July 2020
#2

Joined Newsletter of the *Digital Earth* Project

Contributions from Digital Earth partners presented at the 2nd Annual Meeting

This newsletter shows a selection of results presented in the framework of the *Digital Earth* 2nd Annual Meeting that took place from 26 to 28 May 2020.

Data flow - from services to modules

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

In the context of data flow, sources of data are arranged into types that reflect the purpose, level of integration, and governance of each source. For instance, registered data of a specific method or of uniform measurements, fulfils the purpose of publication, but are not necessarily integrated in a data flow system. A repository provides the first and high level of integration that strongly depends on the standardization of incoming data. Collections gather data from multiple repositories and aggregates metadata to offer harmonized data in a broader scope of institutions, though with less integration than that of a repository. Federations are similar to collections, however in a federated database, control and maintenance remains with the data providers. Starting at the repository level, an integrative set of solutions is required for interoperability among data services and to the end-application. Applications within the showcases of Digital Earth are connected to repositories and federated databases. These applications are built for data flow from services that implement different solutions, including standards of the Open Geospatial Consortium (OGC), multiple protocols and specific service configurations. While at the application side, synergy demands requesting data from different sources, the service side faces the challenge to provide a comprehensive set of functionalities to the applications. One example is the framework Observation to Archive and Analysis (O2A) that is operational and continuously developed at the Alfred-Wegener-Institute, Bremerhaven. For interoperability with other repositories, the O2A uses OGC standards and a representational state transfer (REST) architecture, where all data and operations are openly available. A repository is one of the components of this framework and much of its functionality (e.g. near real-time monitoring) depends on the standardization of the incoming data. Within O2A, a modular approach has been developed to provide the data standardization at ingest and the quality control for monitoring of the ingested data. Two modules are under development to sequentially perform the data standardization and the quality control. First, the driver module executes generic transformation into a standardized format. Second, the quality control module do automatically request the sensor metadata and runs the quality tests on the ingested data. In that concept, the sensor operator and the data scientist interact with both ends of the ingest component within the O2A framework (http://data.awi.de/o2a-doc). The result is the harmonized data of multiple sources that can be accessed via the data web service (https://dashboard.awi.de/data-xxl/). Current quality control is based on tests with reviewed formulation and the result is given by an ordinal flag. A new flagging scheme is under construction to allow for monitoring individual processing steps and for using a quality score. That subsequently leads to a dual approach, where quality can be mapped among different services and be transferred throughout the data flow. In the future, these modules can be used in two contexts; first, for construction of the data repository, and second, for data and quality harmonization at the end-application.

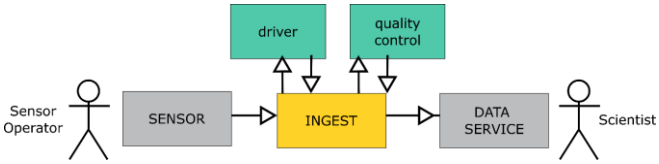

Figure 1: Concept for modules "driver" and "quality control" for construction and harmonization steps in the data flow framework that includes "SENSOR", "INGEST", and "DATA SERVICE" components.

Heat Waves and Myocardial Infarctions in Augsburg

Lennart Marien
GERICS, Helmholtz-Zentrum Geesthacht

At the Digital Earth Annual Meeting we presented the current state of the Bridging Postdoc project "*Machine Learning methods for assessing causal links in heterogeneous data: applied to Climate Change and Health*". This initiative joins researchers from GERICS/HZG and HMGU in applying Machine Learning (ML) and Artificial Intelligence (AI) methods to model the relationship between extreme temperatures and myocardial infarctions (MI) in the Augsburg region, Germany shown in Figure 2. MI is one of the most common causes of death worldwide.
Epidemiological research indicates that high temperatures may contribute to the development of MI. At the same time, heat waves are expected to increase in terms of frequency, intensity and duration in the future due to climate change. The project aims to build data-driven models to estimate the risk of suffering heat-related MI and to project that risk into a future under climate change.

Figure 2: Study region, Data: Copernicus Open Access Hub, © GERICS/Marien

In the first phase of the project leading up to the annual meeting the focus has been on building a diverse database of health, socio-economic, environmental and climatic data as input for the ML and AI algorithms. Compounding factors for MI are manifold and need to be included to allow the distinction between heat-related and other risk factors such as, e.g., air pollution, smoking or age. This database currently includes weather observations, air quality data, vegetation and green spaces, socio-demographic data as well as climate change projections and will be continually expanded throughout the project.
At the heart of the project is data from the KORA ("Kooperative Gesundheitsforschung in der Region Augsburg") project that provides representative cohort data and a registry of roughly 30.000 cases of MI in the Augsburg region since 1985. This provides us with the ground truth and allows the investigation of relationships between different risk factors and health outcomes.
A key challenge has been dealing with the heterogeneous input data. The raw data features differences in file formats, spatio-temporal resolutions and data representations such as station networks, gridded data or aggregated statistics. To make these available as input to ML algorithms we have built a workflow for streamlining the data in a consistent manner illustrated in Figure 3.
The raw data is read, cleaned to account for missing or spurious data, and then processed into time series compatible with the KORA MI registry data. Depending on the source data this can include different steps such as Kriging of station data, temporal interpolation and aggregation over the study region. The result are consistent time series that can readily be used as input to ML algorithms.
In the next step we will focus on the development of ML/AI approaches such as clustering, density estimation techniques and CART methods before moving on to more advanced approaches.

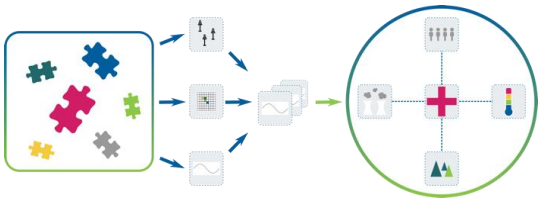

Figure 3: Schematic of the data workflow, © GERICS/Marien

Socio Economic Impacts Workflow
Kai Schröter
GFZ German Research Centre for Geosciences, Potsdam

The Socio-Economic Impact workflow is a part of the Digital Earth Flood Event Explorer focusing on indicators to assess flood impacts. It is being worked out within the Bridging Postdoc Project 'Advanced data integration methods towards large-scale flood impact indicators'.
Floods impact individuals and communities and may have significant social, economic and environmental consequences. Understanding the controls of flood impacts is crucial to mitigate consequences and reduce vulnerability. Therefore, the key question of the socio-economic impact workflow is: What are useful indicators to assess flood impacts? To answer this question, the interactions of complex flood generation and impact processes across the boundaries of 'climate and atmosphere', 'catchment and river network', and 'socio-economy' compartments are investigated and flood impact indicators are devised.
The approach integrates data from the different compartments in a new flood data set, which gives a comprehensive view to flood controls, flood impacts and their interrelationships (Figure 4)

Figure 4: General Concept of a flood system with different compartments, processes and data sources.

Flood controls, or driving factors, are variables that influence the generation and the characteristics of floods, for instance, precipitation, snow cover and soil moisture. Flood impacts comprise the intensity of inundations in terms of the affected area, inundation depth, as well as adverse consequences. The region of interest is the Elbe catchment in Germany. The period considered follows the availability of data and the occurrence of past floods. The data sources comprise of hindcast simulations from the hydrological model mHM, the regional climate downscaling model REMO, and the Regional Flood Model (RFM), which represents hydrological, hydraulic and damage processes. Further data sources are climate and precipitation stations, and water level and discharge gauges. Besides, inundation maps for past floods are derived by fusing data from multiple sensors including in-situ stations, remote sensing, and volunteered geographic information.

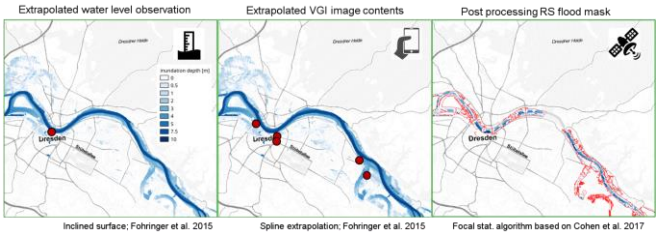

Figure 5: Using diverse data sources for the fusion of inundation depth maps, examples for the June flood 2013 in Dresden, Germany.

The aim of data fusion for inundation depth mapping is to achieve improved information, i.e. with higher quality and reliability. One example of data fusion for inundation depth mapping uses online water level observations, geo-located images posted in social media from which information about the inundation depth can be estimated given the context shown in the image, as well as flood masks derived from satellite images. Each data source has its own processing workflow and produces an inundation depth map by combining information about inundated locations or areas with topographic data, e.g. from a digital elevation model (Figure 5). The fusion of these maps combines inundation depth values in each location, and thus leverages complementary information from the different sources. Further it provides a basis to assess the confidence in the resulting data, e.g. using an agreement index or residuals between different data sources.

Inundation depth maps are one example for flood impact indicators. Within the socio-economic flood impact workflow also flood controls and impacts are explored using data science methods, e.g. clustering, classification and correlation, to derive flood indicators. The flood impact indicators represent individual or aggregated controls and allow for evaluating flood events. For current flood events, the indicators enable a quick classification of expected impacts. For future floods, they provide a means to assess future changes in flood impacts.