ODV format. Data stored in netCDF files, tar files, and others, receive a unique script that downloads, reads, and writes an ASCII file containing all metadata and data variables. Other data sets require a special conversion due to the data structure, without this extra conversion, visualization and analysis possibilities with webODV would be limited. It is also part of our semi-automatic process to thoroughly verify the converted data collection before it is uploaded

to webODV. These data collections are also available for DIVA and Data Cubes exploration in the M-VRE server.

Posters, Demos and Refreshments / 20

The FAIR SAMPLES template for IGSN sample registration

Author: Mareike Wieczorek[1]

Co-authors: Alexander Brauser [2]; Birgit Heim [1]; Kirsten Elger [2]; Linda Baldewein [3]; Simone Christina Frenzel [4]; Ulrike Kleeberg [3]

[1] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung*

[2] *Deutsches GeoForschungsZentrum (GFZ) Potsdam*

[3] *Hereon*

[4] *GFZ*

Corresponding Authors: ulrike.kleeberg@hereon.de, simone.christina.frenzel@gfz-potsdam.de, linda.baldewein@hzg.de, kirsten.elger@gfz-potsdam.de, mareike.wieczorek@awi.de, birgit.heim@awi.de, alexander.brauser@gfz-potsdam.de

The International Generic Sample Number (IGSN) is a unique and persistent identifier, originally developed for samples in the Geosciences. A strategic partnership between IGSN e.V. and DataCite in 2023 led to the integration of all IGSNs as DataCite DOIs. While samples can be registered with the DataCite mandatory metadata, this schema is not designed to the comprehensive description of physical objects.
Within the project "FAIR Workflows to establish IGSN for Samples in the Helmholtz Association (FAIR WISH)", funded by the Helmholtz Metadata Collaboration Platform (HMC), we developed the FAIR SAMPLES template. The template is Excel-based, generic and customizable to fit for various use cases and disciplines in Earth and Environmental sciences. It is designed for individual researchers to submit sample metadata for bulk registration of IGSN, regardless of the scientist's technical background. On the other hand, the Hereon use case within our project demonstrated the use of the template for populating data from an existing sample database.
The template offers flexibility, comprising both few mandatory and many optional variables to describe a sample, the sampling activity, location and more. Users can easily create a personalized template, including only the variables relevant to describe their samples.
The information collected with the template is currently transmitted to GFZ and there converted in XML files to generate IGSN landing pages and DataCite metadata.
The usage of IGSNs for samples bridges one of the last gaps in the full provenance of research results. In this presentation, we showcase the FAIR SAMPLES template, highlighting its user-friendly design and varied applicability across disciplines.

Posters, Demos and Refreshments / 21

IT project management in a scientific environment to transfer knowledge into a public usable software –ValidITy project as an example

Author: Stephan Meyer[1]

Co-authors: Anne Hennke [2]; Flemming Stäbler [3]; Jens Greinert [2]; Josephine Brauer [4]; Torsten Frey [2]; Valentin Buck [1]

[1] *GEOMAR Helmholtz Centre for Ocean Research*

[2] *Geomar Helmholtz Centre for Ocean Research*

[3] *GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel*

[4] *GEOMAR - Helmholtz-Zentrum für Ozeanforschung Kiel*

Corresponding Authors: ahennke@geomar.de, jbrauer@geomar.de, smeyer@geomar.de, tfrey@geomar.de, fstaebler@geomar.de, jgreinert@geomar.de, vbuck@geomar.de

Scientific units nowadays commonly develop their own software in order to solve problems within their research. The planned user base is often limited to the originating unit which sometimes results in limited usability for external people. Oftentimes a software is only being developed to a prototype standard, so that it is usable for the group's research, but not any further. When these stages need to be overcome, ad-hoc self-organizing management approaches can reach their limitations.

With this poster we present a user centric approach to the development of a software product not only to be used by our own group but industry, state organizations and other research institutes. We show how the core team, consisting of three software developers, one scientific advisor and tester, one IT project manager and the scientific project leader, organizes and communicates. A core tool for this is a variant of the agile project management framework Kanban implemented as GitLab issue boards, which serves as the main organizational tool for this project. We also explain how involving potential users through user meetings and early access versions of the software can be key to project success und feedback loops that can be used to ensure software quality.

ValidITy (Validation of Intelligent Terrain Feature Recognition Methods for Hydrographic Data \[https://validity-project.eu/\]) is a GEOMAR project to develop an object annotation and terrain classification software for gridded bathymetric data using machine learning and classification dictionaries.

Session 1: Capacity Building and AI / 22

A.I-based characterization of seafloor habitats and megafaunal species composition: Examples using optical images from the Pacific and tropical Northeast Atlantic

Author: Benson Mbani[1]

Co-author: Jens Greinert [2]

[1] *GEOMAR*

[2] *Geomar Helmholtz Centre for Ocean Research*

Corresponding Authors: bmbani@geomar.de, jgreinert@geomar.de

The characterisation of seafloor habitats and their resident megafaunal communities contributes to our collective understanding of the global ocean health and resilience. Whereas direct sampling e.g using box corers provides physical samples that can be archived and analysed later in the lab, recent advances in optical imaging platforms have enabled the generation of high-resolution images and at high temporal frequency. This makes image-based analysis the prefered approach for exploring and mapping seabed ecosystems at medium-to-large spatial scales. However, manual inspection of these huge volumes of acquired images is a time consuming endeavor that poses a real bottleneck when extracting both qualitative and quantitative actionable insights about marine biodiversity. Here, we automate this process using our A.I-based seafloor classification (AI-SCW) and megafaunal species composition (FaunD-Fast) workflows. We present findings following our application of these workflows to specific case studies in the Pacific and tropical Northeast Atlantic. Based on this, we demonstrate that the integration of A.I and marine sciences significantly expedites the generation of baseline information for objective monitoring of remote benthic ecosystems.

Session 2: Data Quality and HPC for Data Science / 23

Data quality control - an essential prerequisite of science

Author: Inge Grünberg[1]

Co-authors: Brian Groenke [2]; Frederieke Miesner [3]; Julia Boike [2]

[1] *AWI*

[2] *Alfred-Wegener-Institute*

[3] *Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research*

Corresponding Authors: julia.boike@awi.de, brian.groenke@awi.de, inge.gruenberg@awi.de, frederieke.miesner@awi.de

Any model can only be as good as the input and validation data. Models of various scales ranging from single soil columns to the complete earth system need time series of input data. Large scale models often rely on global products derived from satellite data or weather model re-analysis. However, classical on-site measured time series still serve as input of such compiled products and are also used to validate the model output. Furthermore, local effects often cannot be adequately represented in products of limited spatial resolution. In the Arctic, the required local reference data are scarce and raw data series often include corrupted or missing values. Our long-term observatories were installed more than 25 years ago and therefore provide very valuable data on soil and climate characteristics at Bayelva (Svalbard) and Samoylov (Lena River Delta, Siberia). In addition to the tremendous effort to maintain the sites and calibrate the instruments, it is essential to perform a thorough quality control of each measured variable. Our routine does not only cover removing values outside the physical limits, but also includes manual checks for plausibility both within a single time series and in comparison to similar variables. We use soil temperature data of the Samoylov observatory, Lena River Delta, Siberia, to highlight the effect of using raw data, quality-controlled data, and gap filling on estimated soil warming trends. In particular, we consider the timing of missing values in our analysis. We show that removing implausible values and a proper handling of data gaps are essential prerequisites of any data analysis including both physical and statistical modelling efforts.

Posters, Demos and Refreshments / 24

Strategies for good and comprehensive metadata in field-based permafrost research

Authors: Frederieke Miesner[1]; Inge Grünberg[2]

Co-author: Julia Boike [3]

[1] *Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research*

[2] *AWI*

[3] *Alfred-Wegener-Institute*

Corresponding Authors: julia.boike@awi.de, frederieke.miesner@awi.de, inge.gruenberg@awi.de

Comprehensive metadata are key to making data FAIR.
It is therefore essential to collect metadata in an organized and standardized way. For standardized data acquisition, ie. on research vessels, tools are already available and constantly improved. In land-based permafrost expeditions, however, the data and metadata are as diverse as the science questions behind them.

We present an overview of this diversity in (meta)data and (meta)data collection and propose strategies to writing good and comprehensive metadata. We encourage to think about metadata right from the start and work on them steadily during the whole process from field work preparation to data collection and from data analysis to final publication. Easy to adapt templates and only choosing the tools that fit the specific data set increases the participation of the whole team.

Posters, Demos and Refreshments / 25

The ValidITy project –Transferring knowledge about bathymetric feature detection into a polished software project - Demo

Authors: Anne Hennke^{None}; Flemming Stäbler[1]; Jens Greinert[2]; Josephine Brauer[3]; Stephan Meyer[4]; Torsten Frey[2]; Valentin Buck[4]

[1] *GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel*

[2] *Geomar Helmholtz Centre for Ocean Research*

[3] *GEOMAR - Helmholtz-Zentrum für Ozeanforschung Kiel*

[4] *GEOMAR Helmholtz Centre for Ocean Research*

Corresponding Authors: jgreinert@geomar.de, tfrey@geomar.de, ahennke@geomar.de, vbuck@geomar.de, fstaebler@geomar.de, smeyer@geomar.de, jbrauer@geomar.de

We will showcase a live demo version of the software developed in the ValidITy project (Validation of Intelligent Terrain Feature Recognition Methods for Hydrographic Data https://validity-project.eu/). This software aims to integrate all necessary steps of an object and terrain annotation and classification workflow on bathymetric data into one user-friendly product.
In the current version the software features a responsive map view that can compute common geomorphometric derivatives immediately due to efficient use of OpenGL compute shaders. Users can then annotate sections of their data with point-, box- or polygon shapes and train machine learning models on them. The trained model will be used to extend the annotations to the full set of data.
Further workflow components such as implementing additional machine learning approaches, annotation quality control, terrain classification and the development of classification dictionaries are planned for the future.
We hope to use this demo not only to introduce the software to an interested scientific audience, but also to gather further suggestions: which further features would be useful for the analysis of bathymetric data and which application problems exist in the workflow with other software packages that might be simplified by the Validity software.

Posters, Demos and Refreshments / 26

DataPLANT - (Meta)data Annotation with Swate and Ontologies - Demo

Author: Angela Kranz^{None}

Corresponding Author: a.kranz@fz-juelich.de

The DataPLANT consortium, a German National Research Data Infrastructure (NFDI), focuses on creating a resilient and enduring data infrastructure to support plant scientists with Research Data Management (RDM). Various tools and services are provided to assist users in this endeavour.
At the centre of DataPLANT lies the Annotated Research Context (ARC), a FAIR digital object. The ARC serves as a standardized and comprehensive method for researchers to document their experimental designs, protocols, workflows, and data in a structured format. The annotation of metadata within the ARCs is facilitated by ontologies. The DataPLANT Ontology Landscape combines the ISA standard with the semantic capabilities of the metadata annotation tool Swate [https://github.com/nfdi4plants/Swate] and the Terminology-Service [https://github.com/nfdi4plants/nfdi4plants_ontology] provided by DataPLANT. This approach addresses the challenge of harmonizing diverse data sources, enabling researchers to seamlessly collaborate, share, and analyze data while fostering reproducibility and interoperability.
The ISA (Investigation, Study, Assay) data model is a well-established standard for capturing and representing metadata. It provides a structured and extensible framework for describing the experimental design and context. Swate, an Excel Add-In, simplifies metadata annotation by relying on the ISA-Tab format in combination with ontology term search via a Terminology Service. This process not only enhances the accuracy and efficiency of metadata annotation but also ensures that metadata annotation is standardized.
With our approach, we show that standards such as ISA in combination with ontologies can be efficiently used across all life science domains for (meta)data annotation using spreadsheets.

Session 2: Data Quality and HPC for Data Science / 27

Building Digital Twins of the Ocean

Author: Timm Schoening[1]

[1] *GEOMAR*

Corresponding Author: tschoening@geomar.de

What are Digital Twins of the Ocean? Who does benefit from this data science tool? Which kind of new science can we address with them? How can we use Digital Twins to transfer knowledge and synthesise gains across research fields? Which building blocks of Digital Twins are required to step ahead? What can we build upon?

The term Digital Twins is a buzzword connected to many stakeholder interests. Yet, implementations of operational Digital Twins for Ocean Science remain scarce. While this method holds the potential to catalyse the cultural change towards Open Science, fundamental technical challenges remain to be solved. At this very moment, initiatives like the DataHub, HMC and HMC projects or data space designers like the NFDIs, EOSC and GAIA-X all contribute puzzle pieces towards building the interoperable and FAIR systems required for operational Digital Twins of the Ocean. Often without specifically targeting Digital Twin methods as a goal.

In this contribution, the puzzle pieces will be placed on a larger canvas, showing connections, gaps, risks and potentials. A roadmap towards interoperable and FAIR marine data spaces and transfer actions will be sketched along the GEOMAR strategy towards facilitating Digital Twins of the Ocean as one tool of future marine science.

Session 2: Data Quality and HPC for Data Science / 28

Keynote: Earth Observation and Environmental Data Science on modern HPC systems

Author: Jonas Eberle[1]

[1] *DLR*

Session 3: Governmental Data and Transfer / 30

Implementing an Analysis Framework in the Marine Data Portal of the German Marine Research Alliance (DAM)

Authors: Philipp Sebastian Sommer[1]; Robin Hess[2]; Björn Lukas Saß[1]; Angela Schaefer[3]

[1] *Helmholtz-Zentrum Hereon*

[2] *Alfred-Wegener-Institut*

[3] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung*

Corresponding Authors: philipp.sommer@hereon.de, bjoern.sass@hereon.de, robin.hess@awi.de, angela.schaefer@awi.de

The continuous growth of Earth System data, coupled with its inherent heterogeneity and the challenges associated with distributed data centers, necessitates a robust framework for efficient and secure data analysis. This abstract outlines the plans to implement an analysis framework within the marine data portal of the German Marine Research Alliance (Deutsche Allianz Meeresforschung,

DAM) at https://marine-data.de. The proposed framework is based on the Data Analytics Software Framework (DASF), chosen for its decentralized, secure, and publisher-subscriber-based (pub-sub-based) architecture, which enables the execution of data analysis backends anywhere without exposing sensitive IT systems to the internet.

The challenges of analyzing Earth System data on the web are multifaceted. Data heterogeneity arises from the diverse sources, formats, and structures of earth system data, making seamless integration and analysis a complex task. The sheer volume of data compounds the challenge, demanding scalable solutions to handle vast amounts of information efficiently. Additionally, the computational power required for meaningful analysis is often expensive and can become a bottleneck in traditional data processing pipelines. Moreover, the distributed nature of data across multiple centers poses logistical challenges in terms of accessibility, security, and coordination.

To address these challenges, the integration of DASF into the marine data portal presents a comprehensive solution. DASF offers a secure and decentralized pub-sub-based remote procedure call framework, providing a flexible environment for executing data analysis backends. One of the key advantages of DASF is its ability to allow these backends to run anywhere without the need to expose sensitive IT systems to the internet, addressing the security concerns associated with data analysis.

The decentralized nature of DASF also mitigates data heterogeneity challenges by offering a unified platform for data integration and analysis. With DASF, disparate data sources can seamlessly communicate, facilitating interoperability and enabling comprehensive analysis across diverse datasets. The pub-sub mechanism ensures efficient communication between components, streamlining the flow of data through the analysis pipeline.

Security is a critical aspect of implementing a robust data analysis framework. DASF addresses this concern by incorporating an OAuth-based authentication mechanism at the message broker level. This ensures that only authorized users can access and interact with the data analysis functionalities. Additionally, the integration with the Helmholtz AAI empowers the sharing of analysis routines with users from other research centers or the general public.

The cost-effectiveness of DASF further enhances its appeal, as it optimizes the utilization of computational resources. By enabling the deployment of analysis components on diverse hardware environments, organizations can leverage existing infrastructure without significant additional investments.

In conclusion, the integration of DASF into the DAM portal marks a significant step toward overcoming the challenges inherent in analyzing Earth System data on the web. By addressing data heterogeneity, accommodating vast datasets, and providing a secure and decentralized architecture, DASF emerges as a key enabler for efficient and scalable data analysis. The adoption of DASF in the marine data portal promises to enhance the accessibility, security, and cost-effectiveness of data analytics, and finally facilitates open science in the research field Earth and Environment.

Posters, Demos and Refreshments / 31

Using psyplot for visualizing unstructured data and vertical transects - Demo

Author: Philipp Sebastian Sommer[1]

[1] *Helmholtz-Zentrum Hereon*

Corresponding Author: philipp.sommer@hereon.de

This presentation demonstrates the most recent features of Psyplot, a powerful tool for visualizing climate data on unstructured grids. The utilization of UGRID conventions has become paramount in handling unstructured grids effectively. This presentation demonstrates how psyplot can be used to effectively and straightforwardly visualize climate data conforming to UGRID conventions, highlighting its role in enhancing the comprehension of spatial and temporal patterns.

Another key focus of the presentation is on the grid-independent extraction of vertical transects in 4-dimensional data, addressing a critical challenge in climate science. I will showcase the new innovative methodologies within Psyplot that facilitate seamless extraction of vertical profiles across varying grid structures, enabling researchers to analyze and interpret climate variables. By showcasing practical applications and case studies, the presentation aims to demonstrate the usefullness of psyplot for climate data analysis and model development. Attendees will gain valuable insights into the potential of Psyplot and its role in pushing the boundaries of visualizing climate data on unstructured grids.

Posters, Demos and Refreshments / 32

Facilitating the heterogeneous scientific data sharing with the THREDDS Control Center - Demo

Authors: Mostafa Hadizadeh[None]; Philipp Sebastian Sommer[1]; Björn Lukas Saß[1]; Christof Lorenz[2]

[1] *Helmholtz-Zentrum Hereon*

[2] *Karlsruhe Institute of Technology*

Corresponding Authors: christof.lorenz@kit.edu, mostafa.hadizadeh@kit.edu, bjoern.sass@hereon.de, philipp.sommer@hereon.de

Scientific data management is a critical aspect of collaborative research, especially in disciplines reliant on large datasets such as earth system sciences. The THREDDS Data Server (TDS), an open-source, Java-based web application, serves as a powerful tool for managing, sharing, and enabling metadata and data access to heterogeneous scientific datasets. However, its complex configuration may hinder wider adoption, especially among non-technical scientists. In response to this challenge, the open-source Django app, THREDDS Control Center, presents a solution by implementing a user-friendly web interface.

This software demonstration introduces the app, showcasing its capability to allow scientists to efficiently manage providing catalog, metadata, and access services for their datasets on THREDDS without having to deal with complex technical aspects. The application eliminates the need for direct access to the THREDDS data server infrastructure, making it accessible to a broader audience.

Key features include a flexible permission-based user management system that facilitates collaborative resource editing on the THREDDS server. This functionality empowers scientists to collectively contribute to and curate datasets without the need for extensive technical knowledge. Admins of the THREDDS-Server benefit from global server-side configurations, such as OpenDAP, WMS, etc., and an automatic reload of the THREDDS Server after configuration changes. Furthermore, the application incorporates a moderation mechanism managed by a dedicated data management team, ensuring data integrity and quality control.

One noteworthy aspect of the THREDDS Control Center is its integration with the Helmholtz AAI. This integration enables the selective sharing of resources on the THREDDS server with specific user groups or the general public. Scientists can leverage this feature to disseminate their findings to a targeted audience, fostering collaboration and information exchange within the scientific community.

In conclusion, the THREDDS Control Center presents a valuable solution for simplifying the management and sharing of NetCDF files on THREDDS servers. By providing an intuitive web frontend, collaborative editing capabilities, and seamless integration with authentication systems, this software contributes to the advancement of data-driven scientific research.

Posters, Demos and Refreshments / 33

The infancy of MANIDE: Machine learning driven Assessment

# of polymetallic Nodule mining Impacts on Deep-sea Ecosystems

Author: Verena Rubel[1]

Co-authors: Thorsten Stoeck [2]; Christina Bienhold [3]; Felix Janssen ; Massimiliano Molari [4]

[1] *MPI/AWI/RPTU*

[2] *RPTU*

[3] *AWI Helmholtz Centre for Polar and Marine Research*

[4] *MPI/AWI*

Corresponding Authors: felix.janssen@awi.de, mamolari@mpi-bremen.de, christina.bienhold@awi.de, vrubel@mpi-bremen.de, stoeck@rptu.de

Sustainability and the ecological impact of deep-sea mining operations are critical concerns addressed through environmental monitoring. Utilizing environmental DNA (eDNA) sequencing coupled with Machine Learning (ML) has proven effective in accurate monitoring, particularly in coastal environments. Currently, our goal is to broaden the application of this effective approach to encompass deep-sea environments, taking advantage of its speed and reliability.

Our goal is to understand and predict alterations in the environmental quality of the deep-sea ecosystem induced by nodule harvesting. Working with microbial communities identified through eDNA sequencing approaches, we seek to uncover species interactions and reactions to changes in environmental parameters. While Supervised Machine Learning (SML) has proven effective in coastal settings, its applicability in the deep-sea remains uncertain. Tree-based methods, such as Random Forest, emerge as potential tools for the deep sea, given the expected high dimensionality of ecological datasets derived from sequencing data. We also want to explore ML clustering approaches like k-means clustering and network analysis to extract information without prior ecological knowledge of microorganisms, crucial in the largely unexplored deep-sea environment. Overall, the prediction of various objectives, such as microbial community interactions, the prediction of biological responses, and sample categorization are enabled through classification and regression analysis provided by a multitude of ML algorithms.

We here present the MANIDE project, dedicated to this exploration, comprehensively tests sequencing approaches—metabarcoding, metagenomics, and metatranscriptomics—with ML. We want to depict recent finding as well as discuss potential bottlenecks e.g. spatial and temporal heterogeneity as they present a challenge, requiring separation from the essential impact information we aim to extract. Furthermore, the project is committed to transparency, making all data, workflow, and findings available to the scientific community.

Session 2: Data Quality and HPC for Data Science / 34

Enabling high performance DS capacity at the AWI's tier-3 HPC

Authors: Paul Gierz[1]; Miguel Andres-Martinez[1]; Pavan Siligam[1]; Seshadri Basava[1]; Stefan Pinkernell[1]; Sven Harig[1]; Bernadette Fritzsch[1]

[1] *Alfred Wegener Institute*

Corresponding Authors: paul.gierz@awi.de, stefan.pinkernell@awi.de, sven.harig@awi.de, seshadri.basava@awi.de, bernadette.fritzsch@awi.de, miguel.andres-martinez@awi.de, pavankumar.siligam@awi.de

The recent leaps in Data Science (DS) methodologies present a unique opportunity for many scientific topics to yield progress faster than ever, by leveraging increased capacity of analyzing data. However, major breakthroughs in most sciences are only achievable by combining these DS methodologies with large computing facilities, which have both the large computing capacity and the ability to fastly access and store the large data volumes. Many fields in science are lagging behind industry on the use of DS methods at large scale (e.g. in High Performance Computers). To a certain degree, the reason behind this is that the technical difficulties associated to using DS methods in distributed computing

platforms pose a high entry barrier for many scientists who are already capable of training ML models or performing parallel analysis of their data in their local machines.

The HPC & Data Processing group at the Alfred Wegener Institute (AWI) is working to bridge the gap by offering a unified DS infrastructure for accessing and operating AWI's tier-3 High Performance Computer (HPC), Albedo. In this contribution to the Data Science Symposium we will present our approach to tackling the challenge of enabling high performance DS capabilities at our institute. Our strategy includes:

1. The building of modules, environments and containers that include the most-commonly used DS software (e.g. TensorFlow) and software for data processing/postprocessing (numpy, xarray, matplotlib, etc.)
2. The deployment of JupyterHub, with easy access to these environments, and kernels for the most common data processing/postprocessing and DS tools and programming languages, such as Python, R and Julia
3. Enabling Dask (and possibly Dask Gateway) to facilitate the execution of parallel data analysis workflows
4. Explore and train users in the utilization of file formats that allow for distributed high-thoughput I/O (for example, Zarr)
5. Train users in the utilization of the GPU resources and integrate GPUs in our interactive infrastructure

Session 3: Governmental Data and Transfer / **35**

Customised data services for different target groups - experiences from Meereisportal/ SEA ICE PORTAL

Author: Annekathrin Jäkel[None]

Co-authors: Bernadette Fritzsch ; Klaus Großfeld ; Marcel Nicolaus ; Renate Treffeisen

Corresponding Authors: bernadette.fritzsch@awi.de, renate.treffeisen@awi.de, annekathrin.jaekel@awi.de, klaus.grosfeld@awi.de, marcel.nicolaus@awi.de

Sea-ice plays an essential role in the Earth system and its impact on climate change is large. It cools the entire planet, affects ocean currents and offers a habitat for countless species. Nearly 7% of the ocean is covered by sea ice, but due to climate change, sea-ice is increasingly disappearing. Research into sea-ice is therefore an important scientific endeavor in order to understand the growth and melting of sea-ice and its change and feedbacks under climate warming conditions. Research institutes from around the world are collaborating to get a close view about it and share their research results with each other. With SEA ICE PORTAL we want to make our scientific information available for everyone –promptly, scientifically based, understandable, transparent and in accessible language. Working on the realization of new ways of knowledge transfer is one of our goals.
The SEA ICE PORTAL is an information and data-portal in combination. It started in 2013 and since 2023, the joint project of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI), the Helmholtz Climate Initiative REKLIM, and the University of Bremen is available in a completely new format. Within a thorough relaunch process, the sea-ice info-portal got a modern and more accessible interface, which shall address especially users being newcomer to the topic of sea-ice. SEA ICE PORTAL presents daily updated maps of sea-ice for several parameters like sea-ice concentration or extent and additional news updates on latest trends, expeditions and reports of scientists, relined with detailed background information in German as well as in English language. The associated data-portal allows users to access raw data collected from various sources directly and work on it themselves. We also offer a wide range of customized data products, which range from maps and animations of sea-ice related parameters, atmospheric data visualizations and measurement data from deployed buoys. To optimize data flows in an O2A process, we support scientists analyzing their processes in collecting data and bring it from observation to archive in a central framework to secure the data flow for long term. SEA ICE PORTAL is a successful example for sharing a wide range of sea-ice information to the public and offering a lot of different sea-ice

depending data for scientific use. The portal is used by interested public, the media, shipping and scientific disciplines that require these customized sea ice data products.
The presentation will outline the challenges especially posed by the heterogeneous user community and the different solutions we have developed to address the target groups.

Posters, Demos and Refreshments / **36**

User-Customizable Map Visualizations in the Earth & Marine Data Portal - Demo

Author: Robin Heß[1]

Co-authors: Peter Konopatzky [1]; Christopher Krämmer [1]; Andreas Walter [1]; Karen Albers [1]; Roland Koppe [1]

[1] *Alfred-Wegener-Institut*

Corresponding Authors: peter.konopatzky@awi.de, christopher.kraemmer@awi.de, roland.koppe@awi.de, andreas.walter@awi.de, karen.albers@awi.de, robin.hess@awi.de

The Earth Data and Marine Data portals provide user-friendly platforms for exploring earth and environmental data. They function as central hubs for decentralized cross-institutional data, enhancing discoverability through a centralized search function and interactive map visualizations. These portals receive support from the German Alliance for Marine Research (DAM) and the Helmholtz-funded DataHUB project.

Configurable map viewers facilitate the visual exploration of existing data products and the creation of personalized visual and interactive data collections. Users can easily log in through their institutional account via Helmholtz-AAI, eliminating the need for a separate account. This integration allows users to save personalized viewers in their profiles and share them. Users can generate a personal link to send to colleagues or create customized viewers for the public to showcase their research areas or projects.

The map viewers support the integration of existing OGC spatial services like WMS and WFS, a drawing function for geometries on the map, the import of GeoJSON files, and the customization of various filtering options. Different visualization options for metadata are available based on the data type, for example the option to visualize charts or media data as needed.

These features create a flexible and collaborative environment for researchers to effectively share their work, either within their research group or publicly on the internet as a showcase. This presentation will provide a brief introduction and a live demonstration on creating customized viewers.

Posters, Demos and Refreshments / **37**

iLOVE: integrating Long-term Observation & Virtual Evaluation

Authors: Matthias Wietz[1]; Ovidiu Popa[2]; Christina Bienhold[3]; Ellen Oldenburg[2]; Raphael Kronberg[2]

[1] *Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research*

[2] *University of Düsseldorf*

[3] *AWI Helmholtz Centre for Polar and Marine Research*

Corresponding Authors: christina.bienhold@awi.de, matthias.wietz@awi.de, ellen.oldenburg@uni-duesseldorf.de, ovidiu.popa@uni-duesseldorf.de, raphael.kronberg@hhu.de

Deep Learning (DL) helps identifying patterns, interactions and trends among biological communities – an important asset to discern temporal dynamics across ecological and environmental gradients.

Project iLOVE, funded by the AWI DataHub program, will establish a comprehensive analytic framework tailored to heterogeneous, multiannual data; based on data from the FRAM Observatory. This ecological time-series continuously records biodiversity and their environmental drivers in Fram Strait, the major gateway between the Arctic and Atlantic Oceans. Specifically, iLOVE will apply the following modules to bacterial sequence data and contextual abiotic parameters:

Analytical Modules: Co-Occurrence Networks (discerning ecological interactions via bacterial amplicons and metagenomes); Convergent Cross Mapping (empiric dynamic modeling to understand causal relationships); Energy Landscape Algorithm (identifying stable states and transitions within the ecological landscape);

DL Modules: Graph-based Neural Networks (capturing complex relationships); Variational Autoencoders (generative models for feature extraction and representation), and Transformers (sequential dependencies).

The resulting, complementary findings on ecological dynamics, network relationships and temporal changes will be integrated with e.g. zooplankton imaging data to identify temporal connectivity across trophic levels. These insights into ecosystem functioning are essential to assess the current and future Arctic Ocean.

iLOVE will deliver (i) a Python package to allow application of the analytic framework to any time-series dataset; and (ii) an easily accessible Web interface to visualize trends in time-series data. iLOVE closely cooperates with related Helmholtz projects; developing DL methods for deep-sea monitoring (MANIDE) and promoting data interoperability across Helmholtz and beyond (HARMONise). By applying the outcomes to ongoing AWI and GEOMAR collaborations, we foster quantitative time-series analyses across Helmholtz; with a strong emphasis on FAIR science.

Session 1: Capacity Building and AI / 38

autoQC: An AI based online app for ocean data quality control

Author: Sebastian Mieruch-Schnuelle[1]

Co-authors: Gastón Kreps [1]; Mohamed Chouai [1]

[1] *AWI*

Corresponding Authors: sebastian.mieruch@awi.de, mohamed.chouai@awi.de, gaston.kreps@awi.de

Marine data quality control (QC) is crucial to provide robust data products for climate analyses, monitoring, process- and model studies and much more. However, the QC of marine measurements of e.g. temperature, salinity, nutrients (phosphate, nitrate, ⋯), oxygen etc. is challenging. Measurements are prone to errors due to external forcing (sun, wind, currents, ⋯), internal variability (e.g. extremes), biogeochemical processes, instrument errors or failures and more. Ocean data QC is an international effort and large marine data infrastructures, like SeaDataNet (https://www.seadatanet.org/), EMODnet Chemistry (https://emodnet.ec.europa.eu/en/chemistry), Argo (http://www.argodatamgt.org/) or IQuOD (https://www.iquod.org/) have created sophisticated QC processing schemes. Typically, ocean data QC is a semi automatic process, whereas the ocean experts use algorithms to identify potentially "bad"data, which are accordingly often visually inspected to make a final decision and to give the data sample a quality flag, i.e. an indicator such as "good", "bad", "probably bad", etc. One widely used tool for the QC is the Ocean Data View (ODV, https://odv.awi.de) software, which is also available as the online version webODV (https://webodv.awi.de). Because of the diverse nature of errors in the data, fully automated QC without expert visual checks is still less skillful and yields to too many misclassifications. However, visual QC is highly time demanding and skillful algorithmic support is needed, especially with the increase of fully automated sensors like the Argo buoys, where currently more than 3.800 buoys are drifting through the oceans and producing immense amounts of data.

To support the visual QC on marine data we have trained a deep neural network with the knowledge of an ocean QC data expert to mimic the human visual QC. The training of the ML algorithm is based on arctic ocean temperature data from UDASH (Unified Database for Arctic and Subarctic Hydrography). The ML algorithm improves the results of the classical checks significantly, hence increasing the data quality and reducing the experts workload.

For user friendly and easy access we have developed an online app at https://mvre.autoqc.cloud.awi.de/, where users can upload their data and let the data be quality controlled on our servers. The app provides detailed documentation and processed data are exported as simple .csv files or ODV Spreadsheet, which can be used directly in ODV or webODV (https://hifis.webodv.cloud.awi.de). The algorithm is written in Python (using Keras and Sklearn) and we provide two GitHub repositories, one which includes the sources of the algorithm, which can be used for further research or for training on other datasets. The other repository includes the fully trained model and provides an easy way to include it into other processing environments.

Currently the algorithm is limited to arctic temperature data and to two types of errors in the data, the so-called "Spikes"and "Suspect Gradients". Next planned steps are to include salinity as well as another important error type named "Statistical Screening".

Session 4: Data Initiatives / 39

HARMONise –Enhancing interoperability of marine biomolecular (meta)data across Helmholtz Centres

Authors: Christina Bienhold[1]; Till Bayer[2]; Lars Harms[None]; Roland Koppe[None]; Stefan Neuhaus[None]; Christian Sander[3]; Sophie Schindler[2]; Isabell Siebert[None]

[1] *AWI Helmholtz Centre for Polar and Marine Research*

[2] *GEOMAR*

[3] *AWI*

Corresponding Authors: soschindler@geomar.de, stefan.neuhaus@awi.de, isabell.siebert@awi.de, tbayer@geomar.de, lars.harms@awi.de, roland.koppe@awi.de, christian.sander@awi.de, christina.bienhold@awi.de

Biomolecules, such as DNA and RNA, provide a wealth of information about the distribution and function of marine organisms, and biomolecular research in the marine realm is pursued across several Helmholtz Centers. Biomolecular (meta)data, i.e. DNA and RNA sequences and all steps involved in their creation, exhibit great internal diversity and complexity. However, high-quality (meta)data management is not yet well developed and harmonized in environmentally focused Helmholtz Centers. As part of the HMC Project HARMONise, we develop sustainable solutions and digital cultures to enable high-quality, standards-compliant curation and management of marine biomolecular metadata at AWI and GEOMAR to better embed biomolecular science into broader digital ecosystems and research domains. Our approach builds on a relational database that aligns metadata with community standards such as the MIxS (Minimum Information about any (x) sequence) supported by the International Nucleotide Sequence Database Collaboration (INSDC) to promote global interoperability. At the same time, we ensure the harmonization of metadata with existing Helmholtz repositories (e.g. PANGAEA). A web-based hub enables the standardized export and exchange of core metadata, e.g. with the Marine Data Portal (https://marine-data.de/), which will enhance the **findability** and **accessibility** of biomolecular (meta)data within and across research areas. The alignment of HARMONise-hosted metadata with domain-specific standards and the provision of data in the relevant exchange formats will facilitate **interoperability** with the Helmholtz knowledge graph (UNHIDE, https://docs.unhide.helmholtz-metadaten.de/intro.html) and global digital ecosystems (Ocean Info Hub of the UNESCO Ocean Data and Information System, https://oceaninfohub.org/). HARMONise thus specifically targets the advancement of F, A, and I in FAIR for biomolecular (meta)data, and supports Helmholtz researchers in delivering high-quality metadata to international data repositories. HARMONise connects with high-level international projects in the Ocean Biomolecular

Observing Network (OBON) Programme of the UN Decade of Ocean Science, to further align our developments with global strategies and ensure Helmholtz-to-global interoperability. The project HARMONise (ZT-I-PF-3-027) is funded by the Initiative and Networking Fund as part of the Helmholtz Metadata Collaboration Project cohort 2021.

Posters, Demos and Refreshments / 40

HMC Earth and Environment - Developing a Robust Framework for Seamless Semantic Interoperability in Earth and Environmental Research

Authors: Dorothee Kottmeier[1]; Andrea Pörsch[2]; Yousef Razeghi[3]; Stanislav Malinovschii[4]; Emanuel Söding[4]

[1] *AWI Bremerhaven*

[2] *GFZ Potsdam*

[3] *UFZ Leipzig*

[4] *GEOMAR Kiel*

Corresponding Authors: dorothee.kottmeier@pangaea.de, yousef.razeghi@ufz.de, andrea.poersch@gfz-potsdam.de, smalinovschii@geomar.de, esoeding@geomar.de

The HMC Earth and Environment Hub is dedicated to establishing a robust framework for seamless semantic interoperability in the field of earth and environmental research. Our approach involves strategically coordinating processes within the Helmholtz Association. Standardizing and semantically annotating metadata and harmonizing existing semantic resources are crucial for bridging the gap across diverse and complex data sets. Our focus on semantic interoperability encompasses several key aspects: 1. Navigating semantic resources is challenging. We guide in evaluating and using existing vocabularies, terminologies, ontologies, and services, setting criteria for sustainability and recommendability. 2. Identifying and prioritizing crucial metadata for semantic annotation, we assess elements essential for meaningful data connections and explore relevant tools and vocabularies. 3. Building a unified data space requires consensus at national and international levels. We collaborate with networks, aligning processes, and adapting services to community needs. 4. Facilitating community collaboration is vital for harmonization. We moderate processes like agreeing on metadata schema and vocabulary, contributing to governance structure development. Our overarching goal is to create a collaborative environment that fosters effective semantic interoperability, streamlining the integration of diverse data sources within the earth and environmental research community.

Posters, Demos and Refreshments / 41

NFDI4Earth Academy - Your training network to bridge Earth System and Data Sciences

Authors: Effi-Laura Drews[1]; Jonas Kuppler[2]; Kristin Sauerland[3]

Co-authors: Hildegard Gödde [2]; Konstantin Ntageretzis [1]; Gauvain Wiemer [4]

[1] *Forschungszentrum Jülich & Geoverbund ABC/J*

[2] *German Research Centre for Geosciences (GFZ) & Geo.X*

[3] *University Bremen/MARUM & Deutsche Allianz Meeresforschung (DAM)*

[4] *Deutsche Allianz Meeresforschung (DAM)*

Corresponding Authors: ksauerland@marum.de, hildegard.goedde@gfz-potsdam.de, jonas.kuppler@gfz-potsdam.de, wiemer.dam@gmail.com, e.drews@fz-juelich.de, k.ntageretzis@fz-juelich.de

The NFDI4Earth Academy is a network of early career –doctoral and postdoctoral - scientists interested in bridging Earth System and Data Sciences beyond institutional borders. The research networks Geo.X, Geoverbund ABC/J, and DAM offer an open science and learning environment covering specialized training courses and collaborations within the NFDI4Earth consortium with access to all NFDI4Earth innovations and services. Academy fellows advance their research projects by exploring and integrating new methods and connecting with like-minded scientists in an agile, bottom-up, and peer-mentored community. We support young scientists in developing skills and mindset for open and data-driven science across disciplinary boundaries.
The first cohort of Fellows successfully completed their first year in the Academy and look forward to continuing their journey with events including an online Spring School and Think Tank. In addition, the second cohort of Fellows will join the Academy this spring.

Posters, Demos and Refreshments / 42

Data science for understanding physics –modelling ship wake detectability using machine learning

Author: Bjoern Tings[1]

[1] *DLR - Remote Sensing Technology Institute*

Corresponding Author: bjoern.tings@dlr.de

The detectability of wake signatures in satellite-based Synthetic Aperture Radar (SAR) acquisitions is dependent on various physical parameters describing the present situation during the detection. Ship wake signatures in SAR are complex structures consisting of multiple wake components appearing differently depending on the present detection situation. The physical parameters with influence on the detectability of those wake components are in the following called influencing parameters. Although various methods for automatic detection of wakes are being developed since decades, the dependency between detectability of wake components and the influencing parameters is not systematically analyzed. In this study, machine learning is applied to model the dependency between all wake components taking all influencing parameters into account. The composition of the machine learning models is analyzed in order to derive statements on physical relationships between influencing parameters and detectability of wake components. For this type of application, a figure of merit for detectability and a measure for uncertainty of derived statements is introduced. The results are contrasted against literature based on simulations and/or physical deductions on ship wakes in SAR imagery and their detectability.

It is demonstrated that data science is not only useful for solving a specific task, i.e. wake component detection, but also to systematically generate understanding of the task's underlying physics, i.e. wake component detectability. The systematic modelling of underlying physics can finally be applied for improving the specific task.

Session 3: Governmental Data and Transfer / 43

About the the collaboration between research institutions and authorities within the German Marine Research Alliance

Author: Gauvain Wiemer[1]

Co-authors: Susanne Tamm [2]; Kirsten Binder [3]; Henning Gerstmann [4]; Robin Heß [5]; Maximilian Betz [5]; Roland Koppe [5]; Carsten Schirnick [6]; Manuela Köllner [2]

[1] *Deutsche Allianz Meeresforschung (DAM)*

[2] *Bundesamt für Seeschifffahrt und Hydrographie*

[3] *Bundesanstalt für Wasserbau*

[4] *Bundesamt für Naturschutz*

[5] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung*

[6] *GEOMAR - Helmholtz-Zentrum für Ozeanforschung Kiel*

Corresponding Authors: roland.koppe@awi.de, maximilian.betz@awi.de, cschirnick@geomar.de, susanne.tamm@bsh.de, henning.gerstmann@bfn.de, robin.hess@awi.de, kirsten.binder@baw.de, manuela.koellner@bsh.de, wiemer.dam@gmail.com

In 2019, German marine research, together with the federal government and the northern German states of Bremen, Hamburg, Mecklenburg-Western Pomerania, Lower Saxony, and Schleswig-Holstein, established the German Alliance for Marine Research (DAM). With this initiative, Germany has launched one of the world's largest marine research alliances.
The goal of DAM is to strengthen the sustainable management of coasts, seas, and oceans through research and knowledge transfer, data management and digitalization, as well as the coordination of infrastructures. Working together with its member institutions, DAM is dedicated to implementing an integrated and reliable data management concept for the research landscape. For the collaboration between research institutions and authorities within DAM, this entails the joint adoption of guidelines and SOPs, the shared use of data infrastructures, and the joint provision of data products to support open access to marine research data according to the FAIR principles.

Session 1: Capacity Building and AI / 44

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

Authors: Michael Schlundt[None]; Julia Oelker[None]; Robert Kopte[None]; Gauvain Wiemer[None]

Corresponding Authors: mschlundt@geomar.de, robert.kopte@ifg.uni-kiel.de, wiemer@allianz-meeresforschung.de, julia.oelker@uni-oldenburg.de

The Davis-SHIP-system (DSHIP) is designed to store and manage environmental and system data collected during expeditions of German research vessels. These data encompass a wide range of information, including physical and chemical parameters of seawater as well as data on weather conditions and other environmental variables. The DAM project "Underway"research data undertakes a scientific evaluation and provides the quality-controlled data of selected environmental parameters in a FAIR manner, openly accessible for re-use. Scientists can leverage these data to gain new insights into marine ecosystems, climate change, and other vital aspects of marine research. Quality control of the data sets can be time-consuming and subjective, when manual flagging needs to be applied, because "classic"quality-control routines struggle to adequately flag the data. To provide the user with data faster and of higher quality, we explore artificial intelligence (AI) approaches within the "Underway"research data project. As a first step, common features are examined with the help of AI in a wide suit of parameters stored in DSHIP. As a second step, we train an AI to obtain quality-controlled data from "raw"DSHIP data and compare the results to the classically quality-controlled data. These approaches aid in identifying patterns and trends within the data that may be of interest for scientific analysis and are a step to further automation of the quality control process.

Session 1: Capacity Building and AI / 46

Keynote: AI

Author: Guido Große[1]

[1] *AWI*

Session 3: Governmental Data and Transfer / 47

Keynote: Marine authority data flow ~ Linked data - from local data nodes to harmonized services

Author: Michael Räder[1]

[1] *MDI-Niedersachsen*

Session 4: Data Initiatives / 48

DataNord: Empowering Bremen's Research Community in Data Literacy

Authors: Lena Steinmann[1]; Tanja Hörner[2]

Co-authors: Frank Oliver Glöckner [3]; Iris Pigeot [4]; Rolf Drechsler [1]

[1] *Data Science Center, Universität Bremen*

[2] *Universität Bremen, U Bremen Research Alliance*

[3] *Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung*

[4] *Leibniz-Institut für Präventionsforschung und Epidemiologie –BIPS*

Corresponding Author: lena.steinmann@uni-bremen.de

DataNord is an interdisciplinary data competence center for the Bremen region, fostering interdisciplinary collaboration and empowering researchers at all career levels to improve their data skills. Through a comprehensive range of practical training courses, self-learning resources, interactive hackathons, personalised support services and networking opportunities, DataNord is dedicated to promoting data literacy across Bremen's research community.

Funded by the BMBF, DataNord is being established as part of the U Bremen Research Alliance – the cooperation network of the University of Bremen and twelve non-university research institutes –and in close cooperation with other partners. It brings together the expertise of universities, non-university research institutions, state institutes, NFDI consortia, and infrastructure centers. The network's profile areas include (1) environmental and marine sciences, (2) social sciences, (3) material and engineering sciences, (4) health sciences, and (5) humanities.

Central to DataNord are two pivotal pillars: The Data Science Center (DSC) at the University of Bremen and the interdisciplinary doctoral training programme "Data Train - Training in Research Data Management and Data Science"of the U Bremen Research Alliance. The DSC establishes a central help desk offering consultation services and specialized trainings to researchers from all participating institutions. Notably, the "Rent a Data Scientist" service facilitates in-depth support for research endeavours, enabling researchers to integrate specialized data scientists directly into their projects for defined durations.

The Data Train programme is provided annually and is targeted at doctoral researchers (but also open to everyone interested whenever possible). The focus is on basic competencies in research data management and data science. Associated with the NFDI and now embedded into DataNord, the programme's curriculum and additional training components will be further developed under the DataNord umbrella.

The initiative further extends its impact through a citizen science project, "Guardians of the Hedgehogs", and strategic science communication, facilitating knowledge transfer to society, industry, and politics.

Our talk will delve into DataNord's role within the Bremen research landscape and explore how scientists can leverage its diverse data services to enhance their data proficiency and maximize the impact of their research endeavours.

Posters, Demos and Refreshments / 49

Exploring the soundscapes of the world's oceans –A demonstration of OPUS, the Open Portal to Underwater Soundscapes

Author: Karolin Thomisch[1]

Co-authors: Lewin Probst [2]; Olaf Boebel [1]; Robin Heß [1]

[1] *AWI*

[2] *software company emirror-de*

Corresponding Author: karolin.thomisch@awi.de

Facing an era of rapid, anthropogenically induced changes in the global oceans, there is increasing acknowledgement of anthropogenic noise as a marine pollutant. Ocean sound is now considered an essential ocean variable by the Global Ocean Observing System (GOOS) and the need for ambient noise surveillance is emphasized not only by the scientific community, but also by policy-making entities.

Our comprehension of the role of underwater sound in the marine realm and our understanding of long-term trends in anthropogenic sound and its effects on marine life and ecosystem health is greatly fostered by personal experience of these soundscapes. To this end, the OPUS (Open Portal to Underwater Soundscapes, https://opus.aq/) data portal is currently being developed by the Ocean Acoustics Group at the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) in Bremerhaven, Germany, embedded in the Helmholtz Association's DataHub initiative and coordinated by the German Marine Research Alliance (DAM).

Designed as an expeditious discovery tool for archived passive acoustic data, OPUS promotes the use of archived acoustic data collected worldwide by providing open access to stacks of spectrograms, progressively enhancing in temporal resolution. To motivate data provision and use, OPUS adopts the FAIR principles for submitted data while assigning a most permissive CC-BY 4.0 license to all OPUS products (i.e. visualizations of and lossy compressed audio data). This unprecedented opportunity to experience marine soundscapes collected worldwide is intended to address a broad range of stakeholders, from the general public, to artists, journalists, fellow scientists, regulatory bodies, consulting companies, and the marine industry, to learn about and access the data for the respective needs of each stakeholder group.

With data becoming openly accessible, the public and marine stakeholders will be able to easily compare soundscapes from different regions, seasons and environments, with and without anthropogenic contributions. Thereby, OPUS contributes to an improved understanding of the world's oceans soundscapes and anthropogenic impacts thereon over various temporal and spatial scales.

During the 9th Data Science Symposium, held in Bremen in 2024, features and functionalities of OPUS will be demonstrated to all interested ocean enthusiasts eager to explore the sound(scape)s of remote and inaccessible areas such as the polar oceans.