Chapter 1
Data Science and Earth System Science

Wolfgang zu Castell, Roland Ruhnke, Laurens M. Bouwer, Holger Brix, Peter Dietrich, Doris Dransch, Stephan Frickenhaus, Jens Greinert, and Andreas Petzold

Abstract Data-driven science has turned into a fourth fundamental paradigm of performing research. Earth System Science, following a holistic approach in unraveling the complex network of processes and interactions shaping system Earth, particularly profits from embracing data-driven approaches next to observation and modeling. At the end, increasing digitalization of Earth sciences will lead to cultural transformation towards a Digital Earth Culture.

Keywords Data analysis · Data exploration · Earth System Science · Data science · Digitalization · Machine learning

W. zu Castell (✉) · D. Dransch
Helmholtz Centre Potsdam—GFZ German Research Centre for Geosciences, Potsdam, Germany
e-mail: wolfgang.castell@gfz-potsdam.de

R. Ruhnke
Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

L. M. Bouwer
Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany

H. Brix
Helmholtz-Zentrum Hereon, Geesthacht, Germany

P. Dietrich
Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany

S. Frickenhaus
Helmholtz Centre for Polar and Marine Research, Alfred Wegener Institute, Bremerhaven, Germany

J. Greinert
GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

A. Petzold
Forschungszentrum Jülich GmbH, Jülich, Germany

W. zu Castell
Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 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_1

1.1 Introduction

When Al Gore coined the term Digital Earth about thirty years ago (Gore 1998), he envisaged a holistic tool for Earth system understanding, exploration and education. Imagining a child visiting an exhibition, he sketched the idea of a comprehensive framework for data integration and analysis, allowing for an overall perspective of planet Earth to dive into and refine with additionally enriched data wherever interest is leading to. As far as maps and imagery are concerned, services such as Google Earth™ became well established in the meantime. Whereas the ability to dive deeper and deeper into details and to explore ever more datasets using this tool, a real Digital Earth is still a vision to be realized.

1.2 Data Science

Integrative, exploratory data analysis has been established as a fourth paradigm of science, next to theory, experiment and simulation (Gray 2009). Indeed, data-driven analysis has led to new insights into several fields of research, in particular, in those fields which by their very nature are lacking a comprehensive underlying theory. Data science "focuses on the processes and systems that enable the extraction of knowledge or insights from data in various forms, either structured or unstructured" (Berman et al., 2016 p. 2). As such, data science utilizes computer science, statistics, machine learning, visualization and human–computer interaction to collect, clean, integrate, analyze and visualize data, as well as to interact with data to create insight into some problem(s) in the real world.

Data-driven approaches to knowledge discovery have penetrated into almost every field of empirical science. Two major developments have paved the way for this radical transformation: first, through the evolution of the World Wide Web, data sources have become available on an unprecedented scale. Using the Internet, access to data sources has been substantially facilitated with more and more data sources becoming available. At the same time, the parallel development of computing technology allowed the processing of an increasing amount of data, allowing researchers to incorporate more data into their models and ingest huge datasets in an automated way. Both of these prerequisites eventually allowed researchers in artificial intelligence to build models which otherwise would not have been feasible to train due to their large number of parameters. Thus, sufficient computing power and the availability of huge amounts of data enabled a switch of paradigm, leading to models of artificial intelligence, predicting possible patterns of interest without the need of an underlying theory. This is particularly true for deep learning networks, the advancement of which developed closely with massively parallel computing technology reaching a commodity level.

In the end, it seems like Al Gore's vision of Digital Earth is but a fingertip away from becoming reality. However, the complexity of a challenge cannot be assessed without taking the first steps toward the goal.

1.3 Earth System Science

Earth System Science, with its historically subdivided disciplines that are based on the Earth compartments, will significantly benefit from integrative data-driven science. Environmental changes are the result of a complex interaction of natural and anthropogenic processes on a wide variety of temporal and spatial scales. Understanding and quantifying these changes must be based on trustworthy and well-documented observations that capture the entire complexity of the Earth system. This includes the manifold interactions between the atmosphere, land and ocean, including the impacts on all forms of life. Targeted environmental research projects and continuously operating multivariate research infrastructures designed to monitor all components of the Earth system are crucial pillars for environmental scientists in their quest for understanding and interpreting the complex Earth system, together with numerical simulations.

Therefore, data in Earth System Science readily complies with four of the 5 Vs of Big Data: volume, velocity, variety and veracity. Space-based observation systems produce a high volume of data at a speed of change (velocity), which increases with every new mission being started. The variety of geospatial information is relying on specialized infrastructures being capable of honoring the spatio-temporal structure of the data (Schade et al. 2020). Due to the global scale and need for long time series, Earth sciences, in particular climate research, have to deal with uncertainty of data on a regular basis (veracity). However, the fifth V, value, can only be extracted when data is turned into knowledge, helping to answer the pressing questions of society (van Genderen et al. 2020).

Making accurate predictions and providing solutions for current questions related, e.g., to climate change, water, energy, biodiversity, food security and the development of scientifically based mitigation and adaptation strategies in the context of climate change and geo hazards are important requests toward the Earth science community worldwide. In addition to these society-driven questions, Earth System Sciences are still strongly motivated by the eagerness of individuals to understand processes, interrelations and tele-connections within small subsystems, between subsystems and the Earth system as a whole. Understanding and predicting temporal and spatial changes and their inherent uncertainties in the above-mentioned micro- to Earth spanning scales are the key to understanding Earth ecosystems. Reliable, high-quality and high-resolution data across all scales (seconds to millions of years; millimeters to 1,000s of km) has to be utilized in an integrative approach enhancing the ability to integrate data from different disciplines, between Earth compartments, and across interfaces.

1.4 Challenges

While embarking on the adventure toward building Digital Earth, we must not stop at collecting data and providing access to various data sources. Data acquisition needs to resolve issues of metadata standards, referencing datasets as well as providing tools for data conversions and data management. High-quality data also needs to be enriched with information on data acquisition technologies, such as error tolerances of sensors and measuring artifacts. Following an Internet of Things (IoT) paradigm, workflows have to be matured toward SMART monitoring, including anomaly detection methods and spatio-temporal imputation.

Taking into account the substantial role of models in Earth System Sciences, computational challenges follow. Simulations need to be run on a sustainable basis with proper methods of parameter tuning. With computing technology changing at a higher rate, legacy code and model libraries have to be adapted to new computing hardware. Thus, Earth scientists providing highly optimized codes have to work in a co-design manner with computer engineers (Schulthess 2015). Splitting code into a backbone part which is obviously closer to the underlying technology platform and a frontend library including application programming interfaces (APIs) will allow scientists to concentrate on their data analysis tasks. At the same time, application programmers can use descriptive programming languages such as Python, leaving imperative programming to the backend.

In the future, geospatial information infrastructures will have to be adjusted in order to cope with rapid changes in computing technology and at the same time scale with an increasing diversity of applications (Bauer et al. 2021). Closely linking model-based simulation with data-driven analysis and prediction will allow to address questions of increasing complexity as resulting from the incorporation of scientific domains lacking an underlying theoretical foundation. Data-driven approaches may also be used to avoid costly simulation runs on high-end HPC systems or to deal with larger gaps in datasets.

However, within a data-centric approach, dealing with large, distributed datasets by means of programming, is unavoidable. Minimizing data movement in algorithms has to be considered as well as making use of data hierarchies (see Schulthess 2015).

Data alone is not sufficient for gaining new insight and knowledge. Many machine learning methods rely on high-quality, annotated data being available for training. Obtaining high-quality, labeled data typically is a tedious task. In order to scale such tasks to a global and just-in-time level, scientists have to be released from doing repetitive, automatable work. Incremental learning techniques have to be developed, filling gaps in data streams, providing reliable labels, as well as sorting out minor quality measurements. Citizen science projects such as PlanktonID (Christiansen & Kiko, 2016) have proven that getting the public involved, large gains can be obtained in combining machine prediction with human perception. This is just one example showing that successful data science approaches reach beyond classical data analysis. At the end, it is the way we interact with data, which will push us to the next level. Being visual beings, new approaches for visual data exploration, technology will

enable users to explore complex datasets and set off to new exploration journeys. For such technology to be developed, interdisciplinary teams of Earth scientists, AI specialists and visualization experts have to join forces in modeling data exploration workflows and identifying entry points for technological support.

1.5 Digital Earth Culture

Working in cross-domain teams, making use of the diversity of expertise will be a key requirement of realizing Digital Earth. A new culture of scientific cooperation has to be implemented (Dai et al. 2018). From a slightly broader perspective, working toward Digital Earth will become an instantiation of digital transformation in Earth System Sciences. Making use of digitalization in order to release humans from automatable tasks, building on human creativity and supporting new insight by data-driven hypothesis making will transform knowledge extraction in Earth System Sciences.

Co-creative processes and agile cycles will become the new way of pursuing science. Cross-disciplinary cooperation will advance tools for scientific research, and advanced tools will foster creativity in Earth System Sciences. In general, digitalization and open science will cross-fertilize each other. With results of scientific work being shared, scientific progress will be fostered (Helmholtz Open Science Office, 2021). The complexity of System Earth will never be captured by a single domain perspective alone. To understand the interplay of Earth's compartments and to provide insight into consequences of anthropogenic influence, a combined effort of scientific diversity is needed. At the end, a fully operational digital twin of System Earth might result, seamlessly fusing data from various sources and allowing users to interact with the data, to explore, to learn and to admire the wonder of planet Earth.

References

Bauer P, Dueben PD, Hoefler T, Quintino T, Schulthess TC, Wedi NP (2021) The digital revolution of Earth-system science. Nat Comput Sci 1:104–113. https://doi.org/10.1038/s43588-021-00023-0

Berman F, Rutenbar R, Christensen H, Davidson S, Estrin D, Franklin M, Hailpern B, Martonosi M, Raghavan P, Stodden W, Szalay A (2016) Realizing the potential of data science. Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group. National Science Foundation. https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf. Accessed 1 October 2021

Christiansen S, Kiko R (2016) PlanktonID. https://planktonid.geomar.de

Dai Q, Shin E, Smith C (2018) Open and inclusive collaboration in science: a framework. OECD Science, Technology and Industry Working Papers, No. 2018/07. OECD Publishing, Paris. https://doi.org/10.1787/2dbff737-en

Gore A (1998) The Digital Earth: understanding our planet in the 21st century. Speech given at the California Science Center, Los Angeles, California, on January 31, ESRI. http://portal.opengeospatial.org/files/?artifact_id=6210. Accessed 1 October 2021

Gray J (2009) Jim Gray on eScience: a transformed scientific method. In: Hey T, Tansley S, Tolle K (eds) The fourth paradigm: data-intensive scientific discovery. Published by Microsoft Research, October 2009, ISBN: 978-0-9825442-0-4. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Helmholtz Open Science Office (2021) Open Science und Digitaler Wandel gehen Hand in Hand (German). Potsdam. https://gfzpublic.gfz-potsdam.de/pubman/item/item_5008705

Schade S, Granell C, Vancauwenberghe G, Keßler C, Vandenbroucke D, Masser I, Gould M (2020) Geospatial information infrastructures. In: Guo H et al (eds) Manual of Digital Earth. Springer, Singapore. https://doi.org/10.1007/978-981-32-9915-3_5

Schulthess TC (2015) Programming revisited. Nature Phys 11:369–373. https://doi.org/10.1038/nphys3294

van Genderen J, Goodchild MF, Guo H, Yang C, Nativi S, Wang L, Wang C (2020) Digital Earth challenges and future trends. In: Guo H et al (eds) Manual of Digital Earth. Springer, Singapore. https://doi.org/10.1007/978-981-32-9915-3_26