August 2019

Newsletter of the *Digital Earth* Project

Contributions of GEOMAR Helmholtz Centre for Ocean Research to Digital Earth

This newsletter presents some specific efforts of GEOMAR Helmholtz Centre for Ocean Research Kiel in activities related to the Show Cases or Work Packages of Digital Earth.

Digital Earth – A Helmholtz project that connects

Daniela Henkel
GEOMAR Helmholtz Centre for Ocean Research Kiel

Digital Earth or "*Towards SMART Monitoring and Integrated Data Exploration of the Earth System - Implementing the Data Science Paradigm*" is running since June 2018 bringing together natural and data scientists from all eight Helmholtz Centres of the Helmholtz Research Division Earth and Environment.
Digital Earth is a Data Science project that aims to enable an iteratively integration of data acquisition (SMART Monitoring) and data analyses (Data Exploration) by explicitly fostering feedback loops between these two topics. Digital Earth uses and advances Data Science methods and workflows across scientific disciplines, centers, and Earth compartments. Thus it fosters large scale knowledge exchange and transfer, develops software and workflows and intends to define *defacto* 'standards' & promotes to implement best practice in Earth Science.

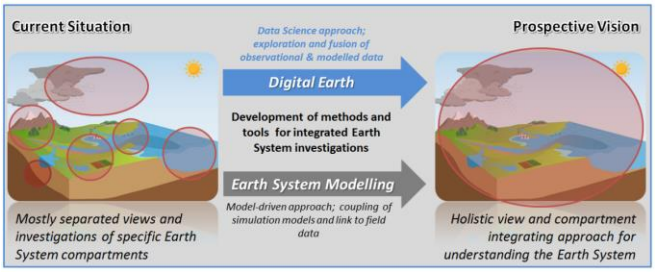

Figure 1 General Concept of Digital Earth to enable a holistic view on the Earth System as a whole.

In order to really live Data Science supported Earth System studies and to achieve the set goals, Digital Earth uses different strategic tools such as Bridging Postdocs with the aim to ensure a flexible and targeted support of emerging scientific questions by using expertise from different Helmholtz Centres. In this context six candidates, together with their individual Digital Earth related projects have been selected in May 2019. The individual PostDocs are employed at one Helmholtz Centre but closely cooperate with at least one other Helmholtz Centre.
Furthermore, Digital Earth financially supports **S**hort **T**erm **S**cientific **M**issions (STSMs) to increase the incentive for visiting other Helmholtz Centres, Institutes, universities or laboratories fostering collaborations, learning innovative techniques, and acquiring new datasets.
Digital Earth is coordinated by the GEOMAR Helmholtz Centre for Ocean Research Kiel and is funded by the Helmholtz Association with five million euros over three years.

Detecting levees with remote sensing data

Patrick Michaelis
GEOMAR Helmholtz Centre for Ocean Research Kiel

Deep learning approaches have revolutionized the field of image recognition in the past few years. Due to the similarity of data formats, the same approaches also work on remote sensing data. We therefore chose this model class to detect levees in remote sensing data as part of the Show Case 'Flood' in Digital Earth. Levees along rivers are essential for protecting our civilisation from flooding and despite one would think that their existence, location and state is well known and digitally available, this is not the case for entire Germany To build a proof-of-concept workflow and train a model for detecting levees, we used publicly available data from North Rhine-Westphalia (https://www.opengeodata.nrw.de/produkte/). The data contains aerial images and a LIDAR (**LI**ght **D**etection **A**nd **R**anging) -based **D**igital **E**levation **M**odel (DEM). Shapefiles with lines for levees are available for North Rhine-Westphalia as well.

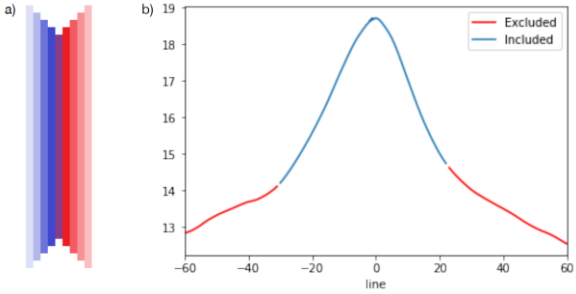

Figure 2 a) The original line segment is in purple, the additional parallel lines are added in red and blue; b) Average heights of the individual lines with the original line segment as line 0.

To detect levees we chose deep neural networks for semantic segmentation. In semantic segmentation every pixel of an image (aerial image and the DEM) is classified as the member of one of predefined class of a set of classes. We therefore had to create masks covering the entire levees from the given lines to use the deep learning algorithms properly.
Our approach begins with the individual line segments. For each segment we add parallel lines to both sides covering one line of pixels each (Figure 2 a)). Next we calculate the average height of each line using the DEM as visualized in Figure 2 b). We then use the typical shape of a levee to determine the endpoints of the levee (basically where it stops being steep) and include only the lines between the endpoints in the mask.
As the neural network topology we chose a U-Net. This class of neural networks works well with small training datasets and delivered good results on various semantic segmentation datasets. Its main building blocks are convolutional and deconvolutional layers.

The input image is transformed to smaller and smaller sizes by the convolutional layers and then expanded by deconvolutional layers. The i-th and (n-i)-th layer are corresponding to each other in their parameterization and the latter one (the deconvolutional layer) also includes data from the former one (the convolutional layer).

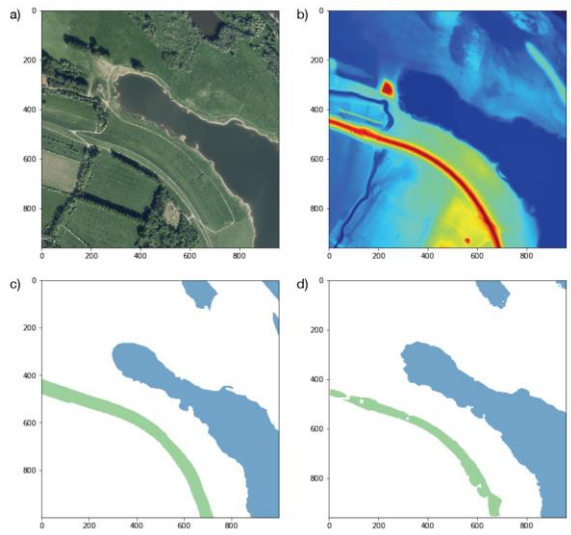

Figure 3 a) Aerial image; b) DEM; c) Ground truth; d) Prediction after post-processing.

These individual pixels are not handled by themselves but in combination with their neighbours. To train the network we use a training dataset with 1 meter resolution and augment the data with rotations and mirroring. We added masks for bodies of water as well.
To get lines from the mask outputs we again use of the typical shape of a levee as a tool. We choose the highest points (pixels) on each mask (typically along the levee crest there are many pixels with nearly the same height). Next we order these pixels using a **P**rincipal **C**omponent **A**nalysis PCA and create a rather complex line with many short line segments. To simplify the line-shape we use standard tools in Python (http://www.python.org).
While the transformations of the line shapes are unique to this application, the general approach can be used in a wide range of situations. Small experiments with aerial images indicate that a classification as body of water is possible with this data alone. The approach might therefore be applicable to measure the extent of floods from aerial or satellite images.

Active learning for marine image data
Patrick Michaelis[1], Everardo González Ávalos[1]
1 GEOMAR Helmholtz Centre for Ocean Research Kiel

Labelling visual data is still often done manually and can be a very time consuming task for large datasets. A number of tools exist to support this task (e.g. BIIGLE; https://annotate.geomar.de/) but the user still has to do the actual work of manual annotation. Massive labelling/annotating of similar features in images is also a typical task for computer vision where deep learning has become an important tool. If an algorithm could support or automate the work, it would speed up the annotation process considerably. We therefore currently develop workflows and tools to include deep learning models in our image annotation tools. The envisioned workflow starts with the user labelling images. Having a number of sample images, the network is trained on this data. The trained network then predicts the labels for some additional images. The user evaluates - and possibly corrects- the labels and the training starts again on the now larger training sample. This process is then repeated.
All this should work without the need to consult an IT and machine learning expert. The users (natural scientists) should be able to set up the tool on their own and label the images however they need it. The idea is to offer a set of pre-selected deep neural network architectures for different tasks to enable the users to use the tool without the need to create their own network architecture. Well performing benchmark architectures exist for many tasks and the literature offers a lot of examples where these architectures are used on different datasets.
There are however a few difficulties. One major issue when training a neural network is that a neural network needs a large number of training samples to learn good representations and reach useful accuracies. Another issue is the time it takes to train a neural network. A solution for both these issues is transfer learning. In transfer learning the training is not starting from zero. Instead the parameters of a network with a similar topology that was trained on a related task are used. This makes the training process faster and the number of additional training samples needed to get good results is lower as well.
We plan on using this tool for marine image data. It is however straightforward to use it for other datasets.

Scalable 4D Data Visualisation Interface for Earth Sciences Observations
Everardo González Ávalos[1], Jens Greinert[1]
GEOMAR Helmholtz Centre for Ocean Research Kiel

A comprehensive study of Earth's environment requires understanding of data acquired with a myriad of different sensors or produced by various and different models. While every observation is intrinsically bound to a spatio-temporal context, their 4D representation varies depending on their different shapes: in ocean sciences, sediment probes and water samples are examples of single point measurements in time and space. Time series from underwater moorings have a constant position but vary over time while bathymetric maps measure spatial variation of terrain but can be assumed to remain constant over a certain time (not on geological time-scales). Finally, CTD profiles and AUV sensor readings are examples of observations variable in time as well as in space. In addition to this, measurements can be higher dimensional vectors, as is the case with currents (which consist of magnitude and direction components) or a bio/geochemical analysis of the change over time of multiple chemical substances.

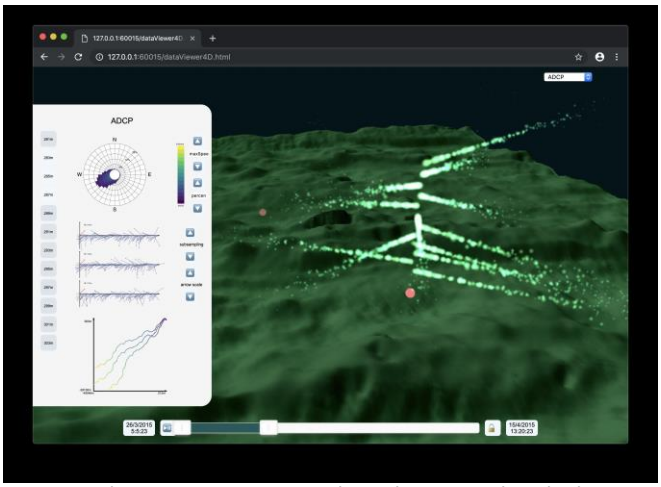

Figure 4 The 4D Data Viewer web application with a bathymetric map and a three-dimensional conceptualisation of ADCP data. Corresponding 2D graphs show on the heads up display (HUD) panel on the left. This is an example of visualisation of multiple three-dimensional vectors changing over time.

Depending on its nature, tools that aid with the visualisation of said data may not be readily available for a desired platform or may come in the form of 'outdated' software that complicates proper interpretation. Even assuming all corresponding visualisation tools as given, the fragmented representation in multiple application windows deters a holistic approach. The *4D Data Viewer* is a component-wise, scalable, and web-based framework for simultaneous visualisation of multiple data sources that helps contextualise mixed observation and simulation data in time and space.
Implemented as client side javascript application, it combines WebGL (via Three.js) to display 3D environments and HTML elements for the graphical user interface (implemented using the *VUE* framework). This modular architecture allows for implementation of new sensor classes and for reusing single HTML elements and javascript functions in stand alone applications.

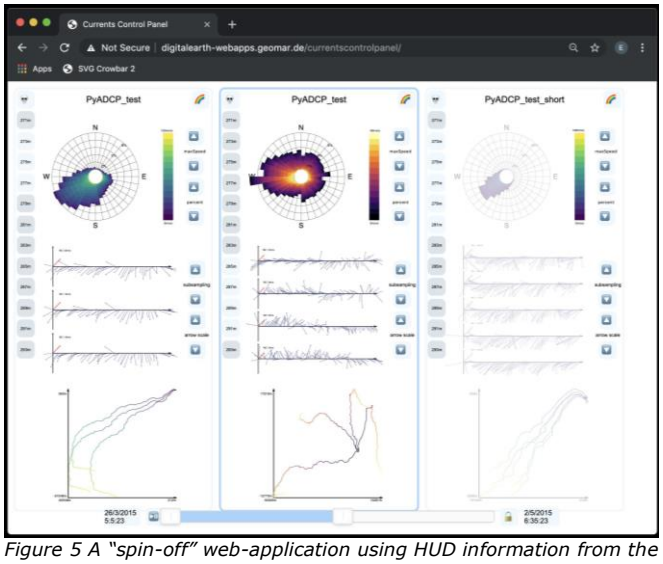

Figure 5 A "spin-off" web-application using HUD information from the ADCP sensor class.

The architecture of this software features a simple component scalability scheme to empower multi-party open- source development in a modular fashion which allows for bidirectional reusability. Lastly, the implementation as a web-application provides cross-platform portability and a familiar set of tools for development.
A following step taken jointly with our colleagues at KIT, FZJ and UFZ, will be to explore the fusion and visualisation of these datatypes using the open source software ParaView.

A sea-going high-performance compute cluster for image analysis
Timm Schoening
GEOMAR Helmholtz Center for Ocean Research Kiel

Marine image analysis faces a multitude of challenges: data set size easily reaches Terabyte-scale; the underwater visual signal often is impaired to the point where information content becomes negligible; interpreters have limited time and can only focus on subsets of the available data due to the annotation effort involved (see above).
Solutions to speed-up the analysis process have been presented in the form of semi-automation with artificial intelligence methods like machine learning. But the algorithms employed to automate the analysis commonly rely on large-scale compute infrastructure. So far, such an infrastructure has only been available on-shore.

Hence, a mobile compute cluster has been developed to bring big image data analysis capabilities out to sea (Figure 1). The **S**ea-going **H**igh-**P**erformance **C**ompute **C**luster (SHiPCC) units are mobile, robustly designed to operate with impure ship-based power supplies and based on off-the-shelf computer hardware. Each unit comprises of up to eight compute nodes with graphics processing units for efficient image analysis and an internal storage to manage the big image data sets. The SHiPCC units are envisioned to generally improve the relevance and importance of optical imagery for the marine sciences.

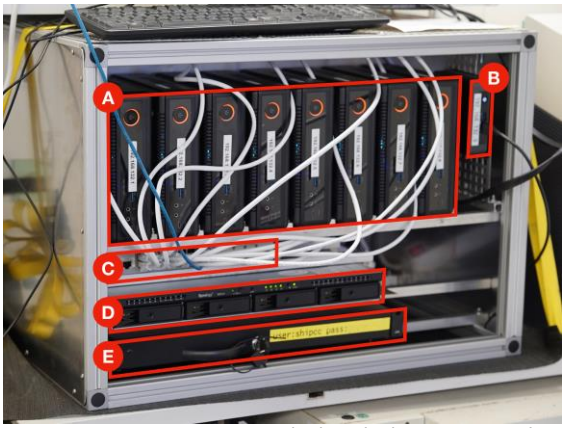

Figure 6 One SHiPCC unit as deployed during research cruise POS526 in July 2018. One unit comprises of up to eight compute nodes accelerated with graphics-processing units (A), a control computer that schedules compute tasks to the nodes and provides centralized software to the ship's network (image annotation software, GIS server, etc.) (B). All nodes are connected through a Gigabit Ethernet switch (C). Data is stored on a Network Attached Storage (NAS) (D). Further equipment (manuals, USB SSDs etc.) can be stored in a drawer (E).

BridgingPostDoc project "Uncertainty quantification of automated machine learning strategies to interpret marine data"
Amir Haroon[1], Hendrik Paasche[2], Sebastian Graber[1], Marion Jegen[1]
1 GEOMAR Helmholtz Centre for Ocean Research Kiel
2 UFZ Helmholtz Centre for Environmental Research

A holistic understanding of the earth and the processes that dictate the interaction of humans with their environment requires a thorough analysis of multivariate data that cross existing Earth science disciplines and earth compartments.
In order to cope with the mass of data that environmental scientists are confronted with, methodologies from Data science offer a promising toolbox that guarantee coherent workflows and processing chains. Yet, a successful integration of these methodologies demands tailored procedures that meet the specific requirements of environmental research.

One of these issues is related to uncertainty quantification of the interpretation based on quantifiable uncertainty, e.g. random (gaussian) noise, unquantifiable uncertainty, e.g. systematic measurement errors, or uncertainties associated with the applied learning kernel of the Data science workflow.

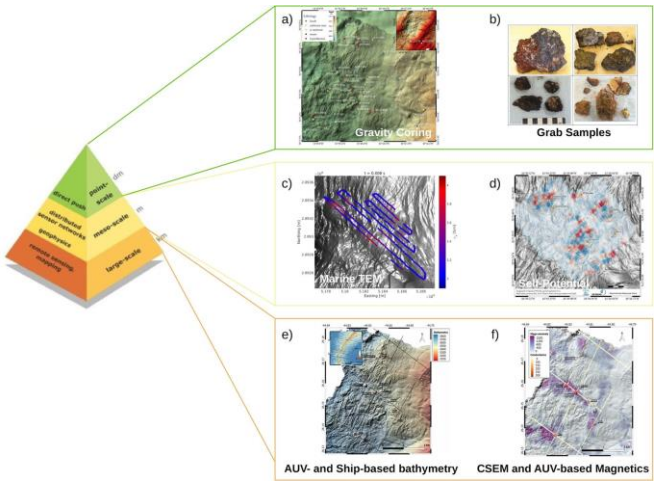

Figure 7 Example of multi-variate marine data at various spatial scales taken from a marine mineral experiment at the TAG hydrothermal field. A framework is needed to effectively integrate all information across various spatial scales with a robust uncertainty prediction. Data made available by S. Graber and S. Petersen.

Together with the Helmholtz Centre for Environmental Research - UFZ we aim to develop a framework for robust uncertainty quantification using multivariate marine data. The research will focus on a reoccurring issue in environmental science, namely the extrapolation of sparsely measured data onto a spatial scale and the inherited uncertainty that arises by doing so. The workflows are meant to benefit all marine research disciplines and will support future experimental design and data analysis procedures.