February 2020

# Newsletter of the *Digital Earth* Project

Contributions of Helmholtz-Zentrum für Umweltforschung GmbH - UFZ

This newsletter presents some general as well as specific UFZ efforts in activities related to the Show Cases or Work Packages of Digital Earth (DE).

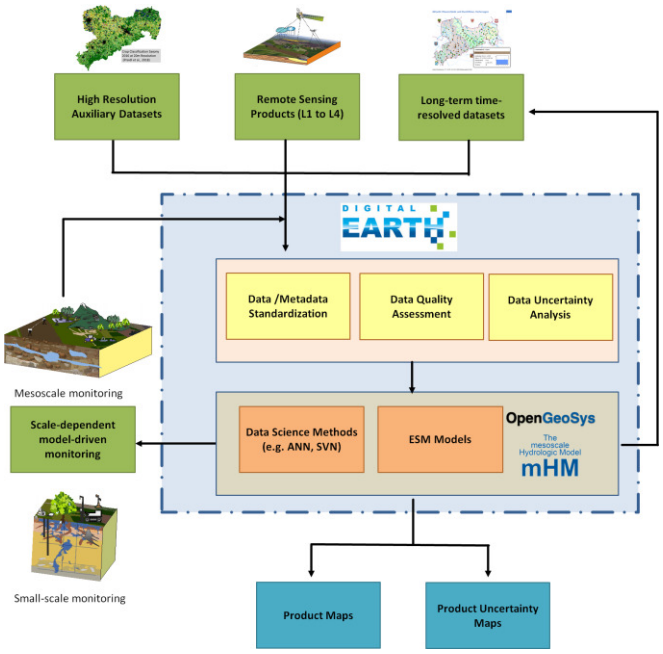


*Figure 1 UFZ activities within Digital Earth*

The work of the UFZ in DE is mainly concentrated in WP1 and WP2. In WP1, the UFZ is concerned with the (1) organization and coordination of the standardization process for the heterogeneous data flows of geoscientific observation scenarios based on a model-driven observation approach, (2) the optimization and standardization of the data flows from the sensor observation to the data archive and (3) the investigation of heterogeneous data structures in the use of cross-compartmental monitoring methods to optimize the adaptive monitoring strategy. In WP2 the tasks are the (1) visual data exploration including, e.g., the development of a visual analytics tool to identify and analyse differences between model data and auxiliary datasets (together with GFZ) and (2) the computational data exploration, e.g., hydrological data acquisition during the DE/MOSES Müglitz Campaign and post-campaign modelling of soil moisture and discharge dynamics.

**SMART Monitoring Tools**

Uta Ködel, Peter Dietrich
*UFZ Helmholtz Centre for Environmental Research*

The term SMART Monitoring has been established within the Digital Earth project defining that measured variables and their values need to be specific, measurable, accepted, relevant, and trackable (SMART) for sustainable use as data. If we look at the real data acquisition, this term can also stand for

- ***scalable*** (Hierarchical monitoring approach; the combination of methods could be cross-scale),
- ***modular*** (combination is based on a broad portfolio of methods so that the methods best suited to the problem can be used),
- ***adaptive*** (target-oriented combination of methods),
- ***robust*** (profound knowledge, e.g., uncertainties; correct selection of procedures requires an excellent expertise regarding the procedures) and
- ***transferable*** (concepts for method combination can be applied to different problems).

Thus, SMART Monitoring could be defined as monitoring with data flow from individual sensors to databases incorporating automated (machine learning) and near real-time interactive data analyses/exploration. In SMART Monitoring, the role of metadata plays a decisive role also to allow joint data analysis. It is essential to apply standard operating procedures within this flow and we derived within WP1 following task to be considered:

- standardized metadata content
- defined data or formats and standardized data format transformation
- standardized routines for Q/A routines to ensure identical data flagging; The standardized routines should correspond to existing Standard Operation Procedures (SOPs), e.g., from ICOS.
- hierarchical data storage structure to combine relevant auxiliary data, calibration protocols and data from intercomparison experiments of devices
- standardized uncertainty analysis of data and Proxy-Transfer Functions
- standardized processing routines in the application of statistical procedures or proxy transfer functions
- standardized reporting routines to ensure that all steps are precisely described and traceable and that all users can assess the data quality of the parameters derived by proxy transfer functions
- standardized visualization tools

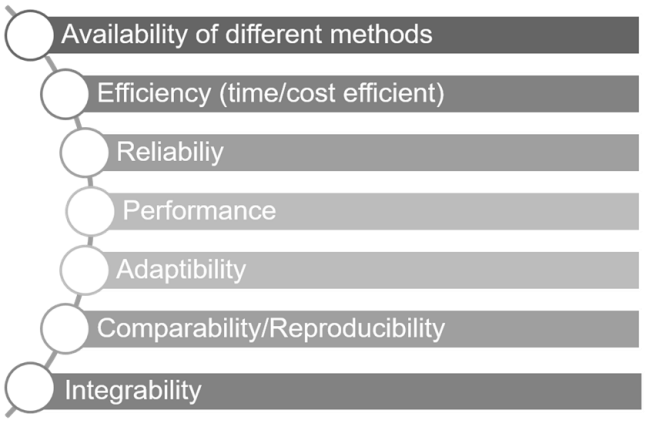


*Figure 2 SMART Monitoring Goals*

Besides, SMART Monitoring refers also to reliable and effective monitoring. Within WP1, we defined several goals of SMART Monitoring showing the significance especially in the FAIR context (Fig.2). Therefore, it is essential to use mathematical and statistical tools to identify reliable sampling points, correlation functions, or even interrelationships between model data, existent monitoring data, and other auxiliary data to derive, e.g., upscaling functions.
As one example, the UFZ team applied the Fuzzy-C Means clustering algorithm developed by H.Paasche to auxiliary data, including categorical land use data and mHM model data to identify areas of common features. Within these areas, sampling points are identified with detailed clustering or weighted conditioned Latin Hypercube Sampling.
Therefore, we want to highlight that SMART monitoring will also help to better adjust sensor settings and monitoring strategies in time and space in iterative feedback also to provide this data needed for modelling runs or validation.

**Model-driven Monitoring**
Uta Ködel, Peter Dietrich
*UFZ Helmholtz Centre for Environmental Research*

Models differ in a multitude of characteristics concerning conception, application area, spatial and temporal resolution, and their computing time as well as in the level of detail of the representation of the different processes to be represented. The choice of a suitable model can vary depending on the objective. Thus, a wide variety of models exist, such as numerical, mathematical, conceptual, or empirical models. The models differ in terms of accuracy (in the process mapping), complexity (level of detail in the process mapping), feasibility (e.g., through computing requirements), and data availability, but the enormous data availability allows the application at different scales. There are two different approaches to derive monitoring, (1) the data driven and (2) the model driven approach (Fig.3). Both approaches have their justification for existence and cannot be used alone because they are interlinked. The interface between models and monitoring lies in the variables, derived parameters or proxies. Monitoring data serve as input as well as validation variables in models and without them models are only mathematical descriptions without any relevance to real heterogeneous conditions.

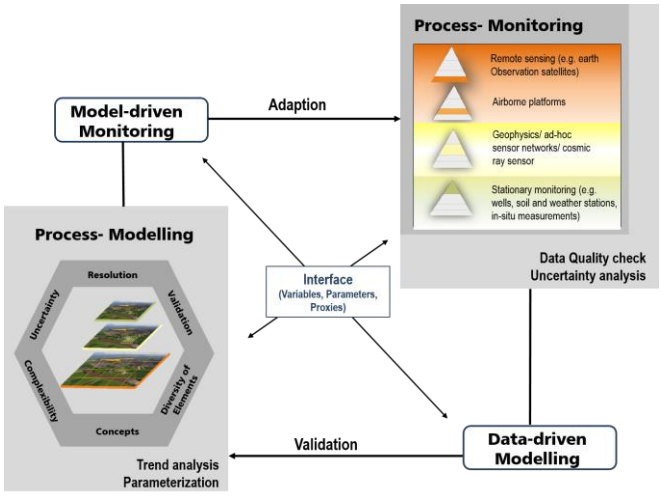


*Figure 3 Two ways of monitoring*

Model results provide temporally and spatially distributed parameters and trend analyses in a defined model grid, which can represent essential decision criteria for field measurements. For example, in models with a spatial reference, areas can be identified where no model data can be calculated or an inconsistency in the model is apparent. Similarly, modellers can identify areas where unrealistic or erroneous model data may occur and where measurements can validate or improve the model. On the other hand, a simplified process representation leads to a less realistic depiction of the process in space and/or time, and monitoring data can help to upgrade the process modelling.
One significant monitoring trade-off lies in the requested spatial coverage and resolution required from models and the feasibility of such measurements in a cost-effective manner. Indirect measurements and the use of proxies might overcome some of the problems. In model-based monitoring, it is crucial that modellers and measurers select the parameters or proxies together that can be captured and quality-checked in the requested resolution.

The following task is essential for modellers regarding model-based monitoring:

- Definition of the research question
- Definition of criteria for performance/progress measurement
- Selection of the parameters to be recorded to solve the problem (including spatial and temporal coverage)

and for field people:

- Selection of suitable (proxies and) measurement technologies
- cost/benefit and uncertainty analysis
- Selection of location/region

**Necessary steps and actions to establish an adequate standard for increased data reliability**
Uta Ködel, Peter Dietrich
*UFZ Helmholtz Centre for Environmental Research*

When it comes to joint scientific activities, standardization is often emphasized as the tool, although no one is aware of how standardization should be carried out. It is clear that standards can ensure data reproducibility, which is a key element of interinstitutional cooperation and joint data analysis. The UFZ did a survey about the requirements and workflows of establishing a standard. The resulting summary was distributed within the DE community.
A standardization process follows clearly defined steps and must be decided upon in consensus with all parties involved. There are many international, regional, and local standards organizations. The most known is the International Organization for Standardization (ISO) and the European Committee for Standardization (CEN) as a regional standards body. The problem is that many of these standards are not public domain and are protected by copyright laws and international treaties. Also, standardization organizations tend to operate at a slower pace, and many standards are not up-to-date in a fast-changing environment anymore.
For a Helmholtz wide research, Standard Operation Procedures seem to be a useful tool to improve data reliability. However, even for such a procedure, it is necessary to follow specifically defined steps: Preparation, announcement of the initiative, kick-off meeting, drafting, consensus, publication and implementation.
We want to highlight the fact that the implementation of written SOP requires a broad commitment of providers as well as users involved for a success. This can be reached by integrating the other key stakeholders outside the Helmholtz Community in the process and guarantee transparency of information, user-friendly description as well as protection of possible copyrights.
The UFZ initiated a questionnaire to gather information regarding standardization initiatives or standardization potentials within the Digital Earth project as well as MOSES measuring campaigns. The sparse participation, however, leads to the assumption that many of us, although emphasizing standardization, are not clear about the requirements for the process of compilation.

**Going beyond FAIR to increase data reliability**

Uta Koedel, Peter Dietrich
*UFZ Helmholtz Centre for Environmental Research*

The FAIR principle is on its way to becoming a conventional standard for all kinds of data. Unfortunately, it is often forgotten that this principle does not consider data quality or data reliability issues. If the data quality is not sufficiently described, a wrong interpretation and use of these data in a common interpretation can lead to false scientific conclusions. Hence, the statement about data reliability is an essential component for secondary data processing and joint interpretation efforts. Information on data reliability, uncertainty, quality as well as information on the used devices are essential and needs to be introduced or even implemented in the workflow from the sensor to a database if data is to be considered in a broader context.

In the past, many publications have shown that the same devices at the same location do not necessarily provide the same measurement data. Likewise, statistical quantities and confidence intervals are rarely given in publications in order to assess the reliability of the data.

Many secondary users of measurement data assume that calibration data and the measurement of other auxiliary variables are sufficient to estimate the data reliability. However, even if some devices require on-site field calibration, which does not mean that data are comparable. Heat, cold, internal processes on electronic components can lead to differences in measurement data recorded with devices of the same type at the same location, especially with the increasingly complex devices themselves.

Besides, many measuring devices measure variables that are not relevant for analysis and are only transferred to the required parameters by established and reported equations or proxy transfer functions (PTFs). It is essential to describe in the metadata which equation or function was used for the conversion and include an uncertainty analysis.

Within the project, the error propagation analysis of different PTFs to derive soil water content from permittivity measurements was carried out and showed that complicated PTFs do not always provide better results, especially when the other input parameter lacks a sure accuracy or also derived by other empirical functions.

The work in the project is mainly focused on the fact, which information in the metadata is needed to guarantee reliable data analysis by third parties even after a more extended period of time.

**Establishing a WebGIS project for hydrological campaign planning and data sharing at the Mueglitz River Basin**

Erik Nixdorf[1], Antonie Haas[2], Andreas Walter[2], Isabel Herrarte[2]
*1 Helmholtz Centre for Environmental Research*
*2 Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research*

Planning of event-driven-monitoring campaigns on the catchment scale requires a comprehensive overview of established measurement locations, previous campaigns, and the distribution of hydrological, geological, and geomorphological features within the study area. In Germany, the required data- and metasets are often distributed among various institutions at the local, state, and federal levels, making them difficult to access. Additionally, retrieved datasets are stored in different file formats (ASCII, MS Excel, MS Access, ESRI Shape), which do mostly neither have the same geographic coordinate system or projection nor a consistent parameter labelling. To optimize data access for campaign planning while minimizing working efforts as well as storage capacities going along with several data requests to external data providers, a centralized data handling, processing, and accessing approach is needed.

For the MOSES intensive test site Mueglitz River Basin, we implemented a WEBGIS project via the AWI GIS infrastructure maps.awi.de. The major components of this established framework are ArcGIS for Server, a PostgreSQL database including Spatial Database Engine (SDE), and desktop GIS software. The WEBGIS project was based on datasets from state authorities, previous measurement campaigns, and results of numerical models as well as environmental datasets that are freely available via online data repositories. Dataset projections, formats, and metadata descriptions have been standardized in close cooperation between the project partners following ISO standards.

The current collection of project data on maps.awi.de is not only for scientific purposes, but it is also open to the public and enables the knowledge transfer to the non-scientific community. All data are Open Geospatial Consortium (OGC) standardized Web Map Services (WMS) or Web Feature Service (WFS), which will facilitate data exchange and data visualization among the project partners. Based on the experience of the joint-work, this WebGIS project will become a part of an overview WebGIS that will be established for the entire Elbe Catchment focusing on near real-time data services (NRT) in order to support the MOSES Elbe campaign 2020, as well as the knowledge and technic transfer of the Digital Earth project.

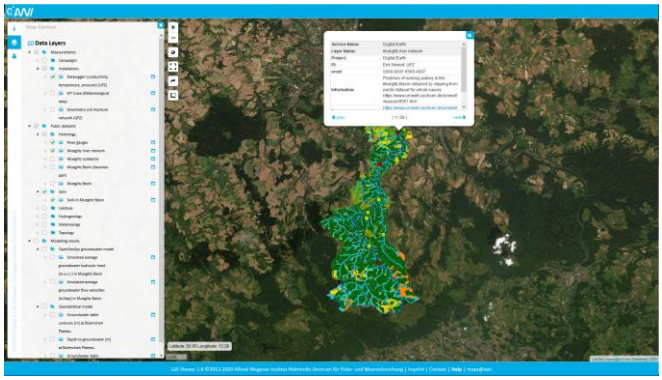

*Figure 4 WebGIS Project*

**Big (geo-social) Data analytics for environmental monitoring: Exploring the potential of microblogs to spatiotemporal characterize floods, droughts, and typhoons in China**

Ching Yin Kwok, Erik Nixdorf, Thomas Kalbacher
*UFZ Helmholtz Centre for Environmental Research*

The monitoring of hydrological events is continuously changing towards a "big data" problem since a growing number of remote sensing information is applied to support traditional hydrological networks. Alternatively, environmental data generated at low-response times within the highly versatile and high-volume data of social media platforms can be applied, particularly if authoritative data are scarce and remote sensing sources are insufficient. Focussing on China, we develop an automatized scheme to retrieve and process microblogs from Sina Weibo. Our approach consists of an API independent web-scraper to retrieve massive amounts of microblogs, a data cleaning and filtering module, a georeferencing module, and a supervised machine learning approach to classify content reliability. The workflow was tested for microblogs related to the topics of floods, droughts, and typhoons during 2018 and 2019. In addition, different sources of satellite data were used to validate the extracted social media data.

For typhoon-related microblogs, the spatial-temporal distribution of microblogs resembled the actual pathway of typhoons, making them a potential tool for trajectory tracking of meteorological events. For droughts, where our study is one of the first investigations of this topic on Sina Weibo,

around three-quarters of the drought-related microblogs are located in areas being classified as dry by the SPE Index. In contrast, limitations of our methodology appeared by analysing floods as the distribution of microblogs was predominantly dominated by population density but not by rainfall patterns and river locations.
Our results clearly emphasize the potential use of geo-social media data, derived and processed by big data analytics, as a proxy to improve environmental monitoring as well as to relocate existing hydrological networks.

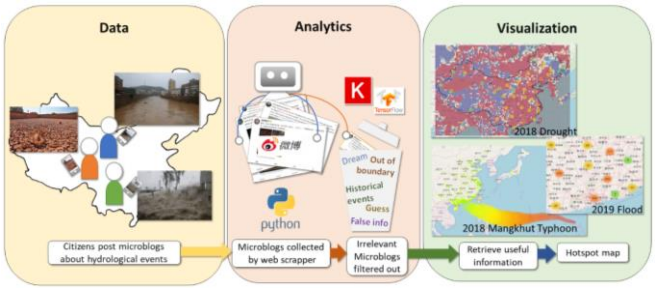

*Figure 5 Workflow*

**Setting up a distributed Hydrological model for the Müglitz River Basin**
Marco Hannemann, Erik Nixdorf, Thomas Kalbacher
*UFZ Helmholtz Centre for Environmental Research*

Understanding hydrological processes in a river basin is a crucial component for predicting catchment characteristics like run-off and soil moisture. Physical based hydrological models are a well-established tool to derive hydrological parameters, but acquiring the required input datasets on a satisfactory resolution is demanding, especially when describing catchments on the mesoscale like the Müglitz River Basin (209 km²). However, continuous progress in available datasets and computing capacity allows modelling even small catchments with challenging starting conditions like fast runoff dynamics and great varieties regarding morphological characteristics.
This study aims to investigate the influence of different sets of input data and the temporal-spatial model discretization on the resulting modelled run-off dynamics and soil moisture distribution in the Müglitz River Basin. The Modelling framework is realized by applying the mesoscale Hydrologic Model Software with a Python-based pre-and post-processing scheme in combination with batch-processing in GIS software, which allows processing high-resolution input datasets, e.g., from remote sensing and weather radar sources including selecting and downloading data from different sources, conversion of file formats, extraction of desired parameters and time scales and finally geographical processing like re-projection, clipping, and resolution resampling. Fig.6 illustrates the workflow from the process of data aggregation to model evaluation.

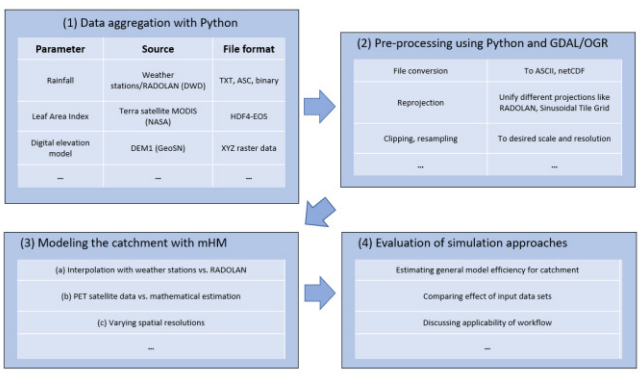

*Figure 6 Developed modelling workflow for Müglitz River Basin*

Different simulation scenarios are set up using varying sets of data input. One approach consists of setting up rainfall data, using a traditional technique by obtaining spatially interpolated data from weather stations, against the usage of the radar-based product RADOLAN by the German Weather Service. Another set-up is realized by comparing the impact of potential evapotranspiration (PET) data obtained from either MODIS sensors or inferred from temperature measurements on the hydrological simulation results. All software products developed within the project will be distributed among the scientific community using infrastructures such as GitLab and GitHub.

**Bridging Postdoc Digital Earth — a data-driven architecture for service-oriented observation methods and in-stream process modelling**
Robert Wagner
*Helmholtz-Zentrum Geesthacht, Institut für Küsten-forschung/ Helmholtz Centre for Environmental Research*

The conservation and long-term protection of our environment requires a better understanding of ecosystems through cross-domain integration of data and knowledge from different disciplines. Current methods used in applied environmental research and scientific surveys are not sufficient to address the heterogeneity and dynamics of ecosystems appropriately. To this end, an urgent need is seen in introducing new technology and methods for a service-oriented and holistic in-situ monitoring with increased spatio-temporal resolution and cutting-edge functionalities. Recent developments in the field of digital information processing, the internet of things (IoT), or the analysis of complex datasets are opening up new possibilities for data-based environmental research. These rapidly developing fields are calling for a disruptive paradigm shift towards a service-oriented earth observation (smart monitoring). To this end, future earth observation approaches will have a much stronger coupling between the modelling and the data acquisition. The development, implementation, and evaluation of such an interface are one of the overall objectives of this project. A basic data model and a distinctive hardware architecture must be defined to achieve this goal. A realistic application scenario will be used to demonstrate the advantages of developing a monitoring strategy that is no longer based on static data collection but on the coupling of modelling and empiricism using integrated sensors for advanced modelling.
Since current methods have so far failed to allow a holistic assessment of varying, large-scale environmental phenomena, there is a corresponding need for capable hardware that is specialized for precisely this purpose. The project aims to introduce an integrated sensor system for advanced modelling in earth sciences. To this end, a data-driven architecture for service-oriented observation methods and in-stream process modelling close to real-time is being developed. In addition to the hardware-related requirements of such a sensor system, the creation of an interface between the physical environments (sensor level) or abstracted model assumption (model level) is a particular focus of the research project. A sampling theorem, the predictive object-specific exposure (POSE), is introduced as an underlying measurement paradigm and data model, which allows considering not only the measured value in the evaluation but also accompanied parameters, which is called the context of measurement. The development and provision of a first adaptive sensor concept resulted in a promising prototype enabling the possibility to record environmental data depending on decision criteria such as location, time, or context. Thus, the project is representing an interesting practical contribution to Digital Earth.