5th Data Science Symposium

22nd January 2021, virtual meeting

Modern digital scientific workflows - often implying Big Data challenges - require data infrastructures and innovative data science methods across disciplines and technologies. Diverse activities within and outside HGF deal with these challenges, on all levels. The series of Data Science Symposia fosters knowledge exchange and collaboration in the Earth and Environment research community.

We invited contributions to the overarching topics of data management, data science and data infrastructures.

The series of Data Science Symposia is a joint initiative by the three Helmholtz Centers:

Helmholtz-Zentrum Geesthacht
Centre for Materials and Coastal Research

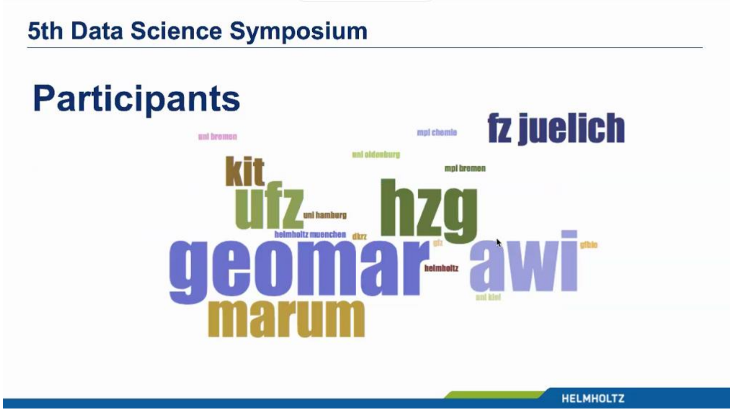
AWI

GEOMAR

The abstracts are given in the order of the program of the day.

Contact/Organization: Hela Mehrtens and Daniela Henkel (GEOMAR)

124 registered participants
97 attendees online at 10:00 am
99 attendees online at 01:40 pm
80 attendees online at 04:00 pm
24 abstracts
3 posters



5th Data Science Symposium

Participants

DAM core area “Data management and Digitalisation”: Pilot project “Underway”-research data

Gauvain Wiemer, DAM Underway Consortium

To meet the challenges of researching the seas and oceans, the scientific and technical capacities of German marine research are bundled in an internationally outstanding network of expertise and institutions (Deutsche Allianz Meeresforschung e.V. - DAM). The alliance addresses the major future issues of marine research at the highest level and strengthens the scientific contribution to knowledge for action for politics, business, and civil society. DAM implements activities in four core areas - research, infrastructure, transfer, and data management and digitization. The non-profit association (DAM) was founded in July 2019, initiated by the federal government and the five northern German states (Free Hanseatic City of Bremen, Free and Hanseatic City of Hamburg, Mecklenburg-Western Pomerania, Lower Saxony and Schleswig-Holstein). In September 2019, DAM began implementing the pilot project "Underway"-Research Data in the core area "Data Management and Digitization". The project "Underway"-Research Data is concerned with the systematic collection, transmission on land, quality control and publication of "Underway"-Data according to the OPEN and FAIR principles (findable, accessible, interoperable, reusable) – a basic requirement for data science. The focus is on the ships MARIA S. MERIAN, METEOR, POLARSTERN and SONNE. “Underway”-Data is defined as data measured by sensors that are permanently on bord of the vessels and collect data relevant for marine science. Devices and sensors that are prioritized in this project include the Acoustic Doppler Current Profiler (ADCP), bathymetry signals, CTD system (Conductivity, Temperature, Depth), the Thermosalinograph (TSG), the Ferrybox and bio-optical sensors. Additionally, the "German marine data portal" is being further developed in close cooperation with the MARE-Hub project to simplify the access to and visualization of marine research data. The DAM office coordinates the networking and coordination process with the participating institutions (AWI, BSH, CAU/KMS, GEOMAR, HZG, ICBM, IOW, MARUM, MPI-C). The presentation will give a short Introduction to the DAM and to the major achievements within the pilot project with regard to potential benefits for the scientific community.

Ingest Framework for scalable raw data publication

Maximilian Betz, Sebastian Immoor, Dana Ransby, Roland Koppe, Peter Gerchow, Antonia Immerz, Stephanie Schumacher, Janine Felden, Frank Oliver Glöckner

The data flow framework Observations to Archives (O2A) developed by the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, is designed for a seamless flow of data from sensors to archives. O2A is currently used as the central data management backbone for MOSAiC - the Multidisciplinary drifting Observatory for the Study of Arctic Climate. Between Sept 2019 and Oct 2020 hundreds of TB of raw data from scientific instruments were created. With over 8000 device operations and 1300 registered sensors and sampling devices it became the O2A’s largest use case. MOSAiC installed research data guidelines to define the flow of data and metadata from research platforms to data storage and archiving. The O2A framework supports the enrichment and harmonisation of distributed heterogeneous data with metadata which is a prerequisite for their publication according to the FAIR (Findable, Accessible, Interoperable, Re-usable) data principles. MOSAiC data is primarily published in the long-term data repository PANGAEA. The ingest submission framework, presented here, implements an automated raw data submission to PANGAEA. Publishing data from a sensor requires a data description. This technical file and storage structure description, in form of a template, is used by the framework to identify the relevant files and to extract available metadata. A sensor is hereby identified by a unique id registered in the sensor information system SENSOR.awi.de. Data becomes interoperable as it is linked to the sensors' metadata, the corresponding scientific event and campaign, contains a file description and a georeference. The goal is that curation work in the repository will be limited to metadata quality checks and obtaining approval for open access publication by data providers. With this, the publication process can be accelerated, metadata is more consistent and the human error is reduced. Publication of subsequent datasets of the same sensor in the future becomes possible with minimal effort by using the initially defined template. This efficient solution scales with increased publication demand.

Marine Inorganic Carbon synthesis data products in the era of big marine data

Nico Lange, Benjamin Pfeil, Steve D. Jones, Kevin M. O'Brien, Björn Fiedler, Toste Tanhua

Entering the era of marine big-data, synthesis data products tailored around specific biogeochemical (BGC) Essential Ocean Variables (EOV) greatly improve today's BGC ocean data usage and implement the FAIR principles in the ocean discipline. Presently, two separate BGC data-products with a different focus on the EOV Inorganic Carbon (IC) and its sub-parameters pH, DIC, TA and pCO2, exist. These prominent products are the Global Ocean Data Analysis Project (GLODAP) and the Surface Ocean CO2 Atlas (SOCAT). While GLODAP contains IC-relevant data from research vessels hydrographic surveys, SOCAT represents pCO2 surface data from moored stations as well as from ships of opportunity (SOOP). Both are used extensively as gold standard for validation of autonomous IC observing platforms, such as the BGC-ARGO fleet. Technical advancements towards modernized data ingestion and data handling have contributed significantly to the success of GLODAP and SOCAT. Foremost, the state-of-the-art automated and streamlined data flow design of SOCAT is of outermost relevance in times of (semi-) automated pCO2 observations and hence increasing amounts of data. Recognizing the importance of these IC synthesis products and identifying the lack of long-term monthly times-series (TS) observations, a ship-based TS product focusing on IC-related observations is in development. Eventually, all products combined enable a more throughout approach in addressing IC-related scientific goals and societal demands. In particular, in respect to national climate-related mitigation actions following the UNFCCC Paris agreement and the UN sustainable development goal (SDG) 14.3 Minimize and address the impact of ocean acidification.

Data flow, standardization, and quality control

Brenner Silva, Computing and Data Centre of the Alfred-Wegener-Institute, Bremerhaven, Germany

Earth system cyberinfrastructures include three types of data services: repositories, collections, and federations. These services arrange data by their purpose, level of integration, and governance. For instance, registered data of uniform measurements fulfil the goal of publication but do not necessarily play a part in a data flow system. The data repository provides the first and high level of integration that strongly depends on the standardization of incoming data. Applications within the Digital Earth showcases connect repositories and federated databases to the end-user, the scientist. One example here is the framework Observation to Archive and Analysis (O2A) that is operational and continuously developed at the Alfred-Wegener-Institute, Bremerhaven. The O2A uses OGC standards and a representational state transfer (REST) architecture, where both data and interface operations are openly available. A data repository is one of the components of the O2A framework and much of its functionality, for instance the near real-time monitoring, depends on the standardization of the incoming data. In this context, we develop a modular approach to provide the standardization and the quality control for monitoring of the ingested data. Two modules are under development. First, the driver module that executes transformation of tabular data into a standardized format. Second, the quality control module that runs the quality tests on the ingested data. Both modules rely on the sensor operator and on the data scientist, two actors that interact with both ends of the ingest component of the O2A framework (http://data.awi.de/o2a-doc). The result is the harmonized data of multiple sources accessible at the end-point, or the web service of the data repository (https://dashboard.awi.de/data-xxl/). Here we focus on the concepts and current development that aim at the enhanced monitoring and scientific workflow with a special focus on the modules driver and quality control.

O3as: an online ozone trend analysis service within EOSC-synergy

Tobias Kerzenmacher, Valentin Kozlov, Borja Sanchis, Ugur Cayoglu, Marcus Hardt, Peter Braesicke

O3as is a service within the framework of the European Open Science Cloud-Synergy (EOSC-Synergy) project, mainly for scientists working on the Chemistry-Climate Model Initiative (CCMI, http://blogs.reading.ac.uk/ccmi/ccmi-phase-two/) for the quadrennial global assessment of ozone depletion (https://www.esrl.noaa.gov/csl/assessments/ozone/2018/). The next assessment will be due in 2022. The recent ozone assessment report consists of six chapters and five appendices with about 25 people actively working on each chapter and a multitude of people working in support of the preparation of the document. The O3as service shall provide an invaluable tool to extract ozone trends from large data volumes created by climate projection models producing figures of stratospheric ozone trends in publication quality, in a coherent way. A web application shall be provided where a user configures his/her requests and performs simple analysis. This request is passed to the O3as service via an O3as REST API call. There, the O3as service processes the request, where the reduced data set is accessed via WebDav and OIDC. In order to produce a reduced data set, regular tasks are run on an HPC to copy primary data and perform data preparation (e.g. data reduction and parameter unification). The O3as service uses EGI Check-In to access certain functionalities of the service, the OIDC-Agent and rclone to mount data servers using WebDAV—HTTP authentication, udocker (a tool to execute simple docker containers in multi-user space without root privileges) to perform data reduction in the HPC environment, and an infrastructure manager to deploy service resources (Kubernetes).

An Integrated Customizable Framework for Data Visualization and Exploration

Robin Heß, Karen Albers, Aarthi Balamurugan, Shahzeib Jaswal, Peter Konopatzky, Roland Koppe, Alireza Mahdavi, Andreas Walter

Map-based viewers for projects of different kinds are often being developed from scratch. As a lot of work has to be done repeatedly, this is inefficient. Existing frameworks are rather fixed, oversized or hard to customize. That is why we are developing a framework for creating easily configurable web viewers in the context of different projects and initiatives like Digital Earth, MareHUB, DataHUB, DAM and O2A. Based on the framework, viewers can be assembled from a multi-level modular system. Modules e.g. for data integration, exploration, visualization can be simply linked together. The core product for end-users is a web viewer for exploration and visualization of scientific data and metadata from different sources [1]. Visualization and data access is supported by a strong modular backend for technical data integration, preparation and harmonization. This backend consists of different specialized server components (e.g. GeoServer, Rasdaman, ESRI). An open access catalogue for services and data layers integrates federated information of service-based data products. This technical framework is completed by a set of evolving standard procedures [2] to transform original scientific data into infrastructure-ready high performance and interactive data products. By using standardized module interfaces, OGC service descriptions and open source license [3] for our developments, we ensure reusability and adaptability of services. As part of the 5th Data Science Symposium, we would like to present the concept and current developments. There will also be an interactive version available for you to try out yourself.

[1] https://marine-data.org/preview/vef/usecases/prototype.html
[2] https://spaces.awi.de/x/fIzFEw
[3] https://spaces.awi.de/x/XqtrF

Distributed data analysis for better scientific collaborations

Philipp Sommer, Viktoria Wichert, Daniel Eggert, Tilman Dinter, Klaus Getzlaff, Andreas Lehmann, Christian Werner, Brenner Silva, Lennart Schmidt, Angela Schäfer

A common challenge for projects with multiple involved research institutes is a well-defined and productive collaboration. All parties measure and analyse different aspects, depend on each other, share common methods, and exchange the latest results, findings, and data. Today this exchange is often impeded by a lack of ready access to shared computing and storage resources. In our talk, we present a new and innovative remote procedure call (RPC) framework. We focus on a distributed setup, where project partners do not necessarily work at the same institute and without access to each other's resources. We present the prototype of an application programming interface (API) developed in Python that enables scientists to collaboratively explore and analyse sets of distributed data. It offers the functionality to request remote data through a comfortable interface, and to share analytical workflows and their results. Our methodology uses the Digital Earth software framework, especially its messaging component. The prototype enables researchers to make their methods accessible as a backend module running on their own servers. Hence researchers from other institutes may apply the available methods through a lightweight python API. This API transforms standard python calls into requests to the backend process on the remote server. In the end, the overhead for both, the backend developer and the remote user, is very low. The effort of implementing the necessary workflow and API usage equalizes the writing of code in a non-distributed setup. Besides that, data do not have to be downloaded locally, the analysis can be executed "close to the data" while using the institutional infrastructure where the eligible data set is stored. With our prototype, we demonstrate distributed data access and analysis workflows across institutional borders to enable effective scientific collaboration, thus deepening our understanding of the Earth system.

Benefits from Managing Biosamples Centrally and Digitally (a GEOMAR case)

Lea Lange, Jan Dierking

With the joint DataHub initiative Helmholtz has started a series of measures to go FAIR. Beginning one year ago GEOMAR aims to take biosample management to the next level. The prototype database currently under development offers a central biosample management and turns it fully digital, linking this important information to already existing data management infrastructures. Present workflows for managing biological samples have grown, been adapted and worked around before. They have proved suitable to get the work done, but not always turned out efficient in the long run. The start of the new Biosample Information System is a major shift, that forces established working routines to adjust. This talk tackles the fear of added workload for users by presenting potential and benefits for your sample work and working routines.

OPUS - An Open Portal to Underwater Soundscapes to explore and study sound in the global ocean

Karolin Thomisch, Michael Flau, Robin Heß, Andy Traumüller, Olaf Boebel

OPUS - An Open Portal to Underwater Soundscapes to explore and study sound in the global ocean Abstract : Facing an era of rapid anthropogenically induced changes in the world oceans, ocean sound is now considered an essential ocean variable (EOV) for understanding, documenting and monitoring long-term trends in anthropogenic sound and its effects on marine life, biodiversity and ecosystem health. The International Quiet Ocean Experiment (IQOE) has identified two major research interests in the context of monitoring the distribution of ocean sound in space and time: i) estimating current levels and distribution of anthropogenic sound in the ocean, and ii) assessing trends in anthropogenic sound levels across the global ocean, and recommendations on ambient noise monitoring are also part of the EU Marine Strategy Framework Directive. To facilitate addressing these research foci by international collaborative research efforts, the OPUS (Open Portal to Underwater Soundscapes) data portal, which is currently being developed by the Ocean Acoustics Group of the Alfred Wegener Institute (AWI) in Bremerhaven, Germany, financially supported by the MAREHUB initiative, is envisioned to promote the use of acoustic data collected worldwide. To this end, an Ocean Sound Software for Making Ambient Noise Trends Accessible (MANTA) is being developed to generate standardized ocean sound level data products from passive acoustic recordings according to IQOE Guidelines, and will be distributed to data owners of underwater passive acoustic data worldwide. OPUS will accept MANTA-processed data (i.e., spectral sound pressure levels at millidecade/minute resolution) together with related metadata as they become available and make them accessible under customized licensing policies via a map- and time-based selection tool and shopping basket functionality. Data products including the compiled MANTA data, parameter-naming conventions, instructions for citing the data, and other information necessary to use the data according to FAIR standards will be regularly produced by OPUS.

THREDDS Data Server a simplified way to discover and access scientific data at GEOMAR

Klaus Getzlaff, Carsten Schirnick, Markus Scheinert, Claas Faber, Franziska Weng, Lisa Paglialonga, Hela Mehrtens, Andreas Lehmann

One of today‘s challenge is the effective access to scientific data either within research groups or across different institutions to allow and increase the reusability of the data. While large operational modeling and service centers have enabled query and access to data via common web services, this is often not the case for smaller research groups. Especially the maintenance of the infrastructure and simple workflows to make the data available is a common challenge for scientists and data management. Here we would like to introduce the updated THREDDS Data Server (TDS) available at GEOMAR to provide, query, access and explore scientific data in netcdf format. This includes a simple and well documented workflow with step-by-step guidelines to provide data to the TDS system. This workflow aims to maximize the use of semi-automated processes, such as data integrity including standard metadata, checksums and persistent identifiers. By doing so, this workflow minimizes extra workload for persons involved in the data provision procedure such as scientists, data stewards and data managers but maximizes data reusibility under the FAIR principles. The TDS is a system developed and maintained by Unidata,a division of the University Corporation for Atmospheric Research (UCAR). The aim of the TDS is 1) to make it simple to enable web service access to existing output files, 2) using free technologies that are easy to deploy and configure, and 3) provide standardized, service-based tools that work in existing research environments. The TDS provides catalog, metadata, and data access services for scientific datasets with remote data access protocols including OPeNDAP, OGC

WCS, OGC WMS, and HTTPS. These standardized services enable reusability and increase the visibility of scientific datasets. We will show examples using viewer technologies to access datasets or directly explore these within common development environments such as Python or MATLAB.

Estimation of atmospheric chemical state from pressure indices

Andrey Vlasenko, Volker Mattias, Ulrich Callies

Pressure indexes explain a significant part of meteorological events. They are estimated either as the difference in maximum pressure variation regions or as the corresponding principal component. Statistical models provide a simple and effective way to predict the atmosphere's state based on pressure indexes. Their disadvantage is the assumption that predictors and forecasts fit a particular statistical model. As an alternative, the neural network approach allows us to estimate the atmospheric chemical state without any preliminary data assumptions. Still, this freedom leads to the network's inexplicable behavior. We propose to combine these approaches to predict the chemical state of the atmosphere based on indices minimizing their caveats. Our approach is based on statistical modeling and neural networks but does not depend on a specific statistical model and is much simpler than typical neural networks. It consists of a single hyperbolic tangential neuron that predicts the chemical concentrations of NO2, SO2, O3, and ethane over Europe from pressure indices.

Investigating Myocardial Infarctions in Augsburg, Germany with Machine Learning

Lennart Marien, Mahyar Valizadeh, Wolfgang zu Castell, Alexandra Schneider, Kathrin Wolf, Diana Rechid, Laurens M. Bouwer

Myocardial infarctions (MI) are a major cause of death worldwide. In addition to well-known individual risk factors, studies have suggested that temperature extremes, such as encountered during heat waves, may adversely affect MI [1]. The frequency and intensity of heat waves is increasing due to global climate change, and will likely increase further, even at levels limited to 1.5° or 2° global warming [2]. The relationship between health impacts and climate is complex, depending on a multitude of climatic, environmental, sociodemographic and behavioral factors. Machine Learning (ML) is a powerful tool for investigating complex and unknown relationships between extreme environmental conditions and their adverse impacts [3]. By combining heterogeneous health, climatic, environmental and socio-economic datasets, this study is a first step towards modelling current and future MI risk due to heat waves with ML. The basis of our data-driven approach is the KORA cohort study [4] and the MI Registry in the Augsburg region of Bavaria, Germany, comprising detailed information on MI and underlying health conditions. Additionally, weather and climate data (observations, and climate projections from the EURO-CORDEX initiative), air pollution data (e.g., PM10, PM2.5, nitrous oxides, and ozone), building characteristics (type and age), and socio-economic data (household income, education) are used for this study. Here, we present work-in-progress towards modelling heat-related health effects in Augsburg based on the MI registry and environmental data. One of the key challenges is to assemble and integrate heterogeneous data from various sources and relate it to the observed MI. We give an overview of our data sources, outline major challenges in combining them, and present ML approaches to build quantitative models from the data. Moreover, we present initial results based on a variety of ML methods such as decision trees, multi-linear regression and support vector machines.

[1] K. Chen, S. Breitner, K. Wolf, R. Hampel, C. Meisinger, M. Heier, W. von Scheidt, B. Kuch, A. Peters, A. Schneider, "Temporal variations in the triggering of myocardial infarction by air temperature in Augsburg, Germany, 1987-2014," *European Heart Journal*, vol. 40, no. 20, pp. 1600–1608, 2019. https://doi.org/10.1093/eurheartj/ehz116

[2] K. Sieck, C. Nam, L.M. Bouwer, D. Rechid, D. Jacob, "Weather extremes over Europe under 1.5 °C and 2.0 °C global warming from HAPPI regional climate ensemble simulations," *Earth System Dynamics Discussions*, 2020. https://doi.org/10.5194/esd-2020-4

[3] D. Wagenaar, J. de Jong, L.M. Bouwer, "Multi-variable flood damage modelling with limited data using supervised learning approaches," *Natural Hazards and Earth System Sciences*, vol. 17, no. 9, pp. 1683–1696, 2017. https://doi.org/10.5194/nhess-17-1683-2017

[4] R. Holle, M. Happich, H. Löwel, H.E. Wichmann, "KORA - A Research Platform for Population Based Health Research," *Gesundheitswesen*, vol. 67, no. S1, pp. 19–25, 2005. https://www.doi.org/10.1055/s-2005-858235

Artificial Intelligence for Cold Regions (AI-CORE) - a Pilot to bridge Data Analytics and Infrastructure Development

Ingmar Nitze, Julia Christmann, Long Phan, Martin Rueckamp, Angelika Humbert, Guido Grosse, Stephan Frickenhaus, Tilman Dinter; Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam/Bremerhaven, Germany

Artificial Intelligence for Cold Regions (AI-CORE) is a collaborative project of the DLR, the AWI, the TU Dresden, and is funded by the Helmholtz Foundation since early 2020. The project aims at developing AI methods for addressing some of the most challenging research questions in cryosphere remote sensing, rapidly changing ice sheets and thawing permafrost. We apply data analytics approaches to discover the data variable from data set simulated with an ice sheet model, observe the migration, and time Series analysis to predict and contrast this to simulated grounding line position. For the data assimilation in simulations of the Greenland ice sheet, we engage a level set method, that allows to derive a continuous function in time and space from discrete information at satellite acquisition time steps. We use an alpha-shape method to derive a seamless product of the margin at each time step to be used in the level set method driving the simulations. We develop AI algorithms and tools that allow scaling of our analyses to very large regions. Here we focus on the detection of Retrogressive Thaw Slumps (RTS), highly dynamic erosion processes caused by rapid permafrost thaw. We apply deep-learning based object detection on dense time-series of high-resolution (3m) multi-spectral PlanetScope satellite images and auxiliary datasets such as digital elevation models. RTS detection is challenging, as they are difficult to define semantically and spatially and are highly dynamic and embedded in different landscape settings. The results will help to understand, quantify and predict RTS dynamics and their landscape-scale impacts in a rapidly warming Arctic. We upgrade the base IT-infrastructure at AWI by integrating new GPU computing hardware into the on-premise IT-infrastructure to speed up the computing, data storage capabilities, and parallel processing, supporting the analytical workflows specifically.

Deep Neural Networks for Total Organic Carbon and Sedimentation Rate Prediction and Data-Guided Sampling

Everardo Gonzales, Ewa Burwicz

Over the past decade deep learning has been used to solve a wide array of regression and classification tasks. Compared to classical machine learning approaches (k-Nearest Neighbours, Random Forests,…) deep learning algorithms excel at learning complex, non-linear internal representations in part due to the highly over-parametrised nature of their underling models; thus, this advantage often comes at the cost of interpretability. In this work we used deep neural network to construct global total organic carbon (TOC) seafloor concentration map. Implementing Softmax distributions on implicitly continuous data (regression tasks) we were able to obtain probability distributions to asses prediction reliability. We used these techniques to create a model information map which is a key element to develop new data-driven sampling strategies for data acquisition. We then used transfer learning techniques to modify the resulting network in order to predict sedimentation rates, a closely related task.

Using machine learning to estimate soil moisture in an agricultural field

Johannes Boog, Pia Ebeling, Timo Houben, Swamini Khurana, Julia Schmid, Lennart Schmidt

Soil moisture influences several natural processes. These include water and nutrient availability to plants and microbes, soil weathering processes as well as water and energy fluxes across the land-atmosphere interface. Therefore, soil moisture is a critical variable for global water balance, ecosystem functioning and agricultural management. Soil moisture depends on various factors such as climatic conditions, topography, land use, land cover and soil properties. Usually, ground-based sensors are deployed to measure soil moisture at the point-scale. In contrast, most applications require highly resolved soil moisture data at the field scale. Interpolation from point to field scale is a challenging task due to the inherent spatio-temporal variability of soil coupled with non-linear dynamics with the afore-mentioned factors. This is especially the case when data is sparsely available. Here, we explored the estimation of soil moisture from its relationship to these factors using data-driven machine learning methods. We compared the performance of Support Vector Regression, Random Forests, Boosted Regression Trees and Neural Networks to estimate soil moisture in a two hectare large area situated in an agricultural catchment in central Germany. The data base consists of daily mean soil moisture records at 30 locations and three different depths across the site from September 2012 to September 2013. Our trained models aim at interpolating the soil moisture in space and time based on climatic factors as well as soil and terrain characteristics. Preliminary results indicate that soil moisture can be interpolated using auxiliary data and machine learning models, yet performance and feature importance highly varies across seasons and algorithms. Our study provides a basis for the features and data density required to estimate spatial and temporal soil moisture in real-time. These estimations can potentially be used by land management practitioners in the future.

A Whole New World: Leveraging the Power of Data Links with Heterogeneous Information Networks

Carola Trahms, Matthias Renz, Martin Visbeck

Datasets, such as physical measurements or particle positions in an experiment, are traditionally stored in tables in a (relational) data base. From a data mining perspective, however, in addition to analysing single data tables, even more interesting information can be found by examining the links between different tables: the context of the data. This context, i.e., the linked data tables, does not necessarily need to be of the same type as the data in the examined table. It can be anything, not solely measurements, but images, text or audio. Analysing the links between such heterogeneous types of data might lead to new and surprising insights. A natural way to represent linked objects are Information Networks. Information Networks containing only links between data of the same type have been studied extensively in recent years and they show great promise to be able to handle links between heterogeneous data as well. Heterogeneous Information Networks (HINs) have been successfully applied for analysing the relationships and patterns of relationships between objects of different types in Social Networks und Recommender Systems. This talk highlights how focusing the analysis of data on the links between different research datasets and their metadata can boost understanding and interpretability of the data. We select a subset of data in the PANGAEA database and show how information already present in the links between these datasets can protrude clearly by viewing the data from a network perspective. Additionally, this change of perspective opens the door to using well established network analytics tools on linked research data that is usually considered tabular. In turn, this can lead to automatically generating insights and discovering new connections in the data and thus, indicating unforeseen points of collaboration between the different scientific domains present at PANGAEA.

A multi-layer web service architecture for HI-CAM Net-Zero-2050 Soil Carbon App

Jianing Sun, Tanja Blome, Juliane El Zohbi, Martin Schultz, Diana Rechid

Within the Helmholtz Climate Initiative's Cluster Net-Zero-2050, focusing on climate mitigation, a "Soil Carbon App" as a prototype application is being developed. The soil carbon app will provide information on soil carbon storage potential for different agricultural land management options under changing climate conditions. The development incorporates the entire workflow from numerical model simulations to specific end user groups and might act as a role model for apps on, for example, reforestation and other land use changes. An interoperable service concept is currently being developed, which will capture the metadata of the Community Land Model simulations during execution of the model, and makes them immediately available for use and parallel large-scale data access. To this end, we propose a multi-layer web service architecture with a raster array database backend and a rich set of web services, which will allow for a flexible, automated exchange of information through standardized REST APIs. These services shall be freely available to the targeted end user groups of the app. This architecture will be a pilot study of utilizing cutting-edge technology for climate research and thus pave the way for new concepts to analyse global monitoring data and evaluate numerical models.

V-FOR-WaTer: A Virtual Research Environment for Environmental Research

Sibylle Hassler, Elnaz Azmi, Mirko Mälicke, Jörg Meyer, Marcus Strobl, Erwin Zehe

The virtual research environment V-FOR-WaTer aims at simplifying data access for environmental sciences, fostering data publications and facilitating pre-processing and data analyses with a comprehensive toolbox. By giving scientists from universities, research facilities and state offices easy access to data, appropriate pre-processing and analysis tools and workflows, we want to accelerate scientific work and facilitate the reproducibility of analyses. The prototype of the virtual research environment consists of a database with a detailed metadata scheme that is adapted to water and terrestrial environmental data. Datasets in the web portal originate from university projects and state offices. We are also finalising the connection of V-FOR-WaTer to GFZ Data Services, a repository for geoscientific data. This will ease publication of data from the portal and in turn give access to datasets stored in this repository. The compliance of the metadata scheme with international standards (INSPIRE, ISO19115) is key to being compatible with established repositories and other initiatives. The web portal is designed to facilitate typical workflows in environmental sciences. Map operations and filter options ensure easy selection of the data, while the workspace area provides tools for data pre-processing, scaling, and common hydrological applications. The toolbox also contains more specific tools, e.g. for geostatistics and soon for evapotranspiration.

It is easily extendable and will ultimately also include user-developed tools, reflecting the current research topics and methodologies in the hydrology community. Tools are accessed through Web Processing Services (WPS) and can be joined and saved as workflows, enabling more complex analyses and reproducibility of the research.

Swarm behavior of German research vessels

Norbert Anselm, Philipp Fischer, Ingeborg Bussmann, Eric Achterberg, Holger Brix, Uta Ködel, Peter Dietrich

Compared to terrestrial environments, where the assessment of environmental conditions over large scales and longer time periods is done routinely e.g., by (satellite) remote sensing or other far distance sensors, the synoptic assessment of environmental parameters over larger areas in the marine environment is still challenging. In a joint approach within the MOSES Project (Modular Observation Solutions for Earth Systems), the three coastal research vessels Mya II/Uthörn (AWI), Littorina (GEOMAR), and Ludwig Prandtl (HZG) were combined to form a "sensor swarm" to synoptically assess the dissolved marine and atmospheric CH4 and CO2 concentrations as well as auxiliary hydrographic parameters in the southern North Sea. Within multiple test-campaigns in the years 2019 to 2021, a meshed ship-to-ship network is developed, installed, and tested, enabling real time high-volume data exchange over up to 50km. Additionally, a broadcasting station is located on the island of Helgoland, allowing continuous internet access. Aboard of each ship in the swarm, a stand-alone decentralized (near-)real time system is established serving as data storage in case a ship is outside the swarm's network range and the Helgoland station, respectively. This across ship streaming networking system facilitates near real time data aggregation within the vessel swarm and the land station. Thus, it allows enhanced online products, such as real time Ocean Data View (ODV) maps of specific target parameters (e.g., CH4 concentrations) available on each ship in the swarm. In conjunction with a mix of centralized (messenger) and decentralized (Team Viewer, VNC) communication approaches, the fleet can be coordinated to (re-)act swarm-like and adapt the so far often rather static sampling plans towards a smart sampling approach including real time measurements of all ships in an adaptive sampling and monitoring strategy.

The German National Research Data Infrastructure NFDI: Current status and perspectives

Barbara Ebert, Frank Oliver Glöckner (AWI Bremerhaven) for NFDI4BioDiversity

In November 2018, the German Science Minister Conference GWK agreed on a 10-year initiative to structure research data services across different scientific domains in Germany. Now, two years later, the National Research Data Infrastructure NFDI has become operational, with a Directorate established in Karlsruhe and a first cohort of nine domain-oriented consortia started on October 1st, 2020. A second and third round of applications are underway, respectively planned. The final NFDI is expected to unite around 30 thematic consortia, covering all research domains in Germany. This keynote informs about the vision and the structure of the NFDI, in which more than 200 stakeholders / services providers are involved so far. Some challenges of the nascent NFDI will be highlighted, based on a recent white paper issued by the NFDI expert commission of the DFG. These challenges include a completely new review and selection process established by DFG, the goal to get good coverage across all research fields in Germany and the many cross-cutting topics which the consortia need to solve collaboratively and which have been subject to many debates. The final part of the presentation will take an international perspective, positioning NFDI within the broader picture of EOSC, GAIA-X and other major data initiatives.

webODV – online data extraction, visualization and analysis

Sebastian Mieruch, Reiner Schlitzer

webODV is based on the Ocean Data View (ODV; https://odv.awi.de) software, which is widely used for the analysis, exploration and visualization of oceanographic and other environmental data. webODV comprises on the server backend the fully functional ODV software additionally equipped with a WebSocket server. On the frontend we have state-of-the-art responsible, interactive browser interfaces. Via the WebSocket technology, which provides a fast bi-directional connection between server (ODV) and client, the browser interface communicates with ODV. webODV is deployed operationally until now in three different projects. It is used in the SDC Virtual Research Environment (VRE; https://vre.seadatanet.org) for data import, extraction, quality control and visualization. For the EMODnet Chemistry project it is deployed at https://emodnet-chemistry.webodv.awi.de and for the GEOTRACES project at https://geotraces.webodv.awi.de, where we have up to now more than 1000 registered users and daily downloads. The webODV suite provides three different services out of the box. We have the data extraction, which offers a simple, intuitive, responsive and fast browser interface for the data selection and extraction without requiring any knowledge about the underlying ODV software. Similar we provide the easy to use import service without the need of prior ODV knowledge. An additional application, which we call ODV-online is provided. Here we mimic the original ODV interface in the browser with approximately covering 99 % of the original features, including the interactive generation of publication ready maps and plots. ODV-online is perfectly suited for the large global ODV community, which is already familiar with the desktop standalone version. Detailed documentation is available at https://odv.awi.de.

Unsupervised Outlier Detection in Large Underwater Image Dataset using Variational Autoencoders

Benson Mbani, Jens Greinert, Timm Schoening, Reinhard Koch

Modern ocean floor observation approaches use towed camera system to record thousands of images in a single deployment. When imaging the seafloor along a transect that e.g. is mostly covered with dark manganese nodules, a sudden rock outcrop or a camera flash malfunction may result in few images that are dark, bluish or, in general, appear significantly different from most other images. We can identify these outlier images, prior to image classification, by fitting a probability distribution over the observations, and identifying the subset of observations that have low density. Because we assume that all observations were sampled from this probability distribution, having a low density implies that such kind of observation is highly unlikely and can be easily flagged. In this work, we aim to classify seafloor images into semantically meaningful categories, and also detect benthic fauna within each image. Before the classification, however, we need to find outlier images and exclude them or account for them in subsequent analysis. We detect the outlier images by assuming all the images were generated by sampling from a probability distribution. In order to model this distribution, we use a variational autoencoder, which uses a convolutional neural network, the encoder, to map an image into parameters of a probability distribution. By sampling from this distribution, and feeding the resulting vector to another network, the decoder, we can reconstruct the original image. Training involves minimizing reconstruction error and constraining the learned distribution to follow a standard normal distribution. We then evaluate the density of each image under this distribution and filter those with low density. On evaluation against an image sequence dataset, with known outliers, we show that this approach efficiently detects outlier images in a completely unsupervised way. After eliminating outliers, we will train a classifier to assign each image to a semantic category.

The Rust programming language for scientific software development

Valentin Buck, Flemming Stäbler, Everardo González Ávalos, Jens Greinert

Originally developed to aid the development of the web browser Firefox, the Rust programming language has been adopted by many software industry giants, such as Amazon, Google, Cloudflare and others. Reasons for adoption range from harnessing the performance increase as compared to interpreted languages such as python or JavaScript, security benefits due its unique memory model to harnessing the modern language environment with dependency management and test integrations. But do these factors also apply to the scientific software development process?

Over the last 10 months we have developed a geospatial data visualization package with a server component written in Rust and have made positive as well as negative experiences with the language.

Here we want to show how Rusts unique type system and memory management features have helped us to avoid common mistakes in software development, such as null references or uninitialized memory. We further demonstrate how both cooperative and multithreaded coprocessing are not only simple and safe, but also somewhat of a default choice in Rust applications. A further component of the presentation is a look at the state of the Rust package ecosystem, especially in the context of scientific programming. We also tell of our experience onboarding a colleague with no prior development experience in this language and conclude by giving an outlook both on the future of Rust as a programming language as well as our project.

The Digital Earth Viewer: An experiment in bringing a new perspective to geospatial data visualization

Valentin Buck, Flemming Stäbler, Everardo González Ávalos, Jens Greinert

Over the last 10 months, we have developed a new program to display geospatial data over time. Choosing a new approach, we display the earth in true 3D and treat time and time ranges as true dimensions. This allows us to display data in a way that has only rarely been used before. Developed as a hybrid application, it may be used both in-situ in a local installation to explore and contextualize new data, as well as in a hosted context to present curated data to a wider audience.

In this presentation, we want to present this software to the community, show its strengths and weaknesses, give insight into the development process and talk about extending and adapting the software to custom use cases.