Helmholtz-Zentrum Hereon

Data Science Symposium No. 7 Program & Book of Abstracts

June 27th & 28th, 2022

Helmholtz-Zentrum
hereon

June 27th, 2022

Online participation link:
https://hereon-de.zoom.us/j/97671200935?pwd=MUxtWTNCT2RnRFhVUS9zbGx5bTBPUT09

13:00 – 13:15 Welcome

13:15 – 15:30 Collaborations, initiatives and data strategies (POF IV, DataHub, etc.)

Convener: Viktoria Wichert

13:15 – 13:45	Data science require appropriate data strategies	Bumberger, Jan (UFZ)
13:45 – 14:00	BELUGA – From Situational Awareness to a coordinated data workflow	Rothenbeck, Marcel (GEOMAR)
14:00 – 14:15	FAIR bathymetry data archiving in PANGAEA - Data Publisher for Earth & Environmental Science	Damaske, Daniel (MARUM)
14:15 – 14:30	Short break	
14:30 – 14:45	First lessons from the implementation of a biosample management system	Eckstein, Jakob, Mittermayer, F. (GEOMAR)
14:45 – 15:00	FAIR WISH - FAIR Workflows to establish IGSN for Samples in the Helmholtz Association	Baldewein, Linda (Hereon)
15:00 – 15:15	Development of a flexible, multi-stage data pipeline for enhanced automation, quality control and observability	Werner, Christian (KIT)
15:15 – 15:30	NFDI4Earth Academy	Wiemer, Gauvain (DAM)

15:30 – 16:00 Coffee and cake break

16:00 – 18:00 Artificial Intelligence / Machine Learning in Earth System Sciences Part 1

Convener: David Greenberg

16:00 – 16:30	Combining machine learning and data assimilation to learn dynamics from sparse and noisy observations	Bocquet, Marc (CEREA)
16:30 – 16:45	Towards physics-based surrogate modeling of two-dimensional mantle convection	Agarwal, Siddhant (DLR)
16:45 – 17:00	Data-driven modeling of coastal groundwater dynamics: Bridging scales and obstacles using the concept of hydrogeological similarity	Nolte, Annika (Hereon)
17:00 – 17:15	A machine learning approach to predict sediment accumulation at the sea floor.	Parameswaran, Naveen Kumar (GEOMAR, Uni Kiel)
17:15 – 17:30	Unsupervised learning on large collections of high-resolution trajectories	Rath, Willi (GEOMAR)
17:30 – 17:45	Improved machine-learning-based open-water–sea-ice–cloud discrimination over wintertime Antarctic sea ice using MODIS thermal-infrared imagery	Paul, Stephan (AWI)
17:45 – 18:00	The AI-CORE Project - Artificial Intelligence for Cold Regions	Baumhoer, Celia (DLR)

18:00 – 20:00 Dinner – BBQ

June 28th, 2022

Online participation link:
https://hereon-de.zoom.us/j/97671200935?pwd=MUxtWTNCT2RnRFhVUS9zbGx5bTBPUT09

09:00 – 09:45 Posters and Live Demos

Convener: Linda Baldewein

Poster	Towards the detection of ocean carbon regimes	Mohanty, Sweety (GEOMAR, Uni Kiel)
Poster	Layerwise Relevance Propagation for Echo State Networks applied to Earth System Variability.	Landt-Hayen, Marco (GEOMAR)
Poster	HARMONise – Enhancing the interoperability of marine biomolecular (meta)data across Helmholtz Centres	Bienhold, Christina, Harms, L., Neuhaus, S. (AWI)
Poster	Fast and Accurate Physics-constrained Learning with Symmetry Constraints for the Shallow Water Equations	Huang, Yunfei (Hereon)
Poster	Data conversion for the MOSAiC webODV	Freier, Julia (AWI)
Poster	Robust Detection of Marine Life with Label-free Image Feature Learning and Probability Calibration	Schanz, Tobias (Hereon)
Poster	Assessing the Feasibility of Self-Supervised Video Frame Interpolation and Forecasting on the Cloudcast Dataset	Lin, Michelle (Mila - Quebec AI Institute)
Poster	Approximation and Optimization of Environmental Simulations in High Spatio-Temporal Resolution through Machine Learning Methods	Azmi, Elnaz (KIT)
Poster	Machine-Learning-Based Comparative Study to Detect Suspect Temperature Gradient Error in Ocean Data.	Chouai, Mohamed (AWI)

Poster	Machine Learning Parameterization for Cloud Microphysics	Sharma, Shivani (Hereon)
Poster	Low-Carbon Routing using Genetic Stochastic Optimization and Global Ocean Weather	Shchekinova, Elena (GEOMAR)
Poster	Learning deep emulators for the interpolation of satellite altimetry data	Georgenthum, Hugo (IMT Atlantique)
Poster	AI4FoodSecurity: Identifying crops from space	Albrecht, Frauke (DKRZ)
Poster	Automatic low-dimension explainable feature extraction of climatic drivers leading to forest mortality	Anand, Mohit (UFZ)
Poster	Model evaluation method affects the interpretation of machine learning models for identifying compound drivers of maize variability	Sweet, Lily-Belle (UFZ)
Live demo	HELMI – The Hereon Layer For Managing Incoming Data	Böcke, Max, Hemmen, J., Leefmann, T. (Hereon)
Live demo	The Coastal Pollution Toolbox – data services and products in support of knowledge for action	Lange, Marcus (Hereon)
Live demo	New approaches for distributed data analysis with the DASF Messaging Framework	Sommer, Philipp Sebastian (Hereon)
Live demo	MOSAiC webODV – An online service for the exploration, analysis and visualization of MOSAiC data	Mieruch-Schnülle, Sebastian (AWI)
Live demo	MuSSeL project data management and outreach	Ryan, Marie (Hereon)
Live demo	Investigating the coastal impacts of riverine flood events with the River Plume Workflow	Wichert, Viktoria (Hereon)

09:45 – 10:45 Artificial Intelligence / Machine Learning in Earth system Sciences Part 2

Convener: David Greenberg

9:45 – 10:00	Unlocking the potential of ML for Earth and Environment researchers	Albrecht, Frauke (DKRZ)
10:00 – 10:15	Making marine image data FAIR with iFDOs	Schoening, Timm (GEOMAR)
10:15 – 10:30	ClimART: A Benchmark Dataset for Emulating Atmospheric Radiative Transfer in Weather and Climate Models	Rühling Cachay, Salva (Mila - Quebec AI Institute)
10:30 – 10:45	Machine Learning applications for impact modelling of climate extremes	Bouwer, Laurens Menno (Hereon)

10:45 – 11: 15 Coffee break

11:15 – 12:45 Towards Digital Twins and other Lighthouse Projects

Convener: Linda Baldewein

11:15 – 11:45	Digital Twins of the Ocean (DITTO) - Opportunities to future-proof Ocean Sustainable Development	Visbeck, Martin (GEOMAR, Uni Kiel)
11:45 – 12:00	Digital hydromorphological twin of the Trilateral Wadden Sea	Plüß, Andreas (BAW)
12:00 – 12:15	Interactive Earth-System Models for Digital-Twins	Claus, Martin (GEOMAR)
12:15 – 12:30	Collaborative Exploration and Annotation of 4D Data with the Digital Earth Viewer	Stäbler, Flemming (GEOMAR)
12:30 – 12:45	A Web Framework for Dynamic Data Presentations in Earth Sciences	Gonzalez, Everardo (GEOMAR)

12:45 – 13:00 Farewell

Collaborations, initiatives and data strategies (POF IV, DataHub, etc.)

1. Keynote: Data science require appropriate data strategies

Bumberger, Jan (Helmholtz Centre for Environmental Research)

Current and future data-driven scientific applications from explorative methods to operational systems require suitable data strategies. The necessary requirements for this are presented in the context of earth system science and their embedding in POF 4 and NFDI. In particular, exemplary examples are presented which stand for consistent data management as a data strategy from data collection to the data product including the data processing steps.

2. BELUGA – From Situational Awareness to a coordinated data workflow

Rothenbeck, Marcel, Leibold, P., Diller, N., Reißmann, S. (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

Those who operate autonomous vehicles under or above water want to know where they are, if they are doing well and if they are doing what they supposed to do. Sometimes it makes sense to interfere in the mission or to request a status of a vehicle. When the GEOMAR AUV team were asked to bring two Hover AUVs (part of the MOSES Helmholtz Research Infrastructure Program; "Modular Observation Solutions for Earth Systems") into operational service, that was exactly what we wanted. We decided to create our own tool: BELUGA.
After almost 3 years BELUGA is a core tool for our operational work with our GIRONA 500 AUVs and their acoustic seafloor beacons. BELUGA allows communication and data exchange between those devices mentioned above and the ship.
Every component inside this kind of ad hoc network has an extended driver installed. This driver handles the messages and their content and decides which communication channels to use: Wi-Fi, satellite or acoustical communication. The driver of the shipborne component has additionally a module for a data model, which allows to add more network devices, sensors and messages in future. The user works on a web based graphical interface, which visualizes all the information coming from the devices. The operator also can interact with the device, start and stop processes or mission, if the integrated BELUGA driver has access to the system of the device or vehicle.

But why this is a topic in a Data Science Symposium? Autonomous devices, either mobile or immobile, carry sensors which gather scientific data. A communication platform like BELUGA could support quality control by receiving and verifying sections or fragments of the data sets from the device according the bandwidth. Vehicles could look for anomalies and allows selective sampling, making data gathering more effective. The user interface could lead the user/operator through the entire, automated and clear coordinated data process until it is handed over to the responsible data management. The above-mentioned check of sent fragments during the mission would be part of this workflow but also an automated data export.

Last year we started a project with our Data Management Department to work on a software tool which download and brings the data sets of each dive of our AUVs into an agreed format. In this way the scientist on board of a research cruise gets the data in a form which is already prepared for a final hand over. It is clearly described where to find raw or already processed data, which metadata are inside and a description of the vehicle mission, the vehicle used and the sensors on it. The export tool is not a part of BELUGA but the user interface allows the triggering and monitoring of the so automated data workflow. We hope this will improve the data handling and ease the work of the team on board. It is still an ongoing project but the first version is going to be used on a research cruise next June.

And there are more aspects in term of data visualization on ongoing campaigns. During former projects an Online-Portal for Monitoring Surface vehicles was created. Those experiences were incorporated into BELUGA. One possible application here is public outreach. The BELUGA user interface is in this case a compact source of information for the interested public. The adventurer Arved Fuchs, for example, uses BELUGA in this way for his OCEAN CHANGE campaign. It shows the track of his ship Dagmar Aaen, visualizes the data of the environmental sensors and it is also used to narrate his travel by linked blocks and social media posts.

BELUGA is our tool to communicate and exchange data and we are looking for more partners to figure out what more is possible.

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

Damaske, Daniel (MARUM - Center for Marine Environmental Sciences, University Bremen)

Compliant with the FAIR data principles, 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 semi-automated processes and standard operating procedures (SOPs) for bathymetry data publishing and simultaneous visualization are currently developed.

This research is part of the "Underway Research Data" project, an initiative of the German Marine Research Alliance (Deutsche Allianz Meeresforschung e.V., DAM). The DAM "Underway Research Data" project, spanning across different institutions, started in mid-2019. The aim of the project is to improve and unify the constant data flow from German research vessels to data repositories like PANGAEA. This comprises multibeam-echosounder and other permanently installed scientific devices and sensors following the FAIR data management aspects. Thus, exploiting the full potential of German research vessels as instant "underway" mobile scientific measuring platforms.

4. First lessons from the implementation of a biosample management system

Eckstein, Jakob, Mittermayer, F. (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

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 is challenging for when creating standardized workflows to provide a uniform metadata collection with complete and meaningful metadata for each sample. As part of the Helmholtz DataHub at GEOMAR the **B** iosample **I** nformation **S** ystem (BIS) has been set up, turning the former decentral sample management into a fully digital and centrally managed long-term sample storage.

The BIS is based on the open-source research data management system CaosDB, which offers a framework for managing diverse and heterogeneous data. It supports fine-grained access permissions and regular backups, has a powerfull search engine, different APIs and an extendable WebUI. We have designed a flexible datamodel and multiple WebUI modules to support scientists, technicans and datamanagers in digitalizing and centralising sample metadata and making the metadata visible in data portals (e.g. https://marine-data.de).

The system allows us to manage a broad variety of sample types ranging from DNA extracts, formaldehyde fixed plankton samples to dried or frozen tissues, and originating from an even larger variety of projects such as cruises, field samplings, experiments and even combinations of the above. We have tested the system and successfully incorporated several "lighthouse" projects representing the diversity of sample types and projects. We will present the current state of the BIS with special emphasis on biosample specific issues addressed during the development, first curated projects and an outlook on upcoming developments.

5. FAIR WISH - FAIR Workflows to establish IGSN for Samples in the Helmholtz Association

Baldewein, Linda[1], Elger, K.[2], Heim, B.[3], Brauser, A.[2], Frenzel, S.[2], Kleeberg, U.[1], Norden, B.[2] ([1] Helmholtz-Zentrum Hereon, [2] GFZ German Research Centre for Geosciences, [3] Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

The International Generic Sample Number (IGSN) is a globally unique and persistent identifier for physical objects, such as samples. IGSNs allow to cite, track and locate physical samples and to link samples to corresponding data and publications. The metadata schemata is modular and can be extended by domain-specific metadata.

Within the FAIR WISH projected funded by the Helmholtz Metadata Collaboration, domain-specific metadata schemes for different sample types within Earth and Environment are developed based on three different use cases. These use cases represent all states of digitization, from hand-written notes only to information stored in relational databases. For all stages of digitization, workflows and templates to generate machine-readable IGSN metadata are developed, which allow automatic IGSN registration. These workflows and templates will be published and will contribute to the standardization of IGSN metadata.

6. Development of a flexible, multi-stage data pipeline for enhanced automation, quality control and observability

Werner, Christian, Lorenz, C. (Karlsruhe Institute of Technology)

Facilitating and monitoring the ingestion and processing of continuous data streams is a challenging exercise that is often only addressed for individual scientific projects and/ or stations and thus results in a heterogeneous data environment.

In order to reduce duplication and to enhance data quality we built a prototypical data ingestion pipeline using open-source frameworks with the goal to a) unify the data flow for various data sources, b) enhance observability at all stages of the pipeline, c) introduce a multi-stage QA/ QC procedure to increase data quality and reduce the lag of data degradation or data failure detection. The system is orchestrated using Prefect , QA/ QC is handled by Great Expectations and SaQC , and the SensorThings API and THREDDS Data Server are used to facilitate data access and integration with other services.

The prototype workflow also features a human-in-the-loop aspect so scientific PIs can act on incoming data problems early and with little effort. The framework is flexible enough so specific needs of individual projects can be addressed while still using a common platform. The final outcome of the pipeline are aggregated data products that are served to the scientists and/ or the public via data catalogues. In the future, we plan to add more data flows at our institute. This will help us to further standardize the processing and QA/ QC - and thus increase data quality and availability - and hopefully also reduce the overall maintenance burden.

7. NFDI4Earth Academy

Gödde, Hildegard[1], Kuppler, J.[1], Ntageretzis, K.[2], Drews, E.-L.[2], Wiemer, G.[3] ([1] GFZ German Research Centre for Geosciences, [2] Forschungszentrum Jülich, [3] Deutsche Allianz Meeresforschung e.V. (DAM))

The NFDI 4 Earth Academy is a network of early career scientists interested in linking 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 that covers specialized training courses, collaborations within the NFDI4Earth consortium and access to all NFDI 4 Earth innovations and services. Fellows of the Academy advance their research projects by exploring and integrating new methods and connect 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.

8. HARMONise – Enhancing the interoperability of marine biomolecular (meta)data across Helmholtz Centres (POSTER)

Bienhold, Christina[1], Harms, L.[1], Neuhaus, S.[1], Bayer, T.[2], Koppe, R.[1] ([1] Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung, [2] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

Biomolecules, such as DNA and RNA, make up all ocean life, and biomolecular research in the marine realm is pursued across several Helmholtz Centres. Biomolecular (meta)data (i.e. DNA and RNA sequences and all steps involved in their creation) provide a wealth of information about the distribution and function of marine organisms. However, high-quality (meta)data management of biomolecular data is not yet well developed in environmentally focused Helmholtz Centres. This impedes every aspect of FAIR data exchange internally and externally, and the pursuit of scientific objectives that depend on this data. In this Helmholtz Metadata Collaboration project between the Alfred-Wegener-Institut Helmholtz Zentrum für Polar- und Meeresforschung and the GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel (with scientific PIs rooted in POFIV Topic 6), we will develop sustainable solutions and digital cultures to enable high-quality, standards-compliant curation and management of marine biomolecular metadata, to better embed biomolecular science in broader digital ecosystems and research domains. The approach will build on locally administered relational databases and establish a web-based hub to exchange metadata compliant with domain-specific standards, such as the MIxS (Minimum Information about any (x) Sequence). To interface with and enhance the Helmholtz digital ecosystem, we aim to link the operations and archiving workflows of our local databases with existing Helmholtz repositories (e.g. the PANGAEA World Data Center) and with systems currently under development (e.g. Sample Management, Marine Data Portal). This will be done through stable, synchronised, and persistent solutions for the export and exchange of (meta)data. By enabling sustainable data stewardship, as well as export and publishing routines, this will support biomolecular researchers in delivering Helmholtz biomolecular data to national European and global repositories in alignment with community standards. Throughout the project, we will establish and cultivate human communication channels, to ensure implementations do not drift apart. Furthermore, we will provide use cases to connect our data holdings with other global interoperability frameworks, such as UNESCO's Ocean Data and Information System. Here we will present a conceptual outline and first steps taken on our road to practically enabling FAIR management of biomolecular (meta)data. 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.

9. The Coastal Pollution Toolbox – data services and products in support of knowledge for action (LIVE DEMO)

Lange, Marcus, Ebinghaus, R., Kleeberg, U., Baldewein, L. (Helmholtz-Zentrum Hereon)

Knowledge transfer requires, first, meaningful approaches and products to transfer knowledge amongst different users and, second, appropriate measures for the creation of knowledge across scientific disciplines. The Coastal Pollution Toolbox (https://www.coastalpollutiontoolbox.org/index.php.en), a central product of the program-oriented funding topic on "Coastal Transition Zones under Natural and Human Pressures", serves as a digital working environment for scientists and knowledge hub and information platform for decision-makers. It supports action and optimisation of scientific concepts to investigate pollution in the land-to-sea continuum.

In order to address demands of various users the toolbox comprises of three compartments: Science Tools provide expert users with information on new methods, approaches or indicators for baseline assessments or for the re-evaluation of complex problems. Synthesis Tools address challenges of global environmental change. They are information-rich products based on consolidated data of different types and origin and provide expert users with knowledge. Management Tools provide usable information and options for action. Ready-to-use tools grounded on evidence-based science are available to those involved in planning and management of coastal and marine challenges.

As part of the development process coastal pollution information services will be created and co-developed with stakeholders and end-users. This will ensure optimal interest and use by a range of actors involved in the direct and indirect impact of coastal and marine pollution. The contribution will highlight the basic approach of the toolbox and some of the products already and planned to be developed.

10. New approaches for distributed data analysis with the DASF Messaging Framework (LIVE DEMO)

Sommer, Philipp Sebastian[1], Eggert, D.[2], Wichert, V.[1], Baldewein, L.[1], Dinter, T.[3], Werner, C.[4] ([1] Helmholtz-Zentrum Hereon, [2] GFZ German Research Centre for Geosciences, [3] Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung, [4] Karlsruhe Institute of Technology)

The Data Analytics Software Framework (DASF, https://doi.org/10.5880/GFZ.1.4.2021.004) supports scientists to conduct data analysis in distributed IT infrastructures by sharing data analysis tools and data. For this purpose, DASF defines a remote procedure call (RPC) messaging protocol that uses a central message broker instance. Scientists can augment their tools and data with this protocol to share them with others or re-use them in different contexts.

Our framework takes standard python code developed by a scientist, and automatically transforms the functions and classes of the scientists' code into an abstract layer. This abstraction, the server stub as it is called in RPC, is connected to the message broker and can be accessed by submitting JSON-formatted data through a websocket in the so-called client stub. Therefore the DASF RPC messaging protocol in general is language independent, so all languages with Websocket support can be utilized. As a start DASF provides two ready-to-use language bindings for the messaging protocol, one for Python and one for the Typescript programming language.

DASF is developed at the GFZ German Research Centre for Geosciences and was funded by the Initiative and Networking Fund of the Helmholtz Association through the Digital Earth project (https://www.digitalearth-hgf.de/). In this talk, we want to present the framework with some simple examples, and present two new approaches for the framework. One is an alternative light-weight message broker based on the python web-framework Django, that supports containerization, user-management and token authentication. The other one is an approach for easily applicable end-to-end-encryption in the messaging framework and user-authentication in the backend module for secure federation of data analysis between research centers.

11. MuSSeL project data management and outreach (LIVE DEMO)

Ryan, Marie (Helmholtz-Zentrum Hereon)

The collaborative project "MuSSeL" investigates various natural and anthropogenic changes, such as climate change, increase of fishing and the development of offshore wind farms, and the effect these changes have on the biodiversity and well-being of benthic communities in the North Sea. A central part of this project is to make any data gathered, easily accessible to stakeholders and the general public alike. To streamline this process, it was decided to use ESRI software solutions, for data management and public outreach. The live demo will demonstrate how data can be visualized, analyzed, and made available on the project website, all using ESRI solutions.

12. Investigating the coastal impacts of riverine flood events with the River Plume Workflow (LIVE DEMO)

Wichert, Viktoria[1], Brix, H.[1], Abraham, N.[1], Rabe, D.[2] ([1] Helmholtz-Zentrum Hereon, [2] GFZ German Research Centre for Geosciences)

The River Plume Workflow is a part of the Digital Earth Flood event explorer (FEE), which was designed to compile different aspects of riverine flood events.

The focus of the River Plume Workflow is the impact of riverine flood events on the marine environment, when, at the end of a flood event chain, an unusual amount of nutrients and pollutants is washed into the coastal waters. The River Plume Workflow provides scientists with tools to detect river plumes in marine data during or after an extreme event and to investigate their spatio-temporal extent, their propagation and impact. This is achieved through the combination of in-situ data from autonomous measuring devices, drift model data produced especially for the observational data and satellite data of the observed area. In the North Sea, we use measurements from the FerryBox mounted on the Büsum-Helgoland ferry to obtain regular in-situ data and offer model trajectories from drift simulations around the time of extreme events in the Elbe River.

The River Plume Workflow helps scientists identify river plume candidates either manually within a visual interface or through an automatic anomaly detection algorithm, using Gaussian regression. Combining the observational data with model trajectories that show the position of a measured water body up to 10 days before and after the measurement allows to investigate the propagation of an anomaly, as well as to check its origin, e.g. the Elbe estuary. This way, scientists can identify regions of interest presumably impacted by riverine flood events. Combining model trajectories with satellite data also provides scientists with time series of parameters, e.g. Chlorophyll-A, along a model trajectory, allowing research on degradation rates and unusual behavior during or after an extreme event.

With the deployment of the River Plume Workflow coming up, I would like to demonstrate the functionalities of the tool and discuss its applications.

Artificial Intelligence / Machine Learning in Earth System Sciences

1. Keynote: Combining machine learning and data assimilation to learn dynamics from sparse and noisy observations

Bocquet, Marc (CEREA, École des Ponts and EdF R&D, Île-De-France, France)

The recent introduction of machine learning techniques in the field of numerical geophysical prediction has expanded the scope so far assigned to data assimilation, in particular through efficient automatic differentiation, optimisation and nonlinear functional representations. Data assimilation together with machine learning techniques, can not only help estimate the state vector but also the physical system dynamics or some of the model parametrisations. This addresses a major issue of numerical weather prediction: model error.
I will discuss from a theoretical perspective how to combine data assimilation and deep learning techniques to assimilate noisy and sparse observations with the goal to estimate both the state and dynamics, with, when possible, a proper estimation of residual model error. I will review several ways to accomplish this using for instance offline, variational algorithms and online, sequential filters. The skills of these solutions with be illustrated on low-order and intermediate chaotic dynamical systems, as well as data from meteorological models and real observations.

Examples will be taken from collaborations with J. Brajard, A. Carrassi, L. Bertino, A. Farchi, Q. Malartic, M. Bonavita, P. Laloyaux, and M. Chrust.

2. Towards physics-based surrogate modeling of two-dimensional mantle convection

Agarwal, Siddhant, Tosi, N. (Deutsches Zentrum für Luft- und Raumfahrt (DLR))

Mantle convection plays a fundamental role in the long-term thermal evolution of terrestrial planets like Earth, Mars, Mercury and Venus. The buoyancy-driven creeping flow of silicate rocks in the mantle is modeled as a highly viscous fluid over geological time scales and quantified using partial differential equations (PDEs) for conservation of mass, momentum and energy. Yet, key parameters and initial conditions to these PDEs are poorly constrained and often require a large sampling of the parameter space to find constraints from observational data.

Since it is not computationally feasible to solve hundreds of thousands of forward models in 2D or 3D, scaling laws have been the go-to alternative. These are computationally efficient, but ultimately limited in the amount of physics they can model (e.g., depth-dependent material properties). More recently, machine learning techniques have been used for advanced surrogate modeling. For example, Agarwal et al. (2020) used feedforward neural networks to predict the evolution of entire 1D laterally averaged temperature profile in time from five parameters: reference viscosity, enrichment factor for the crust in heat producing elements, initial mantle temperature, activation energy and activation volume of the diffusion creep. In Agarwal et al. (2021), we extended that study to predict the full 2D temperature field of a Mars-like planet using convolutional autoencoders and long-short-term memory networks.

Despite producing reasonably accurate and realistic looking temperature fields, these data-driven surrogates require a lot of simulations, are not as accurate as traditional numerical solvers and do not predict the remaining state variables (velocity and pressure). Thus, we are keen on exploring physics-based machine learning algorithms that embed a few or all of the underlying PDEs into the loss function of a learning algorithm and thus, might require none (e.g. as in Wandel et. al 2021) to little data. We discretize the PDEs in space using finite differences, which are implemented as convolutional operators. Preliminary results from an analytical benchmark (Trubitsyn, 2006) show that for a given temperature field, the corresponding velocities can be calculated with a mean absolute error of approximately 5x10^-4. The next step is to see if these velocities can be used to advect the temperature field, solve for the new velocities and so on and so forth to advance the flow in time. This will lead to the next benchmark (Blankenbach et al., 1989), where the flow should converge to a steady-state.

3. Data-driven modeling of coastal groundwater dynamics: Bridging scales and obstacles using the concept of hydrogeological similarity

Nolte, Annika (Helmholtz-Zentrum Hereon)

For some time, large scale analyses and data-driven approaches have become increasingly popular in all research fields of hydrology. Many advantages are seen in the ability to achieve good predictive accuracy with comparatively little time and financial investment. It has been shown by previous studies that complex hydrogeological processes can be learned from artificial neural networks, whereby Deep Learning demonstrates its strengths particularly in combination with large data sets. However, there are limitations in the interpretability of the predictions and the transferability with such methods. Furthermore, most groundwater data are not yet ready for data-driven applications and the data availability often remains insufficient for training neural networks. The larger the scale, the more difficult it becomes to obtain sufficient information and data on local processes and environmental drivers in addition to groundwater data. For example, groundwater dynamics are very sensitive to pumping activities, but information on their local effects and magnitude – especially in combination with natural fluctuations – is often missing or inaccurate. Coastal regions are often particularly water-stressed. Exemplified by the important coastal aquifers, novel data-driven approaches are presented that have the potential to both contribute to process understanding of groundwater dynamics and groundwater level prediction on large scales considering local processes.

4. A machine learning approach to predict sediment accumulation at the sea floor.

Parameswaran, Naveen Kumar[1, 2] , Wallmann, K.[1] , Braack, M.[2] , Gonzalez, E.[1] , Burwicz-Galerne, E.[3] ([1] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, [2] Christian Albrecht University of Kiel, [3] MARUM - Center for Marine Environmental Sciences, University Bremen)

Together, the creatures of the oceans and the physical features of their habitat play a significant role in sequestering carbon and taking it out of the atmosphere. Through the biological processes of photosynthesis, predation, decomposition, and the physical movements of the currents, the oceans take in more carbon than they release. With sediment accumulation in the deep seafloor, carbon gets stored for a long time, making oceans big carbon sinks, and protecting our planet from the devastating effects of climate change.

Despite the significance of seafloor sediments as a major global carbon sink, direct observations on the mass accumulation rates(MAR) of sediments are sparse. The existing sparse data set is inadequate to quantify the change in the composition of carbon and other constituents at the seabed on a global scale. Machine learning techniques such as the k-nearest neighbour's algorithm have provided predictions of sparse sediment accumulation rates, by correlating known features(predictors) such as bathymetry, bottom currents, distance to coasts and river mouths, etc.

In my current work, global maps of the sediment accumulation rates at the seafloor are predicted using the known fea ture maps and the sparse dataset of sediment accumulation rates using multi-layer perceptrons(supervised models). Despite a good model accuracy, the predictions are not reliable, according to expert knowledge. Some of the main problems are the low availability of labelled data, uneven distribution(both spatially and mathematically) of sediment accumulation rates, and low knowledge about feature relevance. To understand the unreliability of predictions and the impact of the problems, model uncertainty is being studied using Bayesian neural networks.

In the presentation, the predictions using the multi perceptron model(in comparison to the previously published results using k-nearest neighbour algorithm) and the model uncertainty using Bayesian neural networks would be shown.

5. Unsupervised learning on large collections of high-resolution trajectories

Rath, Willi[1], Trahms, C.[1, 2], Handmann, P.[1], Wölker, Y.[1, 2] ([1] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, [2] Christian Albrecht University of Kiel)

The Lagrangian perspective on Ocean currents describes trajectories of individual virtual or physical particles which move passively or semi-actively with the Ocean currents. The analysis of such trajectory data offers insights about pathways and connectivity within the Ocean. To date, studies using trajectory data typically identify pathways and connections between regions of interest in a manual way. Hence, they are limited in their capability in finding previously unknown structures, since the person analyzing the data set can not foresee them. An unsupervised approach to trajectories could allow for using the potential of such collections to a fuller extent.

This study aims at identifying and subsequently quantifying pathways based on collections of millions of simulated Lagrangian trajectories. It develops a stepwise multi-resolution clustering approach, which substantially reduces the computational complexity of quantifying similarity between pairs of trajectories and it allows for parallelized cluster construction.

It is found that the multi-resolution clustering approach makes unsupervised analysis of large collections of trajectories feasible. Moreover, it is demonstrated that for selected example research questions the unsupervised results can be applied.

6. Improved machine-learning-based open-water–sea-ice–cloud discrimination over wintertime Antarctic sea ice using MODIS thermal-infrared imagery

Paul, Stephan (Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

The frequent presence of cloud cover in polar regions limits the use of the Moderate Resolution Imaging Spectroradiometer (MODIS) and similar instruments for the investigation and monitoring of sea-ice polynyas compared to passive-microwave-based sensors. The very low thermal contrast between present clouds and the sea-ice surface in combination with the lack of available visible and near-infrared channels during polar nighttime results in deficiencies in the MODIS cloud mask and dependent MODIS data products. This leads to frequent misclassifications of (i) present clouds as sea ice or open water (false negative) and (ii) open-water and/or thin-ice areas as clouds (false positive), which results in an underestimation of actual polynya area and subsequently derived information. Here, we present a novel machine-learning-based approach using a deep neural network that is able to reliably discriminate between clouds, sea-ice, and open-water and/or thin-ice areas in a given swath solely from thermal-infrared MODIS channels and derived additional information. Compared to the reference MODIS sea-ice product for the year 2017, our data result in an overall increase of 20 % in annual swath-based coverage for the Brunt Ice Shelf polynya, attributed to an improved cloud-cover discrimination and the reduction of false-positive classifications. At the same time, the mean annual polynya area decreases by 44 % through the reduction of false-negative classifications of warm clouds as thin ice. Additionally, higher spatial coverage results in an overall better subdaily representation of thin-ice conditions that cannot be reconstructed with current state-of-the-art cloud-cover compensation methods.

7. The AI-CORE Project - Artificial Intelligence for Cold Regions

Baumhoer, Celia[1] , Dietz, A. J.[1] , Heidler, K.[1] , Zhu, X. X.[2] , Scheinert, M.[3] , Loebel, E.[3], Nitze, I.[4] , Dinter, T.[4] , Frickenhaus, S.[4] ([1] Deutsches Zentrum für Luft- und Raumfahrt (DLR), [2] TUM, [3] TU Dresden, [4] Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

Funded by the Helmholtz Foundation, the aim of the Artificial Intelligence for COld REgions (AI-CORE) project is to develop methods of Artificial Intelligence for solving some of the most challenging questions in cryosphere research by the example of four use cases. These use cases are of high relevance in the context of climate change but very difficult to tackle with common image processing techniques. Therefore, different AI-based imaging techniques are applied on the diverse, extensive, and inhomogeneous input data sets.

In a collaborative approach, the German Aerospace Center, the Alfred-Wegener-Institute, and the Technical University of Dresden work together to address not only the methodology of how to solve these questions, but also how to implement procedures for data integration on the infrastructures of the partners. Within the individual Helmholtz centers already existing competences in data science, AI implementation, and processing infrastructures exist but are decentralized and distributed among the individual centers. Therefore, AI-CORE aims at bringing these experts together to jointly work on developing state of the art tools to analyze and quantify processes currently occurring in the cryosphere. The presentation will give a brief overview of the geoscientific use cases and then address the different challenges that emerged so far in this still on-going project. Moreover, we will give an overview of the status of this implementation and demonstrate the already available functionalities.

8. Unlocking the potential of ML for Earth and Environment researchers

Albrecht, Frauke, Arnold, C., Caus, D., Grover, H., Vlasenko, A., Weigel, T. (Deutsches Klimarechenzentrum (DKRZ))

This presentation reports on support done under the aegis of Helmholtz AI for a wide range of machine learning based solutions for research questions related to Earth and Environmental sciences. We will give insight into typical problem statements from Earth observation and Earth system modeling that are good candidates for experimentation with ML methods and report on our accumulated experience tackling such challenges with individual support projects. We address these projects in an agile, iterative manner and during the definition phase, we direct special attention towards assembling practically meaningful demonstrators within a couple of months. A recent focus of our work lies on tackling software engineering concerns for building ML-ESM hybrids.

Our implementation workflow covers stages from data exploration to model tuning. A project may often start with evaluating available data and deciding on basic feasibility, apparent limitations such as biases or a lack of labels, and splitting into training and test data. Setting up a data processing workflow to subselect and compile training data is often the next step, followed by setting up a model architecture. We have made good experience with automatic tooling to tune hyperparameters and test and optimize network architectures. In typical implementation projects, these stages may repeat many times to improve results and cover aspects such as errors due to confusing samples, incorporating domain model knowledge, testing alternative architectures and ML approaches, and dealing with memory limitations and performance optimization.

Over the past two years, we have supported Helmholtz-based researchers from many subdisciplines on making the best use of ML methods along with these steps. Example projects include wind speed regression on GNSS-R data, emulation of atmospheric chemistry modeling, Earth System model parameterizations with ML, marine litter detection, and rogue waves prediction. The presentation will highlight selected best practices across these projects. We are happy to share our experience as it may prove useful to applications in wider Earth System modeling. If you are interested in discussing your challenge with us, please feel free to chat with us.

9. Making marine image data FAIR with iFDOs

Schoening, Timm (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

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 and to provide a standardised interface to image (meta) data for data science applications such as annotation, visualization, digital twins or machine learning.

10. ClimART: A Benchmark Dataset for Emulating Atmospheric Radiative Transfer in Weather and Climate Models

Rühling Cachay, Salva[1, 2], Ramesh, V.[1], Cole, J.[3], Barker, H.[3], Rolnick, D.[1] ([1] Mila - Quebec AI Institute, [2] TU Darmstadt, [3] Environment and Climate Change Canada)

Numerical simulations of Earth's weather and climate require substantial amounts of computation. This has led to a growing interest in replacing subroutines that explicitly compute physical processes with approximate machine learning (ML) methods that are fast at inference time. Within weather and climate models, atmospheric radiative transfer (RT) calculations are especially expensive. This has made them a popular target for neural network-based emulators. However, prior work is hard to compare due to the lack of a comprehensive dataset and standardized best practices for ML benchmarking. To fill this gap, we introduce the ClimART dataset, which is based on the Canadian Earth System Model, and comes with more than 10 million samples from present, pre-industrial, and future climate conditions.

ClimART poses several methodological challenges for the ML community, such as multiple out-of-distribution test sets, underlying domain physics, and a trade-off between accuracy and inference speed. We also present several novel baselines that indicate shortcomings of the datasets and network architectures used in prior work.

11. Machine Learning applications for impact modelling of climate extremes

Bouwer, Laurens Menno (Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany)

The impacts from anthropogenic climate change are directly felt through extremes. The existing research skills in assessing future changes in impacts from climate extremes is however still limited. Despite the fact that a multitude of climate simulations is now available that allows the analysis of such climatic events, available approaches have not yet sufficiently analysed the complex and dynamic aspects that are relevant to estimate what climate extremes mean for society in terms of impacts and damages. Machine Learning (ML) algorithms have the ability to model multivariate and nonlinear relationships, with possibilities for non-parametric regression and classification, and are therefore well-suited to model highly complex relations between climate extremes and their impacts.

In this presentation, I will highlight some recent ML applications, focussing on monetary damages from floods and windstorms. For these extremes, ML models are built using observational datasets of extremes and their impacts. Here I will also address the sample selection bias, which occurs between observed moderate impact events, and more extreme events sampled in current observed and projected future data. This can be addressed by adjusting weighting for such variable values, as is demonstrated for extreme windstorm events.

Another application focusses on health outcomes, in this case the occurrence of myocardial infarctions (MI). Several ML algorithms are tested to better predict MI events under changing environmental and demographic conditions, using data from the city of Augsburg (Germany) between 1998 and 2015. Multivariable predictors include weather (air temperature, relative humidity), air pollution (particulate matter, nitrogen oxide, nitrogen dioxide, sulphur dioxide, and ozone), surrounding vegetation, as well as demographic data.

Finally, I will suggest some further applications that could be developed for predicting climate impacts as well as impacts from policy planning and adaptation.

12. Towards the detection of ocean carbon regimes (POSTER)

Mohanty, Sweety[1, 2], Kazempour, D. D.[2], Patara, D. L.[1], Kröger, P. D. P.[2] ([1] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, [2] Christian-Albrechts Universität zu Kiel)

In the context of global climate change and environmental challenges, one research question is how different ocean regions take up carbon dioxide and which bio-physical drivers are responsible for these patterns. The carbon uptake at the sea surface is different in different areas. It depends on several drivers (sea surface temperature, the salinity of the water, alkalinity, dissolved inorganic carbon, phytoplankton, etc.), which enormously vary on both a spatial and seasonal time scale. We name a carbon regime a region having common relationships (on a seasonal and spatial scale) between carbon uptake and its drivers (sea surface temperature, etc.).

We are using the output of a global ocean biogeochemistry model providing surface fields of carbon uptake and its drivers on a monthly time scale. We aim to use spatial and seasonal correlations to detect the regimes. We take advantage of both supervised and unsupervised machine learning methodologies to find different carbon states. The aim is to determine individual local correlations in each carbon state. We build a top-down grid-based algorithm that incorporates both regression and clustering algorithms. The technique divides the entire ocean surface into smaller grids. The regression model detects a linear relationship between carbon uptake and other ocean drivers in each grid box and over each of the twelve months in a year. The correlation clustering model provides clusters of carbon states that have a distinct connection between carbon uptake and different ocean drivers. While the detection of clusters that exhibit correlations relies on static data, here, the aim is to include both the spatial and temporal dimensions, which will reveal temporal trajectories of changes in correlations.

13. Layerwise Relevance Propagation for Echo State Networks applied to Earth System Variability. (POSTER)

Landt-Hayen, Marco (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

Artificial neural networks (ANNs) are known to be powerful methods for many hard problems (e.g. image classification or timeseries prediction). However, these models tend to produce black-box results and are often difficult to interpret. Here we present Echo State Networks (ESNs) as a certain type of recurrent ANNs, also known as reservoir computing. ESNs are easy to train and only require a small number of trainable parameters. They can be used not only for timeseries prediction but also for image classification, as shown here: Our ESN model serves as a detector for El Nino Southern Oscillation (ENSO) from sea-surface temperature anomalies. ENSO is actually a well-known problem and has been widely discussed before. But here we use this simple problem to open the black-box and apply layerwise relevance propagation to Echo State Networks.

14. Fast and Accurate Physics-constrained Learning with Symmetry Constraints for the Shallow Water Equations (POSTER)

Huang, Yunfei, Greenberg, D. S. (Helmholtz-Zentrum Hereon)

The shallow water equations (SWEs) are widely employed for governing a large-scale fluid flow system, for example, in the coastal regions, oceans, estuaries, and rivers. These partial differential equations (PDEs) are often solved using semiimplicit schemes that solve a linear system iterativelyat each time step, resulting in high computational costs. Here we use physics constrained deep learning to train a convolutoinal network to solve the SWEs, while training on the discretized PDE directly without any need for numerical simulations as training data data. To improve accuracy and stability over longer integration times, we utilise group equivariant convolutional networks, so that the the learned model respects rotational and translational symmetries in the PDEs as hard constraints at every point in the training process. After training, our networks accurately predict the evolution of SWEs for freely chosen initial conditions and multiple time steps. Overall, we find that symmetry constraints signficantly improve performance compared to standard convolution networks.

15. Robust Detection of Marine Life with Label-free Image Feature Learning and Probability Calibration (POSTER)

Schanz, Tobias, Möller, K. O., Rühl, S., Greenberg, D. S. (Helmholtz-Zentrum Hereon)

Advances in imaging technology for in situ observation of marine life has significantly increased the size and quality of available datasets, but methods for automatic image analysis have not kept pace with these advances. On the other hand, knowing about distributions of different species of plankton for example would help us to better understand their lifecycles, interactions with each other or the influence of environmental changes on different species.

While machine learning methods have proven useful in solving and automating many image processing tasks, three major challenges currently limit their effectiveness in practice. First, expert-labeled training data is difficult to obtain in practice, requiring high time investment whenever the marine species, imaging technology or environmental conditions change. Second, overconfidence in learned models often prevents efficient allocation of human time. Third, human experts can exhibit considerable disagreement in categorizing images, resulting in noisy labels for training. To overcome these obstacles, we combine recent developments in self-supervised feature learning based with temperature scaling and divergence-based loss functions. We show how these techniques can reduce the required amount of labeled data by ~100-fold, reduce overconfidence, cope with disagreement among experts and improve the efficiency of human-machine interactions. Compared to existing methods, these techniques result in an overall 2 % to 5 % accuracy increase, or a more than 100-fold decrease in the human-hours required to guarantee semiautomated outputs at the same accuracy level as fully supervised approaches. We demonstrate our results by using two different plankton image datasets collected from underwater imaging systems at the coast of Helgoland and from a research vessel cruise in front of Kap Verde.

16. Assessing the Feasibility of Self-Supervised Video Frame Interpolation and Forecasting on the Cloudcast Dataset (POSTER)

Lin, Michelle (Mila - Quebec AI Institute)

Cloud dynamics are integral to forecasting and monitoring weather and climate processes. Due to a scarcity of high-quality datasets, limited research has been done to realistically model clouds. This proposal applies state-of-the art machine-learning techniques to address this shortage,using a real-life dataset,CloudCast.

Potential techniques, such as RNNs and CNNs paired with data augmentations are explored. Preliminary results show promise for the task of supervised video frame interpolation and video prediction. High performance is achieved with a supervised approach.

These video techniques demonstrate a potential to lower the cost for satellite capture, restoration, and calibration of errors in remote sensing data. Future work is proposed to develop more robust video predictions on this and other similar datasets. With these additions, climate scientists and other practitioners could successfully work at a higher frequency.

17. Approximation and Optimization of Environmental Simulations in High Spatio-Temporal Resolution through Machine Learning Methods (POSTER)

Azmi, Elnaz, Meyer, J., Strobl, M., Streit, A. (Karlsruhe Institute of Technology)

Environmental simulations in high spatio-temporal resolution consisting of large-scale dynamic systems are compute-intensive, thus usually demand parallelization of the simulations as well as high performance computing (HPC) resources. Furthermore, the parallelization of existing sequential simulations involves potentially a large configuration overhead and requires advanced programming expertise of domain scientists. On the other hand, despite the availability of modern powerful computing technologies, and under the perspective of saving energy, there is a need to address the issues such as complexity and scale reduction of large-scale systems' simulations. In order to tackle these issues, we propose two approaches: 1. Approximation of simulations by model order reduction and unsupervised machine learning methods, and 2. Approximation of simulations by supervised machine learning methods.
In the first method, we approximate large-scale and high-resolution environmental simulations and reduce their computational complexity by employing model order reduction techniques and unsupervised machine learning algorithms. In detail, we cluster functionally similar model units to reduce model redundancies, particularly similarities in the functionality of simulation model units and computation complexity. The underlying principle is that the simulation dynamics depend on model units' static properties, current state, and external forcing. Based on this, we assume that similar model units' settings lead to similar simulation dynamics. Considering this principle in the use case of a hydrological simulation named CAOS [1], we clustered the model units, ran the simulation model on a small representative subset of each cluster, and scaled the simulation output of the cluster representatives to the remaining cluster members. Experiments of this approach resulted in a balance between the simulation uncertainty and its computational effort. For evaluation of the quality of our approach, we used the proximity of the test simulation output to the original simulation, and to show the computational complexity of the approach, we measured the speedup of test simulation run time to the original simulation. Applying this approach to the CAOS use case results in a Root Mean

Square Error (RMSE) of 0.0049 and a 1.8x speedup compared to the original simulation.
In the second method, we approximate simulations through supervised machine learning methods focusing on deep neural networks. In this ongoing approach, we input multidimensional time series data into a Long Short-Term Memory network (LSTM). The LSTM model learns long-term dependencies and memorizes the information of previously seen data to predict the future data. In our use case simulation ICON-ART [2], the atmosphere is divided into cells with several input variables in which the concentration of trace gases is simulated. This simulation is based on coupled differential equations. The goal of this approach is to replace the compute-intensive chemistry simulation of about two million atmospheric cells with a trained neural network model to predict the concentration of trace gases at each cell and to reduce the computation complexity of the simulation.

[1] E. Zehe, et al. 2014. HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. HESS 18: 4635–4655. doi: 10.5194/hess-18-4635-2014
[2] D. Rieger, et al. 2015. ICON–ART 1.0 – a new online-coupled model system from the global to regional scale. GMD 8: 1659–1676. doi: 10.5194/gmd-8-1659-2015

18. Machine-Learning-Based Comparative Study to Detect Suspect Temperature Gradient Error in Ocean Data. (POSTER)

Chouai, Mohamed, Reimers, F., Vredenborg, M., Pinkernell, S., Mieruch-Schnülle, S. (Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

Thousands of ocean temperature and salinity measurements are collected every day around the world. Controlling the quality of this data is a human resource-intensive task because the control procedures still produce many false alarms only detected by a human expert. Indeed, quality control (QC) procedures have not yet benefited from the recent development of efficient machine learning methods to predict simple targets from complex multi-dimensional features. With increasing amounts of big data, algorithmic help is urgently needed, where artificial intelligence (AI) could play a dominant role. Developments in data mining and machine learning in automatic oceanographic data quality control need to be revolutionized. Such techniques provide a convenient framework to improve automatic QC by using supervised learning to reduce the discrepancy with the human expert evaluation.

This scientific work proposes a comparative analysis of machine learning classification algorithms for ocean data quality control to detect the suspect temperature gradient error. The objective of this work is to obtain a very effective QC classification method from ocean data using a representative set of supervised machine learning algorithms. The work to be presented consists of the second step of our overall system, in which the first is based on a deep convolutional neural network to detect good/bad profiles, and the second is to locate bad samples. For this reason, the dataset used to train the used benchmarking models is composed only of bad profiles.

The following algorithms are used in this study (with a hyperparameters optimisation): Multilayer Perceptron (MLP), Support Vector Machine (SVM) with different kernels, Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA). Optimization of the hyper-parameters using Grid-Search is required to ensure the best classification results.

The results obtained on the Unified Database for the Arctic and Subarctic Hydrography (UDASH) dataset are promising, especially with the MLP algorithm, in which we had an accuracy of 86.64% in the detection of good samples and 88.84% in the detection of the bad samples, where room for improvement exists. This system could have the potential to be used as a semi-automatic quality control system.

19. Machine Learning Parameterization for Cloud Microphysics (POSTER)

Sharma, Shivani, Greenberg, D. S. (Helmholtz-Zentrum Hereon)

In weather and climate models, physical processes that can't be explicitly resolved are often parameterized. Among them is cloud microphysics that often works in tandem with the convective parameterization to control the formation of clouds and rain.

Existing parameterization schemes available for cloud microphysics suffer from an accuracy/speed trade-off. The most accurate schemes based on Lagrangian droplet methods are computationally expensive and are only used for research and development. On the other hand, more widely used approaches such as bulk moment schemes simplify the particle size distributions into the total mass and number density of cloud and rain droplets.

While these approximations are fairly realistic in many cases, they struggle to represent more complex microphysical scenarios.
We develop a machine learning based parameterization to emulate the warm rain formation process in the Super droplet scheme (a type of Lagrangian scheme) in a dimensionless control volume. We show that the ML based emulator matches the Lagrangian simulations better than the bulk moment schemes, especially in the cases of very skewed droplet distributions. Compared to previous attempts in emulating warm rain, our ML model shows a better performance. The ML model inference runs fast thereby reducing the computational time otherwise needed for Lagrangian schemes. Thus, we have developed an ML based emulator that is more accurate than the commonly used schemes with only a small computational overhead, hence, making it possible to indirectly use Lagrangian schemes in operational weather models.

20. Low-Carbon Routing using Genetic Stochastic Optimization and Global Ocean Weather (POSTER)

Shchekinova, Elena[1], Rath, W.[1], Biastoch, A.[1], Amann, N.[2], Hennemann, F.[3], Kraus, P.[3], Myagotin, A.[3], Renz, M.[2], van der Wulp, S.[3] ([1] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, [2] Christian-Albrechts Universität zu Kiel, [3] TrueOcean GmbH)

Introduction
Existing marine technology introduces the capability for every marine vehicle to integrate in a timely manner a large collection of global ship positioning data along with instructed higher safety and reduced emission intensity routes. New AI-assisted navigational devices could quickly integrate unsupervised routing predictions based on assimilated and forecasted global ocean and weather.
Description of goals
Here we design a route optimization algorithm for taking advantage of current predictions from ocean circulation models. We develop validation scenarios for a marine vehicle and show how it can propagate in a safer environment. Our optimization is designed to fulfill emission goals by achieving a lower fuel use.
Results and Achievements
The weather data from the European Observational Marine Copernicus Center allows leveraging both satellite observations and high resolution wind, wave and current predictions at any position in the global ocean. We propose a low fuel consumption, carbon emission ship routing optimization that employs these real time high resolution data in a stochastic optimization algorithm.

The proposed optimization method is based on local continuous random modifications subsequently applied to an initial shortest-distance route between two points. It is parallelised using a genetic approach. The model is validated using both vessel noon reports and data from a global automated identification system records.
We show that for the performed routing scenarios across the north Atlantic the achieved fuel could save about 10% of fuel for slow-steaming scenarios and hence lead to an additional reduction of carbon emissions even for already fuel-optimized operation.

21. Learning deep emulators for the interpolation of satellite altimetry data (POSTER)

Georgenthum, Hugo, Fablet, R. (IMT Atlantique)

Over the last few years, a very active field of research has aimed at exploring new data-driven and learning-based methodologies to propose computationally efficient strategies able to benefit from the large amount of observational remote sensing and numerical simulations for the reconstruction, interpolation and prediction of high-resolution derived products of geophysical fields. These approaches now reach state-of-the performance for the reconstruction of satellite-derived geophysical fields. In this context, deep emulators emerge as new means to bridge model-driven and learning-based frameworks. Here, we focus on deep emulators for reconstruction and data assimilation issues, and more specifically on 4DVarNet schemes. These schemes bridge variational data assimilation formulation and deep learning schemes to learn 4DVar models and solvers from data. Here, we present an application of 4dVarNet schemes to the reconstruction of sea surface dynamics. More specifically, we aim at learning deep emulators for the interpolation from altimetry data. Similarly to a classic optimal interpolation, we leverage a minimization-based strategy but we benefit from the modeling flexibility of deep learning framework to embed non-linear and multi-scale priors and learn jointly a gradient-based solver for the underlying variational cost. Overall, the proposed 4dVarNet scheme defines an end-to-end neural architecture which use irregularly-sampled altimetry data as inputs and outputs a gridded and gap-free fields. We report numerical experiments within a data challenge dedicated to the benchmarking of SSH (Sea Surface Height) mapping algorithms. This data challenge relies on an observation system

simulation experiments (OSSE) setting in a Gulf Stream region with nadir and wide-swath SWOT satellite altimetry data.
Our numerical experiments demonstrate that we can train neural interpolation schemes with very large missing data rates (between 90% and 95%) in a supervised manner. The proposed approach outperforms state-of-the-art schemes, including model-driven ones, and significantly improves the resolved space and time scales compared to the operational optimally-interpolated SSH product. We further discuss the extensions of the proposed scheme especially towards the multi-scale reconstruction of sea surface dynamics from multi-source data.

22. AI4FoodSecurity: Identifying crops from space (POSTER)

Albrecht, Frauke[1], Arnold, C.[1, 2], Bukas, C.[2] ([1] Deutsches Klimarechenzentrum (DKRZ), [2] Helmholtz Zentrum München)

The European Space Agency (ESA) launched the AI4EO initiative to bridge the gap between the artificial intelligence (AI) and the Earth observation (EO) communities [1]. In the AI4FoodSecurity challenge [2], the goal is to identify crops in agricultural fields using time series remote sensing data from different satellites. In the first challenge track, predictions were made for a region in South Africa, including a spatial domain shift. In the second challenge track, predictions were made for a region in Germany (Brandenburg), including a spatio-temporal domain shift.
We here present our contribution to the AI4FoodSecurity challenge. As data sources we selected both radar wavelength images from the Sentinel-1 satellite, as well as visual and infrared images from the Planet Fusion Monitoring satellites. We implemented a Pixel-Set Encoder with Lightweight Temporal Attention (PseLTae) [3]. Samples are constructed by randomly selecting pixels from a given agricultural field. We train separate encoders for each data source. Attention heads are used to extract characteristic changes of the distinct crop types throughout the growing season. At the decoder stage, both sources are combined to yield a prediction for the crop type. We used data augmentation by oversampling the agricultural fields, as well as cross validation.
The quality of the predictions is evaluated by a multi-class classification score. We finished the challenge in second place in both tracks. We found the model performs better in the South African region, where only the spatial domain changed, compared to the German region that was evaluated in the next vegetation period. Our code is available on Github [4].

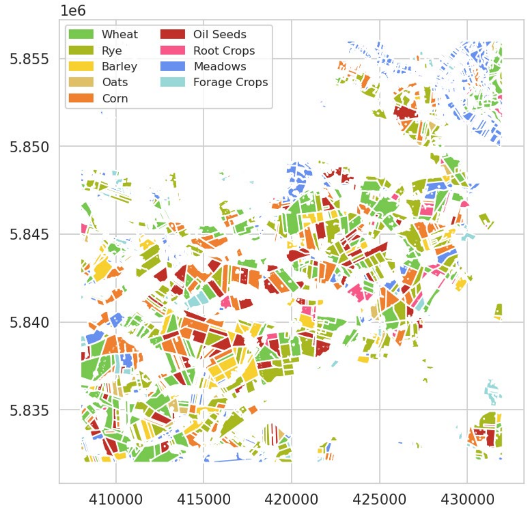

Figure: Colored map of crop predictions for the region in Germany, including nine crop types.

References:

[1] A.-K. Debien et al, 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 251-253 (2021).

[2] AI4FoodSecurity Challenge: https://ai4eo.eu/ai4eo-ai4foodsecurity-challenge

[3] V. Sainte Fare Garnot et al, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12325-12334 (2020)

[4] https://github.com/crlna16/ai4foodsecurity

23. Automatic low-dimension explainable feature extraction of climatic drivers leading to forest mortality (POSTER)

Anand, Mohit[1], Camps-Valls, G.[2], Zscheischler, J.[1] ([1] Helmholtz Centre for Environmental Research, [2] Image Processing Laboratory (IPL), Universitat de València)

Forest mortality is a complex phenomenon because of the interaction of multiple drivers over a long period. Understanding these interactions and their relevant time scale is important for forest management. Unlike climate data which are continuous (daily or hourly resolutions), forest mortality related observations are discrete (e.g. the number of trees, mortality fraction) with lower frequencies (once/twice a year). Forests also have persistent memory with a large buffering capacity. All the above-mentioned reasons make the analysis of forest mortality difficult with the conventional tools. Deep learning is well suited for modelling multivariate time series with persistent nonlinear interactions. In this study, we generate 200,000 years of hourly climate data using a weather generator (AWE-GEN). We aggregate the hourly data to daily values and feed it to a process based forest model (FORMIND). The forest model gives us mortality fractions per year, in line with the forest mortality related observations. For the method development phase, we use these simulated data. First, we use a variational autoencoder to extract climatic features and use that for the prediction of forest mortality. In the second stage, we do the prediction of forest mortality and feature extraction together and illustrate the difference between the extracted features. Further different approach to disentangle the extracted feature are tested. Finally, we present the analysis of performance versus explainability for different approaches.

24. Model evaluation method affects the interpretation of machine learning models for identifying compound drivers of maize variability (POSTER)

Sweet, Lily-Belle, Zscheischler, J. (Helmholtz Centre for Environmental Research)

Extreme impacts can be caused by the compounding effects of multiple drivers, such as weather events that might not individually be considered extreme. An example of this is the phenomenon of 'false spring', where a combination of a warm late winter or early spring, followed by a frost once the plants have entered a vulnerable stage of development, results in severe crop damage. The relationships between growing-season climate conditions and end-of-season crop yield are complex and nonlinear, and improving our understanding of such interactions could aid in narrowing the uncertainty in estimates of climate risk to food security. Additionally, data-driven methods that are capable of identifying such compounding effects could be useful for the study of other sectoral impacts.

Machine learning is an option for capturing such complex and nonlinear relationships for yield prediction. In order to extract these relationships, explainable or interpretable machine learning has been identified as a potential tool. However, the usefulness of those extracted interpretations is dependent on the assumption that the model has learned the expected relationships. One prerequisite for this assumption is that the model has sufficient predictive skill. The method chosen for measuring model performance is therefore an important methodological decision, but as yet the 'best practice' when handling spatiotemporal climate data is not clearly defined.

In this study we train machine learning models to predict maize yield variability from growing-season climate data, using global climate reanalysis data and corresponding driven process-based crop model output. We assess the impact of the cross-validation procedure used for model skill measurement on each step of the modelling process: hyperparameter tuning, feature selection, performance evaluation and model interpretation. We show that the method of evaluating model skill has significant impacts on results when using interpretable machine learning methods. Our results suggest that the design of the cross-validation procedure should reflect the purpose of the study and the qualities of the data used, which in our case are highly-correlated spatiotemporal climate and crop yield data.

Towards Digital Twins and other Lighthouse Projects

1. Keynote: Digital Twins of the Ocean (DITTO) - Opportunities to future-proof Ocean Sustainable Development

Visbeck, Martin (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, Christian-Albrechts Universität zu Kiel)

The ocean economy is growing and the pressures on our seas and the ocean, including from over exploitation, pollution and climate change, have asserted significant stresses on the marine system. Digital twins are rich, virtual representations of objects and systems, in this case the ocean system, or a part of it. They allow us to track how and why the things we care about are changing and simulate what their futures could be, including by exploring 'what if?' scenarios. They can provide critical knowledge to plan and guide human activities in the ocean and coastal spaces to safeguard a healthy ocean and support a sustainable green-blue economy.

Digital twins depend upon: an integrated, and sustainable, ocean observing system; well -managed, accessible and interoperable data and software; predictive processes or data-driven models with which users can interact, to support their needs; sharing of good and best practice and training, education and outreach.

The connection between a digital twin and its real-world counterpart requires a well-formulated interface between the digital twin, environmental and societal data, and the user. User interaction is an essential function embedded in the design of digital twins, to ensure maximum information value for investment in ocean observations. This may include user driven development of visualisation, user-driven data transformation and data-science tools or predictive modelling.

Digital twins thus enhance our ability to make informed operational, scientific management and policy decisions about the systems they represent. They can play an essential role in planning for future uses of the ocean and thus support ecosystem based ocean management or marine spatial planning. They also enable impactful communication that brings data to life.

The scale of effort to support the delivery of ocean information required to understand the climate system, address negative impacts of human activities, improve large-scale marine ecosystem management and guide the development of a more sustainable ocean economy, is beyond the capabilities of any single nation. Global trans-basin investment and coordination are required across the value chain to ensure fit-for-purpose Digital Twins of the Ocean. This is what the global programme DITTO of the UN Decade of Ocean Science for Sustainable Development pursues. More information about the DITTO Programme can be found at https://ditto-oceandecade.org/

2. Digital hydromorphological twin of the Trilateral Wadden Sea

Plüß, Andreas (Bundesanstalt für Wasserbau – Hamburg)

The project "Digital hydromorphological twin of the Trilateral Wadden Sea" focuses
on cooperation of cross-border data innovations between The Netherlands, Germany and Denmark,
the provision / harmonization of data together with a new digital geodata and analysis infrastructure for the trilateral Wadden Sea World Heritage Site These data and information are linked via Web portals and services to form a versatile assistance system.

Different demands, requirements and restrictive environ-mental legislation pose major
challenges for the planning and maintenance of transport infra-structure in the marine environment. TrilaWatt aims at developing and implementing a powerful
spatial data and analysis infra-structure on a homogenized database comprising
the Trilateral Wadden Sea area of the Netherlands, Germany and Denmark.

High-resolution hydrographic data series are locally available for the German Bight.
These can be combined into spatial models. Due to the high mapping effort, area-wide
morphological and especially sedimentological surveys can only be conducted at
intervals of several years or decades. On the other hand, current issues
require much higher temporal resolution for the analysis and assessment of

environmental impacts in the entire Wadden Sea. Pressures are addressed in the MSFD Annex III among others as descriptors "seabed integrity" (D6) and "hydrographic alterations" (D7). These can be sufficiently classified with modelling data as physical loss or physical disturbance, which is a crucial distinction &nbsp;for the approval of new transport and infrastructure projects.

Today, the cost of maintaining infrastructure for transport is very high. A solid and consistent database can help to find optimization potentials for partially conflicting goals such as economic efficiency, environmental interests, navigability, acceptance, etc. A digital planning assistance system based on comprehensive processed data and documented by meaningful metadata will aid the evaluation process.

Maintenance of seaport approaches and port facilities is strongly determined by the seaward sediment input. An accurate description of hydromorphology allows to identify sources, sinks, and transport paths of sediment. For a heuristic and synoptic modelling set up of the German Bight the adjacent regions must be considered. Therefore, planning and assessment for the southern North Sea should be done on a transnational basis.

Reproducing the complex physical processes in coastal and especially in tidal mudflat regions depends on an accurate data situation, which unfortunately is often heterogeneous, patchy and not harmonized across borders. TrilaWatt will provide quality-assured spatial data of geomorphology, sedimentology, and hydrodynamics together with extensive analyses obtained from numerical simulations (2005, 2010, 2015 and 2020). The data is available free of charge via geoportals such as MDI-DE, mCLOUD and GOVDATA according to the Open Data directive and the FAIR-Principle (findable, accessible, interoperable and reusable).

The novel service-based assistance system for plan-ning and reporting using generic documentation components implements Web-Processing Services (WPS) on OGC standards. This service can be used in various target systems, e. g., for transport and environmental management or the classification of Outstanding Universal Values (OUV) in the Wadden Sea. It is already being used as part of the EU reporting obligations for the Water Framework Directive (WFD) and the Marine Strategy Framework Directive (MSFD) as well as for detailed scientific studies.

Project partners are the Federal Waterways Engineering and Research Institute (BAW), the Wadden Sea Forum e. V., smile consult GmbH and planGIS GmbH.

Funding is provided by the Federal Ministry for Digital and Transport (BMDV) in its mFUND funding line.

3. Interactive Earth-System Models for Digital-Twins

Claus, Martin[1], Gundlach, S.[2], Hasselbring, W.[2], Jung, R.[2], Rath, W.[1] ([1] GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, [2] Christian-Albrechts Universität zu Kiel)

Today's Earth-System Models (ESM) are not designed to be interactive. They follow a configure, setup, run, and data analysis scheme. Thus, researchers often have to wait for hours or days until they can inspect model runtime diagnostic data. This causes time-consuming round trips between setup and analysis and limits interaction and insights into an ESM at runtime.

We aim to overcome this static scientific modeling process with interactive exploration of ESMs. It will allow to monitor the state of the simulation via dashboards presenting real-time diagnostics within a digital twin world. It will support to halt simulations, move back in time, and explore divergent setup at any given point in the simulation. Therefore, we have to include code in ESMs to access, store, and change data in every part, and make it available to interactive visualization dashboards.

In today's monolithic implementation of ESMs and and other scientific models, we have to modularize models and discover or recover interfaces between these modules. The modularization does not only help with restructuring existing ESMs, it also allows to integrate additional scientific domains into the interactive simulation environment.

We apply a domain-driven modularization approach utilizing reverse engineering techniques combining static and dynamic analysis of ESMs, as well as, restructuring methods to support scientists and developers in efforts to realize a modularization. Static analysis parses the program code extracting operation calls, e.g., subroutine and function calls in Fortran, as well as, data flow and derives the modular structure from this information. This structure is called an architectural model of the ESM. While the dynamic analysis is based on observations of operation calls at runtime utilizing compiler instrumentation functionality. The latter provides insights in the number of calls to an operation and allows also to observe access to program libraries. Both information are then combined to produce an architecture model comprising static and dynamic information. We then apply optimization methods to propose improved modularizations of the ESM that can be used to restructure ESM code, improve program comprehension, visualize dependencies, and allow to integrate interfaces to support interactive ESMs as part of digital twins.

4. Collaborative Exploration and Annotation of 4D Data with the Digital Earth Viewer

Stäbler, Flemming, Buck, V., Gonzalez, E. (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

The Digital Earth Viewer is a visualisation platform capable of ingesting data from heterogeneous sources and performing spatial and temporal contextualisation upon them. Its web-based nature enables several users to access and visualise large geo-scientific datasets. Here we present the latest development of this viewer: collaborative capabilities that allow parallel, live exploration and annotation of 4 dimensional environments by multiple remote users.

5. A Web Framework for Dynamic Data Presentations in Earth Sciences

Gonzalez, Everardo (GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel)

A comprehensive study of the Earth System and its different environments requires understanding of multi-dimensional data acquired with a multitude of different sensors or produced by various complex models. Geoscientists use state of the art instruments and techniques to acquire and analyze said data, which is in stark contrast with the outdated means that are often selected to present the resulting findings: today's most popular presentation software choices (PowerPoint, Keynote, etc.) were developed to support a presentation style that has seen slim to none development over the last 70 years. Here I present a software framework for creating dynamic data presentations. This combination of different web based resources enables a new paradigm in data visualisation for scientific presentations.

6. Data conversion for the MOSAiC webODV (POSTER)

Freier, Julia, Mieruch-Schnülle, S., Schlitzer, R. (Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

This contribution is an introduction to the data management process in the MOSAiC - Virtual Research Environment (M-VRE) project. The M-VRE project aims to make the unique and interdisciplinary data set of MOSAiC easily accessible to the field of scientists from different research areas [ADD22]. In addition, a virtual environment is available to analyze and visualize them directly online. This supports the research in improving transparency, traceability, reproducibility and visibility.
One tool that is incorporated within M-VRE is webODV, the online version of Ocean Data View (ODV) [Sch22]. ODV is a software for visualization of oceanographic data in oceanography since almost 30 years. Given its software structure, it is equally suitable for data of the atmosphere, on land, on ice. Yet, there are requirements of ODV regarding the format of the data set which is why a conversion of the data is required.
In the following, the workflow of data from archive to webODV is described. First of all, the data source needed to be defined. As part of the MOSAiC project, an agreement was reached through the MOSAiC Data Policy to upload the data to the long-term archive PANGAEA [Imm+19]. For this reason, PANGAEA is used as the data source for the webODV implementation in M-VRE.

Secondly, the automated query and download of the MOSAiC data is applied. The search of entries with tag "mosaic20192020" is automated. The PANGAEA Request Results service [PAN22b] is used to access the metadata.
The third step is the conversion of the data format. It is based on the code pangaea2odv written by R. Koppe (AWI Bremerhaven). It is a Python script to convert the PANGAEA .tab format to an ASCII format executable by ODV. The target format is a .txt file consisting in header and data in tab-separated columns. The following meta variables are defined: Basis, Cruise, Event, Station, Project, URL, RIS and BibTeX citation, Version, Last modified, Scientists, main scientist, Contact, Method, Bot. Depth [m], Original file URL, Longitude and Latitude. The data variables include all the data variables defined in PANGAEA. Depending on the data types of the collection, the primary variable is selected. Furthermore, the collections are supposed to resemble the PANGAEA data sets as closely as possible. However, it is necessary that similar measurements are combined in the same. For instance, 89 data sets were uploaded by [Aka+21]. Each record is an event and the variables and many metadata are identical. A python routine generates collection names based on the titles of the PANGAEA entries. Among other things, dates, leg numbers, etc. are removed. Finally, to build collections readable by webODV the spreadsheet files first have to be imported into ODV and then saved as a collection (consisting of .odv file and .data folder). This is automated using the terminal.
The deployment of the M-VRE webODV is accessible through the M-VRE [ADD22] project website or directly through the URL https://mvre.webodv.cloud.awi.de/ [AWI22]. However, the MOSAiC data policy [Imm+19] established that the public release will be on 01/01/2023. Until then, the login requires an AWI account and membership in the MOSAiC consortium. The data structure in which the collections are embedded is based on the structure of the science teams during the Expedition.

[Imm+19] Immerz et al. MOSAiC Data Policy. 2019
[Aka+21] Akansu et al. Tethered balloon-borne measurements of turbulence during the MOSAiC expedition from December 2019 to May 2020. 2021
[AWI22] AWI. MOSAiC webODV. 2022
[ADD22] AWI, DKRZ, and DLR. MOSAiC –Virtual Research Environment.2022
[PAN22a] PANGAEA. Data Publisher for Earth & Environmental Science. 2022
[PAN22b] PANGAEA. OAI 2.0 Request Results. 2022
[Sch22] Schlitzer. ODV 5.6.2. 2022

7. HELMI – The Hereon Layer For Managing Incoming Data (LIVE DEMO)

Böcke, Max, Hemmen, J., Jacobsen, C., Leefmann, T., Listing, O., Plewka, J. (Helmholtz-Zentrum Hereon)

The Hereon operates a larger number of continuously measuring sensors on mobile and stationary platforms outside the Hereon Campus. Transferring the data from the sensors to the internal network is a critical step , as data is often required to be accessible for researchers in near-real-time (NRT) and needs be retrieved from the outside in a secure way that does not pose a threat to the internal infrastructure.
Therefore, we developed the He reon L ayer for M anaging I ncoming data (HELMI). Using HELMI, data from external sensor systems is moved securely via a Virtual Private Network solution (Wireguard) to the Hereon internal infrastructure. The Wireguard client can be either installed directly on the sensor system or on a piece of hardware dedicated for data transfer that is connected to the sensor. Data is transferred as files via the RSYNC or as NRT data via the Message Queuing Telemetry Transport (MQTT) protocol, respectively. After transfer researchers can retrieve their files and access telemetry from an internal endpoint. NRT-Data can be automatically visualized using web applications and parameters at the client can be remotely controlled .
The automatic data transfer via HELMI minimizes the risk of data loss, reduces memory requirements on the measuring system and allows to provide data in near real-time and thus to speed up the publication process of data.

8. MOSAiC webODV – An online service for the exploration, analysis and visualization of MOSAiC data (LIVE DEMO)

Mieruch-Schnülle, Sebastian, Freier, J., Schlitzer, R. (Alfred-Wegener-Institut - Helmholtz-Zentrum für Polar- und Meeresforschung)

Introduction
MOSAiC (https://mosaic-expedition.org/) has been the largest polar expedition in history. The German icebreaker Polarstern was trapped in the ice from October 2019 to October 2020, and rich data have been collected during the polar year. The M-VRE project (The MOSAiC Virtual Research Environment, https://mosaic-vre.org/) has the aim to support the analysis and exploitation of the MOSAiC data by providing online software tools for the easy, interdisciplinary and efficient exploration and visualization of the data. One service provided by

M-VRE is webODV, the online version of the Ocean Data View Software (ODV, https://odv.awi.de/).

Setup

The MOSAiC webODV is available via https://mosaic-vre.org/services/ or dirctly at https://mvre.webodv.cloud.awi.de/. Due to a moratorium until the end of 2022, the data can only be accessed by the MOSAiC consortium. From 2023 on, MOSAiC data and thus webODV will be available for the science community and the general public. In the webODV configuration, datasets as well as the ODV software reside and run on a server machine, not on the client computer. The browser client communicates with the server over the Internet using secure websockets. Up to now we provide two webODV services, which are described in the following.

Data Extraction

The Data Extraction service is based on intuitive web elements like buttons, dropdowns, date widgets and a drag & drop zoom function for the map. The aim is to provide data sub-setting as easy and fast as possible. A pager on the top of the site guides the user through the pages, which includes mainly the selection of stations, variables and finally the download function. Data can be finely granulated selected e.g. by zooming into the map, defining time windows and restrictions to specific variables. Finally the selected data can be downloaded as text files, ODV collections or netCDF files for later use with the standalone ODV or other tools.

Data Exploration

The Data Exploration service or ODV-online aims to provide the look-and-feel and functionality of the desktop ODV in the browser window for creating maps, surface plots, section plots, scatter plots, filtering data etc. Here the browser window resembles the ODV desktop application window consisting of canvas with station map and data windows and metadata, sample data and isosurface value lists on the right side. As in the desktop ODV, left mouse clicks on stations or data points select these items. Right-clicks bring up context menus providing the familiar ODV functionality. Users can download image files of the entire canvas or individual windows and can export the data of the current station set or of individual data windows.

Data provider traceability

A “cite” button has been implemented, which provides all dataset DOIs or citations (.txt, .bib or .ris) involved in the current visualization.

Reproducibility

Analyses and visualizations which have been created with ODV-online can be saved as so-called xviews, which are essentially XML files, which include instructions for ODV to create visualizations. Via our xview manager users can download, upload and delete personal xviews. These files can be used to fully reproduce the analyses and visualizations and can be shared e.g. with colleagues. Additionally, the xview files can be added e.g. as supplementary material to publications, and together with the link to the MOSAiC data collection in webODV, provide full reproducibility.

Sharing

A quick share functionality has been implemented, i.e. a link to a visualization, which can be retrieved by a single click and send to a colleague and is valid for 72h. If a colleague opens the link in the browser he/she is immediately send to the just created visualization to quickly discuss e.g. important findings or to continue own research.

Spitzenforschung
für eine Welt im Wandel

Helmholtz-Zentrum Hereon
Max-Planck-Straße 1
21502 Geesthacht

www.hereon.de