Data Science Symposium 2023

Thursday, June 8th 2023 - Friday, Juni 9th 2023

GEOMAR East Shore Campus

Book of abstracts

Group photo of the participants of the 8th Data Science Symposium at GEOMAR.

Edited by Timm Schoening, Ulrike Kleeberg and Stephan Frickenhaus

Contents

Talks / 24

Recent Highlights from Helmholtz AI Consulting in Earth & Environment

Autor Caroline Arnold[1]

Co-Autoren: Adeniyi Mosaku ; Danu Caus [2]; Harsh Grover [2]; Paul Keil [2]; Tobias Weigel [2]

[1] *Deutsches Klimarechenzentrum*

[2] *DKRZ*

Corresponding author: arnold@dkrz.de

As Helmholtz AI consultants we support researchers in their machine learning and data science projects. Since our group was founded in 2020, we have completed over 40 projects in various subfields of Earth and Environmental sciences, and we have established best practices and standard workflows. We cover the full data science cycle from data handling to model development, tuning, and roll-out. In our presentation, we show recent highlights to give an impression of how we support researchers in the successful realisation of their machine learning projects and help you to make the most of your scientific data.
In particular, we will present our contributions to:
* Detecting sinkhole formation near the Dead Sea using computer vision techniques (Voucher by D. Al-Halbouni, GEOMAR)
* Integrating trained machine learning algorithms into Earth system models (Voucher by S. Sharma, HEREON)
* Encoding ocean model output data for scientific data sonification (Voucher by N. Sokolova, AWI)
* Operational machine learning: automated global ocean wind speed predictions from remote sensing data
Our services as Helmholtz AI consultants are available to all Helmholtz researchers free of charge. We offer support through so-called vouchers, which are up to six-month long projects with clearly defined contributions and goals. Beyond individual support, computational resources and training courses are available through Helmholtz AI. If you would like to get in touch, please contact us in person during the symposium, via consultant-helmholtz.ai@dkrz.de, or at

https://www.helmholtz.ai/themenmenue/you-helmholtz-ai/ai-consulting/index.html.

Talks / **27**

Beyond ESGF - Regional climate model datasets in the cloud on AWS S3

Autor Lars Buntemeyer[1]

[1] *Helmholtz-Zentrum Hereon*

Corresponding author: lars.buntemeyer@hereon.de

The Earth System Grid Federation (ESGF) data nodes are usually the first address for accessing climate model datasets from WCRP-CMIP activities. It is currently hosting different datasets in several projects, e.g., CMIP6, CORDEX, Input4MIPs or Obs4MIPs. Datasets are usually hosted on different data nodes all over the world while data access is managed by any of the ESGF web portals through a web based GUI or the ESGF Search RESTful API. The ESGF data nodes provide different access methods, e.g., https, opendap or globus.
Beyond ESGF, there has been the Pangeo / ESGF Cloud Data Working Group that coordinates efforts related to storing and cataloging CMIP data in the cloud, e.g., in the Google cloud and in the Amazon Web Services Simple Storage Service (S3) where a large part of the WCRP-CMIP6 ensemble of global climate simulations is now available in analysis-ready cloud-optimized (ARCO) zarr format. The availibility in the cloud has signifcantly lowered the barrier for users with limited resources and no access to an HPC environment to work with CMIP6 datasets and at the same time increases the chance for reproducibilty and reusability of scientific results.
We are now in the process of adapting the Pangeo strategy for publishing also regional climate model datasets from the CORDEX initiative on AWS S3 cloud storage. Thanks to similar data formats and meta data conventions in comparison to the global CMIP6 datasets, the workflows require only minor adaptations. In this talk, we will show the strategy and workflow implemented in python and orchestrated in github actions workflows as well as a demonstration of how to access CORDEX datasets in the cloud.

Talks / 9

Forecast system for urban air quality in the city of Hamburg, Germany

Autor Rehan Chaudhary[1]

Co-Autor: Matthias Karl

[1] *Hereon*

Corresponding author: rehan.chaudhary@hereon.de

Air quality in urban areas is an important topic not only for science but also for the concerned citizen as the air quality affects personal wellbeing and health. At Hereon in the project SMURBS (SMART URBAN SOLUTIONS) a model to forecast urban air quality at high spatial and timely resolution was developed (CityChem) with the aim to carry out exposure studies and future scenario calculations and provide current air quality information to the citizen at the same time. The daily model output is presented in the 'Urban Air Quality Forecast' Tool, which is based on an ESRI and ArcGIS application. The tool helps citizens and decision makers to investigate forecasted 24 hour pollutant data for 7 pollutants including NO_2, NO, O_3, SO_2, CO, PM10, PM25 along with the AQI Index. The model output is generated in the form of network Common Data Form (netCDF) files which are brought into the GIS environment using ArcGIS API for JavaScript. The netCDF files are converted into multidimensional rasters to display air pollutants information over a 24-hour period. These rasters are then shared as web services over Hereon's Geohub Portal from where they are consumed to be visualized over the Urban Air Quality Forecast Tool. The data on the tool is updated daily via cron jobs which run the complete implemented workflow of fetching the newly generated model outputs and updating shared web services. This tool offers many different opportunities to investigate the current air quality in Hamburg.

Talks / 29

Data collaboration without borders: bringing together people, data, and compute in the ZMT datalab

Autor Arjun Chennu[1]

[1] *ZMT*

Corresponding author: arjun.chennu@leibniz-zmt.de

Creating data collaborations that are both effective and equitable is challenging. While there are many technical challenges to overcome, there are also significant cultural issues to deal with, for example, communication styles, privacy protection and compliance to institutional and legal norms. At ZMT Bremen, we had to address the central question of "How to build a good data collaboration with our tropical partners?" while also dealing with the restrictions of the pandemic, which laid bare the digital divides that separate us. In 2020, the ZMT datalab (www.zmt-datalab.de) was born out of the search for an answer. The datalab is a data platform that brings together people, data and compute resources, with a strategy to not just share data but to share work, and without borders. Here we present the design principles and considerations to navigate the conflicting incentives that go towards building a datalab that is both open and effective.

Poster session / 5

FAIR bathymetry data archiving in PANGAEA - Data Publisher for Earth & Environmental Science

Autor Daniel Damaske[1]

[1] *PANGAEA*

Corresponding author: daniel.damaske@pangaea.de

Compliant with the FAIR data principles [1], long-term archiving of bathymetry data from multibeam echosounders – a highly added value for
the data life cycle - is the challenging task in the data information system PANGAEA. To cope with the increasing amount of data
("bathymetry puzzle pieces") acquired from research vessels and the demand for easy "map-based" means to find valuable data, new semiautomated
processes and standard operating procedures (SOPs) for bathymetry data publishing and simultaneous visualization are currently
developed and implemented, anchored in the O2A - Data Flow Framework from Sensor Observations to Archives [2]. This research is part of
the "Underway Research Data" project, an initiative of the German Marine Research Alliance (Deutsche Allianz Meeresforschung e.V., DAM).

Talks / 13

OSIS: From Past to Present and Future

Autoren FDI Digital Research Services[1]; FDM Digital Research Services[1]

Speaker Claas Faber[1]

[1] *GEOMAR*

Corresponding author: fdi@geomar.de

The Ocean Science Information System (OSIS) was first established in 2008 at GEOMAR as a metadata-centric platform for ongoing marine research collaborations. It has become a vital platform for collecting, storing, managing and distributing ocean science (meta)data. We will discuss how OSIS has enabled multidisciplinary ocean research and facilitated collaborations among scientists from different fields.

Furthermore, we will present the major refactoring that is currently underway to move OSIS to a new microservice architecture. This architecture will not only update the OSIS website but will also expose all of the information stored in the OSIS database as a RESTful API webservice. This move will make it easier for researchers and other stakeholders to access and use the data, promoting transparency, and open data practices.

Additionally, we will discuss how OSIS will continue to evolve to meet the changing needs of ocean research and the DataHub community in the future.

Join us for this exciting event to learn more about the history and future of OSIS and how the move to microservices and APIs will enhance the platform's capabilities and impact, and how it will continue to add scientific visibility and reusability to your research's data output.

Poster session / 14

Unlocking the Power of APIs: An Interactive Demonstration of a Federated Workspace for Marine (Meta)Data

Autoren FDI Digital Research Services[1]; DLS Data Division[2]

Speaker Carsten Schirnick[1]

[1] *GEOMAR*

[2] *AWI*

Corresponding author: fdi@geomar.de

In an interactive demonstration, we will showcase how already available APIs can be used to unlock the power of a federated workspace for marine research, allowing for collaboration and data sharing across multiple services and applications.

Through the use of APIs, distributed metadata-services, such as the AWI Sensor Registry and GEOMARs Ocean Science Information System, DSHIP systems aboard research vessels and onland can be connected with data streams and integrated into a single workspace, i.e. for example a Jupyter Notebook or a customised website. During the demonstration, attendees will learn how to use APIs of various services within MareHub - DataHub - DAM to enable collaboration across organisations and to streamline data workflows.

Attendees will have the opportunity to ask questions and provide feedback on the use and their needs of APIs in scientific data analysis and provisioning of up-to-date metadata. By the end of the demonstration, attendees will have a better understanding of how APIs can be used to unlock the power of a federated workspace, and how such a workspace can enhance collaboration and (meta)data sharing across multiple services and applications.

Talks / 6

The NFDI4Earth Academy – Your training network to bridge Earth System and Data Science

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

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

[2] *Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum GFZ*

[3] *MARUM - Center for Marine Environmental Sciences; GFBio e.V.*

Corresponding author: e.drews@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 has successfully completed their first six months within the Academy and is looking forward to a number of exciting events this year, such as a Summer School, a Hackathon and a Think Tank event. Furthermore, this autumn the second call for fellows will be released.

Talks / **7**

The Renewed Marine Data Infrastructure Germany (MDI-DE)

Autor Henning Gerstmann[1]

Co-Autoren: Jochen Krause [1]; Johannes Melles [2]; Kirsten Binder [3]

[1] *Bundesamt für Naturschutz*

[2] *Bundesamt für Seeschifffahrt und Hydrographie*

[3] *Bundesanstalt für Wasserbau*

Corresponding author: henning.gerstmann@bfn.de

The processes of search, access and evaluation of geospatial datasets is often a challenging task for researches, managers, administrations and interested users due to highly distributed data sources. To provide a single access point to governmentally provided datasets for marine applications, the project "Marine Spatial Data Infrastructure Germany" (MDI-DE) was launched in 2013. Its geoportal has been completely renewed on open-source software to follow the approaches of the FAIR-principles and compliance to international standards regarding web services and metadata.

The MDI project is driven by a cooperation of 3 federal and 3 regional authorities, which represent the coastal federal states of Germany. Each of the signatory authorities operates a proprietary spatial data infrastructure based on different technical components to publish web services of, among others, the marine domain. So, one core component of the MDI-DE is a metadata search interface across all participants' SDI's. The provided marine-related services are regularly harvested into the own MDIs' metadata catalogue using OGC CSW-services. The interface enables users to find datasets using keywords and filters. Besides of formal and legal information, it provides the possibility to visualise the services and to download the underlying datasets.
Second, a web mapping application is part of MDI-DE. Here, a large number of categorised view services, originating from the consortium and from external institutions, is available. Moreover, the portal enables any user to import own geodata from disc or from OGC-conformant web services. Further, a map server is deployed within the MSDI to publish harmonised WMS and WFS services based on distributed data sources.
MDI-DE and the German Marine Research Alliance (DAM) plan to intensify their cooperation and to promote their respective public visibility by referencing each other's portals, integrate their metadata catalogues and jointly develop new products.

Poster session / 30

Shark out of water: Augmenting classification of imagery from caught sharks with machine learning

Autor Fridolin Haag[1]

Co-Autor: Arjun Chennu [1]

[1] *Leibniz Centre for Tropical Marine Research (ZMT)*

Corresponding author: fridolin.haag@leibniz-zmt.de

The incorporation of machine learning (ML) into wildlife monitoring has the potential to revolutionize data collection and conservation initiatives. We outline a human-in-the-loop ML approach for the identification of sharks caught along Tanzanian coasts. The image sources for this monitoring come from marketplaces or landing sites, resulting in a vast range of image quality, the state and poses of sharks, camera angles, and background and foreground clutter. The ML integration aims to augment an existing manual identification process, which is time consuming and requires specialized training. Our strategy includes mapping the current workflow, identifying areas for ML optimization, and adapting existing classifiers to the context and imagery. The proposed approach includes an object detection model that identifies shark objects in images and a hierarchical classifier that classifies shark images down to genus and species level, both using deep learning frameworks. Key challenges include class imbalance, the image variability, and the effective use of multiple views for single specimen identification. A project interest is to integrate the ML approach into the species identification workflow of human experts to determine the degree of automation that is feasible and beneficial. This study provides a step towards combining technological advances with marine conservation efforts to improve data quality and operational efficiency.

Talks / 17

DS2STAC: A Python package for harvesting and ingesting (meta)data into STAC-based catalog infrastructures

Autor Mostafa Hadizadeh[1]

Co-Autoren: Christof Lorenz [1]; Sabine Barthlott [1]; Romy Fösig [1]; Uğur Çayoğlu [1]; Robert Ulrich [1]; Felix Bach [2]

[1] *Karlsruher Institut für Technologie*

[2] *FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur*

Corresponding author: mostafa.hadizadeh@kit.edu

Despite the vast growth in accessible data from environmental sciences over the last decades, it remains difficult to make this data openly available according to the FAIR principles. A crucial requirement for this is the provision of metadata through standard catalog interfaces or data portals for indexing, searching, and exploring the stored data.

With the release of the community-driven Spatio-Temporal Assets Catalog (STAC), this process has been substantially simplified as STAC is based on highly flexible and lightweight GeoJSONs instead of large XML-files. The number of STAC-users has hence rapidly increased and STAC now features a comprehensive ecosystem with numerous extensions. This is also why we have chosen STAC as our central catalog framework in our research project Cat4KIT, in which we develop an open-source software stack for the FAIRification of environmental research data.

A central element of this project is the automatic (meta)data and service harvesting from different data servers, providers and services. This so-called DS2STAC-module hence contains tailormade harvesters for different data sources and services, a metadata validator and a database for storing the STAC items, collections and catalogs. Currently, DS2STAC can be used for harvesting from THREDDS-Server, Intake-Catalogs, and SensorThings APIs. In all three cases, it creates and manages consistent STAC-items, -catalogs and -collections which are then made openly available through the pgSTAC-database and the STAC-FastAPI to allow for a user-friendly interaction with our environmental research data.

In this presentation, we want to demonstrate DS2STAC and also show its functionalities within our Cat4KIT-framework. We also want to discuss further use-cases and scenarios and hence propose DS2STAC as a modular tool for harvesting (meta)data into STAC-based catalog infrastructures.

36

Data Stories - Data Science Unit Storytelling Techniques

Autor Anne Hennke[1]

[1] *GEOMAR*

Corresponding author: ahennke@geomar.de

Tell the world the story of your data with ArcGIS storymaps - present them and their value to an even broader audience and possibly reach overarching disciplines.
As part of the Data Science Unit (DSU) portfolio at GEOMAR we want to create Storymaps that can give other researchers, as well as nonscientific interested parties, a quick initial overview of a particular topic. This poster gives insights into first DSU storymap products and you are invited to scroll through the Storymap about the Boknis Eck Time Series Station.

Talks / 33

The Earth Data Portal for Finding and Exploring Research Content

Autor Robin Heß[1]

Co-Autoren: Andreas Walter [1]; Karen Albers [1]; Peter Konopatzky [1]; Roland Koppe [1]

[1] *Alfred-Wegener-Institut*

Corresponding author: robin.hess@awi.de

Digitization and the Internet in particular have created new ways to find, re-use, and process scientific research data. Many scientists and research centers want to make their data openly available, but often the data is still not easy to find because it is distributed across different infrastructures. In addition, the rights of use, citability and data access are sometimes unclear.

The Earth Data Portal aims to provide a single point of entry for discovery and re-use of scientific research data in compliance with the FAIR principles. The portal aggregates data of the earth and environment research area from various providers and improves its findability. As part of the German Marine Research Alliance and the Helmholtz-funded DataHub project, leading German research centers are working on joint data management concepts, including the data portal.

The portal offers a modern web interface with a full-text search, facets and explorative visualization tools. Seamless integration into the Observation to Analysis and Archives Framework (O2A) developed by the Alfred Wegener Institute also enables automated data flows from data collection to publication in the PANGAEA data repository and visibility in the portal.

Another important part of the project is a comprehensive framework for data visualization, which brings user-customizable map viewers into the portal. Pre-curated viewers currently enable the visualization and exploration of data products from maritime research. A login feature empowers users to create their own viewers including OGC services-based data products from different sources.

In the development of the portal, we use state of the art web technologies to offer user-friendly and high-performance tools for scientists. Regular demonstrations, feedback loops and usability workshops ensure implementation with added value.

Poster session / 32

A Flexible yet Sustainable Spatial Data Infrastructure for the Integration of Distributed Research Data

Autor Peter Konopatzky[1]

Co-Autoren: Andreas Walter [1]; Robin Heß [1]; Roland Koppe [1]

[1] *AWI*

Corresponding author: peter.konopatzky@awi.de

The need for discoverability and accessibility of research data and metadata is huge, driven both by the FAIR principles and user requirements regarding research data portals, repositories and search engines. Interactive, visual and especially map-based exploration of research data is becoming increasingly popular. Bringing together technical reality and custom user vision in the design and provision of services for interactive map viewers without sacrificing sustainability can be challenging.

Thanks to our O2A Spatial framework, the classic Spatial Data Infrastructure (SDI) components — storage, database, geo web server, catalogue — can be (re)deployed and configured quickly and with low effort. This includes the creation and curation of data products which can be compiled from differently-sourced data. The list of currently supported data sources contains the PANGAEA repository, the Observations to Archives and Analysis (O2A) pipeline, Sensor Observation Services (SOS) and data provided by scientists directly. Simple metadata harmonisation is possible. Public available Standard Operating Procedures (SOPs) and data exchange specifications document the ways in which scientists and institutes can have their desired products hosted.

The modular, scalable, flexible and highly automated SDI has been developed and operated at Alfred Wegener Institute (AWI) for more than a decade, continuously improving and providing map services for GIS clients and portals including the Marine Data and Earth Data Portals.

Long-term maintainability is ensured through the use of common open-source technologies, established geodata standards, containerisation and the high degree of automation. The modularity of O2A Spatial and SDI components ensures flexibility and future expandability. Being embedded into O2A, SDI development and operation is financially and staff-wise secured in the long run.

Talks / 3

Data-driven Attributing of Climate Events with a Climate Index Collection based on Model Data (CICMoD)

Autor Marco Landt-Hayen[1]

Co-Autoren: Willi Rath [1]; Sebastian Wahl [1]; Nils Niebaum [1]; Martin Claus [2]; Peer Kröger [2]

[1] *GEOMAR Helmholtz Centre for Ocean Research*

[2] *Christian-Albrechts-Universität zu Kiel*

Corresponding author: mlandt-hayen@geomar.de

Machine learning (ML) and in particular artificial neural networks (ANNs) push state-of-the-art solutions for many hard problems e.g., image classification, speech recognition or time series forecasting. In the domain of climate science, ANNs have good prospects to identify causally linked modes of climate variability as key to understand the climate system and to improve the predictive skills of forecast systems. To attribute climate events in a data-driven way with ANNs, we need sufficient training data, which is often limited for real world measurements. The data science community provides standard data sets for many applications. As a new data set, we introduce a collection of climate indices typically used to describe Earth System dynamics. This collection is consistent and comprehensive as we use control simulations from Earth System Models (ESMs) over 1,000 years to derive climate indices. The data set is provided as an open-source framework that can be extended and customized to individual needs. It allows to develop new ML methodologies and to compare results to existing methods and models as benchmark. Exemplary, we use the data set to predict rainfall in the African Sahel region and El Niño Southern Oscillation with various ML models. We argue that this new data set allows to thoroughly explore techniques from the domain of explainable artificial intelligence to have trustworthy models, that are accepted by domain scientists. Our aim is to build a bridge between the data science community and researchers and practitioners from the domain of climate science to jointly improve our understanding of the climate system.

Talks / 20

The Coastal Pollution Toolbox: A collaborative workflow to provide scientific data and information on marine and coastal pollution

Autor Marcus Lange[1]

Co-Autoren: Linda Baldewein ; Louis Celliers ; Ralf Ebinghaus ; Ulrike Kleeberg [1]

[1] *Hereon*

Corresponding author: marcus.lange@hereon.de

Interaction between pollution, climate change, the environment, and people is complex. This complex interplay is particularly relevant in coastal regions, where the land meets the sea. It challenges the scientific community to find new ways of transferring usable information for action into actionable knowledge and into used information for managing the impact of marine pollution. Therefore, a better understanding is necessary of the tools and approaches to facilitate user engagement in co-designing marine pollution research in support of sustainable development.

This presentation highlights a collaborative workflow to provide scientific data and information on marine and coastal pollution. The Coastal Pollution Toolbox provides innovative tools and approaches by marine pollution science to the scientific community, the public and agencies and representatives from policymaking and regulation. The framework provides a suite of scientific communication, science and synthesis, and management products. Essential components for the provision of the tools are the underlying data infrastructures, including data management technologies and workflows.

The presentation concludes by establishing the scope for a ‘last mile’ approach supporting engagement and co-designing by incorporating scientific evidence of pollution into decision-making.

Keywords: Marine pollution, participatory approaches, co-design, engagement, tools, and approaches, HCDC infrastructure

Talks / 16

Rapid segmentation and measurement of an abundant mudsnail: Applying superpixels to scale accurate detection of growth response to ocean warming

Autor Liam MacNeil[1]

Co-Autoren: Léa J. Joly [1]; Marco Scotti [1]; Maysa Ito [1]

[1] *GEOMAR*

Corresponding author: lmacneil@geomar.de

Advances in computer vision are applicable across aquatic ecology to detect objects from images, ranging from plankton and marine snow to whales. Analyzing digital images enables quantification of fundamental properties (e.g., species identity, abundance, size, and traits) for richer ecological interpretation. Yet less attention has been given to imaging fauna from coastal sediments, despite these ecosystems harboring a substantial fraction of global biodiversity and whose members are critical for nitrification, cycling carbon, and filtering water pollutants. In this work, we apply a superpixel segmentation approach— grouped representations of underlying colors and other basic features— to automatically extract and measure abundant hydrobiid snails from images and assess size-distribution responses to experimental warming treatments. Estimated length measurements were assessed against >4500 manually measured individuals, revealing high accuracy and precision across samples. This method expedited hydrobiid size measurements tremendously, tallying >40k individuals, reducing the need for manual measurements, limiting human measurement bias, generating reproducible data products that could be inspected for segment quality, and revealed species growth response to ocean warming.

Talks / **15**

webODV - Interactive online data visualization and analysis

Autor Sebastian Mieruch[1]

Co-Autoren: Ingrid Linck Rosenhaim [1]; Reiner Schlitzer [1]

[1] *AWI*

Corresponding author: sebastian.mieruch@awi.de

webODV is the online version of the worldwide used Ocean Data View Software (https://odv.awi.de). During the recent years **webODV** (https://webodv.awi.de) has been evolved into a powerful and distributed cloud system for the easy online access of large data collections and collaborative research. Up to know more than 1000 users per months (worldwide) are visiting our **webODV** instances at

- *GEOTRACES*: https://geotraces.webodv.awi.de
- *EMODnet Chemistry*: https://emodnet-chemistry.webodv.awi.de
- *Explore*: https://explore.webodv.awi.de
- *MOSAiC*: https://mvre.webodv.cloud.awi.de
- *HIFIS*: https://hifis.webodv.cloud.awi.de
- *EGI-ACE*: https://webodv-egi-ace.cloud.ba.infn.it

webODV facilitates the generation of maps, surface plots, section plots, scatter plots, and much more for the creation of publication ready visualizations. Rich documentation is available (https://mosaic-vre.org/docs), as well as video tutorials (https://mosaic-vre.org/videos/webodv). **webODV** online analyses can be shared via links and are fully reproducible. Users can download so-called “xview” files, small XML files, that include all instructions for **webODV** to reproduce the analyses. These files can be shared with colleagues or attached to publications, supporting fully reproducible open science.

Talks / 25

Streamlining Remote Sensing Data into Science-Ready Datacubes with rasdaman

Autor Dimitar Mishev[1]

Co-Autor: Peter Baumann [1]

[1] *Constructor University*

Corresponding author: dmisev@constructor.university

Remote sensing data is generally distributed as individual scenes, often with several variations based on the level of preprocessing applied. While there is generally good support for filtering these scenes by metadata, such as intersecting areas of interest or acquisition time, users are responsible for any additional filtering, processing, and analytics. This requires users to be experienced in provisioning hardware and deploying various tools and programming languages to work with the data. This has led to a trend towards services that consolidate remote sensing data into analysis-ready datacubes (ARD). Users can do analytics on this data through the service directly on the server and receive the precise results they are interested in.

The rasdaman Array Database is one of the most comprehensive solutions for ARD service development. It is a classical DBMS specializing in the management and analysis of multidimensional array data, such as remote sensing data. On top of its domain-agnostic array data model, rasdaman implements support for the OGC/ISO coverage model on regularly gridded data, also known as georeferenced datacubes. It offers access to the managed datacubes via standardized and open OGC APIs for coverage download (Web Coverage Service or WCS), processing (Web Coverage Processing Service or WCPS), and map portrayal (Web Map Service or WMS, and Web Map Tile Service or WMTS). Developing higher-level spatio-temporal services using these APIs is similar to developing Web applications with an SQL database.

In this talk, we discuss the current status of rasdaman and share our experiences in building ARD services across a wide range of scales, from nanosats in orbit with less computing power than modern smartphones to data centers and global federations.

Poster session / 28

From Planning to Publication: The BIS (Biosample Information System) workflow

Autor Felix Mittermayer[1]

Co-Autoren: Dirk Sarpe ; Hela Mehrtens [1]

[1] *GEOMAR*

Corresponding author: fmittermayer@geomar.de

In an ongoing effort within the Helmholtz association to make research data FAIR, i.e. findable, accessible, interoperable and reusable, we also need to make biological samples visible and searchable. To achieve this, a first crucial step is to inventory already available samples, connect them to relevant metadata and assess the requirements for various sample types (e.g. experimental, time series, cruise samples). This high diversity of sample types is challenging for creating standardized workflows and providing a uniform metadata collection with complete and meaningful metadata for each sample. As part of the Helmholtz DataHub at GEOMAR, the Biosample Information System (BIS) has been developed, turning the former decentral sample management into a fully digital and centrally managed long-term sample storage and metadata system.
The BIS is based on the open-source research data management system CaosDB in LinkAhead from IndiScale Data Services, which offers a framework for managing diverse and heterogeneous data. We have designed a flexible datamodel and multiple WebUI modules to support scientists, technicians and data managers in digitalizing and centralising sample metadata. To show the availability of samples and metadata to the scientific community, they are also visible in central research data infrastructures like OSIS (https://portal.geomar.de/osis) and in a harmonized way with other sample types and repositories in the data viewer (https://marine-data.de).We manage sampling metadata along with publication relevant data to increase the FAIRness of registered samples.
After a year of working on so called lighthouse samples, we have now entered a new chapter in the developmental phase during which we include ongoing sampling campaigns and further adapt the BIS system based on new technical and user requirements. We present a generalized work flow from the planning of a sampling campaign until data publication.

Poster session / 31

Detection and Tracking of Ocean Carbon Regimes Through Machine Learning

Autor Sweety Mohanty[1]

Co-Autoren: Daniyal Kazempour [2]; Lavinia Patara [1]; Peer Kröger [2]

[1] *GEOMAR Helmholtz Centre for Ocean Research Kiel*

[2] *CAU Kiel*

Corresponding author: smohanty@geomar.de

Our research focuses on detecting and tracking ocean carbon regimes, which are useful tools for understanding the impacts of climate change on ocean carbon uptake. Geoscientific datasets in Earth System Sciences often contain local and distinct statistical distributions at a regional scale. This poses a significant challenge in applying conventional clustering algorithms for data analysis. Based on the observed limitations of prominent methods, in our study, we propose a framework that enhances well-established unsupervised machine-learning methods tailored to applications on geoscientific datasets. We define a carbon uptake regime as a region characterized by common relationships between the carbon uptake and its drivers, as simulated by a multi-annual hydrodynamic model simulation. As a first step, we compute multivariate linear regressions capturing local spatial relations between carbon dioxide uptake and its drivers to discover such regimes. This is followed by an agglomerative hierarchical clustering constructed upon the collection of regional multivariate linear regression models. To overcome the emerging limitations of a global cut for partitioning, which is inadequate to capture the local statistical distributions, we present a novel, straightforward and adaptive approach to detect and visualize ocean carbon uptake regimes in this work. This method relies on the distance-variance selection technique and detects multiple local cuts on the dendrogram by considering both the compactness and similarity of the clusters. Detecting meaningful and well-defined carbon uptake regimes is vital for their tracking over time. The tracking is performed through a simple yet effective approach where summary structures derived from the clusters are traced over time. Applied over longer time scales, this novel method will enable marine scientists to effectively monitor the impacts of climate change on the ocean carbon cycle more.

Talks / 18

Tracing Extended Internal Stratigraphy in Ice Sheets using Computer Vision Approaches.

Autor Hameed Moqadam[1]

Co-Autoren: Adalbert Wilhelm [2]; Olaf Eisen [1]

[1] *Alfred Wegener Institute*

[2] *Constructor University*

Corresponding author: hameed.moqadam@awi.de

Polar ice sheets Greenland and Antarctica play a crucial role in the Earth's climate system. Accurately determining their past accumulation rates and understanding their dynamics is essential for predicting future sea level changes. Ice englacial stratigraphy, which assigns ages to radar reflections based on ice core samples, is one of the primary methods used to investigate these characteristics. Moreover, modern ice sheet models use paleoclimate forcings which are derived from ice cores and stratigraphy contribute to improvement and tuning of such models.

However, the conventional semi-automatic process for mapping internal reflection horizons is prone to shortcomings in terms of continuity and layer geometry, and can be highly time-consuming. Furthermore, the abundance of unmapped radar profiles, particularly from the Antarctic ice sheet, underscores the need for more efficient and comprehensive methods.

To address this, machine learning techniques have been employed to automatically detect internal layer structures in radar surveys. Such approaches are well-suited to datasets with different radar properties, making them applicable to ice, firn, and snow data. In this study, a combination of classical computer vision methods and deep learning techniques, including Convolutional Neural Networks (CNN), were used to map internal reflection horizons.

The study's methodology, including the specific pre-processing methods, labeling technique, and CNN architecture and hyperparameters, is described. Results from more promising CNN architectures, such as U-net, are presented and compared to those from image processing methods. Challenges, including incomplete training data, unknown numbers of internal reflection horizons in a profile, and a large number of features in a single radargram, are also discussed, along with potential solutions.

Poster session / 35

Automatic classification of coastline and prediction of change - an exemplary study for the North Sea and Baltic Sea

Autor David Pogorzelski[1]

Co-Autoren: Peter Arlinghaus [1]; Wenyan Zhang [1]

[1] *Helmholtz-Zentrum Hereon*

Corresponding author: david.pogorzelski@hereon.de

Automatic coastline classification based on machine learning is proven to be robust for sandy beaches on regional and global scales. However sandy beaches only make around one third of the world's ice-free shoreline. The rest consists of mudflats, cliffs, different types of vegetation and human constructions. Classification of these features is
more challenging. For instance, mild foreshore slopes resulting in large horizontal tidal excursions and high water content impede shoreline identification and classification in mudflats. Seasonal growth cycles pose difficulties in classification of vegetation.

Here we present a classification method based on an existing ground type classifier developed for Sentinel 2 data, which is able to identify the mentioned features including sandy beaches. In a later stage of the project the developed model will be used to identify the changes of the different coast types in the North and Baltic Sea within the past four decades. Generating the labelling data is a crucial task to achieve the required model performance. In order to minimize the effort to label the huge amount of available data most efficiently (Landsat and Sentinel) an active learning approach was chosen which will be presented together with some preliminary classification results.

The presented results will ultimately be used to generate a map, indicating hotspots of coastline erosion and deposition in the North and Baltic Sea. Combining these results with climate data (winds, storms, waves, surges, currents) utilizing explainable AI can provide further insights into the drivers of coastline evolution. In a final step possible coastline development until the end of the century may be predicted. The latter will not be part of the presented results.

Poster session / 23

Sample Management System: SAMS

Autor Maren Rebke[1]

Co-Autoren: Roland Koppe [1]; Tilman Dinter [1]

[1] *AWI*

Corresponding author: maren.rebke@awi.de

Sample management is an integral part of research data management. The aim of the sample management system SAMS is to provide a centralized user management solution for the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. It represents a repository for metadata accompanying samples on sample collection and sample storage with a web-based interface. SAMS includes the generation of persistent identifiers for samples.

Poster session / 8

Greenland Drone Explorer – A browser-based platform for openly accessing and exploring high resolution, 3D drone imagery using ESRI solutions

Autor Marie Ryan[1]

Co-Autoren: Daniel Frazier Carlson [1]; Rehan Chaudhary [1]; Max Boecke [1]

[1] *Hereon*

Corresponding author: marie.ryan@hereon.de

Monitoring environmental changes in terrestrial, coastal, and marine areas in remote locations, such as Greenland, can prove challenging due to harsh weather conditions and lack of infrastructure. It is therefor of great importance that any data gathered in these locations, is made freely available for viewing and reuse. To address this challenge, we developed the Greenland Drone Explorer using ESRI tools, in particular ArcGIS Experience Builder. This platform provides an interactive and user-friendly interface for exploring 3D drone imagery from various locations in Greenland, enabling scientists and researchers to gain valuable insights into the environmental changes that are happening there. To promote data sharing and collaboration, the platform also provides users with links to where the open source datasets are stored on Zenodo. The creation of this tool was a multistep process that began by importing drone data, which were in the form of orthorectified images (TIFF) and Digital elevation models (TIFF), into the GIS environment by creating web services through ArcGISPro. Web scenes were then created out of these webservices, so that the 3D data could be explored and navigated. Finally, the web scenes were used to create the final application in ArcGIS Experience Builder.

Talks / 10

Automation of Workflows for Numerical Simulation Data and Metadata

Autor Peter Schade[1]

Co-Autoren: Frank Kösters [1]; Mohammad Shafi Arif [1]

[1] *Bundesanstalt für Wasserbau*

Corresponding author: peter.schade@baw.de

Recording the workflow of numerical simulations and their postprocessing is an important component of quality management as it, e.g., makes the CFD result files reproducible. It has been implemented in an existing workflow in coastal engineering and has proven its worth in practical application.
The data history recording is achieved by saving the following information as character variables together with the actual results in CF-compliant NetCDF files:
• content of the steering files
• parameterization of simulation and postprocessing
• name of initial and boundary datasets
• name and version of the used simulation and analysis programs
Each program inherits this information from its predecessor, adds its own information, e.g., its steering file, and writes both old and new information in its result file. As a consequence, a researcher can, by downloading just the result file of the last program in the processing chain, access the recorded metadata of the whole chain.

In order to make use of these data, a second automated workflow has been established to archive the results and store the metadata in a Metadata Information System (https://datenrepository.baw.de). A Java-based postprocessor has been extended to evaluate the above mentioned NetCDF files and convert the information into ISO compliant files in XML format. The postprocessor starts copying the numerical results into a long-term storage via a REST interface and, in case of a successful copy, imports the XML metadata into the Metadata Information System (MIS) via a CSW-T interface. The MIS offers interactive search and download links to the long-term storage.

Both workflows follow the FAIR principles, e.g., by making numerical results and their metadata available to the public. The automation eliminates the error-prone manual editing and processing, relieves the user and assures quality.

Poster session / 1

Making marine image data FAIR

Autor Timm Schoening[1]

[1] *GEOMAR*

Corresponding author: tschoening@geomar.de

Underwater images are used to explore and monitor ocean habitats, generating huge datasets with unusual data characteristics that preclude traditional data management strategies. Due to the lack of universally adopted data standards, image data collected from the marine environment are increasing in heterogeneity, preventing objective comparison. the extraction of actionable information thus remains challenging, particularly for researchers not directly involved with the image data collection. Standardized formats and procedures are needed to enable sustainable image analysis and processing tools, as are solutions for image publication in long-term repositories to ascertain reuse of data. The FaIR principles (Findable, accessible, Interoperable, Reusable) provide a framework for such data management goals. We propose the use of image FaIR Digital Objects (iFDOs) and present an infrastructure environment to create and exploit such FAIR digital objects. We show how these iFDOs can be created, validated, managed and stored, and which data associated with imagery should be curated. the goal is to reduce image management overheads while simultaneously creating visibility for image acquisition and publication efforts.

Poster session / 34

Seeing the forest for the trees: mapping cover and counting trees from aerial images of a mangrove forest using artificial intelligence

Autor Daniel Schuerholz[1]

Co-Autoren: Gustavo A. Castellanos-Galindo [2]; Elisa Casella [3]; Juan C. Mejía-Rentería ; Arjun Chennu [1]

[1] *Leibniz Center for Tropical Marine Research*

[2] *Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany*

[3] *Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari, University of Venice, Italy.*

Corresponding author: daniel.schuerholz@leibniz-zmt.de

Mangrove forests are threatened by multiple anthropogenic stressors, urging researchers to create improved monitoring methods for their conservation and management. Recent remote sensing efforts have found some success using high resolution imagery of mangrove forests with sparse vegetation. In this study we focus on stands of mangrove forests with dense vegetation, consisting of the endemic Pelliciera rhizophorae and the more widespread Rhizophora mangle mangrove species, located in the remote Utria National Park in the Colombian Pacific coast. Our workflow used consumer-grade Unoccupied Aerial System (UAS) imagery of the mangrove forests, from which large orthophoto mosaics and digital elevation models are built. We apply convolutional neural networks (CNNs) for instance segmentation, to accurately delineate (33% AP) individual tree canopies for the Pelliciera rhizophorae species. We also apply neural networks to accurately identify (97% precision and 87% recall) the area coverage of the Rhizophora mangle mangrove tree species, as well as the area coverage of surrounding mud and water land-cover classes. We provide a novel algorithm for merging predicted instance segmentation tiles of trees, to recover tree shapes and sizes in overlapping and bordering regions of tiles. Using the segmented ground areas we interpolate their height from the digital elevation model to generate a digital terrain model, significantly reducing the effort for ground pixel selection. Finally, we calculate a canopy height model from the digital elevation and terrain models, and combine it with the inventory of Pelliciera rhizophorae trees to derive the height of each individual mangrove tree. The resulting inventory of mangrove trees, with individual P. rhizophorae tree height information, as well as crown shape and size descriptions, enables the use of allometric equations to calculate important monitoring metrics, such as above ground biomass and carbon stocks.

Talks / 26

The MareHub initiative: A basis for joint marine data infrastructures for Earth System Sciences

Autor Angela Schäfer[1]

Co-Autor: Team MareHub

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

Corresponding author: angela.schaefer@awi.de

The MareHub is a cooperation to bundle the data infrastructure activities of the marine Helmholtz centers AWI, GEOMAR and Hereon as a contribution to the German Marine Research Alliance and its complementary project “Underway”-Data.

The MareHub is part of the DataHub, a common initiative of all centers of the Helmholtz Association in the research field Earth and Environment. The overarching goal is to build a sustainable distributed data and information infrastructure for Earth System Sciences.

Common topics of the MareHub include the provision of standardized interfaces and SOPs, as well as solutions for the long-term storage and dissemination of marine observations, model data, and data products. Besides providing central and easy findability and accessibility of marine research data in the portal marine-data.de, the MareHub works on the aggregability and visualization of specific as well as standardized and curated data collections. Visualisation takes place in the form of thematic data viewers with content from oceanographic and seabed observation variables, video and image data, hydro acoustic data, bathymetric seabed maps, remote sensing data and sample data. These thematic viewers are developed on a sustainable generic or project basis.

In this context, standardized machine-readable data services and web map services will be established to enable the dissemination of marine data in international networks and to support advanced data science in the scientific community.

Poster session / 19

Boosting the virtual research environment V-FOR-WaTer to facilitate data science across scales

Autor Marcus Strobl[1]

Co-Autoren: Achim Streit ; Alexander Dolich ; Ashish Manoj J ; Balazs Bischof ; Elnaz Azmi ; Erwin Zehe ; Jörg Meyer [1]; Mirko Mälicke ; Sibylle K. Hassler

[1] *KIT*

Corresponding author: marcus.strobl@kit.edu

Complex interdisciplinary research challenges in environmental sciences require integration of data from different sources, different domains, and different scales of measurement. Despite the growing amount and diversity of available data, they are often stored in varying formats with inadequate metadata, making them challenging to access and integrate. The virtual research environment V-FOR-WaTer aims to facilitate access and integration of these diverse data originating from university projects and state offices. The metadata scheme complies with international standards like INSPIRE and ISO19115, enabling easy publication of data through established repositories such as the GFZ Data Services.

The V-FOR-WaTer web portal features a workspace area that already offers tools for data pre-processing, scaling, and common hydrological applications. Additionally, the toolbox contains more specific tools, e.g. for geostatistics and evapotranspiration, and as the tools are added through the OGC Processes API, the toolbox can be easily extended to include user-developed tools. Users can access tools through a graphical user interface and combine, save, and share them as workflows, facilitating complex analyses and reproducibility of results.

The portal is designed to streamline typical workflows in environmental sciences, offering map operations and filter options to help with data selection. Ultimately, the portal aims to provide a virtual research environment to enable data science across scales and disciplines in environmental science.

Talks / 22

A progress report on OPUS - The Open Portal to Underwater Soundscapes to explore global ocean soundscapes

Autoren Karolin Thomisch[1]; Robin Hess[1]; Lewin Probst[None]; Olaf Boebel[1]

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

Corresponding author: karolin.thomisch@awi.de

In an era of rapid anthropogenically induced changes in the world's oceans, ocean sound is considered an essential ocean variable (EOV) for understanding and monitoring long-term trends in anthropogenic sound and its effects on marine life and ecosystem health.
The International Quiet Ocean Experiment (IQOE) has identified the need to monitor the distribution of ocean sound in space and time, e.g. current levels of anthropogenic sound in the ocean.
The OPUS (Open Portal to Underwater Soundscapes) data portal, is currently being developed at the Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) in Bremerhaven, Germany, in the framework of the German Marine Research Alliance (DAM) DataHUB/MareHUB initiative.
OPUS is designed as an expeditious discovery tool for archived passive acoustic data, envisioned to promote the use of archived acoustic data collected worldwide, thereby contributing to an improved understanding of the world's oceans soundscapes and anthropogenic impacts thereon over various temporal and spatial scales. 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 (visualizations of and lossy compressed audio data). OPUS provides direct access to underwater recordings collected world-wide, allowing a broad range of stakeholders (i.e., the public, artists, journalists, fellow scientists, regulatory agencies, consulting companies and the marine industry) to learn about and access this data for their respective needs. 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.
Here, an update on achieved and upcoming milestones and developments of OPUS will be presented, together with ongoing and future case studies and use cases of the portal for different user groups.

Poster session / 11

Verification of ESM-ML hybrids

Autor Tobias Weigel[1]

[1] *DKRZ / Hereon*

Corresponding author: weigel@dkrz.de

The transformative power of ML unleashed in Earth System modelling is taking shape. Recent advances in building hybrid models combining mechanistic Earth system models grounded in physical understanding and machine learning models trained from huge amounts of data show promising results and are in the focus of international research initatives. However, the ongoing implementation of such models poses not only technical challenges, but also raises questions of trustworthiness and verifiability. The acceptance of hybrid approaches and full replacement models critically depends on the means scientists have for validating them. For further adoption of ML methods, it is crucial to establish the credibility of ML-enhanced Earth system models through statistical reproducibility. This contribution will structure the challenges and requirements, highlight emerging community approaches and provide a germination point for discussions between ESM and ML developers and downstream users.

Talks / 12

The project “underway”-data and the MareHub initiative: A basis for joint marine data management and infrastructure

Autoren MareHub & DAM-Underway-Data-Team

Speaker Gauvain Wiemer[1]

[1] *Deutsche Allianz Meeresforschung*

Corresponding author: wiemer@allianz-meeresforschung.de

Coordinated by the German Marine Research Alliance (Deutsche Allianz Meeresforschung (DAM)) the project “Underway”–Data is supported by the marine science centers AWI, GEOMAR and Hereon of the Helmholtz Association research field “Earth and Environment”. AWI, GEOMAR and Hereon develop the marine data hub (Marehub). This MareHub initiative is a contribution to the DAM. It builds a decentralized data infrastructure for processing, long-term archiving and dissemination of marine observation and model data and data products. The “Underway”-Data project aims to improve and standardize the systematic data collection and data evaluation for expeditions with German research vessels. It supports scientists with their data management duties and fosters (data)science through FAIR and open access to marine research data. Together, the project “Underway”-Data and the MareHub initiative form the base of a coordinated data management infrastructure to support German marine sciences and international cooperation. The data and infrastructure of these joint efforts contribute to the NFDI and EOSC.
This talk will present the above mentioned connections between the different initiatives and give further detail about the “underway”-data project.

Talks / 21

LightEQ: On-Device Earthquake Detection with Embedded Machine Learning

Autor Tayyaba Zainab[1]

[1] *GEOMAR*

Corresponding author: tzainab@geomar.de

The detection of earthquakes in seismological time series is central to observational seismology. Generally, seismic sensors passively record data and transmit it to the cloud or edge for integration, storage, and processing. However, transmitting raw data through the network is not an option for sensors deployed in harsh environments like underwater, underground, or in rural areas with limited connectivity. This paper introduces an efficient data processing pipeline and a set of lightweight deep-learning models for seismic event detection deployable on tiny devices such as microcontrollers. We conduct an extensive hyperparameter search and devise three lightweight models. We evaluate our models using the Stanford Earthquake Dataset and compare them with a basic STA/LTA detection algorithm and the state-of-the-art machine learning models, i.e., CRED, EQtransformer, and LCANet. For example, our smallest model consumes 193 kB of RAM and has an F1 score of 0.99 with just 29k parameters. Compared to CRED, which has an F1 score of 0.98 and 293k parameters, we reduce the number of parameters by a factor of 10. Deployed on Cortex M4 microcontrollers, the smallest version of \project-NN has an inference time of 932 ms for 1 minute of raw data, an energy consumption of 5.86 mJ, and a flash requirement of 593 kB. Our results show that resource-efficient, on-device machine learning for seismological time series data is feasible and enables new approaches to seismic monitoring and early warning applications.
Next step:
We can have a more in-depth analysis of seismic data such as detection of P and S waves on the sensors.

*This paper is already accepted and peer-reviewed *