Chapter 6
The Digital Earth Smart Monitoring Concept and Tools

Uta Koedel, Peter Dietrich, Philipp Fischer, Jens Greinert, Ulrich Bundke, Ewa Burwicz-Galerne, Antonie Haas, Isabel Herrarte, Amir Haroon, Marion Jegen, Thomas Kalbacher, Marcel Kennert, Tobias Korf, Ralf Kunkel, Ching Yin Kwok, Christoph Mahnke, Erik Nixdorf, Hendrik Paasche, Everardo González Ávalos, Andreas Petzold, Susanne Rohs, Robert Wagner, and Andreas Walter

Abstract Reliable data are the base of all scientific analyses, interpretations and conclusions. Evaluating data in a smart way speeds up the process of interpretation and conclusion and highlights where, when and how additionally acquired data in the field will support knowledge gain. An extended SMART monitoring concept is introduced which includes SMART sensors, DataFlows, MetaData and Sampling approaches and tools. In the course of the Digital Earth project, the meaning of SMART monitoring has significantly evolved. It stands for a combination of hard- and software tools enhancing the traditional monitoring approach where a SMART monitoring DataFlow is processed and analyzed sequentially on the way from the sensor to a repository into an integrated analysis approach. The measured values itself, its metadata, and the status of the sensor, and additional auxiliary data can be made available in real time and analyzed to enhance the sensor output concerning

U. Koedel (✉) · P. Dietrich · T. Kalbacher · C. Y. Kwok · E. Nixdorf · H. Paasche
Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany
e-mail: uta.koedel@ufz.de

P. Fischer · A. Haas · I. Herrarte · A. Walter
Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany

J. Greinert · A. Haroon · M. Jegen · E. González Ávalos
GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

U. Bundke · M. Kennert · T. Korf · R. Kunkel · C. Mahnke · A. Petzold · S. Rohs
Forschungszentrum Jülich GmbH, Jülich, Germany

R. Wagner
Leibniz Institute for Baltic Sea Research Warnemünde IOW, Rostock, Germany

E. Nixdorf
Department of Groundwater and Soil, Federal Institute for Geosciences and Natural Resources (BGR), Hannover, Germany

E. Burwicz-Galerne
University of Bremen, Bremen, Germany

L. M. Bouwer et al. (eds.), *Integrating Data Science and Earth Science*,
SpringerBriefs in Earth System Sciences,
https://doi.org/10.1007/978-3-030-99546-1_6

accuracy and precision. Although several parts of the four tools are known, technically feasible and sometimes applied in Earth science studies, there is a large discrepancy between knowledge and our derived ambitions and what is feasible and commonly done in the reality and in the field.

Keywords Adaptive · Prediction · Monitoring · Sensors · Metadata · FAIR · DataFlow · SMART concept · SMART tools

6.1 Challenges

The understanding of the Earth system with all its different habitats, processes, connections between spheres and feedback loops demands the repeated observation of high number of parameters in high spatial and temporal resolution over long time. Such kind of monitoring is e.g. the base of our understanding of climate change, changes in biodiversity or on shorter time scales e.g. the development of a sediment plume in the deep sea during deep sea mining. The base of all this knowledge gain is monitoring of specific parameters in the field, be it in the atmosphere, the oceans or on land. Conservation and long-term protection of the environment requires a better understanding of the ecosystem through cross-domain integration of data and knowledge from different disciplines. Current methods used in applied environmental research and scientific surveys are often not sufficient to appropriately address the heterogeneity and dynamic of ecosystem changes.

Thus, new technologies and methods for integrated in-situ and near real-time monitoring with increased spatiotemporal resolution and adaptive functionalities are needed. Recent developments in digital information processing, the internet of things (IoT) or the improved analysis of complex datasets are opening up new possibilities for data-based environmental research. Moreover, these rapidly developing fields call for a paradigm shift towards a SMART monitoring concept that even stronger couples modelling and data acquisition in the field. Having the none achievable goal of "measuring everything, everywhere at any time" in mind sets some challenges to the task of twenty-first-century environmental monitoring which are:

- SMART sensors: Advancing and developing sensors that have real-time data (pre)processing capacities and are linked in a self-organizing sensor network is still a challenging technological task. Automated event-detection, drift correction and failure detection are possible but still rarely done. Real-time data connections and centralized visualization and analyses are more and more established, but the real challenge is that such SMART sensors and sensor networks become easy to use and the standard way of acquiring multiparameter data in the field.
- SMART DataFlow: An easy to use, scalable and adaptable way of receiving data from sensors and re-distributing them through various channels and means also in real time is the challenge for an efficient SMART monitoring DataFlow. Standardized and largely automated procedures are needed to obtain reliable data.

As an essential part of the live cycle of data is the DataFlow crucial for acquiring high quality data at the right time and location

- SMART MetaData: Columns of numbers of a time series alone are not useful without the context these numbers have been generated. The suitable description of data is a prerequisite for any secondary use of data. Apart from FAIR descriptions, the data trustworthiness also needs to be assessed and described to allow a correct evaluation of the data. Compiling these data in a complete manner and raise the awareness again, that MetaData are crucial for the correct use of data, is the real challenge for SMART MetaData.
- SMART Sampling: Objectively finding the best possible sample location in space and time (most informative information for the respective research question), ideally in an automated and adapting way is a challenging task. SMART sampling strategies are supporting this challenge. Applying state-of-the-art statistic and AI methods jointly with interactive visualization and analyses is increasing in the community. The challenge is to spread the knowledge about these methods and present easy ways of using them to lower the hurdle of their application.

Addressing these challenges was the main objective of the SMART monitoring efforts within the Digital Earth project. The involved research centres started, iterated and further developed the idea of a SMART monitoring concept, that finally integrates four conceptual groups of tools, each tackling one of the above stated challenges.

6.2 SMART Monitoring Concept

6.2.1 An Expanded SMART Monitoring Concept

SMART monitoring typically refers to "Self-Monitoring, Analysis, and Reporting Technology" which implies that sensors utilizes e.g. artificial intelligence and big data analysis capabilities to provide an automated data acquisition, simultaneous processing, standardized storage and retrieval of multiple data (Ullo & Sinha 2020; Zhang et al. 2015; Spencer et al. 2004; Thakur et al. 2019; Alharbi & Soh 2019, Lombard et al., 2019). We would like to *expand* the meaning of SMART monitoring in such a way that measured environmental parameters and their values need to be:

- **Specific/Scalable**—Specific relates to accurate and precise and means also that something is clearly defined or identified. Scalable refers to a hierarchical monitoring approach combining multiple sensors measuring across scales and parameters.
- **Measurable/Modular**—Measurable values imply that quantitative information can be measured. Modular refers to a portfolio of different independent methods available to measure a specific parameter to eliminate specific methods' disadvantages/shortcomings.

- **Accepted/Adaptive**—Accepted values are internationally defined in UNESCO standards for the target parameters. Adaptive refers to an easy application to new research questions and the combination of various sensors in a network.
- **Relevant/Robust**—Relevant values are commonly accepted as representative of a specific measurement. Robust refers to self-repair calibration mechanisms in operation in case of sensor failure and profound knowledge of the accuracy and precision of the sensor data.
- **Trackable/Transferable**—Trackable data can be tracked by specific hardware and software tools conveying information where the data's status is at any point in time. Transferable refers to concepts of method combination applied to different problems.

To meet these SMART monitoring criteria, we suggest an iterative SMART monitoring concept with methods, approaches and tools for *SMART sensors*, *Data-Flow*, *Metadata* and *Sampling* technologies. Figure 6.1 shows this concept with its four overlapping groups of tools.

1. SMART sensor tools enable the interconnection of a large number of sensors, automated data access via standardized and well-documented interfaces, and remote adaptation of measuring schemes to the prevailing measuring conditions. They meet the SMART criteria specific, measurable, robust and trackable.

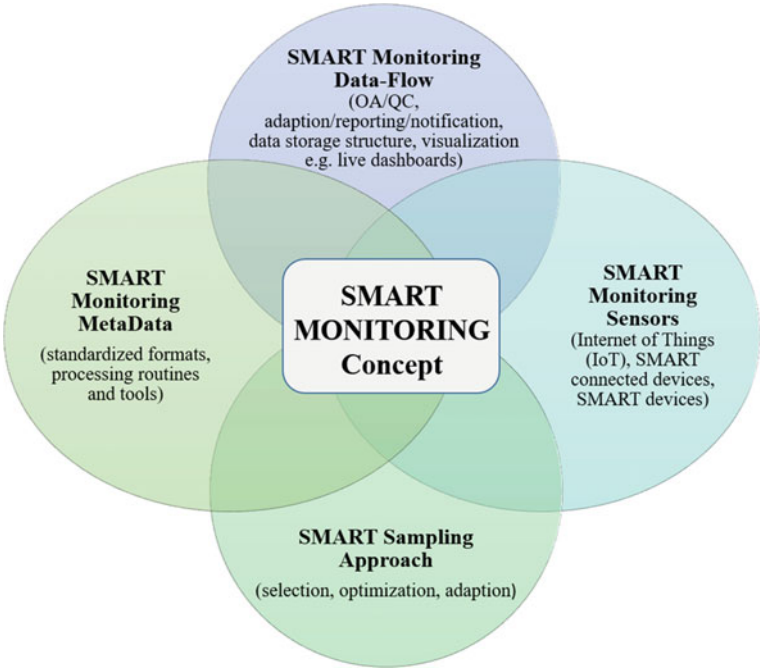

Fig. 6.1 The SMART monitoring concept consisting of four overlapping groups of tools

2. The SMART DataFlow approaches incorporate a variety of standardized data flows within the data lifecycle. Analytical tools and methods provided in automated analysis workflows allow data exploration and analysis to determine a suitable monitoring strategy; this includes methods to classify data, determine outliers, fill data gaps, recognize patterns in data and adapt data to different spatial and temporal resolutions. SMART DataFlow approaches meet the SMART criteria scalable, modular, accepted, adaptive and relevant.
3. SMART MetaData approaches enable a comprehensive description of newly acquired data increasing their reliability. Standardized data descriptions are important for data fusion, joint analysis and interpretation as well as for the creation of training data sets for machine learning. SMART MetaData approaches meet the SMART criteria specific, accepted, relevant and trackable.
4. SMART Sampling approaches. They, for example, support selecting the most representative sampling points based on auxiliary or prior measured data. They meet the SMART criteria scalable, modular, adaptive and transferable.

6.2.2 Pre-Conditions for SMART Monitoring

6.2.2.1 SMART Monitoring and Technological Advancements of Sensors

SMART monitoring of the environment incorporates SMART sensors such as the internet of things (IoT), SMART connected sensors, and smart devices, playing an essential role within SMART monitoring. Sensor technology is improving from a simple measurement sensor to a SMART sensor with local intelligence, decentralized data pre-processing, digital output, and near real-time communication options. Microsystems technology with micro- and nano-electronics low-power computing capabilities, high data volume storage and better batteries are driving this development. An ideal SMART sensor is a sensor in which the complete signal conditioning and signal processing are combined in one unit in addition to the actual measurement acquisition. Such sensors usually include a microprocessor or microcontroller and provide standardized interfaces for communication, e.g. via field bus systems or sensor networks. Thereby, such sensors' complete sophisticated task is to be fulfilled without an external computer to meet miniaturization demands, decentralization, increasing reliability, reducing costs and improved flexibility. The characteristics of a SMART sensor are (Sauerer, 2013):

- provides a digital output signal, often via a standardized interface; in stand-alone systems also via a wireless data connection
- can be addressed via an address and has a bidirectional digital interface
- data transmission and information exchange among different devices and different domains

- executes commands and logical functions (complex measurement value processing up to measurement evaluation) to allow (near) real-time decision-making, service supporting and management
- existence of extensive calibration and diagnostic capabilities
- in self-sufficient systems existence of data memory
- self-sustaining sensor systems with wireless data transmission and energy harvesting eliminate the need for cabling or battery replacement, allowing almost unlimited operating time in hard-to-access locations

Therefore, a SMART sensor is suitable to meet the increasing demands on reliable monitoring tools such as miniaturization, the realization of high data sampling rates with higher accuracy and reliability, acquisition and decentralized real-time processing of spatially distributed measurement data, sensor fusion allowing the combination of different sensor data, ease of integration, (wireless) self-sufficient networking, higher reliability and less maintenance due to self-maintenance opportunities, low power and low latency (data transmission with minimal delay), the possibility of mobile edge computing (MEC), especially for mobile crowdsensing and cost reduction.

A vital facet addresses the monitoring of the sensor function itself. A service reduced and reliable sensor operation, especially in long-term remote-controlled applications, requires modern communication procedures between the sensor and the control unit. Today, even the most straightforward I.T. equipment like printers have fully automatic reconnection, self-analysis, and, if required, also calibration procedures. Any necessary information like driver updates or serial number-related information is available in repositories. The sensor should automatically connect in case of malfunctions or even routinely checks if updates or improvements are available. Unfortunately, this is not the case in most environmental sensors, which often also do not even have the most basic plug-in connection procedures. Therefore, significant technological innovations in sensor development are needed to provide smart monitoring technologies with self-repair mechanisms if the control software fails and reliable alerting functions in the event of contact failure (Fischer 2020). We need to implement state-of-the-art I.T. technology in the field, working based on plug-and-play technology, including fully automated transmission, verification, storage, accessibility of sensor metadata and sensor actions such as deployment or maintenance so that human interaction in sensor operation can be significantly reduced.

6.2.2.2 SMART Monitoring and FAIR Principles

The availability of sensor data represents a prerequisite for parametrization and model validation. Such data fusion and integration demonstrate the importance of implementing FAIR principles for sustainable data management and allows interoperability among different data services (Wilkinson et al., 2016). These principles primarily aim to generate the maximum benefit from data and their metadata by

supporting machine-actionable data infrastructure processes making the data findable and accessible by machines and humans. However, it needs to be mentioned that the FAIR principles do not require a detailed description of data quality and do not cover content-related quality aspects. An assessment of data quality and the data provenance is essential to preclude the possibility of inaccurate, incomplete or even unsatisfactory data analysis applying and avoiding poorly derived, misleading or wrong conclusions. For our SMART monitoring approach, the following information needs to be part of any FAIR compliant data:

- available auxiliary or prior measured data
- essential information on smart sensor networks, including timing, ambient conditions and data aggregation issues
- existent, general information regarding both sensor and measurement uncertainties and calibration of sensors in a traceable way
- information on executed processing and analysis steps with their assumptions, e.g. information on quality control/assurance steps or applied Proxy–Transfer Functions to derive parameters of interest
- information on used methods and their assumptions (e.g. test and training data set) of supervised and unsupervised machine learning generated data products
- information on gridding algorithm and their specific assumptions/parameter settings especially for larger scale sensor data that have been converted into a derived data product (grid/raster; correlation)

6.2.2.3 SMART Monitoring and Standardization

As crucial part for joint scientific activities SMART monitoring Data-Flow and SMART Metadata tools will need to support scientific cooperation and data integration through standardized workflows and metadata schemes. Standardization is often emphasized as important process, although there is little awareness about how standardization should be carried out. Standards as expressions of consensus enable in general safety, allow to control processes, increase transferability and also support creativity. Standards can be established by geographical extent or reach (e.g. international and national), by scope, by strength (e.g. regulation versus recommendation) and by subject (e.g. devices, procedures and workflows).

Standards can ensure data reproducibility, a key element of interinstitutional cooperation and joint data analysis. Many scientists mention standardization as the means to make data interoperable, but many are not clear about the requirements for the respective standardization process. Such a process requires a transparent input into the corresponding workflows and daily work to those involved to achieve high acceptance and develop the best possible standard. The smaller the affiliated group, the easier it is to define standardized procedures and apply them within the group/consortium. Therefore, there is a seemingly infinite number of standards for different data workflows which makes a reliable standardization difficult. The Vienna agreement from 1991 allows cooperative standardization efforts between European

Committee for Standardization (CEN) and the International Organization for Standardization (ISO). Many standardization activities are going on in so called technical committees at CEN or ISO. Another faster option is to prepare a CEN Workshop Agreement (CWA) at the European Committee for Standardization (CEN) to reach a broad acceptance internationally (CEN, 2021). This kind of agreement must follow specifically defined steps: preparation, the initiative's announcement, kick-off meeting, drafting, consensus, publication and implementation (CEN, 2021). The advantage of such a standardization effort is that such "lower" standard is being reviewed every 3 or 5 years allowing the consideration of novel developments.

Even with the understanding that standardization is essential within SMART monitoring tools, we could not prepare and start such a CWA process within the Digital Earth project lifetime for our SMART monitoring approach. Such standardization efforts would need to include all steps from data acquisition, data processing and data storage, all described as clear and reproducible as possible. Both, standardized procedures in the data acquisition (monitoring set-up, acquisition, calibration, data cleaning,) and standardized metadata, decide on the usability and trustworthiness of monitoring data. Cooperation with existing initiatives and infrastructures such as DataFlow Framework from Sensor Observations to Archives (O2A at AWI) and Modular Observation Solutions for Earth Systems (MOSES) were intensified to improve existing tools and bundle competencies (UFZ 2021; Koppe et al. 2015; Gerchow et al. 2015).

6.2.2.4 SMART Monitoring and Data Quality

Data quality is a crucial, although not explicitly mentioned, requirement for data FAIRness; it is essential to ensure reusability of data. A documented data quality is required to enable meaningful data selection for data fusion and reuse. Due to the versatile application of low-cost sensors in environmental science, information on data quality has become increasingly important and scientists who acquire and use monitoring data must be aware of the importance of data quality and their trustworthiness. Even though data may be FAIR in terms of availability, the data are not necessarily "good" with regards to accuracy and precision. Unfortunately, there is still considerable confusion in science about what good or trustworthy data are (e.g. Dorgio et al. 2021). Trustworthy data may be achieved by simple/automated data flagging algorithms, ensuring that data are plausible with respect to specific criteria (e.g. threshold). But real trustworthy data imply more when considering accepted standards for scientific data that have an uncertainty value for each measurement. An accepted approach in providing an uncertainty range for single data points/measurements is by providing accuracy and precision as defined in ISO standard 5725–1(1994; accuracy evaluates the proximity of measurement results to an accepted reference value, precision considers repeatability or reproducibility of measurements). This approach allows for a numerical expression of how close a measured value lies with a certain statistical probability to the real value (e.g. 90%).

Even though the uncertainty assessment is commonly accepted and good scientific practice in most experimental studies and measurements, it is not yet a must-have in many monitoring approaches where single sensor data or a time series is often provided without assumptions of the associated uncertainty. Such assessments are of significant importance when classifying data as relevant or good for a specific application or scientific question. Discriminating, e.g. different water masses based on temperature and salinity requires high accuracy and precision. Estimating the presence or absence of a specific fish species based on the concept of fundamental and realized ecological niches using the same two parameters, temperature and salinity, allows for a much larger accuracy and precision range of the data. Therefore, the same data cannot be used by e.g. an oceanographer but by the behavioural ecologist. Each scientist must be enabled to decide if data are good or have to be rejected as probably unfit for a specific scientific question; without an assumption of the data uncertainty, this is practically impossible.

Modern data science methods such as machine learning can blend diverse datasets even with lower quality to gain valuable information. However, to assess and interpret such results, knowledge of data quality is required. Traditional repositories hosting data from scientific or regulatory monitoring initiatives as well as from scientific field campaigns could usually rely on more or less rigid quality assurance chains. According to the international organization ASQ, quality assurance can be defined as "part of quality management focused on providing confidence that quality requirements will be fulfilled." Quality control as the "part of quality management focused on fulfilling quality requirements" is essentially the inspection component of quality assurance (ASQ, 2021). Quality control (QC) is distributed across data acquisition, data management and data curation tasks and should be discussed as such and jointly executed by the involved people scientist check the validity of the data values themselves whereas data manager and data curators possibly focus more on metadata quality (completeness, standardized terminology etc.). Similar quality control issues occur in different monitoring domains, scientific disciplines and involve many different data types. Recently, many projects and initiatives have begun to harmonize data quality control efforts (e.g. ENVRI-FAIR and NFDI) and develop software tools to assist in quality control across various environmental research domains (Schultz et al., 2019; see also Sect. 6.4.1 for examples). Such initiatives may only affect the data processing steps but have also considerable effects on entire monitoring set-ups.

6.2.3 Future Tasks to Further Increase Smart Monitoring Efforts

The development of new and faster machine learning and generally AI tools will undoubtedly bring new possibilities for advancing tools of SMART monitoring. Despite the integration of these new methods we see a number of essential tasks as important for future applications of SMART monitoring (Table 6.1). These must go

Table 6.1 Future needs and challenges for SMART Monitoring Tools

SMART Sensors	• fully automated transmission, verification, storage, accessibility of sensor data and their metadata • automated sensor maintenance and set-up • powerful synchronization and data management system for wireless sensor networks (real-time clock; data versioning, …)
SMART DataFlow	• standardized approaches, e.g. documentation/reporting, QA/QC, data gap filling, statistical analysis, quality assessment and visualization • quality assurance routines and metadata description for supervised and unsupervised machine learning algorithm to achieve reproducibility of the results • near real-time visualization with incorporated analysis tools
SMART MetaData	• agreement on a shared vocabulary for the metadata elements and values as an interdisciplinary approach • automatic filling of metadata (e.g. electronic and paper documents, software) and • automatic error checking to reduce human errors automating metadata management
SMART Sampling	• automatic grabbing of all available web data for a selected area with various resolution to save and visualize this data in a standardized format • automated bundling of all available data (former monitoring data, satellite data, auxiliary data) from a selected area and • determine the representative sampling area or points accordingly

hand in hand with additional IT security when distributed dataFlows and IoT sensors and data repositories become the new standard for monitoring the Earth environment.

6.3 SMART Monitoring approaches and tools

In Digital Earth, we developed several approaches and tools to enable the expanded SMART monitoring concept we set up in Digital Earth.

6.3.1 Hard-and Software Tools for a Modern Communication between Sensor and Control to Enhance Traditional Monitoring Efforts

Environmental research is changing towards using monitoring strategies that are no longer based on static data collection, but on the coupling of prediction and empirical data that integrate sensors near real-time data stream for continuous modelling. The challenge, that in practice, requires a sophisticated implementation of a decision and control basis at sensor level. While sensors used to be rather one-dimensional and stupid data suppliers, nowadays complex sensor systems where several sensors are

interconnected are being used increasingly. An important aspect in this respect is the implementation of a data model according to a holistic data processing e.g. the Lambda architecture (e.g. Kiran et al., 2015). In this way, sensors do not just deliver measured values but entire message packages (e.g. JSON) via defined protocols (e.g. MQTT) and interfaces (e.g. HTTP). In addition to the measured value, these message packages contain various descriptive metadata about the sensor and the situation of the measurement.

This procedure allows complex algorithms and analyses to be applied directly to the data stream based on the message packets. Thus, the approach in applied environmental monitoring shifts from pure data collection to adaptive measurement in the context of an application or a specific scenario. The measured sensor value is embedded in a context and thus receives a clear space–time reference as well as a context-related allocation. The fusion even of heterogeneous data streams is thus considerably simplified, since connecting descriptive parameters are available within the metadata, which allow linking different message packets.

The main innovation of the process flow is that data collected in the course of monitoring can be directly related to a-priori information. It is irrelevant whether the context is based on modelling or accompanying measurements. Since the infrastructure and the underlying data model represent an always existing and complete solution space, the monitoring efforts can be constantly optimized, similar to a machine learning approach, because the set of rules for data sampling/sensor measurement (sampling interval, additional sensors on/off, …) are subject to constant proving. This approach allows a broader or rather holistic assessment of varying, large-scale environmental phenomena. To do so, there is a corresponding need for capable hardware and software tools that are specialized to execute such an assessment in a tailored way.

Figure 6.2 gives an illustration of a data stream architecture with real-time data processing and IoT-capable sensors. Starting with the sensor technology, sensors must not only transfer the results of the measurement conversion (e.g. calculating turbidity from a voltage signal of a turbidity sensor based on light backscatter) but obviously also needs to provide information about the context (e.g. calibration, application conditions). In the next step, a gateway is needed to collect the sensor data and harmonize them according to the data model and assign them to a reference system with the information about place, time and sensor ID. The gateway also serves to define global parameters such as the sampling rate or to ensure time synchronization. The time base is the foremost quality assurance criterion for the implementation of such a sampling paradigm.

Once the message packets are in the data stream, the downstream processing steps are borrowed from other domains such as logistics or business informatics. Low-code programming for event-driven applications can be achieved, for example, by using Node-Red as a powerful and versatile platform (https://nodered.org/) for connecting hardware devices, APIs and online services. Time series databases have proven to be an efficient and robust solution for storing sensor data, e.g. influx DB (https://www.influxdata.com/). Thanks to their own syntax and robust architecture, the query and storage of even large amounts of data is very fast and allows establishing complex

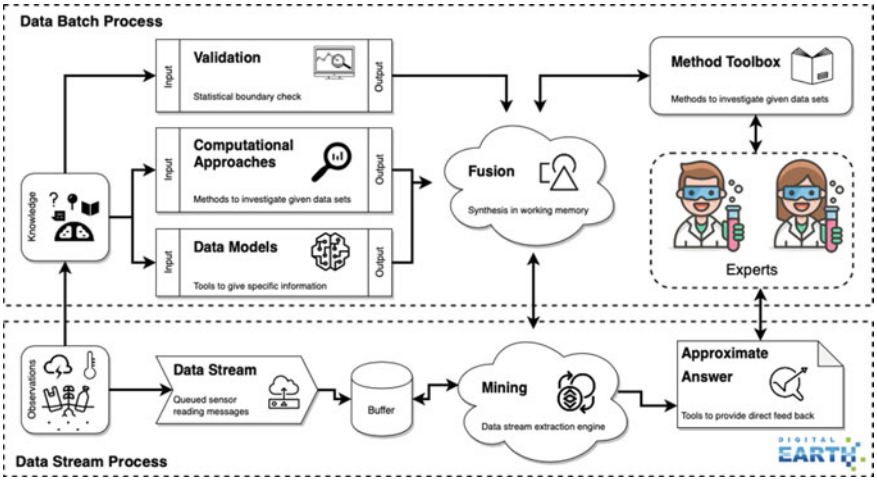

Fig. 6.2 Data stream architecture for processing real-time data based on IoT-capable sensor systems and considering server-side services for the valorization and contextualization of environmental monitoring data

query and processing procedures. As such, a data-driven architecture for service-oriented observation methods and in-stream process modelling close to real time is ready for use, available as open-source with powerful capabilities thanks to a large user community. This makes such an approach rather low-cost with low overhead for not directly necessary tasks.

6.3.2 SMART DataFlow

A SMART DataFlow from the sensor to the database is a central part of the SMART monitoring concept as highlighted in Chapter 6. This DataFlow represents parts of the data life cycle and plays an important role in data acquisition, data handling and data management (Fig. 6.3). Within a SMART DataFlow standardized and automated procedures are needed to obtain reliable data for subsequent analysis and application.

To gain reliable and trustworthy data, it is essential to develop and apply standard operating procedures within the DataFlow from the individual sensor to the repository. The following prerequisites or conditions have been identified as important:

- Availability of a portfolio of different independent methods measuring a specific parameter
- Combination of various SMART sensors in a network allowing self-repair / calibration mechanisms during operation

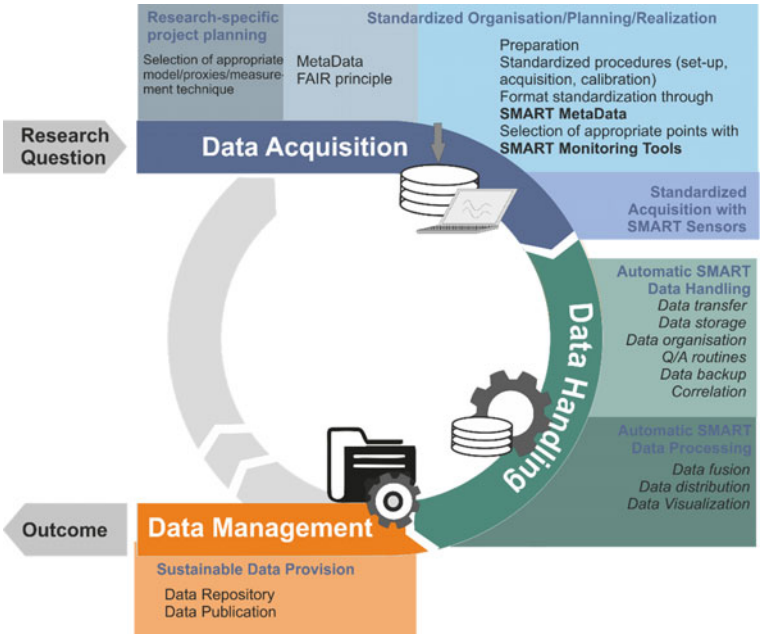

Fig. 6.3 SMART DataFlow as backbone of the SMART Monitoring concept

- Possibility of applying a hierarchical monitoring approach across different scales and parameters
- Possibility to transfer concepts of combined methods to further research questions
- Capacities for automatic data handling using standardized data transfer, data storage, Q/A routines and data backup rules and routines in, e.g. traditional repositories; important are:

  o defined data or formats and standardized data format transformation procedures

  p standardized routines for Q/A to ensure identical data flagging; standardized routines should correspond to existing internationally accepted and applied Standard Operating Procedures (SOPs), e.g. from ICOS, ARGO, the World Meteorological Organisation Global Atmosphere Watch programme or more informal at GO-SHIP from the marine science (https://www.go-ship.org/HydroMan.html)

  q standardized, automated and interactive uncertainty analysis tools for data and proxy transfer functions

  r standardized processing routines for integrated data analysis or proxy transfer functions

- Capacities for automated SMART data processing that allows data fusion, distribution and visualization of all available data such as dashboards. Processing should include:
  - o hierarchical data storage structure to combine and integrate relevant auxiliary data, calibration protocols and data from intercomparison experiments of the respective device
  - p standardized data analysis and visualization tools and software
  - q automated (near) real-time application of visual and interactive tools to: (1) process various (near) real-time data; (2) apply multiple filters for data analysis; (3) combine different data sets; (4) connect to other software packages; (4) enable joint data analysis by different users
- Standardized reporting routines which ensure that all processing steps are precisely described and traceable and that all users can assess the data quality of the parameters derived by proxy transfer functions, needed are:
  - o standardized metadata vocabulary and schema

In the following, we present a number of approaches and tools we have developed in Digital Earth to address certain aspects of the SMART DataFlow.

6.3.2.1 Automated QA/QC pipelines (Quality Assurance/Quality Check)

During the Digital Earth project, automated workflows for data processing have been developed, focusing on the near real-time quality assessment and quality control of the collected data. One major design criterion for the workflows was their composability with existing workflows of other users and their scalability to be easily adaptable to other requirements. In this subchapter, two examples of the successful implementation of the developed workflows in the European Research Infrastructures IAGOS and TERENO are introduced

IAGOS: The European Research Infrastructure IAGOS (In-service Aircraft for a Global Observing System; www.IAOGS.org) operates a global-scale observing system for atmospheric composition and essential climate variables by deploying automated instruments on passenger aircraft during their commercial flights. To handle the immense DataFlow from the fleet of aircraft collecting data, IAGOS has implemented an automatic workflow for data management, organizing the DataFlow starting at the sensor towards the central data portal located in Toulouse, France. The workflow is realized and documented using the web-based Django framework with a model-based approach using Python (Fig. 6.4). In Fig. 6.1, the overall sketch is shown.

A permanent active cronjob called Task Manager (outer box) activates an individual task instance (dotted box) of a task class describing the complete data handling process. This includes the following steps: (1) The Transfer Handler checks for new

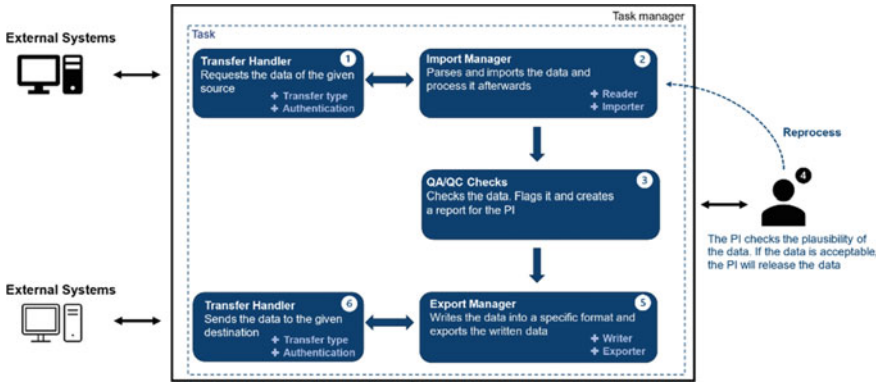

Fig. 6.4 *Scheme of the IAGOS process chain including a QA/QC pipeline*

data of the task-specific data type by using a RESTful API (Application Programming Interface) operated at the data centre in Toulouse. The API handles the necessary authentication by using an individual ssh256 encrypted token generated with a pre-shared passphrase and unique token (timestamp).

If new data is transferred, the data is passed to the Import Manager (2). The Import Manager reads and parses the raw files (pandas toolset) and processes the raw data to meaningful values. In the end, the Import Manager stores the processed time series to the instrument database for further processing. As the next step (3), the advanced QA/QC Handler performs checks, flags the data and produces a report for the PI who has to release the data for Level 1 and Level 2. (see Table 6.2).

In principle, step 2 and 3 are the same for all different data levels. In the end, the data-level reached depends only on predefined requirements e.g. the availability of the post-flight calibration. In step 4, the Export Manager writes the data to a specific transport format (e.g. NetCDF, or API specific format) and passes it to the Transfer Handler (6), which finally handles the transfer for a specific data type towards the data centre, including the authentication process already described for the Transfer Handler (1). The Task Manager tracks the status of all tasks even if they are terminated by reaching Level 2 or stopped by the PI. All information on tasks, including decisions of the PI, is stored for later reprocessing if needed.

Table 6.2 Definitions of the level of maturity for IAGOS data

Level 0	Raw data checking
NRT	Fully automated upload of processed raw data including automated QA/QC checks and flagging within 3 days (Near Real Time)
Level 1	NRT grade data, approved by PI
Level 2	Fully reprocessed scientific grade data, including post flight calibration, automated QA/QC and flagging approved by PI

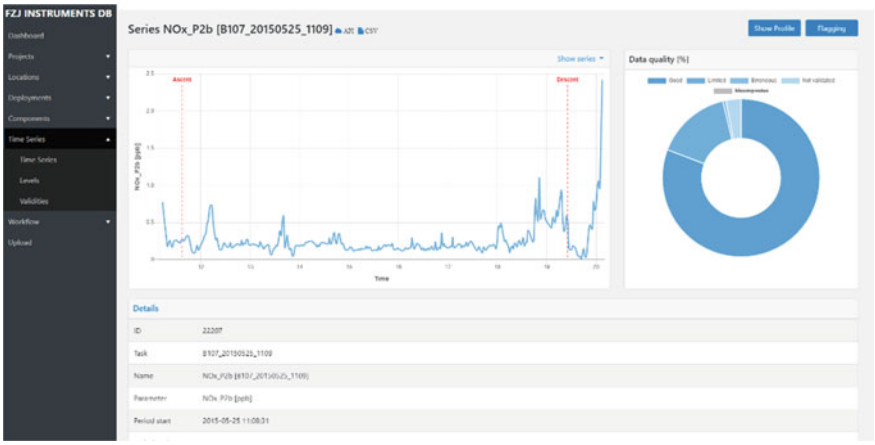

Fig. 6.5 Screenshot of the IAGOS Instrument Database application for visual quality check

This workflow performs all necessary data processing and QA/QC tests to automated upload NRT processed data and serves the PI as a basis for approval decisions. This includes repeated cycles for different stages of data maturity. The PI can monitor the status of all tasks through web-based reports produced by the Task Manager. An automated reprocessing is possible by storing metadata on all steps as well as decisions of the PI. Implementing the workflow is one big step to making IAGOS data handling compliant with the FAIR principles (Fig. 6.5).

The automated QA/QC tests are accessed inside the workflow using the Python framework **Autom8QC** developed and used by the DE community. It fulfils the following prerequisites:

- Application of probabilistic approaches
- Easy adaptation of test strategies
- Adaptation to different environments
- Application of different test measures (e.g. logical-, statistical-measures)
- Combination of tests in groups
- Combination of tests in sequences

The following QA/QC tests are already implemented for common use (Table 6.3):

The described framework was successfully integrated into the IAGOS workflow to create a QA/QC pipeline that generates an automated test report and automatically flags the measured and processed data. The PI uses these reports for the final approval decision (data maturity Level 1 and Level 2).

TERENO is an initiative funded by the Helmholtz Association to create observation platforms to facilitate the investigation of the consequences of global change on terrestrial ecosystems and the socioeconomic implications of these (Zacharias et al., 2011). Four observatories have been set up in 2008, each operated by one

Table 6.3 List of implemented test modules for the Auto8QC framework

Test type	Test name	Description
Value	Flatline test	Identifies repeated occurrence of one value in a time series
	NaN test	Test for "Not a Number" values
	Specific Value test	Test for a specific value
Time	ITT	Increasing time test (monotonic timeline)
	Time Gap test	Test for unexpected gaps in the timeline
	Time Range test	Test for specific timeline range
Limit	Global Minimum test	Test for values below specific minimum
	Global Maximum test	Test for values above specific maximum
	Global Range test	Test for values outside of specific range
Outlier	ESD (extreme Studentized deviate) test	Detects one or more outliers in a univariate data set
	LOF (Local Outlier Factor) test	Compares the density of any given data point to the density of its neighbours
	IQR (Inter-Quartile Range) test	Detects outliers using Inter-Quartile Range
	MAD (Median Absolute Deviation) test	Detects outliers using Median Absolute Deviation
	OutlierZ test	Detects outliers using the Z-score
Peak	ScipyPeak test	Uses the Python library SciPy to detect peaks

Helmholtz Centre, which maintains its local data infrastructure. The individual infrastructures are interconnected into the distributed TERENO Online Data RepOsitORy (TEODOOR), supporting the acquisition, provision, and management of observations via SWE specifications and several other OGC web services (Kunkel et al., 2013).

For the Eifel observatory, operated by the Research Centre Jülich, about 180 mio. mostly meteorological, aquatic and terrestrial observations are collected each year, from which about 90 Mio. (54%) have to be quality checked. Each observation is, among others, attributed to a processing status and a data quality flag. Following the IOC of UNESCO (2013), we adopted a two-level scheme to assign the data quality for each observation. The first level defines the generic data quality flags, while the second-level complements the first level by providing the justification for the quality flags based on validation tests and data processing history. In TERENO, the second-level flags are specified by the domain experts. The processing status describes the data type and determines the workflow for data editing and publication (Table 6.4).

Observation data are imported into the observational database and managed with our time series management system (TSM 2.0) (Kunkel et al., 2013). The system includes a highly configurable file parser, a data processor as well as a task manager

Table 6.4 Data processing levels for TERENO data

Level	Descriptions	Data Source	QC	Data Editing	Availability
Level 1	Raw data	Automatic importing or manual upload	No	Not allowed	Internal (on request)
Level 2a	Externally quality controlled data; an approval is pending	Level 1 data (manual upload)	Yes	Not allowed, flagging only (except human observations)	Internal (on request)
Level 2b	Quality controlled data with automatic QC procedures	Level 1 data (automatic upload)	Yes	Not allowed, flagging only	Internal (on request)
Level 2c	Externally quality controlled data with an expert approval	Level 2a data	Yes	Not allowed, flagging only	Public
Level 2d	Quality controlled data with semi-automatic QC procedures (automatically and by human)	Level 2b data	Yes	Not allowed, flagging only	Public
Level 3	Derived data	One or more Level 2 data	Yes	Allowed	Public

for internal and external procedures. It allows automated data pre-validation such as transmission and threshold checks and flagging of the data along with the importing process. After a visual examination by the responsible scientist or technician, data of Level 2 or higher are made available online automatically via Sensor Observation Services, which provide the data, the processing levels and the data quality flags (Devaraju et al., 2015). The full process of data collection, transmission, processing, QA/QC and management is fully documented and certified according to ISO 9001.

However, several issues and limitations arise with this QA/QC workflow, like:

- QA/QC is performed inside the TSM during the data import with the advantage of very fast processing of the data under consideration of parameters and sensors. In practice, however, the system is limited to basic QA/QC routines like threshold checks.
- Implementing additional, more complex and/or site-specific QA/QC routines and its application to specific data processing workflows requires significant programming skills, making it almost impossible for scientists to develop these routines by themselves.
- For these reasons, the scientists will usually download the data to develop and perform their own QA/QC routines on their computer systems. In most cases, processed and/or QA flagged data will not get back into the infrastructure.

To overcome these limitations, we extended the workflows within the DE project for automated data processing by Auto-QA. This browser-based flow editor that allows users to develop, test and run their own processing chains to add their own procedures to access the data infrastructure using standardized interfaces and to run, visualize and upload the results. As a basis, we used the Node-Red software, a flow-based development tool for visual programming. Node-RED, built on Node.js, provides a web browser-based flow editor, which can be used to create JavaScript functions. Elements of applications can be saved or shared for reuse using JSON. Red-Node uses "nodes", which can be interconnected graphically to processing chains. Each node can be characterized by individual properties, which can be used for process control.

In order to use Node-Red for TERENO, we developed a library of modules for importing, visualization and exporting data and included the QA/QC routines from the Python framework Autom8QC. The framework was extended by a module for automated conversion of PYTHON code into JAVASCRIPT to allow the incorporation of custom PYTHON modules in Auto-QA.

Figure 6.6 shows a simple example of a global range test of an observed parameter. The process will be initiated by a trigger, which may be a manual start or, for instance, a GET request via HTTP to this particular trigger. The trigger initiates the data to be read from TEODOOR by an OGC-SOS client (CLISOS). The output produced by this node will be sent to the variables and min_max node. "variables" will then print all available variables to the debug panel. "min_max", which is a global range filter, flags the data according to the settings made in the node. The output will be formatted in the "clisos_format" and then be outputted to the debug panel. The data sent to the debug_panel can be visualized graphically, written to a "file" node or uploaded directly into the infrastructure using a SOS-T. The whole process can be parametrized by the GET parameters of the trigger and stored.

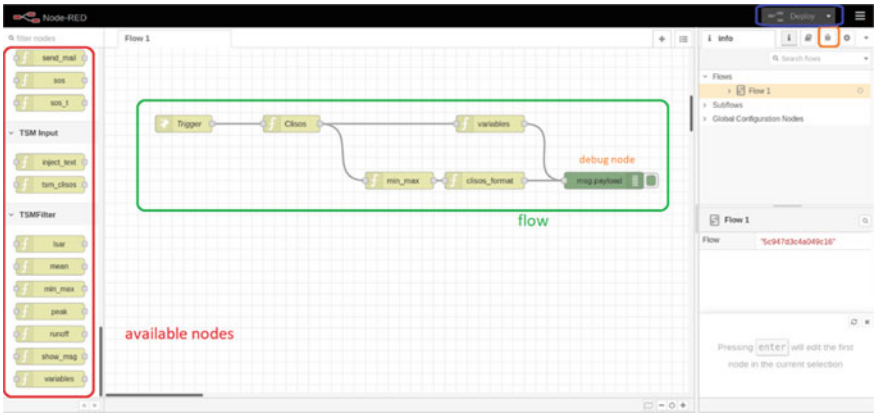

Fig. 6.6 Screenshot of a simple QA/QC workflow for a simple global range test of an observed parameter

This makes is possible to use the developed workflows as templates for the application to other parameters and/or sites. Moreover, the Auto-QA can easily be initiated automatically through the TSM, by calling the trigger through the CRON module of the TSM followed by an automated import of the flagged and/or produced data. Finally, Auto-QA is not exclusively used for QA/QC, but also for the development of workflows for extended data processing, e.g. calculating runoff data from water levels on surface water gauging stations or the calculation of soil moisture data from Cosmic Ray stations.

6.3.2.2 Analytics of Big Geo-Social Data for Environmental Monitoring: Exploring Microblogs to Spatiotemporally Characterize Floods, Droughts and Typhoons in China

In order to better understand and respond to the occurrence and impacts of extreme hydrological events, SMART monitoring should also integrate data and metadata related to citizens' perceptions of floods and droughts.

As the trend towards social media data has increased rapidly over the last decade due to the simplicity and accessibility of social media platforms, it is obvious to develop a web scraper and data filtering tool for an environmental-hazards-analysis as well, to automatically filter out their spatially and temporally distributed occurrence from microblogging services such as Twitter. Citizen Science could thus serve, for example, as first-hand information on the hydrological situation in communities.

Although more and more remote sensing data are being used to monitor hydrologic events in support of traditional hydrologic monitoring networks, observation is limited by the coverage of monitoring networks and the spatiotemporal resolution of satellite imagery, and thus its informative value. Alternatively, the highly diverse and rich data from social networks contain valuable meta-information on hydrological events and can be extracted using keyword-based sentiment analysis techniques (Olteanu et al., 2015; Wang et al., 2016).

An automated approach that focuses on China in our study retrieves and processes microblogs from Sina Weibo. This approach consists of an API-independent web scraper to retrieve large volumes of microblogs, a data cleaning and filtering module, a georeferencing module and a supervised machine learning approach to classify content reliability (Fig. 6.7). The workflow was applied to analyze more than 700,000 microblogs on typhoons, droughts, and floods from 2018 to 2019. For validating the extracted information, remote sensing data from International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al., 2010 and 2018), standardized precipitation evapotranspiration index (SPEI) drought index (Vicente-Serrano et al., 2010) and NASA Global Precipitation Measurement (Huffmann et al., 2014) were utilized.

In addition, a data collection system was developed to capture large social media data from Sina Weibo, a popular microblogging website in China. Microblog details, including content, authors, time of publication and geotags, were captured by a Python script that used packages to log in and parse HTML scripts from websites

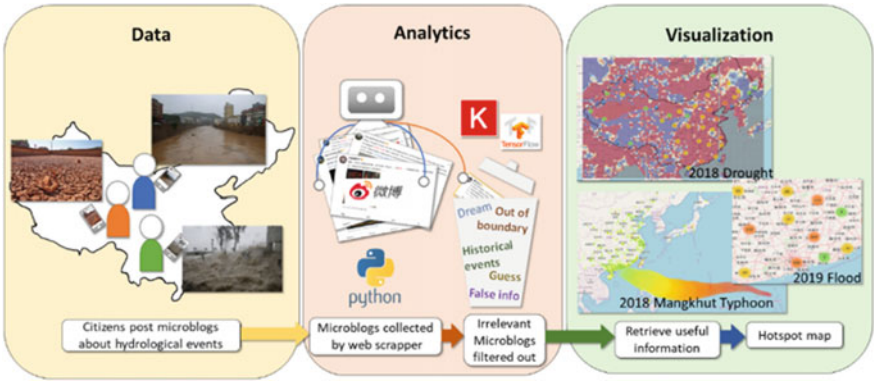

Fig. 6.7 Simplified process chain for utilizing microblogs to predict floods

to capture targeted data. Keywords related to floods, droughts, and typhoons were prepared and sent as search queries to the Sina Weibo search engine, and the collected data were stored.

In the analysis of microblogs related to typhoons, the method works qualitatively very well and the spatiotemporal distribution of microblogs reflects the actual course of typhoons. This makes this type of workflow a potential tool for tracking the trajectory of meteorological events with sufficient social attention. In the case of droughts, where our study is one of the first researches on this topic on Sina Weibo, about three-quarters of the microblogs are located in areas classified as dry according to the SPEI drought index, which shows that this approach could provide meaningful information here as well. Interestingly, however, limitations were encountered in the analysis of flooding, as the distribution of microblogs was highly dependent on population density, and furthermore could not be correlated with precipitation patterns and river flows over a wide area. Future improvements can be made by including additional social media data sources to achieve even higher data density.

In conclusion, while not all microblogs were useful for our data analysis, and their content needs to be better cleaned for analysis (e.g. removal of assumptions, past events, advertisements, and reposts), the integration of social media into monitoring approaches is a big step towards more data-based environmental research.

6.3.2.3 Establishing a Webgis Project for Hydrological Campaign Planning and Data Sharing at the Mueglitz River Basin

Planning event-driven monitoring campaigns on spatial catchment scale requires a comprehensive overview of existing measurement/monitoring locations, previous campaigns, and the distribution of hydrological, geological, and geomorphological features within the study area. In Germany, data and related meta-data are often distributed among various institutions at the local, state and federal level, making

them difficult to access. Additionally, datasets are stored in different file formats (ASCII, MS Excel, MS Access, ESRI Shape) and typically have neither the same geographic coordinate system or projection nor a consistent parameter labelling. To optimize data access for campaign planning while minimizing working efforts and storage capacities (several data requests to external data providers), better centralized data handling, processing, and access approaches are needed.

For the MOSES intensive test site Mueglitz River Basin, a WebGIS project was established via the AWI GIS infrastructure maps.awi.de. The framework's significant components are an ArcGIS Server, a PostgreSQL database including a Spatial Database Engine (SDE), and a desktop GIS software. The WebGIS project was based on datasets from state authorities (Sächsisches Landesamt für Umwelt, Landwirtschaft und Geologie), measurement campaigns, results of numerical models as well as environmental datasets that are freely available via online data repositories. The links to the original datasets can be found within the AWI WebGIS. Dataset projections, formats and metadata descriptions have been standardized in close cooperation between the project partners following ISO standards. The data and metadata provided by the WebGIS include:

- Locations of weather stations, monitoring gauges and drilling logs
- Modelling results from the OpenGeoSys groundwater flow model (Ver. 5.7; https://www.opengeosys.org/ogs-5/) and the DIFGA rainfall-runoff model (Schwarze et al., 1991)
- Maps of soil, geology, rainfall and land use patterns in the study area
- Locations of sampling and installations during the MOSES campaign 2019

Rich metadata information is given for the "MOSES/Digital Earth Müglitz Campaign" project on https://maps.awi.de/awimaps/catalog/ (Fig. 6.8). 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 provided through Open Geospatial Consortium (OGC) standardized Web Map Services (WMS) or Web Feature Service (WFS), which facilitate data exchange and data visualization among project partners. Based on the experience of the joint-work during related MOSES campaigns, the framework of such a WebGIS project will support upcoming campaigns and serves as a good blueprint for an "almost seamless" DataFlow for campaign planning.

6.3.3 SMART MetaData: Without Trustworthy Descriptions, Data can be Un-FAIR

Documented information about why and how data were collected, their structure, quality assurance, confidentiality, access possibilities and terms of use are typical metadata information. The identification and tracking of different datasets versions are essential in order to find, use, share and manage data especially over longer

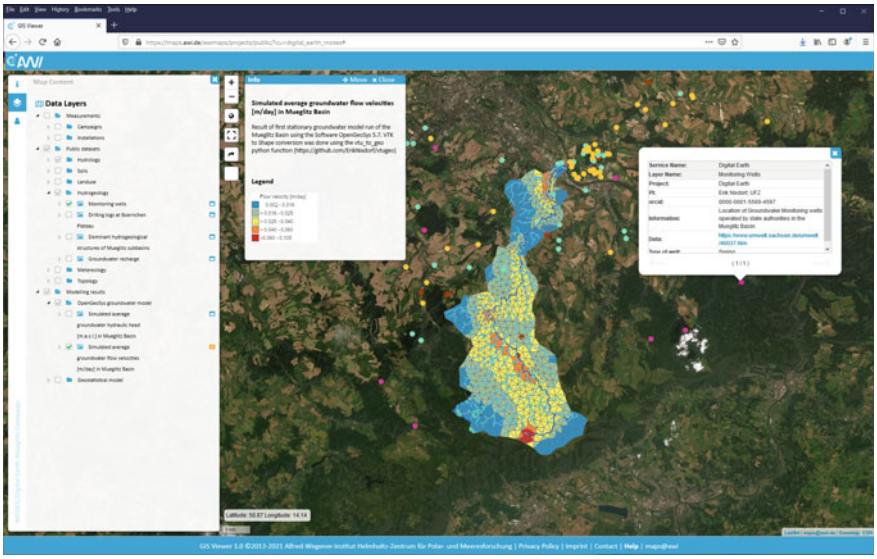

Fig. 6.8 Screenshot of the Müglitz campaing at the maps.awi webpage

time in a sustained way. Metadata thus describe other data or information about an object, be it physical or digital, to facilitate search, evaluation, acquisition and use of resources (Duval 2001). Metadata can be descriptive and consist of information about the data content and context (e.g. title, keywords, abstract); they can be structural provide information on the relationship between different datasets and they can be administrative which is essential to manage the data with respect to e.g. ownership and rights management. Accurate and complete metadata are a prerequisite for data sharing and interoperability across different data types. However, the process of describing and documenting scientific data has remained a tedious, manual process even when data collection is fully automated. Researchers are often reluctant to share data with good metadata information even with close colleagues because creating documentation takes much effort and time.

In SMART monitoring, the availability and incorporation of metadata and auxiliary environmental data play an important role for data and knowledge improvement. Below we show an example of how good metadata information can be structured and what might happen if metadata description is lacking and data become not usable anymore.

The COVID-19 pandemic shows the importance of mathematical models for understanding the spread of the disease which is at the base for deriving mitigation measures (Bjørnstad et al. 2020; Dehning et al. 2020). In this process, it is essential to determine and validate relevant model parameters and data integration from different sources is a prerequisite for parameterization and validation of predictive tools or models. The needed data integration calls for having FAIR principles in place for a reliable analysis and interpretation of the data (Wilkinson et al. 2016).

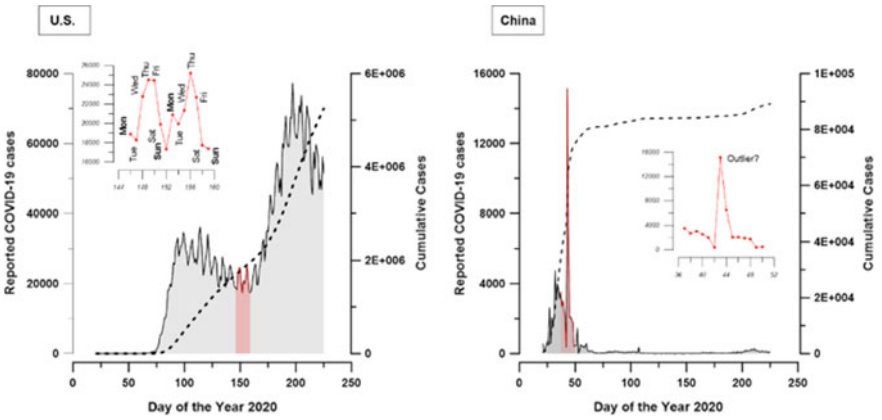

Fig. 6.9 *Daily and cumulative cases for the U.S. with an impressive example of a weekly periodicity and for China with an example of a rapid increase of daily numbers due to changes in counting (*Dong et al. 2020*)*

COVID-19 data illustrate the importance of data description with respect to origin and data quality to interpret data accurately. Figure 6.9 shows an example of reported daily COVID-19 cases for two different countries, here the US and China. Visual analysis of the curve of the daily cases reveals an approximately seven-day overlap of a slow trend process and a higher frequency process of weekly periodicity. Today, we know that the higher frequency process represents an artefact of data acquisition as a result of that laboratories and authorities that collect the board's data do not work on weekends. The behaviour of such an artefact can change over time if the data acquisition and communication processes changes/adjusts over time. A specific kind of adjustment is a change in counting as visible in the curve provided for China. The evaluation of the total number of covid cases with such sudden changes could lead to doubts concerning the completeness of the data.

An assessment of data quality and data origin is essential to preclude the possibility of inaccurate, incomplete or even unsatisfactory data analysis particularly when automated methods are applied that may lead to misleading or incorrect conclusions. The term "trustworthiness" of data summarizes all the aspects related to this potential problem. The fundamental features of trustworthiness are validity, provenience/provenance and reliability (Fig. 6.10). Using secondary data demand an additional detailed description and assessment of their reliability and validity which causes an increase of data collection methods. Validity assessments apply defined procedures to check for the accuracy of the observation findings. Data reliability analyses evaluate the quality of research by indicating the observed data's consistency and stability (repeatability and reproducibility).

The data reliability evaluation, especially in environmental science, should assess how precise and accurate individual measurements and with which uncertainty the measured value reflects the real value. It is evident that all acquired data come with an inherent uncertainty (Paasche et al., 2020). It is close to impossible to measure any

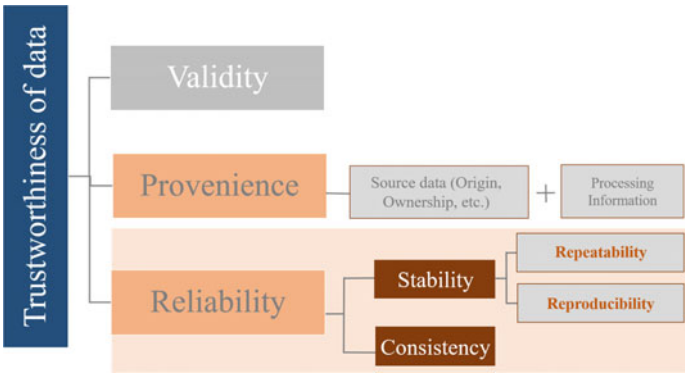

Fig. 6.10 Components of data trustworthiness, which should be considered to avoid wrong analysis and incorrect conclusions. A detailed description of the topic can be found in Koedel et al. (2022)

environmental variable with 100% certainty due to numerous limitations such as non-representative sampling schemes, improper sample handling, limited expertise of the data collector, unspecified effects of environmental conditions on the observed data, effects of specific electrical components of the measuring device on the data or data collector's bias. Even if this uncertainty cannot be assessed complete a description of the "where, how, who" provides valuable information for further data analysis.

In the case of COVID-19 example, the information about the overall test number to a country's population, delivery time to laboratories, quality aspects of the test centres (experience, analytics devices, number of confirmed invalid or incorrect tests), the daily processing capacity of laboratories, laboratory operating hours, reporting lag time and known test uncertainties is all important to understand and interpret the data in a consistent and comparable way.

Especially for these data and the derived measures, a comprehensive data analysis, including the analysis of stationary trends (7-day trend) and possible anomalies and uncertainties, is required to maintain the population's support. For all kinds of data collection, international efforts should be made to assess the data reliability in a standardized way.

Data provenience and provenance supply important information on the data source (provenience) and the applied processing steps (provenance), e.g. in the laboratory, to understand the data and its trustworthiness. For the given example it makes a difference if the data come from a certified laboratory that has consistently followed standardized routines or not.

It should be all scientists' task to request essential information on data quality from the data provider and to support the authorities through internationally accepted standardized workflows and metadata schemes. However, for most disciplines and their observations there are not yet (universally) accepted standards such as Standard Operating Procedures or suitable representation of uncertainty or other of indicators of validity. All this detailed information on data trustworthiness allows scientists to find appropriate tools and methods for FAIR data handling and a more accurate

data interpretation. There is need for an ongoing discussion among scientists, data managers and other data users to establish standardized criteria for trustworthiness assessment.

6.3.4 SMART Sampling Approaches

Another important development concerning SMART monitoring is the adaptation of statistical and AI-related methods to evaluate the location and time of data acquisition for implementing an adaptive sampling approach. A still very typical approach is that sampling strategies of physical samples or sensor-based measurements in the field are pre-determined following e.g. a random or targeted design (sampling at dedicated locations of interest) or a temporal- and/or spatial grid-like sampling scheme. Applying a SMART sampling approach means a pre-designed sampling strategy is adapted based on the specific and additional/external continuously accumulating data and knowledge during the sampling procedure itself.

When trying, e.g. to locate and quantitatively sample a plume of any substance in a specific volume (typically the ocean or atmosphere), it will be SMART to adapt sampling locations and sampling interval and density based on the measured concentration gradients during the sampling itself. Following an arbitrary but regular grid covering the assumed volume the substance might disperse in, can result in incomplete data. Adaptive sampling requires both auxiliary and real-time data that combined with advanced statistical or AI supported procedures and algorithms define better sampling locations or times for a specific monitoring task. Such an advancement in monitoring will not only support decisions about sensor locations but also sensor settings and the monitor strategy in time and space in an iterative way. For this kind of SMART Sampling approach additional supporting tool need to consider other essential data like: previously measured data in the area, auxiliary data for characterizing that area (land use, geology) potentially from remote sensing data, real-time sensor measurements if exiting and ideally modelled data. The aim is to apply mathematical and statistical methods and tools for deciding on where, when and how often sampling should happen, and how reliable sampling points, correlation functions and interrelationships of processes can be derived.

The resolution of the result depends mainly on the input parameters. The resolution of the inhomogeneous data input (remote sensing data, soil map data, land use data,) must be identical to run the clustering algorithm and the output only occurs on the common, coarsest grid size.

Therefore, a cascading coupling of the algorithm should be aimed at. For example, remote sensing data, representing data for large areas, often have a 10–50 km resolution and e.g. a sensing depth of 0.01–0.05 m for soil moisture depending on the measurement principle, frequency, and polarization direction. Within these data representative areas can be determined but downscaling of such data is possible but comes with an increase uncertainty during analysis and interpretation. Mesoscale

data (10 to 1000 m) such as geophysical or hydrological data can be used to determine representative sampling points in areas of interest and/or areas derived from clustering of remote sensing or model data. If predictions on representative depths are necessary, as for soil moisture, direct sampling data or other depth-oriented measurement methods (e.g. direct push methods; Dietrich and Leven, 2006) must be included. At which step model data are added depends on the resolution and representatives of the modelled data. Three studies that aim at determining sampling points, extrapolate metal resources in the deep sea and predict organic carbon deposition in the oceans give examples of typical SMART Sampling applications as part of SMART monitoring efforts.

6.3.4.1 How to Determine Representative Sampling Points at Meso-Scale

The determination of representative sampling points regarding the scientific question and regarding already existing data is a prerequisite to save time and workforce and to collect the best suitable data. The challenge is to achieve this in easy, quick and reliable manner, even in the field for a quasi-instant use.

Machine learning tools can be applied to determine representative sampling points in a specific area of interest. For example, clustering mechanisms, as the Fuzzy C means (FCM) algorithm or weighted conditioned Latin Hypercube Sampling are used for exploring multidimensional data with no prior knowledge of possible data relations (Hoeppner et al. 2000, Paasche et al. 2006). Therefore, these methods are applied to identify representative areas for sampling within the much larger area of interest. The application of Fuzzy C-Means Clustering algorithm by Paasche & Tronicke (2007) allows to identify areas of common features and to cluster multidimensional data by assigning each point a membership in each cluster centre. FCM is based on iteratively minimizing an object function for a defined number of clusters and provides the optimum locations of the cluster centres and the degree of partial membership of the clustered data points to the clusters (Paasche et al. 2006). The normalized classification entropy (NCE) indicates the optimum number of clusters by analyzing the membership distribution (Paasche et al. 2010). Previous experiments showed that fuzzy c-means (FCM) with spherically shaped clusters and Gustafson-Kessel clustering (GK) provides good clustering results. However, land use data as categorical variables represent a challenge in implementing this cluster algorithm. Heterogeneous data of this type are not directly usable to clustering methods. First, grouping in larger groups was necessary. Then, the Gowers generalized coefficient of dissimilarity was applied to calculate the distances with L1 (city block) (Gower, 1971; Gower & Legendre, 1986).

The area around Dittersdorf in the Müglitz river catchment area is a focus area of the MOSES hydrological extremes campaigns and as part of this, well-informed sampling locations were needed. To identify areas of common features, the Fuzzy C-Means clustering algorithm was applied to two data sets describing large areas with reasonable low resolution. These datasets were (1) digital elevation model, slope and

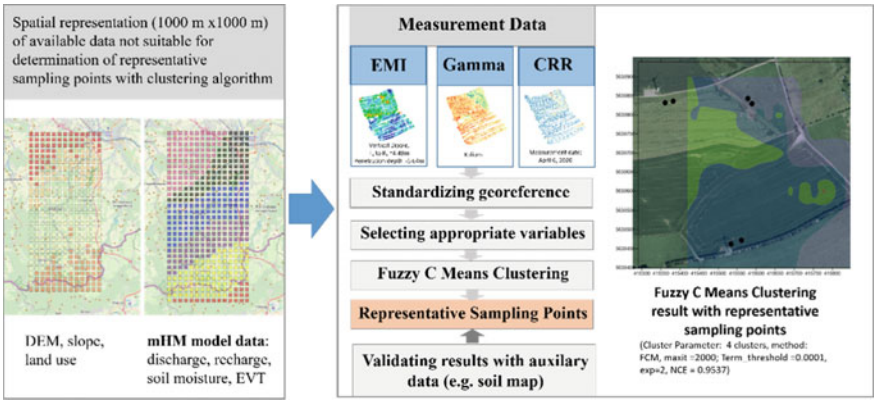

Fig. 6.11 Cascading coupling of clustering algorithm with selection of optimal sampling points at a medium scale

land use (as a categorical variable), and (2) mHM model data such as discharge, recharge, soil moisture and evapotranspiration from 1950–2009 (Fig. 6.11). Spatial representations of these clustering results are not suitable for the determination of representative sampling points because the grid-scale was ~ 1000 m x ~ 1000 m for all input data. Therefore, two meso-scale geophysical data (gamma ray, electromagnetic induction) and cosmic ray measurements were used to determine representative sampling points in Dittersdorf. Gamma ray measurements detect the decay rates of radionuclides with long decay times in soil using a scintillation detector with single sodium iodide crystals to determine the potassium, thorium and uranium concentrations and the natural gamma dose rate. The electromagnetic induction measurement is a highly adaptable none-invasive technique that measures the apparent bulk electrical conductivity of soil (ECa) to get information about field heterogeneity of soil texture and soil water content (Schmidhalter et al., 2008, Viscarra Rossel et al., 2011). Soil moisture content on a horizontal scale of hectometers and at depths of decimeters can be inferred from measurements of low-energy cosmic ray neutrons that are generated within the soil, moderated mainly by hydrogen atoms, and diffused back to the atmosphere (Zreda et al., 2012). Cosmic Ray Neutron Sensing's(CRNS) mobile application is a promising approach to measure field soil moisture noninvasively by surveying large regions with a ground-based vehicle (Schrön et al. 2018).

After standardization of existing georeferenced measurements, appropriate variables were defined and applied to the Fuzzy C-Means Clustering Algorithm. Finally, representative sampling points could be chosen for the meso-scale area. Such a cascading clustering allows the application of heterogeneous data scales and selecting representative areas or points at different scales.

Such a cascading approach can also be called a hierarchical approach. A combination of the most representative areas and then most representative sampling points and further most representative sampling depth allows an effective monitoring approach

and saves costs and workforce. Eventually, the provision and use of additional category variables such as passability or entry permits in the cluster algorithm also improve the adaptive monitoring and sampling approach.

6.3.4.2 Using Machine Learning For Automated Site Detection of Massive Seafloor Sulphides

Many current research questions in marine sciences are related to understanding the complex processes that govern resource occurrences. Relationships between the driving forces and response functions are complex, multi-faceted and usually non-linear. Additionally, multivariate and multi-disciplinary data acquired in the marine realm span disparate spatial scales and cross traditional geoscientific domains like geophysics, geochemistry or geology. Integrative data assessment approaches will play an essential role for amalgamating these cross-disciplinary interpretations, for redefining acquisition procedures to close existing data gaps, and for optimizing information extraction using image processing methodologies to make the most accurate prognoses of where to find previously undiscovered natural resources. The key task to establishing a foundation for multivariate data-driven analyses relies on developing integration concepts tailored to the available data on various spatial scales and linking these within established data science workflows. The challenge is to acquire all the needed data in a spatially correct context and apply ML on this small training data set.

Here, we present an example from seafloor massive sulphide (SMS) detection at the Trans-Atlantic Geotraverse (TAG) hydrothermal field. Various sources of marine data including autonomous underwater vehicle (AUV) bathymetry and magnetics (Petersen, 2019), and seafloor conductance data derived from Controlled-Source Electromagnetic (CSEM) inversion models are used (Gehrmann et al., 2019).

SMS indicators include a distinct bathymetric manifestation, magnetic low, and high electrical conductance. The latter is likely most indicative of mineral accumulations on the seafloor but only exists along 2D profiles crossing the measurement area due to the logistical expense of acquiring such data. As a result, robust extrapolation of sparsely sampled conductance data onto a regional scale seems efficient for predicting further occurrences of SMS by integrating the acquired bathymetric and magnetic data into a data science framework. This can help improve current predictions of available SMS on the seafloor and provide high-priority site predictions for future validation and sampling campaigns.

The available data allows us to use both unsupervised and supervised machine learning strategies to (a) classify the seafloor based on the spatially distributed bathymetry and magnetic data helping to identify regions with similar characteristics as the known SMS sites; (b) extrapolate conductance data to predict possible SMS sites outside of the 2D CSEM profile lines. Additionally, the high-resolution bathymetry data allows us to enhance our spatial feature matrix through image processing techniques, i.e. edge detection, circular Hough-transforms, and Gabor filtering to improve our spatial understanding of bathymetric features and feed these

into a sequential application of fuzzy clustering with random forest regression. The spatially sampled maps are used to create a segmented map of the seafloor combining regions of similar behaviour into common clusters. The pixel fuzziness, which describes the affiliation of each pixel to the corresponding cluster, is then used in a random forest regression approach at the defined locations of the sparsely sampled conductivity data to derive a model that allows us to extrapolate the sparsely sampled conductance data onto a regional scale (Fig. 6.12). Such two-step strategy deprives the ML kernel any physical meaning and relates its predictions solely to the learned patterns.

The results of this pilot study show that unsupervised and supervised machine learning strategies can be used to not only classify the seafloor into regions with similar behaviour, but also identify and predict known and unknown SMS sites in almost real-time. Thus, machine learning provides a robust framework to integrate multivariate data based solely on data-driven analyses, which will be of benefit to marine sciences to (a) optimize marine sampling campaigns through targeted point-scale measurements at regions of greatest interest defined through spatially

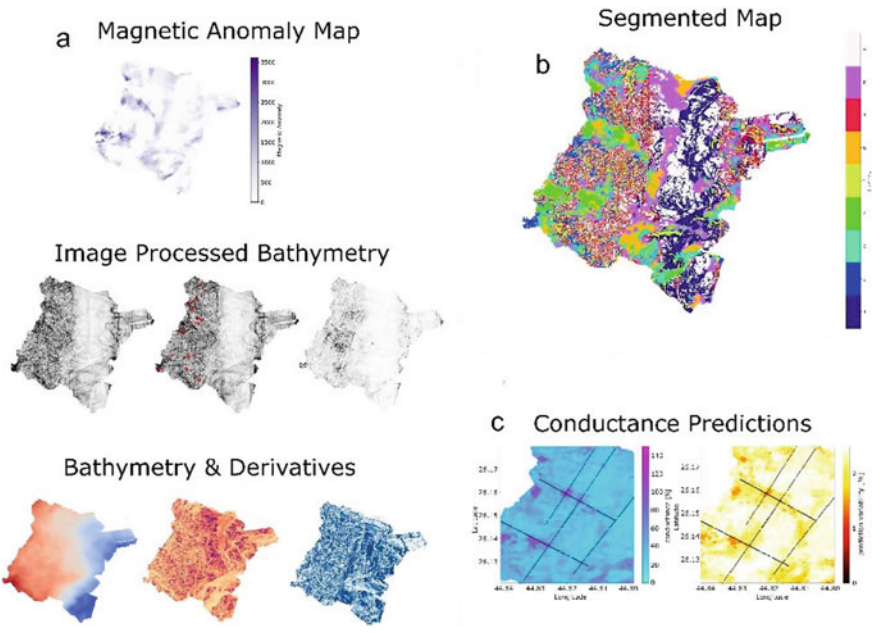

Fig. 6.12 Schematic of ML workflow for predicting SMS sites using multivariate spatial data. (a) Enhanced input features for unsupervised fuzzy clustering, consisting of regional bathymetry data and its derivatives (e.g. slope, ruggedness, aspect), feature enhanced image processed bathymetry (i.e. edge detection, circular Hough Transform and Gabor filtering), and physical property data (magnetic anomaly map). (b) Segmented output map imaging the main contributing component of each pixel. (c) Extrapolated CSEM conductance derived from random forest regression and the corresponding prediction variability assuming 5% Gaussian error. The black markers denote the actual profiles of the CSEM conductance data

distributed geophysical, geochemical, geomorphological, oceanographic or geological data; (b) update first-order predictions of available strategic metals on the seafloor through guided geophysical interpretations (Galley et al. 2021).

6.3.4.3 Deep Neural Networks for Total Organic Carbon Prediction and Data-Driven Sampling

The oceans comprise about 72% of the earth surface and due to its size and available technology, direct seafloor samples collected so far are sparse in space. The existing data sets on sediment composition are inadequate to quantify the fluxes of carbon and other seawater constituents across the seabed globally. Sediment and ocean models are strongly relying on these fluxes to simulate the uptake of atmospheric CO_2 and the biogeochemical cycles in the ocean. Moreover, sampling campaigns are often restricted by ship time, funds, and the lack of consistent methodologies to collect and process the data. Thus, the challenge is to find methods that allow to predict the total organic carbon (TOC) content everywhere in the ocean and show the uncertainty of this prediction with it.

To approach this problem, machine learning methods were adapted to marine sciences to approximate the seafloor physical and biogeochemical properties without the need of direct sampling. Some of these methods (e.g. k-Nearest Neighbours) provide a sophisticated averaging tool to estimate the seafloor property based on the data points nearest in space. However, this approach performs better in more homogeneous environments, which does not apply to global-scale problems.

Over the past decade, deep learning has been used to solve various regression and classification tasks (LeCun et al., 2015). Compared to classical machine learning approaches (k-Nearest Neighbours, Random Forests, etc.), deep learning algorithms excel at learning complex, non-linear internal representations in part due to the highly over-parameterized nature of their underlying models. This advantage often comes at the cost of interpretability. Exemplarily we used deep neural networks (DNN) to assess the TOC content of the global seafloor surface (Fig. 6.13). Implementing Softmax distributions on implicitly continuous data (regression tasks), we obtain probability distributions which can be used to quantify the model's intrinsic information. A variation of the Dropout method, i.e. the Monte Carlo Dropout, is used during the inference step providing a tool to model prediction reliability. Using transfer learning techniques, the resulting model was modified to also make sedimentation rate predictions; sedimentation rate ultimately relates to the problem related of calculating seafloor TOC.

We used these techniques to create model information maps that are a key element in developing new data-driven sampling strategies for data acquisition. Mapping prediction probabilities provide a quantitative analysis of the model information and allows us to define geographical locations that are under-sampled. By acquiring new information at these selected coordinates during upcoming research cruises potentially as part of new global sampling programmes will overall strongly and quickly improve global predictions.

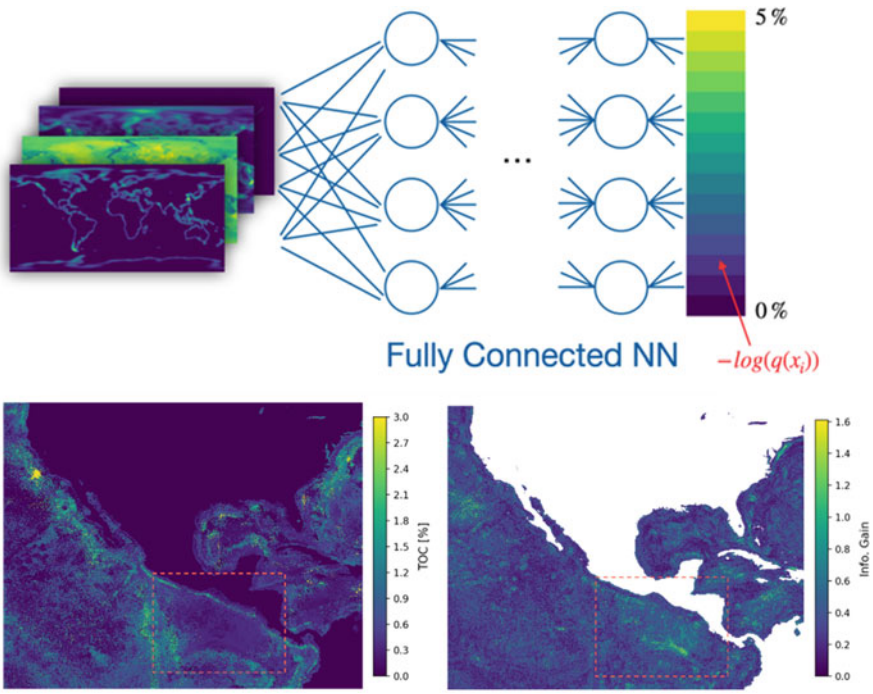

Fig. 6.13 Top: A fully connected neural network with a Softmax activation layer outputs a probability distribution of which the maximum corresponds to the predicted regression value. Bottom: Each prediction point is accompanied by an expected information gain sampling. Often times unexpected/counterintuitive prediction values are tied to a higher expected information gain values (here: low TOC patch in the Pacific Coast of Central America). This points to a higher model uncertainty for the region

Using the prediction probabilities to calculate the information gain from sampling, we were able to generate global maps that can aid data-driven sampling in the future. These information gain maps might be used by scientist to derive the most beneficial decisions on next sampling locations and potentially supports scientific research vessels of opportunity to collect data "during transit" when they pass one these important locations.

References

Alharbi N, Soh B (2019) Roles and Challenges of Network Sensors in Smart Cities, 2019 International Conference on Smart Power & Internet Energy Systems, IOP Conf. Series: Earth and Environmental Science322 (2019) 012002, https://doi.org/10.1088/1755-1315/322/1/012002.

ASQ (2021) Quality Assurance & Quality control, retrieved from https://asq.org/quality-resources/quality-assurance-vs-control on February 19, 2021.

Bjørnstad ON, Shea K, Krzywinski M et al (2020) Nat Methods 17:455–456. https://doi.org/10.1038/s41592-020-0822-z

Bruns T, Eichstädt S (2018) A smart sensor concept for traceable dynamic measurements. J Phys: Conf Ser 1065: 212011. https://doi.org/10.1088/1742-6596/1065/21/212011

CEN (2021) CEN Workshop Agreement , retrieved from https://boss.cen.eu/developingdeliverables/CWA/Pages/ on February 19, 2021.

Dehning J, Zierenberg J, Spitzner F P, Wibral M, Neto J P, Wilczek M, Priesemann V (2020) Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions, Science, 369: eabb9789.

Devaraju N, Bala G, Modak A (2015) Effects of large-scale deforestation on precipitation in the monsoon regions: Remote versus local effects. Proc Natl Acad Sci USA 112:3257–3262 https://doi.org/10.1073/pnas.1423439112

Dietrich P, Leven C (2006) Direct Push Technologies. In: Kirsch R (ed) Groundwater Geophysics. Springer, pp 321–340

Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Dorigo W, Himmelbauer I, Aberer D, Schremmer L, Petrakovic I, Zappa L, Preimesberger W, Xaver A, Annor F, Ardö J, Baldocchi D, Blöschl G, Bogena H, Brocca L Calvet JC, Camarero J J, Capello G, Choi M, Cosh M C, Demarty J, van de Giesen N, Hajdu I, Jensen KH, Kanniah KD, de Kat I, Kirchengast G, Rai PK, Kyrouac J, Larson K, Liu S, Loew A, Moghaddam M, Martínez Fernández J, Mattar Bader C, Morbidelli R, Musial JP, Osenga E, Palecki MA, Pfeil I, Powers J, Ikonen J, Robock A, Rüdiger C, Rummel U, Strobel M, Su Z, Sullivan R, Tagesson T, Vreugdenhil M, Walker J, Wigneron JP, Woods M, Yang K, Zhang X, Zreda M, Dietrich S, Gruber A, van Oevelen P, Wagner W, Scipal K, Drusch M, and Sabia R The International Soil Moisture Network: serving Earth system science for over a decade, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-2, accepted, 2021.

Duval E (2001) Metadata Standards: What, Who & Why. J. UCS. 7:591–601

Fischer P (2020) "Intelligent Sensor Technology: A 'Must-Have' for Next-Century Marine Science," In AI Technology for Underwater Robots, (eds.), KF, SS, KD & HN: Springer, 19–36. https://doi.org/10.1007/978-3-030-30683-0_2

Galley C, Lelièvre P, Haroon A, Graber S, Jamieson J, Szitkar F, et al. (2021) Magnetic and Gravity Surface Geometry Inverse Modelling of the TAG Active Mound. J Geophys Res Solid Earth, 126, e2021JB022228 https://doi.org/10.1029/2021JB022228

Gehrmann R, North LJ, Graber S, Szitkar F, Petersen S, Minshull TA, Murton BJ (2019) Marine mineral exploration with controlled source electromagnetics at the TAG hydrothermal field, 26°N Mid-Atlantic ridge. Geophys Res Lett 46:5808–5816 https://doi.org/10.1029/2019GL082928

Gerchow P, Koppe R, Macario A, Haas A, Schäfer-Neth C. and Pfeiffenberger H. (2015): O2A: A Generic Framework for Enabling the Flow of Sensor Observations to Archives and Publications, European Geosciences Union, Vienna, 12 April 2015—17 April 2015

Gower JC (1971) A General Coefficient of Similarity and Some of Its Properties. Biometrics 27:857–871 https://doi.org/10.2307/2528823

Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3:5–48 https://doi.org/10.1007/BF01896809

Green S (2003) Metadata: Essential Standards for Management of Digital Libraries, ALI Digital Library Workshop Linda Cantara, Metadata Librarian Indiana University, Bloomington, downloaded from https://slideplayer.com/slide/7641133/ on November 12, 2020

Higgins S (2007) DCC Standards Watch 1: What are Metadata Standards?, downloaded from https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/using-metadata-standards on November 16, 2020

Höppner F, Klawonn F, Kruse R, Runkler T (2000) Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. J Oper Res Soc 51 https://doi.org/10.2307/254022

Huffman GD, Bolvin D, Braithwaite K, Hsu R, Joyce Xie P (2014) Integrated Multi-satellitE Retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center

Kiran M, Murphy P, Monga I, Dugan J (2015) and Baveja SS "Lambda architecture for cost-effective batch and speed big data processing." IEEE International Conference on Big Data (big Data) 2015:2785–2792 https://doi.org/10.1109/BigData.2015.7364082

Knapp KR, Diamond HJ, Kossin JP, Kruk MC, Schreck CJ (2018) International Best Track Archive for Climate Stewardship (IBTrACS) Project, version 4. NOAA National Centers for Environmental Information https://doi.org/10.25921/82ty-9e16

Knapp KR, Kruk MC, Levinson DH, Diamond HJ, Neumann CJ (2010) The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone bst track data. Bull Am Meteor Soc 91:363–376 https://doi.org/10.1175/2009BAMS2755.1

Koedel U, Schuetze C, Fischer FP, Bussmann I, Sauer PK, Nixdorf E, Kalbacher T, Wiechert V, Rechid D, Bouwer LM, Dietrich P (2022) Challenges in the evaluation of observational data trustworthiness from a data producers viewpoint (FAIR+). Front Environ Sci 9:art. 772666 https://doi.org/10.3389/fenvs.2021.772666

Koppe R, Gerchow P, Macario A, Haas A, Schäfer-Neth C, Pfeiffenberger H (2015) O2A: A generic framework for enabling the flow of sensor observations to archives and publications, *OCEANS 2015 - Genova.* Genova, Italy 2015:1–6 https://doi.org/10.1109/OCEANS-Genova.2015.7271657

Kunkel KE, Karl TR, Easterling DR, Redmond K, Young J, Yin X, Hennon P (2013) Probable maximum precipitation and climate change. Geophys Res Lett 40:1402–1408 https://doi.org/10.1002/grl.50334

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444 https://doi.org/10.1038/nature14539

Lombard et al. (2019). Globally Consistent Quantitative Observations of Planktonic Ecosystems. Front. Mar. Sci., https://doi.org/10.3389/fmars.2019.00196

Olteanu A, Vieweg S, Castillo C, 2015 February: What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pp. 994–1009. Association for Computing Machinery https://doi.org/10.1145/2675133.2675242

Paasche H, Paasche K, Dietrich P (2020) Nature and Culture 15:1–18 https://doi.org/10.3167/nc.2020.150101

Paasche H, Tronicke J (2007) Cooperative inversion of 2D geophysical data sets: A zonal approach based on fuzzy c-means cluster analysis. Geophysics 72:A35–A39 https://doi.org/10.1190/1.2670341

Paasche H, Tronicke J, Dietrich P (2010) Automated integration of partially colocated models: subsurface zonation using a modified fuzzy c-means cluster algorithm. Geophysics 75:P11–P22 https://doi.org/10.1190/1.3374411

Paasche H, Tronicke J, Holliger K, Green AG, Maurer H (2006) Integration of diverse physical-property models: Subsurface zonation and petrophysical parameter estimation based on fuzzy c-means cluster analyses. Geophysics 71:H33–H44

Petersen S, (2019): Bathymetric data products from AUV dives during METEOR cruise M127 (TAG Hydrothermal Field, Atlantic). GEOMAR - Helmholtz Centre for Ocean Research Kiel, PANGAEA, https://doi.org/10.1594/PANGAEA.899415

Sauerer J, (2013) Smart Sensors, Sensorik für erneuerbare Energien und Energieeffizienz : Beiträge zum Workshop vom AMA Fachverband für Sensorik e.V. und vom Forschungs Verbund Erneuerbare Energien am 12. und 13. März 2013 in Berlin-Adlershof Berlin, 2013, https://www.fvee.de/fileadmin/publikationen/Workshopbaende/ws2013/ws2013_03_02.pdf, downloaded on November 14, 2020

Schmidhalter U, Maidl F-X, Heuwinkel H, Demmel M, Auernhammer H, Noack P, Rothmund M (2008) Precision Farming—Adaptati on of land use management to small scale heterogeneity. In: Schröder P, Pfadenhauer J, Munch JC (eds) Perspectives for agroecosystem management, Elsevier, pp 121–199

Schrön M, (2017) Cosmic-ray neutron sensing and its applications to soil and land surface hydrology (PhD thesis). Potsdam, Germany: University of Potsdam

Schrön M, Zacharias S, Womack G, Köhli M, Desilets D, Oswald SE, Bumberger J, Mollenhauer H, Kögler S, Remmler P, Kasner M, Denk A, Dietrich P (2018) Intercomparison of cosmic-ray neutron sensors and water balance monitoring in an urban environment. Geosci. Instrum. Method. Data Syst. 7(1):83–99

Schultz MG, Kunkel R, Petzold A, (2019). New perspectives on quality assurance and quality control of environmental observation data, D.E. Newsletter December 2019, downloaded on https://www.digitalearth-hgf.de/storage/379/Newsletter_DE_2019_12_FZJ.pdf November 28, 2020

Schwarze R, Herrmann A, Münch A ,Grünewald U, Schöniger M (1991) Rechnergestützte Analyse von Abflusskomponenten und Verweilzeiten in kleinen Einzugsgebieten. 35:143-184

Spencer B, Ruiz Sandoval M, Kurata N (2004) Smart Sensing Technology: Opportunities and Challenges. Struct Control Health Monit 11:349–368 https://doi.org/10.1002/stc.48

Steinacker A, Ghavam A, Steinmetzt R, (2001) Metadata Standards for web-based Resources. IEEE Multimedia, downloaded at http://ivizlab.sfu.ca/arya/Papers/IEEE/Multimedia/2001/Jan/Metadata%20Standards%20for%20Web-based%20Resources.pdf on November, 17, 2020

Thakur D, Kumar Y, Kumar A, Singh P (2019) Applicability of Wireless Sensor Networks in Precision Agriculture: A Review. Wireless Pers Commun 107 https://doi.org/10.1007/s11277-019-06285-2

UFZ (2021). Modular Observation Solutions for Earth Systems (MOSES), retrieved from https://www.ufz.de/moses/ on February 19, 2021

Ullo SL, Sinha GR (2020) Advances in Smart Environment Monitoring Systems Using IoT and Sensors. Sensors (basel, Switzerland) 20(11):3113 https://doi.org/10.3390/s20113113

Vicente-Serrano SM, Beguería S, López-Moreno JI, (2010) A Multi-scalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index - SPEI. J Clim , 23 (7):1696–1718 https://doi.org/10.1175/2009JCLI2909.1

Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, Lobsey C (2011) Proximal soil sensing: An effective approach for soil measurements in space and time. In: Donald L. Sparks, (ed) Advances in Agronomy, Vol. 113, Burlington: Academic Press, pp 237–282. ISBN: 978-0-12-386473-4, Elsevier Inc. Academic Press

Wang Y, Wang T, Ye X, Zhu J, Lee J (2016) Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm. Sustainability 8(1):25 https://doi.org/10.3390/su8010025

Wilkinson M, Dumontier M, Aalbersberg I et al (2016) Sci Data 3:160018 https://doi.org/10.1038/sdata.2016.18

Zacharias S, Bogena H, Samaniego L, Mauder M, Fuß R, Pütz T, Frenzel M, Schwank M, Baessler C, Butterbach-Bahl K, Bens O, Borg E, Brauer A, Dietrich P, Hajnsek I, Helle G, Kiese R, Kunstmann H, Klotz S, Munch JC, Papen H, Priesack E, Schmid HP, Steinbrecher R, Rosenbaum U, Teutsch G, Vereecken H (2011) A network of terrestrial environmental observatories in Germany. Vadose Zone Journal 10(3):955–973 https://doi.org/10.2136/vzj2010.0139

Zhang D, Eng B, Prof S, Connor NEO, Regan PF, Ph.D. Thesis. Dublin City University; Dublin, Ireland: 2015. Multi-Modal Smart Sensing Network for School of Electronic Engineering

Zhang AB, Gourley D, (2009) Creating metadata, In Chandos Information Professional Series

Zreda M, Shuttleworth WJ, Zeng X, Zweck C, Desilets D, Franz T, Rosolem R (2012) COSMOS: The COsmic-ray soil moisture observing system. Hydrol Earth Syst Sci 16:4079–4099 https://doi.org/10.5194/hess-16-4079-2012