



Chen, Y., & Han, D. (2016). On Big Data and Hydroinformatics: 12th International Conference on Hydroinformatics (HIC 2016) - Smart Water for the Future. *Procedia Engineering*, 154, 184-191.
<https://doi.org/10.1016/j.proeng.2016.07.443>

Publisher's PDF, also known as Version of record

License (if available):
CC BY-NC-ND

Link to published version (if available):
[10.1016/j.proeng.2016.07.443](https://doi.org/10.1016/j.proeng.2016.07.443)

[Link to publication record on the Bristol Research Portal](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via Elsevier at <http://www.sciencedirect.com/science/article/pii/S187770581631832X>. Please refer to any applicable terms of use of the publisher.

University of Bristol – Bristol Research Portal

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/>

12th International Conference on Hydroinformatics, HIC 2016

On big data and hydroinformatics

Yiheng Chen^a, Dawei Han^{a*}^aWater and Environment Management Research Centre, Department of Civil Engineering, University of Bristol, Bristol BS8 1TR, UK

Abstract

Big data is an increasingly hot concept in the past five years in the area of computer science, e-commerce, and bioinformatics, because more and more data has been collected by the internet, remote sensing network, wearable devices and the Internet of Things. The big data technology provides techniques and analytical tools to handle large datasets, so that creative ideas and new values can be extracted from them. However, the hydroinformatics research community are not so familiar with big data. This paper provides readers who are embracing the data-rich era with a timely review on big data and its relevant technology, and then points out the relevance with hydroinformatics in three aspects.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the organizing committee of HIC 2016

Keywords: Big data; Hydroinformatics

1. Introduction

The inevitable trend of big data along with the growing capability to handle huge datasets is reshaping how we understand the world. The International Data Corporation (IDC) report has estimated that the data size of the world will grow from 130 exabytes (10¹⁸ bytes) in 2005 to 40 zettabytes (10²¹ bytes) in 2020, at a 40% annual increase [1]. New datasets are continuously being collected from the internet, the Internet of Things, the remote sensing network and e-commerce, wearable devices, etc. Unfortunately, only 3% of all data is properly tagged and ready for use, and only 0.5% of data is analyzed, which yields a large potential market for data utilization [2]. A famous early

* Corresponding author. Tel.: +44 117 9289768; fax: +44 117 9289770.

E-mail address: yiheng.chen@bristol.ac.uk

attempt of big data application was Google Flu Trend (GFT) that monitored health-seeking behavior in the form of online web search queries by millions of users around the world every day. The methodology was to find the best matches among 50 million search terms to fit 1152 flu data points from Central Disease Control. GFT estimated the level of weekly influenza activity with a one day reporting lag, much shorter than the Central Disease Control with two week reporting lag [3]. GFT predicted the influenza activity responding much faster than CDC, but suffered its problematic performance. In 2009, its poor underestimation of the influenza-like illness in the United States of the swine flu pandemic forced Google to modify its algorithm. GFT overestimated flu prevalence in 100 out of 108 weeks from 21 August 2011 to 1 September 2013 [4]. In December 2012, it overestimated more than double the doctor visits for influenza-like illness (ILI) than the Central Disease Control [5]. Google stopped publishing flu trend data and started to pass the data to specialized organizations to empower their research in summer 2015 [6]. Another application of big data is precision marketing, i.e. the online movie subscription rental service provider Netflix has its recommendation system based on hundreds millions of accumulated anonymous movie ratings to improve the probability that the users rent the movies recommended by Netflix [7].

Although the popularity of big data is related with its commercial value, we believe that the idea of big data can benefit the hydroinformatics research for multiple reasons. First, the big data analysis encourages the utilization of multiple datasets from various sources to discover the big trend. Secondly, the computing tools developed for the big data analysis, e.g. parallel computing and distributed data storage, can help tackle the data-intensive jobs in the field of hydroinformatics. Thirdly, the novel correlation found by mining various large datasets has the potential to lead to new scientific exploration. Apart from the companies in the internet industry working closely with the data from the internet, the scientists have collected substantial amount of data for hydrology, meteorology and earth observation with a history much longer than that of the internet. The development of internet and the movement of open data significantly accelerate the data sharing and improve the accessibility of the archived data. The hydroinformatics community will benefit from the active combination of a huge amount of data and the data processing technologies for knowledge discovery and management. Precipitation is one important part of the water cycle in hydrology. The accumulated precipitation datasets from heterogeneous sources, e.g., rain gauges, weather radars, satellite remote sensing and numerical weather models, have reached tens of terabytes in size, with different characteristics, i.e., spatial and temporal coverage, resolution, and uncertainties. Data fusion is a possible method to utilize the accumulated datasets to produce a better result with enhanced resolution and minimized uncertainty.

This paper aims to provide readers who are not so familiar to big data with a timely review on its concept and the relevant technology, starting from the explanation of the concept of big data, then introduction to the popular Apache Hadoop family to handle large amount of data. After that, the relevance of big data with hydroinformatics is explained in three dimensions, the natural dimension, the social dimension and the business dimension, for the purpose of encouraging more researchers in the hydroinformatics community to attempt novel research based on big data.

2. Big data and the relevant technology

2.1. Make the concept clear

The fashionable term of ‘Big Data’ is sometimes so hot that many people attempt to embrace it in this data-rich era without a clear understanding. The concept of big data originated from the extreme large datasets that have been collected but cannot be processed in tolerable elapsed time with traditional data processing methods. The term ‘big data’ is simple but its meaning is ambiguous. It is commonly used to describe data sets with quantity and complexity beyond the capacity of normal computing tools to capture, curate, manage, and process with a tolerable speed [8]. Another explanation of Big Data refers to developing new insights or creating new values at a large scale instead of a smaller one [9]. A formal definition of big data is the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value, based on investigation of 14 existing definitions of big data [10]. This definition can be subdivided into three groups: the characteristics of the data sets, the specific technologies and analytical methods to manipulate the data, and the ideas

to extracts insights from the data and creation of new values. Therefore, big data is not just about massive amounts of data. In general, the goal of big data analysis is knowledge discovery from massive data sets, which is a challenging systematic problem. The data analysis systems should utilize the existing hardware platform with distributed and parallel computing, accommodate a variety of data formats, models, loss functions and methods, be highly customizable for users to specify their data analysis goals through an expressive but simple language, provide useful visualizations of key components of the analysis, communicate with other computational platforms seamlessly, and provide many of capabilities familiar from large-scale databases [11].

2.2. *The MapReduce parallel computing*

The MapReduce parallel computing is the new computing model featuring parallel data processing to speed-up the data I/O efficiency, developed in the big data era. The motivation of such a computing method is that more emphasis has been put on data I/O apart from the computing process itself. The concern is whether the existing computing system can handle the increasingly large data within tolerable time. The data storage capacity increased dramatically in the past decades. In 2014, Western Digital shipped the 8 TB hard drive and announced the world first 10 TB hard drive [12]. The unit cost of data storage will drop down from \$2.00 per GB to \$0.20 per GB from 2012 to 2020 [1]. The storage of data should no longer be a big problem owing to the massive storage technologies such as Direct Attached Storage (DAS), Network Attached Storage (NAS) and Storage Area Network (SAN), as well as the cloud data storage. However, the I/O speed of the hard disk grows slowly limited by the hard disk mechanism. Solid state disk (SSD) has a much higher I/O rate and negligible seek time, in the meantime, the cost per unit storage is much higher than that of the hard disk. Regardless of the cost, the SSD has a lower storage capacity of single device. The I/O speed of the data storage devices is the bottleneck of extreme large data processing rather than the data storage capacity.

An appropriate software system is essential to dealing with extremely large datasets apart from the development of the hardware system. As the improvement of I/O speed of the hardware system did not catch the speed of the expansion of data storage, the time required to process the data dramatically increased without an appropriate algorithm. The parallel computing and distributed storage were developed to encounter this issue. MapReduce is a distributed programming model for processing and generating large datasets developed by Google. The idea of MapReduce is to specify a Map and a Reduce function which are suitable for parallel computing, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks. As the size of datasets is extremely large for big data problems, a cluster of machines connected in a network are used to overcome the limit of computing power and data storage of a single machine, but the network bandwidth becomes the bottleneck as it is a rare resource. Thus, the MapReduce system is optimized targeting at reducing the data transfer across the network through sending the code to the local machine and writing the intermediate data to local disk. The MapReduce system minimized the impact of slow machines, and can handle machine failures and data loss by redundant execution [13]. The Hadoop is an open-source version of the MapReduce framework developed by Apache, freely available for scientific community. The Hadoop contains the Hadoop Distributed File System (HDFS) working together with MapReduce after Google published the technical details of the Google File System [14], apart from which the Apache Hadoop also contains Hadoop Common, the common utilities that support the other Hadoop modules; and Hadoop YARN, a framework for job scheduling and cluster resource management. There are many other projects in Apache which are related to Hadoop, including HBase (a scalable, distributed database that supports structured data storage for large tables), Hive (a data warehouse infrastructure that provides data summarization and ad hoc querying), Mahout (a scalable machine learning and data mining library), Pig (a high-level data-flow language and execution framework for parallel computation) and ZooKeeper (a high-performance coordination service for distributed applications), etc.[15].

Hadoop MapReduce has a weakness during iterative data analysis that the intermittent datasets are stored on the local hard disk. As the iterative data analysis requires multiple read and write of local intermittent data, which will

dramatically slow down the analysis. This happens to most machine learning algorithms, e.g., gradient decent. Apache Spark is the latest programming model in the big data world featuring its lightning fast data processing speed for iterative jobs [16]. The Spark achieved its lightning fast speed by the implementing Resilient Distributed Datasets (RDDs), a distributed memory abstraction that lets the programmers perform in-memory computation [17]. The Spark outperforms Hadoop by 20 times in speed by utilizing the RAM instead of hard disk to store the intermittent data.

3. Relevance to hydroinformatics

Hydroinformatics, originated from the computational hydraulics, comprises the application of information and communications technologies (ICTs) to the understanding and management of the waters of the world [18], addressing the increasingly serious problems of the equitable and efficient use of water for different purposes. Once the term hydroinformatics was defined, it meant to integrate artificial intelligence to the numerical simulation and modelling, and to shift the computational-intensive analysis to information-based research. The two main lines of hydroinformatics, data mining for knowledge discovery and knowledge management [19], are strongly dependent on information of which data, both textual or non-textual, is the major carrier. Data from smart meters, smart sensors and smart services, remote sensing, earth observation systems, etc., will prompt hydroinformatics into the inevitable big data era. The challenge of big data and data mining for environmental projects is the most pressing one in the near future [20]. In general, the water-related problems are quite complex due to the lurking interrelationships between water-related environmental, social and business factors. The data being generated and collected relevant to hydroinformatics features huge volumes and multiple types. For the purpose of simplification, the data sources for the hydroinformatics, without loss of generality, can be classified into three dimensions, i.e., the natural dimension, the social dimension, and the business dimension.

3.1. The natural dimension

The natural dimension is about water as one important component of the natural environment. Understanding the water cycle, the temporal and spatial distribution of water and the interaction of water and the environment is part of the objectives of hydroinformatics for improving the water resource management, flood and drought management. The water-related data includes the measurements of precipitation (rainfall, snow and hail), river flow, water quality, soil moisture, soil characteristic, ground water condition, air temperature and humidity, solar flux, etc. The observation methods developed from local station for point measurement to remote sensing - radar and satellites, and drone. The earth observation satellites are generating huge volume of data including weather and water-related information. ESA has launched SMOS for soil moisture observation in 2009, and will launch ADM-Aeolus for Atmospheric Dynamics observation in 2017 [21]. NASA launched SMAP to map soil moisture and determine the freeze or thaw state in 2015 [22]. The GPM mission launched in 2015 aims to provide global rain and snow observation based upon the success of TRMM launched in 1997 [23]. EUMETSAT has two generations of active METEOSAT satellites in geostationary orbit and a series of three polar orbiting METOP satellites for weather nowcasting and forecasting and understanding the climate change. Without doubt, the increasing amount of earth observation data, including precipitation, soil moisture and wind speed etc., will improve the understanding of the global water cycle, and benefit the weather forecasting, flood and drought prediction. Unfortunately, although many satellites were launched or to be launched, the huge amount of available data is rarely used, only three to five percent of data is used on daily average, while billions of dollars have been invested annually [24]. Apart from the earth observation data, reanalysis data is another important information source with high data quality. In other words, the information source is not limited to the observation of the current situation and the archived past situation, the model generated data cannot be neglected. Reanalysis of archived observations is achieved by combining advanced forecast models and data assimilation systems to create global data sets of the atmosphere, land surface, and oceans, as an operational analysis dataset will suffer from inconsistency due to the frequent improvements of the forecast models. The NCEP Climate Forecast System Reanalysis includes over 80 variables, goes back to 1948 and is continuing [25]. ECMWF has series of ERA projects for global atmospheric reanalysis tracing back to 1957 [26]. The Japan Meteorological Agency conducted the JRA-55 project for a high-quality homogeneous climate dataset

covering the last half century [27]. The model generated data is four dimensional, three dimensions in space and one in time, and of high spatial and temporal coverage and resolution, resulting in huge volume of data, which means the hydroinformatics is entering a data-intensive era. Utilization of the currently available data is challenging due to the uncertainties of the data, the challenges of processing and the lack of ideas of data utilization. In the big data era, it is encouraged to make the best of the huge amount of data with tolerance of the uncertainties. The processing of large amount of datasets is becoming easier with the development of computing tools. The lack of creative ideas is the main limitation of the utilization of data. A frontier application example is a prototype software that automatically finds ideal location for hydro-power based on over 30 freely remote sensing and environmental datasets in UK [28].

3.2. The social dimension

The social dimension is about the interaction of water environment and the human society. With the digitalization of textual information available online and the explosion of social media, textual mining technologies enable the new research area of the public attitude towards certain issue. For instance, 5 million scientific articles have been analyzed to explore the impact of the Fukushima disaster on the media attitude towards nuclear power [29]. Similar ideas can be migrated to discover water-related issues, e.g., the social attitude towards climate change, water saving, water policy, etc. Apart from the discovery of public attitude, the internet is logging the activities of internet users, which can be potentially valuable for discover real world situations demonstrated by the example of Google Flu Trend mentioned in the previous section. The Twitter data is now attracting many researchers to dig into for water environment related research. It was found that Twitter content could infer daily rainfall rates in five UK cities, which revealed the online textual features in Twitter were strongly related to the topic with significant inference [30]. Two Dutch organizations, Deltares and Floodtags, have developed a real-time flood-extent maps based on tweets about floods for Jakarta, Indonesia [31]. This method gives the disaster management a real-time view of the situation with a wide coverage. The enrichment of the new media data on the internet enables a new model for scientific research. The new model gathers information from what the internet users post online. The users are actually acting a role of information collector, and they deposit the information about what they observe about the environment to the internet. The internet is like a boundless ocean of data that records how the internet users interact with the internet. The data ocean has a valuable potential for scientists to discover novel correlations between real world situations. The fundamental data mining techniques behind the big data application, such as Google Flu Trend, estimating precipitation from Twitter, etc., are the same, i.e. to dig out the correlation between the information and the targeted result. The distinction of these analyses is that the social network data application is based on people's mental reaction to certain events while the nature scientific research is mainly based on the physically interpretable model. As the behavior of people is ambiguous to interpret and predict, the big data analysis of social network data is dominated by the machine learning or statistical approaches.

3.3. The business dimension

The business dimension covers but not limited to water extraction, water treatment, water supply, waste water collection and treatment. IBM has been a pioneer in utilizing data and computing tools collaboration with NOAA to explore the business of weather. They built one of the first parallel processing supercomputers for weather modelling in 1995, named as Deep Thunder Project. Deep Thunder creates 24- to 48-hour forecasts at 1 - 2 km resolution with a lead time of three hours to three days and combines with other data customized for business purposes such as to help a utility company prepare for the after effects of a major storm or to help airlines and airports manage the weather-generated delays by rearranging or combining flights more efficiently [32]. Another possibility is that, as inspired by the big data application in e-commerce that utilize the accumulated user activity logs for recommendation system, the smart metering data can be integrated with end-user water consumption data, wireless communication networks and information management systems in order to provide real-time information on how, when and where water is being consumed for the consumer and utility [33]. The information from the combination of data will be valuable to architects, developers and planners, seeking to understand water

consumption patterns for future water planning. Smarter metering is one example of the ambitious idea of the Internet of Things as a global infrastructure for the information society, enabling advanced services by interconnecting things based on existing and evolving interoperable information and communication technologies [34]. Furthermore, the operation data collected by companies in the water industry also has potential values for data mining for optimizing the system and providing more information for decision making.

3.4. The trend of open data

The increasing number of openly available data sources will benefit the research community as data is the basic material for data-based research. Open data means data that can be freely used, modified, and shared by anyone for any purpose [35]. Open data is the further development of free data that data is freely licensed for limited purposes and certain users, while closed data is usually restricted by copyright, patents or other mechanisms. The goals of the open data movement are similar to those of other "Open" movements such as open source, open hardware, open content, and open access. The data owner may not have the appropriate ideas and techniques to produce extra values from the data, while, on the other hand, people with innovative ideas and ability of processing the data may find it difficult to find and access the data they need. The open data movement will activate the combination of data, data mining methods and new ideas to create additional values by removing the barrier between the data providers and the data users. Thus, the research data and its products can achieve the full value and accelerate the future research only when being open. Multiple national governments created web sites for the open delivery of their data for transparency and accountability, e.g., Data.gov for the US government, Data.gov.uk for the UK government, European Union Open Data Portal (<http://open-data.europa.eu/>) and Canada's Open Government portal (<http://open.canada.ca/en>) etc. For open data in science, the World Data System (WDS) of the International Council for Science was created based on the legacy of the World Data Centres in 2008 to ensure the universal and equitable access to quality-assured scientific data, data services, products and information. National Climatic Data Center, containing huge amount of environmental, meteorological and climate data sets, is the world's largest archive of weather data. SWITCH-ON is a European project that works towards sustainable use of water resources, a safe society and advancement of hydrological sciences based upon Open Data. The project aims to build the first one-stop shop portal of open data, water information and its users in one place [36]. EarthCube is a project launched in 2011 that develops a common cyberinfrastructure for the purpose of collecting, accessing, analyzing, sharing and visualizing all forms of data and related resources for understanding and predicting a complex and evolving solid Earth, hydrosphere, atmosphere, space environment systems, through the use of advanced technological and computational capabilities [37]. The on-going movement of open data can boost the data-based research and the data usage by removing the legal restriction on the data use. Many data portals are being created for data sharing through web service with much powerful data search tools where users can find data by location, time, and data types, etc.

3.5. Boosts from cloud computing

The tools developed in the big data era, such as Hadoop MapReduce, Apache Spark, can handle extremely large datasets within tolerable runtime, but the knowledge and technique to set up and manage the tools are required. The commercial cloud computing service is available to scientists as an alternative, where data storage and processing can be done in the cloud, such as Microsoft Azure, Amazon Elastic Compute Cloud, Google Compute Engine, Rackspace, Verizon and GoGrid. The commercial cloud has a usage based price policy, making the computing job cost effective than implementing local clusters. The cloud computing is scalable to suit the job, and does not require extensive knowledge on configuring local clusters. US National Oceanic and Atmospheric Administration (NOAA) has launched its Big Data Project collaborating with Amazon Web Service, Google Cloud Platform, IBM, Microsoft, and the Open Cloud Consortium [38]. The NOAA data will be brought to the cloud platform together with big data processing services such as Google BigQuery and Google Cloud Dataflow, to explore, and create new findings. NOAA's Big Data Project indicated a coming trend of combining the tremendous volume of high quality data hold by the government and the industrial vast infrastructure and technical capacity of data management and analysis.

4. Conclusion

The big data era is an upcoming trend that no one can escape from. Scientists are expected to embrace the big data era rationally without being blurred by the overwhelming trend. The concept of big data originated from the popularization of internet as digitalizing of the information among the world becomes much easier and cheaper for future data mining purpose. The commercial value, e.g., precision marketing, data-based decision making, behind the expanding datasets makes the term ‘big data’ extremely trendy. The idea of big data is very adaptable, and can be valuable for academic purpose as well. Hydroinformatics can benefit from the expending amount of data collected, generated and opened to the research community. Data from smart meters, smart sensors and smart services, remote sensing, earth observation systems, Internet of Things, etc., will prompt hydroinformatics into the inevitable big data era. The data usage can be categorized into three dimensions, the natural dimension, analyzing the climate change, flood and drought management and the global water cycle; the social dimension, focusing on the interaction between water environment and the human society; and the business dimension, using data-based decision making system for optimizing the water resource management system and future water planning. The data processing tools like parallel computing, distributed storage have been developed to help users to handle the large datasets in hundreds GBs or TBs in tolerable time to make real-time application possible and interactive human-computer analysis feasible. The cloud computing platforms will make it unnecessary to download the data to local machine or run the model locally but provide superior computing efficiency in the future cloud computing era. The real challenge in the near future is how to make the best use of the available data, as currently there is little done about big data relevant to hydroinformatics. Thus, the purpose of the paper is to encourage the research community to develop new ideas for the big data era.

References

- [1].Gantz, J. and D. Reinsel, *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east*. IDC iView: IDC Analyze the Future, 2012. **2007**: p. 1-16.
- [2].Burn-Murdoch, J. *Study: less than 1% of the world's data is analysed, over 80% is unprotected*. 2012 [cited 2015 June 18]; Available from: <http://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume>.
- [3].Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data*. Nature, 2009. **457**(7232): p. 1012-4.
- [4].Lazer, D., et al., *Big data. The parable of Google Flu: traps in big data analysis*. Science, 2014. **343**(6176): p. 1203-5.
- [5].Butler, D., *When Google got flu wrong*. Nature, 2013. **494**(7436): p. 155.
- [6].Team, T.F.T. *The Next Chapter for Flu Trends*. 2015 August 20, 2015 [cited 2016 March 22]; Available from: <http://googleresearch.blogspot.co.uk/2015/08/the-next-chapter-for-flu-trends.html>.
- [7].Bennett, J. and S. Lanning. *The netflix prize*. in *Proceedings of KDD cup and workshop*. 2007. New York: ACM.
- [8].Snijders, C., U. Matzat, and U.-D. Reips, *Big data: Big gaps of knowledge in the field of internet science*. International Journal of Internet Science, 2012. **7**(1): p. 1-5.
- [9].Mayer-Schönberger, V. and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. 2013: Houghton Mifflin Harcourt.
- [10].De Mauro, A., M. Greco, and M. Grimaldi. *What is Big Data? A Consensual Definition and a Review of Key Research Topics*. in *4th International Conference on Integrated Information, Madrid*. doi. 2014.
- [11].Council, N.R., *Frontiers in massive data analysis*. 2013, Washington, D.C.: The National Academies Press.
- [12].Hartin, E. and K. Watson. *Announces New Innovations that Set the Standard for Performance, Reliability, Capacity, Agility and Efficiency for Helping Companies Harness the Power of Data*. HGST Storage 2014 [cited 2015 May 10]; Available from: <http://www.hgst.com/press-room/press-releases/HGST-unveils-intelligent-dynamic-storage-solutions-to-transform-the-data-center>.
- [13].Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters*. Communications of the ACM, 2008. **51**(1): p. 107-113.
- [14].Ghemawat, S., H. Gobioff, and S.-T. Leung. *The Google file system*. in *ACM SIGOPS operating systems review*. 2003. ACM.
- [15].Apache. *What Is Apache Hadoop?* 2015 [cited 2015 June 22]; Available from: <https://hadoop.apache.org/>.
- [16].Zaharia, M., et al. *Spark: cluster computing with working sets*. in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. 2010.
- [17].Zaharia, M., et al. *Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing*. in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. 2012. USENIX Association.
- [18].Abbott, M.B., *Hydroinformatics: information technology and the aquatic environment*. 1991: Averbury Technical.
- [19].Abbott, M., *Introducing Hydroinformatics*. Journal of hydroinformatics, 1999. **1**: p. 3-19.
- [20].Pierson, L. *Civil Engineer Turned Environmental Data Scientist Harnesses Big Environmental Data at UNESCO-IHE*. 2014 [cited 2015 June 15]; Available from: <http://www.statisticsviews.com/details/feature/7136441/Civil-Engineer-Turned-Environmental-Data-Scientist-Harnesses-Big-Environmental-D.html>.
- [21].ESA. *Earth Explorers overview*. 2016 [cited 2016 Feb 26]; Available from: http://www.esa.int/Our_Activities/Observing_the_Earth/Earth_Explorers_overview.
- [22].SMAP. *SMAP Overview*. 2015 [cited 2015 July 5]; Available from: <http://smap.jpl.nasa.gov/observatory/overview/>.

- [23].NASA. *Global Precipitation Measurement (GPM) Mission Overview*. 2011 [cited 2015 May 11]; Available from: <http://pmm.nasa.gov/GPM>.
- [24].Selding, P.B.d. *U.S. Government-leased Satellite Capacity Going Unused*. 2012 [cited 2015 June 20]; Available from: <http://spacenews.com/32581us-government-leased-satellite-capacity-going-unused/>.
- [25].National Centers for Environmental Prediction, N.W.S.N.U.S.D.o.C., *NCEP/NCAR Global Reanalysis Products, 1948-continuing*. 1994, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory: Boulder, CO.
- [26].ECMWF. *ECMWF Climate Reanalysis*. 2015 [cited 2015 July 16]; Available from: <http://www.ecmwf.int/en/research/climate-reanalysis>.
- [27].Kobayashi, S., Y. Ota, and Y. Harada, *The JRA-55 Reanalysis: General Specifications and Basic Characteristics*. Journal of the Meteorological Society of Japan, 2015. **93**(1): p. 5-48.
- [28].Leicester, U.o. *Big data technology finds ideal river locations to generate hydro-power*. 2015 [cited 2015 July 28]; Available from: <http://www.sciencedaily.com/releases/2015/04/150413075144.htm>.
- [29].Lansdall-Welfare, T., et al. *On the coverage of science in the media: A big data study on the impact of the Fukushima disaster*. in *Big Data (Big Data), 2014 IEEE International Conference on*. 2014. IEEE.
- [30].Lampos, V. and N. Cristianini, *Nowcasting events from the social web with statistical learning*. ACM Transactions on Intelligent Systems and Technology (TIST), 2012. **3**(4): p. 72.
- [31].Eilander, D. *Twitter used to create real-time flood maps*. Deltares 2015 [cited 2015 April 27]; Available from: <https://www.deltares.nl/en/news/twitter-used-to-create-real-time-flood-maps/>.
- [32].IBM. *IBM100 - Deep Thunder*. 2015 [cited 2015 14 Dec]; Available from: <http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deeptunder/>.
- [33].Stewart, R.A., et al., *Web-based knowledge management system: linking smart metering to the future of urban water planning*. Australian Planner, 2010. **47**(2): p. 66-74.
- [34].ITU. *Internet of Things Global Standards Initiative*. 2015 [cited 2015 July 28]; Available from: <http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx>.
- [35].Opendefinition. *Defining Open in Open Data, Open Content and Open Knowledge*. 2015 [cited 2015 July 5]; Available from: <http://opendefinition.org/od/>.
- [36].SWITCH-ON. *About SWITCH-ON*. 2015 [cited 2015 June 15]; Available from: <http://www.project.water-switch-on.eu/>.
- [37].EarthCube. *About EarthCube*. 2015 [cited 2015 July 15]; Available from: <http://earthcube.org/info/about>.
- [38].Commerce, D.o. *U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data*. 2015 [cited 2015 02 Dec]; Available from: <https://www.commerce.gov/news/press-releases/2015/04/us-secretary-commerce-penny-pritzker-announces-new-collaboration-unleash>.