From Slide Rule to Big Data: How Data Science is Changing Water Science and Engineering

Janet G. Hering

Director, Swiss Federal Institute for Aquatic Science and Technology (Eawag), CH-8600 Dübendorf, Switzerland; Professor, Institute of Biogeochemistry and Pollutant Dynamics, Swiss Federal Institute of Technology Zürich, CH-8092 Zürich, Switzerland; Professor, School of Architecture, Civil and Environmental Engineering, Swiss Federal Institute of Technology Lausanne, CH-1015 Lausanne, Switzerland. Email: janet.hering@eawag.ch

Forum papers are thought-provoking opinion pieces or essays founded in fact, sometimes containing speculation, on a civil engineering topic of general interest and relevance to the readership of the journal. The views expressed in this Forum article do not necessarily reflect the views of ASCE or the Editorial Board of the journal.

https://doi.org/10.1061/(ASCE)EE.1943-7870.0001578

Introduction

My university cohort was one of the first to be allowed to use handheld calculators (replacing the slide rules that had been used previously) in our exams (B.A. 1979) and to create our figures and write our doctoral dissertations using graphics software and word processing programs (Ph.D. 1988). I distinctly remember going to the library to consult the printed version of *Chemical Abstracts* as well as the period when the online version of *Chemical Abstracts* went back only to the mid-1980s (resulting in a steep decline in referencing of older literature). For most of my career, it seemed that these developments were incremental, and my colleagues and I adjusted to them without major changes in our approaches and expectations. Over time, however, the developments in information technology (IT) and data science have reached the point where the field of water science and engineering (like many others) is confronted with a bewildering array of options and opportunities. This is challenging our fundamental approaches and assumptions about how to do our science and bringing about cultural changes in our expectations regarding the roles of individuals and institutions in the production and sharing of knowledge.

I started to pay serious attention to these issues a few years ago in my capacity as Director of the Swiss Federal Institute of Aquatic Science and Technology (Eawag). In addition to my own personal struggles to keep abreast of the exploding amount of information relevant to Eawag's mandate and positioning, I also have to make budgetary decisions regarding investments in IT infrastructure, research data management, and open-access publications and to respond to pleas from our researchers for scientific IT services. I used an invitation to write a book chapter to engage two of my colleagues (from our IT department and library) in addressing issues related to knowledge management. In that chapter, we were able to make some inroads in addressing issues relating to research data management and open access and to lay out the special challenges posed by experiential and practical knowledge, which are highly context-dependent (Hering et al. 2018). We stopped well short, however, of grappling with the complexities inherent in the volumes of heterogeneous data with which we are increasingly confronted and which I address in this article.

Here, I highlight the opportunities and challenges associated with

- rapidly increasing availability of voluminous, high-resolution data on water systems,
- web-based access to information and the consequent opportunities to contribute to online data sets and/or to develop models and software collaboratively,
- applications of computational science (especially machine-learning) to environmental data, and
- emerging challenges associated with open data and open science.

Although this is not a review, I have tried to reference the literature that addresses big data challenges in water science and engineering, including some of the broader literature on environmental applications. I follow the 4V concept of defining big data by volume, variety, veracity, and velocity (Farley et al. 2018). Data can be big with regard to one or more of these aspects (Fig. 1). Volume and heterogeneity (i.e., variety) of data are the most commonly considered aspects, but challenges also arise from the quality, reliability, and uncertainty of data (veracity) as well as the rates at which data are acquired or must be processed for particular applications (velocity). With this background, I illustrate some ways in which individual scientists and academic research institutions are taking advantage of new data-driven opportunities and accommodating the demands that accompany them. I also hope to be able to endorse some further steps we could take to promote the "move from data to

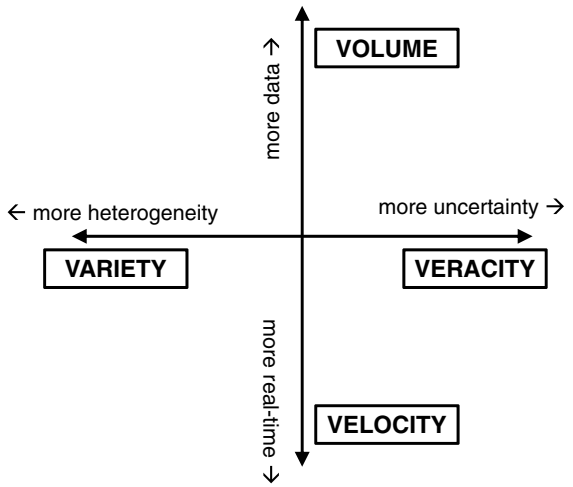

Fig. 1. Four axes along which big data can be defined. For a given (big) data set, a spider graph can be used to illustrate which of the 4Vs contributes the most to the bigness of the data set, whether this is simply the amount of data (volume), their heterogeneity (variety), uncertainty, and related aspects of quality and reliability (veracity), and/or the rates at which data are acquired or must be processed for particular applications (velocity). (Adapted from Farley et al. 2018.)

information to knowledge and, ultimately, to action for . . . sustainability and human well-being" (Ramaswami et al. 2016).

Data Fire Hose

For most of my career, the environmental sciences, particularly in the water domain, were rather data-poor. Aquatic scientists looked enviously at atmospheric scientists, who benefited from continuous online measurements of gases conducted from aircraft or balloons as well as ground-based (and later satellite) spectroscopic measurements that integrate over a column of air. Today, aquatic scientists and engineers are being flooded with data (pun intended). This flood has three main sources: omics, online and remotely deployed sensors, and remote sensing (Table 1). What these three sources have in common is the sheer volume of data; temporal and spatial resolution are additional challenges of the latter two sources. Omics data have expanded well beyond their origins in genomics to include high-throughput analyses of proteins (proteomics) and metabolites (metabolomics). Analysis of omics data (as well as other high-volume data) requires the development of data pipelines that automate the processes of extracting, transforming, combining, validating, and loading data for further analysis and visualization (Alley 2018). As the frequency of monitoring and/or the scale of experiments increases, data sets that have traditionally been analyzed manually also require automated pipelines for data handling and analysis (Durden et al. 2017; Farley et al. 2018; Pennekamp et al. 2017, 2018; Thomas et al. 2018a, b). Satellite observations, which at previous levels of spatial resolution were relevant mainly for marine systems, are, with improved resolution, increasingly relevant for lakes (Matthews and Odermatt 2015; Odermatt et al. 2018). Other spatially explicit data sources include remote sensing from drones and information collected by citizen scientists using mobile devices (McCabe et al. 2017). The spatial and temporal resolution of data from remotely deployed and online sensors and from remote sensing from aircraft and satellites pose additional challenges related to linking data to their time and location as well as to visualizing data, for example, in animated maps.

In engineering practice, water-treatment and wastewater-treatment plants are becoming more highly automated, and remote monitoring is increasingly used in distribution and/or conveyance systems, resulting in a substantial increase in the amount of data generated during system operation. These developments offer opportunities for performance optimization (Corominas et al. 2018; Ingildsen and Olsson 2016). They may also allow for novel management strategies, such as using excess sewer capacity to reduce overflows at wastewater-treatment plants (Zhang et al. 2018). Risks associated with vulnerability to cyber-attacks may, however, be increased (Taormina and Galelli 2018; Taormina et al. 2017).

Web-Based Collaboration

Web-based access to observational databases builds on a long historical tradition of monitoring data curated by (often governmental) institutions. The incorporation of well-defined data into online databases has been relatively straightforward, but even governmental agencies face challenges of curating and conserving legacy data. This challenge has been addressed by programs to preserve "data at risk" (Griffin 2015; USGS n.d.). Formal and/or informal scientific consortia have also formed to contribute to these efforts. The Force 11 consortium works to establish norms and standards, specifically the findable, accessible, interoperable, and reusable (FAIR) data principles (Wilkinson et al. 2016). The Research Data Alliance provides a neutral space where its more than 7,000 members

Table 1. Illustrative, noncomprehensive list of relevant online resources, with entries ordered alphabetically by name

Name	Description	URL
BioTIME	The BioTIME database contains raw data on species identities and abundances in ecological assemblages through time.	https://zenodo.org/record/1095627
CAMEL	Comprehensive Assessment of Models and Events using Library Tools (CAMEL) Framework is an integrated and flexible framework allowing users to seamlessly compare space weather and space science model outputs with observational data sets.	https://ccmc.gsfc.nasa.gov/camel/
Colaboratory	Tool for machine-learning education and research based on the Jupyter notebook.	https://colab.research.google.com/
DataJoint	A hub for developing, sharing, and publishing scientific data pipelines.	https://datajoint.io/
DRYAD	Online repository for data underlying scientific publications. It is curated and makes the data freely reusable and citable.	https://datadryad.org/
Earth System Data Lab (ESDL)	ESDL provides access to a series of highly-curated data cubes containing preprocessed data that are ready for analysis. A framework is provided to map user-defined functions to a data cube.	https://www.earthsystemdatalab.net/
Envidat	A portal to publish, connect, and search across existing data generated by the Swiss Federal Institute for Forest, Snow and Landscape (WSL).	https://www.envidat.ch/ui/#/
Fluxdata	Data portal for FLUXNET (https://fluxnet.ornl.gov/), a global network for eddy covariance flux measurements of carbon, water vapor, and energy exchange.	http://fluxnet.fluxdata.org/
freshwaterecology.info	Autecological characteristics, ecological preferences, and biological traits as well as distribution patterns of more than 20,000 European freshwater organisms belonging to fish, macro-invertebrates, macrophytes, diatoms, and phytoplankton.	https://www.freshwaterecology.info/
FreshWaterWatch	A platform for citizen science monitoring of freshwater ecosystems.	https://freshwaterwatch.thewaterhub.org/
GAP	Groundwater Assessment Platform (GAP) facilitates the exchange of data and information and supports predictive modeling of geogenic contaminants in groundwater.	https://www.gapmaps.org/

Table 1. (*Continued.*)

Name	Description	URL
GenBank	The NIH genetic sequence database is an annotated collection of all publicly available DNA sequences.	https://www.ncbi.nlm.nih.gov/genbank/
GEO	Gene Expression Omnibus (GEO) repository stores curated gene expression DataSets, as well as original Series and Platform records. DataSet records contain additional resources including cluster tools and differential expression queries.	https://www.ncbi.nlm.nih.gov/gds
GEOSS Portal	Data portal for the Group on Earth Observations (GEO) (http://www.earthobservations.org/index2.php), an intergovernmental organization working to improve the availability, access, and use of Earth observations for the benefit of society.	http://www.geoportal.org/
GitHub	Platform for code developers.	https://github.com/
GitLab	Platform for code developers (based on an open-core development model).	https://about.gitlab.com/
Global Reservoir and Dam (GRanD) Database	Compilation of existing dam and reservoir data sets with the aim of providing a single, geographically explicit, and reliable database for the scientific community.	http://www.gwsp.org/products/grand-database.html
Globus	Management service for research data allowing file transfer and sharing, data publication, and workflow development.	https://www.globus.org/
Google Earth Engine	Google Earth Engine combines a multipetabyte catalog of satellite imagery and geospatial data sets with planetary-scale analysis capabilities and makes it available for scientists, researchers, and developers.	https://earthengine.google.com/
HydroShare	CUAHSI's online collaboration environment for sharing data, models, and code.	https://www.hydroshare.org/
LTER Network Data Portal	Long Term Ecological Research (LTER) Network Information System Data Portal contains ecological data packages contributed by previous and present LTER sites.	https://portal.lternet.edu/nis/home.jsp
Map of Life	Map of Life is built on a scalable web platform designed for large biodiversity and environmental data and endeavors to provide best-possible species range information and species lists for any geographic area.	https://mol.org/
Metabolomics Workbench	Supports the development of next-generation technologies, provides training and mentoring opportunities, increases the inventory and availability of high-quality reference standards, and promotes data sharing and collaboration.	http://www.metabolomicsworkbench.org/
Meteolakes	Platform providing output from a three-dimensional lake model applied to three Swiss lakes.	http://meteolakes.ch
PRIDE	Proteomics Identifications (PRIDE) database is a centralized standards-compliant public data repository for proteomics data, including protein and peptide identifications, posttranslational modifications, and supporting spectral evidence.	https://www.ebi.ac.uk/pride/archive/
RENKU	Platform to facilitate the sharing and reuse of data and algorithms.	https://datascience.ch/solutions/
Simstrat	Platform providing output from one-dimensional lake model applied to 54 Swiss lakes open to updating by contribution of users' lake observations for calibration.	https://simstrat.eawag.ch/
Swiss Data Cube (SDC)	SDC contains 33 years of Landsat 5,7,8 (1984–2017) and 3 years of Sentinel-2 (2015-2018) Analysis Ready Data corresponding to more than 6,500 scenes.	https://www.swissdatacube.org/
TERENO Data Discovery Portal	Access to data from environmental observatories.	http://teodoor.icg.kfa-juelich.de/ddp/
USGS Water Services	Provides automated access to USGS water data (https://waterdata.usgs.gov/nwis)	https://waterservices.usgs.gov/

Note: NIH = National Institutes of Health; CUAHSI = Consortium of Universities for the Advancement of Hydrologic Science; and USGS = U.S. Geological Survey. For additional resources, particularly relating to water data portals, see CUAHSI (2019).

can "come together to develop and adopt infrastructure that promotes data-sharing and data-driven research" (RDA n.d.). A large consortium of researchers from almost 200 institutions acquired funding from a variety of sources to assemble the BioTIME database, which includes over 8 million species abundance records (Dornelas et al. 2018). Web-based collaboration can also facilitate citizen science initiatives (Shen et al. 2018).

In the omics and remote-sensing domains, data have been produced in a context in which the need for online storage and access quickly became obvious. With support from the NIH, Genbank (Table 1) was established in 1982. Today, sequence data deposition is a routine aspect of publication in the molecular biology community, although questions have been raised recently about how this may be affected by the Nagoya Protocol (Deplazes-Zemp et al. 2018). With satellite data, national and international space agencies have a vested interest in improving the accessibility and usability of their data and downstream data products.

Through such online resources, individual scientists or scientific consortia have the opportunity both to contribute to and exploit the wealth of web-accessible data. Models and tools for modeling are also increasingly available through online platforms (Table 1). Online platforms provide support for collaborative and/or participatory modeling (Basco-Carrera et al. 2017; Langsdale et al. 2013) (Gaudard et al. 2019), although platforms and models may be less important than trustful interpersonal interactions and adequate governance structures (Parrott 2017).

Applications of Data Sciences

Increasingly, the analysis of big environmental data (in the sense of one or more of the 4Vs in Fig. 1) relies on data science methods, particularly machine learning. In some approaches, hypotheses are generated and then tested using big data (Peters et al. 2018, 2014), which can also provide useful benchmarking for mechanistic models. Other approaches employ machine learning to extract trends or even elucidate hypotheses or model structures from data that are not biased by expectations (Ilie et al. 2017; Shen 2018; Thomas et al. 2018a). Although this approach can be compromised by spurious correlations in the data (N. Schuwirth, "How to make ecological models useful for environmental management," submitted, Eawag, Dübendorf, Switzerland), this problem can be minimized if sampling is informed by knowledge about the system (Strobl et al. 2008) and/or if appropriate tests are applied (Broadhurst and Kell 2006). Potential problems have been illustrated by the prevalence of false positives in a study investigating the possible use of variance and/or autocorrelation as early warning indicators for the abundance of aquatic taxa (Burthe et al. 2016).

Application of data science methods is necessitated when multiple types of data inputs must be combined (e.g., data from remote sensing and high-throughput DNA analysis) and interpreted using multiple modeling frameworks, especially when there is a goal of producing near-real-time predictions as the basis for decision making (Bush et al. 2017; Dafforn et al. 2016). Real-time data analysis can also support adaptive operation of the data acquisition system, as illustrated by a recent study of turbidity currents (Paull et al. 2018). Even the sheer size of environmental data sets may preclude conventional statistical analysis and necessitate data analysis based on machine learning, which does not require assumptions regarding data distributions, shape, and covariance structure (Cox 2015). The assumptions of common statistical methods (e.g., linearity and independence of variables) are unlikely to be applicable to large, multidimensional environmental data sets (McGowan et al. 2017; Sugihara et al. 2012; Ye and Sugihara 2016).

One recognized limitation of machine-learning approaches is their lack of interpretability (Pearl 2018; Shen 2018; Shen et al. 2018), which raises important questions of accountability when decision making is based on such approaches (EPFL IRGC 2018). This issue is a topic of intensive research in the data science community, although it has only begun to be addressed in the environmental research application area (Shen 2018; Shen et al. 2018). In this domain, integration of mechanistic models and/or inclusion of prior knowledge may offer insights into patterns derived from computational data analysis. Methods such as gene expression programming (GEP) generate explicit model structures from a specified set of operators applied to predictor variables and can be used in a reverse engineering approach (Ilie et al. 2017). Visualization of network activations can help to identify key forcing inputs triggering specific responses (Shen 2018; Shen et al. 2018). The requirement of machine-learning approaches for sufficient data also constitutes a limitation that has been addressed by using generative adversarial networks (GANs) to generate training data sets (Li et al. 2018).

A few examples clearly demonstrate the value of the analysis and interpretation of big data on aquatic systems. At the level of process understanding, the combination of remote-sensing data on temperature and chlorophyll with three-dimensional lake modeling allows surface biomass variations to be interpreted in relation to wind-driven transient upwelling and basin-scale internal waves (Bouffard et al. 2018). Analysis of historical records has demonstrated the legacy effects of deforestation (with consequent increases in discharge and infiltration) on wetland development (Woodward et al. 2014). Improved estimates of global river runoff have indicated that rivers play a larger role in the exchange of carbon dioxide between the land surface and the atmosphere than had previously been realized (Allen and Pavelsky 2018). Concerted efforts to compile and harmonize data on dams and their impacts have provided important insights into the aggregate impacts of dams on surface freshwater storage, run-off, nutrient and sediment transport, and sea-level rise as well as the consequences for aquatic ecosystems (Chao et al. 2008; Doell et al. 2009; Grill et al. 2015; Kondolf et al. 2014; Lehner et al. 2011; Maavara et al. 2015). With the planned and anticipated increases in dam construction, such an evidence base is needed to inform decision making (Fan et al. 2015; Zarfl et al. 2015).

Open Data and Open Science

The preceding discussion was based on the presumption that there is a common understanding of what data should be deposited online. This makes sense in the context of historical monitoring data or supporting data for journal publications but becomes blurred in the emerging context of open science, which incorporates the entire research cycle (Bueno de la Fuente n.d.). The caching of intermediate results, such as outputs of simulation runs, has been explicitly recommended (Peters et al. 2014), although this is widely considered to be impracticable. Although the depositing of genomic data is well-established, the increasing trend toward resequencing (from which the DNA of a specific individual can be compared against a composite reference genome) raises the question of what data must be stored: the full resequenced genome or a compressed version based on the reference genome (Pinho et al. 2012). At the other extreme, data produced by detectors at the Large Hadron Collider (LHC) at CERN are subjected to real-time analysis to reduce data volumes by factors of 1,000–10,000 before data storage (Gligorov 2015). The demands for data storage and speed of data transmission

are two of the most visible challenges for academic research institutions.

Institutional Challenges and Opportunities

There is no shortage of papers promising that big data will provide the basis for a profound improvement in our understanding of environmental systems and our capacity to manage them (Dafforn et al. 2016; Durden et al. 2017; Farley et al. 2018; Peters et al. 2018, 2014). Activities in synthesis centers such as The National Center for Ecological Analysis and Synthesis (NCEAS) and The National Socio-Environmental Synthesis Center (SESYNC) have demonstrated the power of data sharing in posing and answering previously intractable questions (Farley et al. 2018). The caveat is the level of investment that will be needed to capture these benefits. Needs for data storage and transmission will require upgrading of IT infrastructure. Support from informatics and data science experts will be needed for environmental scientists to apply computational methods to their data and models. But cultural changes in the attitudes and expectations of environmental scientists will also be needed to support the sharing of data as well as their collaborative use, interpretation, and presentation (Dafforn et al. 2016; Durden et al. 2017; Peters et al. 2018, 2014). Application of data sciences further imposes the need to share code and workflows, which requires proper annotation to support reproducibility (Hutton et al. 2016).

Research institutions must be aware of how their incentive systems (i.e., hiring, promotion, and tenure) may bias against data sharing and collaborative activities, issues that are particularly problematic for junior researchers (Gewin 2016). Even decisions about using proprietary or open-source software, which are often made at the level of an individual investigator or research group, can have important implications for further collaborative use of research products. At the same time, institutions have the capacity to support platforms for collaboration (such as the Swiss Data Science Center (SDSC n.d.) and to promote collaborative activities as exemplified by the July 2018 call for a biodiversity knowledge alliance (GBIF n.d.). Simply keeping abreast of all these developments poses its own challenges. Here, institutions can promote the FAIR data principles (Wilkinson et al. 2016) and encourage cross-referencing, harmonization, and (when appropriate) consolidation of platforms (Hering and Vairavamoorthy 2018). Funding agencies, in particular, should pay attention to the inherently transient nature of project-based platforms and take steps to ensure that successful platforms are embedded in an institutional structure. In general, successful platforms could be considered as small wins (Termeer and Dewulf 2018) whose aggregation could help to increase the visibility, accessibility, and reuse of environmental data.

I am convinced that the ability to access big data on water systems, combine these data with modeling, and update models (i.e., data assimilation) will dramatically expand our understanding of these systems and provide a robust basis for real-time prediction and systems control and/or management. The water sector is well-known for its long time horizons (i.e., accompanying major infrastructure investments) and consequent inflexibility. The ability to monitor and model water systems more accurately and respond more quickly to observed changes could provide a basis for adaptive management. Allowing for more variance in water systems could help to improve their resilience (Carpenter et al. 2015). The effective use of big data could also provide the basis for balancing trade-offs in integrated land and water management (Davis et al. 2015) and for adaptive management in the restoration of aquatic ecosystems (Geist and Hawkins 2016). Big data offer an exciting opportunity to make our management of water systems more sustainable. As the capstone of my professional journey through the evolving landscape of data science, I hope to foster the cooperation and focus on outcomes and impacts that will be needed to realize this promise.

Acknowledgments

I am, by no means, an expert in this field and am obviously much too old to be considered a digital native. For their constructive comments on this manuscript, I would like to thank my younger and/or more expert colleagues at Eawag (additional affiliations in parentheses): Carlo Albert, Florian Altermatt (University of Zurich), Damien Bouffard, Juan Pablo Carbajal, Francesco Pomati, Peter Reichert, Nele Schuwirth, Jonas Šukys, Kris Villez, and A. Johny Wüest (EPFL). I also thank Miguel Mahecha (MPI Biogeochemistry) and two anonymous reviewers for their helpful comments.

References

Allen, G. H., and T. M. Pavelsky. 2018. "Global extent of rivers and streams." *Science* 361 (6402): 585–588. https://doi.org/10.1126/science.aat0636.

Alley, G. 2018. "What is a data pipeline?" Accessed December 12, 2018. https://www.alooma.com/blog/what-is-a-data-pipeline.

Basco-Carrera, L., A. Warren, E. van Beek, A. Jonoski, and A. Giardino. 2017. "Collaborative modelling or participatory modelling? A framework for water resources management." *Environ. Modell. Software* 91 (May): 95–110. https://doi.org/10.1016/j.envsoft.2017.01.014.

Bouffard, D., I. Kiefer, A. Wuest, S. Wunderle, and D. Odermatt. 2018. "Are surface temperature and chlorophyll in a large deep lake related? An analysis based on satellite observations in synergy with hydrodynamic modelling and in-situ data." *Remote Sens. Environ.* 209 (May): 510–523. https://doi.org/10.1016/j.rse.2018.02.056.

Broadhurst, D. I., and D. B. Kell. 2006. "Statistical strategies for avoiding false discoveries in metabolomics and related experiments." *Metabolomics* 2 (4): 171–196. https://doi.org/10.1007/s11306-006-0037-z.

Bueno de la Fuente, G. n.d. "What is open science? Introduction." Accessed December 13, 2008. https://www.fosteropenscience.eu/content/what-open-science-introduction.

Burthe, S. J., et al. 2016. "Do early warning indicators consistently predict nonlinear change in long-term ecological data?" *J. Appl. Ecol.* 53 (3): 666–676. https://doi.org/10.1111/1365-2664.12519.

Bush, A., et al. 2017. "Connecting Earth observation to high-throughput biodiversity data." *Nat. Ecol. Evol.* 1 (7): 0176. https://doi.org/10.1038/s41559-017-0176.

Carpenter, S. R., W. A. Brock, C. Folke, E. H. van Nes, and M. Scheffer. 2015. "Allowing variance may enlarge the safe operating space for exploited ecosystems." *Proc. Natl. Acad. Sci. U.S.A.* 112 (46): 14384–14389. https://doi.org/10.1073/pnas.1511804112.

Chao, B. F., Y. H. Wu, and Y. S. Li. 2008. "Impact of artificial reservoir water impoundment on global sea level." *Science* 320 (5873): 212–214. https://doi.org/10.1126/science.1154580.

Corominas, L., M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortes, and M. Poch. 2018. "Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques." *Environ. Modell. Software* 106 (Aug): 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023.

Cox, D. R. 2015. "Big data and precision." *Biometrika* 102 (3): 712–716. https://doi.org/10.1093/biomet/asv033.

CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science). 2019. "*Data portals*." Accessed May 3, 2019. https://www.cuahsi.org/data-models/portals/.

Dafforn, K. A., E. L. Johnston, A. Ferguson, C. L. Humphrey, W. Monk, S. J. Nichols, S. L. Simpson, M. G. Tulbure, and D. J. Baird. 2016. "Big data opportunities and challenges for assessing multiple stressors across

scales in aquatic ecosystems." *Mar. Freshwater Res.* 67 (4): 393–413. https://doi.org/10.1071/MF15108.

Davis, J., et al. 2015. "When trends intersect: The challenge of protecting freshwater ecosystems under multiple land use and hydrological intensification scenarios." *Sci. Total Environ.* 534 (Nov): 65–78. https://doi.org/10.1016/j.scitotenv.2015.03.127.

Deplazes-Zemp, A., S. Abiven, P. Schaber, M. Schaepman, G. Schaepman-Strub, B. Schmid, K. K. Shimizu, and F. Altermatt. 2018. "The Nagoya Protocol could backfire on the global South." *Nat. Ecol. Evol.* 2 (6): 917–919. https://doi.org/10.1038/s41559-018-0561-z.

Doell, P., K. Fiedler, and J. Zhang. 2009. "Global-scale analysis of river flow alterations due to water withdrawals and reservoirs." *Hydrol. Earth Syst. Sci.* 13 (12): 2413–2432. https://doi.org/10.5194/hess-13-2413-2009.

Dornelas, M., et al. 2018. "BioTIME: A database of biodiversity time series for the Anthropocene." *Global Ecol. Biogeogr.* 27 (7): 760–786. https://doi.org/10.1111/geb.12729.

Durden, J. M., J. Y. Luo, H. Alexander, A. M. Flanagan, and L. Grossmann. 2017. "Integrating "big data" into aquatic ecology: Challenges and opportunities." *Limnol. Oceanogr. Bull.* 26 (4): 101–108. https://doi.org/10.1002/lob.10213.

EPFL IRGC (Ecole polytechnique fédérale de Lausanne International Risk Governance Center). 2018. "The governance of decision-making algorithms." Accessed December 13, 2018. https://irgc.epfl.ch/wp-content/uploads/2018/12/IRGC-2018-The-Governance-of-Decision-Making-Algorithms-Workshop-report.pdf.

Fan, H., D. M. He, and H. L. Wang. 2015. "Environmental consequences of damming the mainstream Lancang-Mekong River: A review." *Earth-Sci. Rev* 146 (Jul): 77–91. https://doi.org/10.1016/j.earscirev.2015.03.007.

Farley, S. S., A. Dawson, S. J. Goring, and J. W. Williams. 2018. "Situating ecology as a big-data science: Current advances, challenges, and solutions." *Bioscience* 68 (8): 563–576. https://doi.org/10.1093/biosci/biy068.

Gaudard, A., L. Råman Vinnå, F. Bärenbold, M. Schmid, and D. Bouffard. 2019. "Toward an open-access of high-frequency lake modelling and statistics data for scientists and practitioners. The case of Swiss Lakes using Simstrat v2.1." *Geosci. Model Dev. Discuss*. https://doi.org/10.5194/gmd-2018-336.

GBIF (Global Biodiversity Information Facility). n.d. "An alliance for biodiversity knowledge." Accessed May 8, 2019. https://www.biodiversityinformatics.org/.

Geist, J., and S. J. Hawkins. 2016. "Habitat recovery and restoration in aquatic ecosystems: Current progress and future challenges." *Aquat. Conserv. Mar. Freshwater Ecosyst.* 26 (5): 942–962. https://doi.org/10.1002/aqc.2702.

Gewin, V. 2016. "Data sharing: An open mind on open data." *Nature* 529 (7584): 117–119. https://doi.org/10.1038/nj7584-117a.

Gligorov, V. V. 2015. "Real-time data analysis at the LHC: Present and future." Accessed May 8, 2019. http://proceedings.mlr.press/v42/glig14.pdf.

Griffin, R. E. 2015. "When are old data new data?" *Geo. Res. J.* 6 (Jun): 92–97. https://doi.org/10.1016/j.grj.2015.02.004.

Grill, G., B. Lehner, A. E. Lumsdon, G. K. MacDonald, C. Zarfl, and C. R. Liermann. 2015. "An index-based framework for assessing patterns and trends in river fragmentation and flow regulation by global dams at multiple scales." *Environ. Res. Lett.* 10 (1): 015001. https://doi.org/10.1088/1748-9326/10/1/015001.

Hering, J. G., L. Nunnemacher, and H. von Waldow. 2018. "Perspectives from a water research institute on knowledge management for sustainable water management." In *Handbook of knowledge management for sustainable water systems*, 13–33. Edited by M. Russ. New York: Wiley.

Hering, J. G., and K. Vairavamoorthy. 2018. "Harvesting experience to support sustainable urban water management." In *Assessing global water megatrends*, 61–75. Edited by A. K. Biswas, C. Tortajada, and P. Rohner. Berlin: Springer.

Hutton, C., T. Wagener, J. Freer, D. Han, C. Duffy, and B. Arheimer. 2016. "Most computational hydrology is not reproducible, so is it really science?" *Water Resour. Res.* 52 (10): 7548–7555. https://doi.org/10.1002/2016WR019285.

Ilie, I., et al. 2017. "Reverse engineering model structures for soil and ecosystem respiration: The potential of gene expression programming." *Geosci. Model Dev.* 10 (9): 3519–3545. https://doi.org/10.5194/gmd-10-3519-2017.

Ingildsen, P., and G. Olsson. 2016. *Smart water utilities: Complexity made simple*, 304. London: IWA Publishing.

Kondolf, G. M., Z. K. Rubin, and J. T. Minear. 2014. "Dams on the Mekong: Cumulative sediment starvation." *Water Resour. Res.* 50 (6): 5158–5169. https://doi.org/10.1002/2013WR014651.

Langsdale, S., A. Beall, E. Bourget, E. Hagen, S. Kudlas, R. Palmer, D. Tate, and W. Werick. 2013. "Collaborative modeling for decision support in water resources: Principles and best practices." *J. Am. Water Resour. Assoc.* 49 (3): 629–638. https://doi.org/10.1111/jawr.12065.

Lehner, B., et al. 2011. "High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management." *Front. Ecol. Environ.* 9 (9): 494–502. https://doi.org/10.1890/100125.

Li, J., K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson. 2018. "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images." *IEEE Rob. Autom. Lett.* 3 (1): 387–394. https://doi.org/10.1109/lra.2017.2730363.

Maavara, T., C. T. Parsons, C. Ridenour, S. Stojanovic, H. H. Durr, H. R. Powley, and P. Van Cappellen. 2015. "Global phosphorus retention by river damming." *Proc. Natl. Acad. Sci. U.S.A.* 112 (51): 15603–15608. https://doi.org/10.1073/pnas.1511797112.

Matthews, M. W., and D. Odermatt. 2015. "Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters." *Remote Sens. Environ.* 156 (Jan): 374–382. https://doi.org/10.1016/j.rse.2014.10.010.

McCabe, M. F., et al. 2017. "The future of Earth observation in hydrology." *Hydrol. Earth Syst. Sci.* 21 (7): 3879–3914. https://doi.org/10.5194/hess-21-3879-2017.

McGowan, J. A., E. R. Deyle, H. Ye, M. L. Carter, C. T. Perretti, K. D. Seger, A. de Verneil, and G. Sugihara. 2017. "Predicting coastal algal blooms in southern California." *Ecology* 98 (5): 1419–1433. https://doi.org/10.1002/ecy.1804.

Odermatt, D., O. Danne, P. Philipson, and C. Brockmann. 2018. "Diversity II water quality parameters from ENVISAT (2002-2012): A new global information source for lakes." *Earth Syst. Sci. Data* 10 (3): 1527–1549. https://doi.org/10.5194/essd-10-1527-2018.

Parrott, L. 2017. "The modelling spiral for solving 'wicked' environmental problems: Guidance for stakeholder involvement and collaborative model development." *Methods Ecol. Evol.* 8 (8): 1005–1011. https://doi.org/10.1111/2041-210X.12757.

Paull, C. K., et al. 2018. "Powerful turbidity currents driven by dense basal layers." *Nat. Commun.* 9 (1): 4114. https://doi.org/10.1038/s41467-018-06254-6.

Pearl, J. 2018. "Theoretical impediments to machine learning with seven sparks from the causal revolution." Preprint, submitted January 11, 2018. https://arxiv.org/abs/1801.04016.

Pennekamp, F., et al. 2018. "Biodiversity increases and decreases ecosystem stability." *Nature* 563 (7729): 109–112. https://doi.org/10.1038/s41586-018-0627-8.

Pennekamp, F., J. I. Griffiths, E. A. Fronhofer, A. Garnier, M. Seymour, F. Altermatt, and O. L. Petchey. 2017. "Dynamic species classification of microorganisms across time, abiotic and biotic environments-A sliding window approach." *PLoS ONE* 12 (5): e0176682. https://doi.org/10.1371/journal.pone.0176682.

Peters, D. P. C., et al. 2018. "An integrated view of complex landscapes: A big data-model integration approach to transdisciplinary science." *Bioscience* 68 (9): 653–669. https://doi.org/10.1093/biosci/biy069.

Peters, D. P. C., K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales. 2014. "Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology." *Ecosphere* 5 (6): 1–15. https://doi.org/10.1890/ES13-00359.1.

Pinho, A. J., D. Pratas, and S. P. Garcia. 2012. "GReEn: A tool for efficient compression of genome resequencing data." *Nucleic Acids Res.* 40 (4): e27. https://doi.org/10.1093/nar/gkr1124.

Ramaswami, A., A. G. Russell, P. J. Culligan, K. R. Sharma, and E. Kumar. 2016. "Meta-principles for developing smart, sustainable, and healthy cities." *Science* 352 (6288): 940–943. https://doi.org/10.1126/science.aaf7160.

RDA (Research Data Alliance). n.d. "Research Data." Accessed May 8, 2019. https://www.rd-alliance.org/.

SDSC (Swiss Data Science Center). n.d. "Data Science." Accessed May 8, 2019. https://datascience.ch/.

Shen, C. 2018. "A transdisciplinary review of deep learning research and its relevance for water resources scientists." *Water Resour. Res.* 54 (11): 8558–8593. https://doi.org/10.1029/2018WR022643.

Shen, C., et al. 2018. "HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community." *Hydrol. Earth Syst. Sci.* 22 (11): 5639–5656. https://doi.org/10.5194/hess-22-5639-2018.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. "Conditional variable importance for random forests." *BMC Bioinf.* 9 (1): 307. https://doi.org/10.1186/1471-2105-9-307.

Sugihara, G., R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. 2012. "Detecting causality in complex ecosystems." *Science* 338 (6106): 496–500. https://doi.org/10.1126/science.1227079.

Taormina, R., and S. Galelli. 2018. "Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems." *J. Water Resour. Plann. Manage.* 144 (10): 04018065. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000983.

Taormina, R., S. Galelli, N. O. Tippenhauer, E. Salomons, and A. Ostfeld. 2017. "Characterizing cyber-physical attacks on water distribution systems." *J. Water Resour. Plann. Manage.* 143 (5): 04017009. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000749.

Termeer, C. J. A. M., and A. Dewulf. 2018. "A small wins framework to overcome the evaluation paradox of governing wicked problems." *Policy Soc.* 1–17. https://doi.org/10.1080/14494035.2018.1497933.

Thomas, M. K., S. Fontana, M. Reyes, M. Kehoe, and F. Pomati. 2018a. "The predictability of a lake phytoplankton community, over time-scales of hours to years." *Ecol. Lett.* 21 (5): 619–628. https://doi.org/10.1111/ele.12927.

Thomas, M. K., S. Fontana, M. Reyes, and F. Pomati. 2018b. "Quantifying cell densities and biovolumes of phytoplankton communities and functional groups using scanning flow cytometry, machine learning and unsupervised clustering." *PLoS ONE* 13 (5): e0196225. https://doi.org/10.1371/journal.pone.0196225.

USGS. n.d. "Data at risk project." Accessed December 13, 2008. https://apps.usgs.gov/ldi/data-at-risk-project.

Wilkinson, M. D., et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Sci. Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Woodward, C., J. Shulmeister, J. Larsen, G. E. Jacobsen, and A. Zawadzki. 2014. "The hydrological legacy of deforestation on global wetlands." *Science* 346 (6211): 844–847. https://doi.org/10.1126/science.1260510.

Ye, H., and G. Sugihara. 2016. "Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality." *Science* 353 (6302): 922–925. https://doi.org/10.1126/science.aag0863.

Zarfl, C., A. E. Lumsdon, J. Berlekamp, L. Tydecks, and K. Tockner. 2015. "A global boom in hydropower dam construction." *Aquat. Sci.* 77 (1): 161–170. https://doi.org/10.1007/s00027-014-0377-0.

Zhang, D., N. Martinez, G. Lindholm, and H. Ratnaweera. 2018. "Manage sewer in-line storage control using hydraulic model and recurrent neural network." *Water Resour. Manage.* 32 (6): 2079–2098. https://doi.org/10.1007/s11269-018-1919-3.