Big data and hydroinformatics

Yiheng Chen and Dawei Han

ABSTRACT

Big data is popular in the areas of computer science, commerce and bioinformatics, but is in an early stage in hydroinformatics. Big data is originated from the extremely large datasets that cannot be processed in tolerable elapsed time with the traditional data processing methods. Using the analogy from the object-oriented programming, big data should be considered as objects encompassing the data, its characteristics and the processing methods. Hydroinformatics can benefit from the big data technology with newly emerged data, techniques and analytical tools to handle large datasets, from which creative ideas and new values could be mined. This paper provides a timely review on big data with its relevance to hydroinformatics. A further exploration on precipitation big data is discussed because estimation of precipitation is an important part of hydrology for managing floods and droughts, and understanding the global water cycle. It is promising that fusion of precipitation data from remote sensing, weather radar, rain gauge and numerical weather modelling could be achieved by parallel computing and distributed data storage, which will trigger a leap in precipitation estimation as the available data from multiple sources could be fused to generate a better product than those from single sources.

Key words | big data, data fusion, hydroinformatics, precipitation estimates

Yiheng Chen (corresponding author)
Dawei Han
Water and Environment Management Research Centre, Department of Civil Engineering,
University of Bristol,
Bristol BS8 1TR,
UK
E-mail: *yiheng.chen@bristol.ac.uk*

INTRODUCTION

The inevitable trend of big data along with the growing capability to handle huge datasets is reshaping how we understand the world. According to Google Scholar, the number of publications containing the phrase 'big data' in the title and the number of publications about big data and water are shown in Figure 1, revealing that the interest in big data has dramatically risen since 2010; however, the research on big data in hydroinformatics is still at a very early stage. This is a very simple example of the so-called big data analysis, as the result is based on searching a vast number of academic publications powered by Google Scholar. Google Scholar indexed academic publications provide internet users with a very efficient way to find academic publications. The value of the online search engine is its lightning fast speed which enables the user to get the result from the ocean of online information in merely milliseconds. 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 of millions of accumulated anonymous movie ratings to improve the probability that users rent the movies recommended by Netflix (Bennett & Lanning 2007).

Although the popularity of big data is related to its commercial value, we believe that the idea of big data can benefit hydroinformatics research for multiple reasons. First, the big data analysis encourages the utilization of multiple datasets from various sources to discover the big trends. Second, the computing tools developed for big data analysis, e.g., parallel computing and distributed data storage, can help tackle data-intensive jobs in the field of hydroinformatics. Third, 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, scientists have collected substantial amounts of data for hydrology, meteorology and earth observation with a history much longer than that of

doi: 10.2166/hydro.2016.180

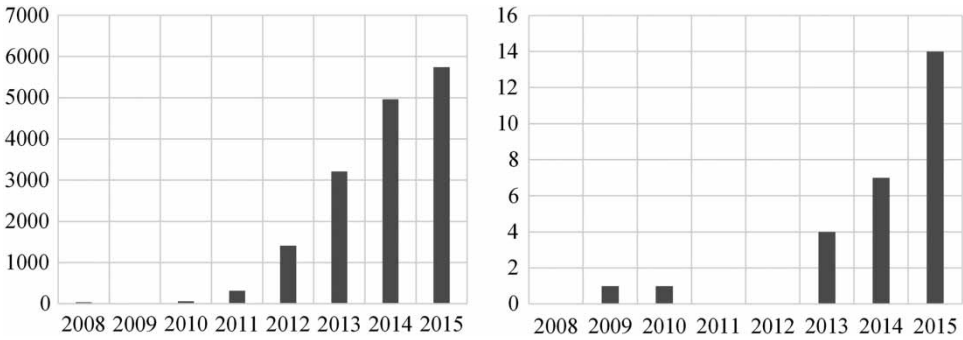

Figure 1 | Left: the number of publications about big data. Right: the number of publications about big data and water by Google Scholar.

the internet. The development of the internet and the movement of open data significantly accelerates data sharing and improves the accessibility of archived data. The hydroinformatics community will benefit from the active combination of a huge amount of data and 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 consists of three parts. The first part starts with an explanation of the concept of big data, then introduces the popular Apache Hadoop family to handle large amounts of data and seven classes of data analysis models, and discusses important ideas developed from the big data era. The second part discusses the impact of big data on hydroinformatics with the focus on the issues of data sharing. Then, the third part emphasizes the future of precipitation data fusion as one promising big data utilization in the area of hydroinformatics.

BACKGROUND

This section aims to introduce the popular term 'big data' starting with the example of Google Flu Detector (GFD), followed by the explanation of the concept of 'big data'. Once we get huge amounts of data, how to physically store and process the data becomes tricky. The conflict between the boom of big data and the data storage hard system, where the I/O speed is limited by the physical mechanism of hard disk, stimulated the development of parallel computing and distributed data storage. After being able to effectively manage large datasets, seven types of data modelling algorithms are summarized. Furthermore, when the correlation between datasets is successfully modelled, whether to only utilize the correlation or to discover more scientific knowledge is discussed.

Google Flu Detector

Currently, the concept of big data is popular in the analysis of sociology, public health, business and bioinformatics. The increasingly expanding internet is attracting people's attention as one major data source. The data on the internet, especially new media, are generated by individuals, reflecting their daily life, emotions, shopping preferences, etc. Without doubt, these types of data can be easily utilized in the field of online business, public health, sociology, as these topics mainly focus on individual behaviours. In fact, big data analysis opens a new way for researchers in these areas to find out what is actually happening from the recorded online behaviours of individuals.

Google developed a flu detector that monitors health-seeking behaviour 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 1,152 flu data points from Central Disease Control (CDC). By analysing the large numbers of search queries, Googler found 45 search terms, when used in a mathematical model, were strongly correlated with the

percentage of physician visits for influenza-like symptoms, based on which the GFD estimates the level of weekly influenza activity with a 1-day reporting lag (Ginsberg *et al.* 2009). From the perspective of the Google users, they tend to consult the accessible internet rather than immediately consulting the doctor, when they feel a bit ill. The GFD predicts the influenza activity from user query logs, though with some noises, responding much faster than the CDC with a 2-week reporting lag, which gives Google an advantage over the traditional disease control method. However, the GFD does not always perform well. In 2009, its poor underestimation of the influenza-like illness (ILI) in the United States of the swine flu pandemic forced Google to modify its algorithm as people's search behaviour changed due to the exceptional nature of the pandemic. In December 2012, it overestimated by more than double the doctor visits for ILI than the CDC (Butler 2013).

Despite the advantage of the quick response and reasonable accuracy of the GFD, the uncertainty from human behaviour searching that led the model to departure from the CDC data cannot be ignored. This type of uncertainty is embedded within the mechanism of the analysis, which may only be overcome by an improved algorithm. Regardless of the weakness of GFD, the point is that the apparent value of the data may only be the tip of the iceberg. Google started its business by providing an online searching service for internet users without the purpose of predicting the outbreak and threats of influenza, but the search query logs become extremely valuable after being accumulated for several years. The reason for this is that Google effectively collected the information that the search engine users want to know at a certain time and certain location. The big information pattern, contributed by millions of users around the world, showed additional big value behind search query data. To summarize, the GFD shows two features of the big data analysis, crowdsourcing and by-product.

What is big data?

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 term 'big data' is simple but makes its meaning ambiguous; it is commonly used to describe datasets with quantity and complexity beyond the capacity of normal computing tools to capture, curate, manage and process with a tolerable speed (Snijders *et al.* 2012). Another explanation of big data refers to developing new insights or creating new values at a large scale instead of a smaller one (Mayer-Schönberger & Cukier 2013). De Mauro *et al.* (2014) investigated 14 existing definitions of big data, and proposed a formal definition as:

> Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.

This definition can be subdivided into three groups: the characteristics of the datasets, 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 datasets, 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 the capabilities familiar from large-scale databases (Council 2013).

The expanding data vs. the developing computing power

The typical big data characteristics include high volume (the quantity of data generated), high velocity (the speed of collecting data), and high variety (the category of data) (Laney 2001). The concern is whether the existing computing system can handle the increasingly large data. An International Data Corporation (IDC) report has estimated that the data size of the world will grow from 130 exabytes (10^{18} bytes) in 2005 to 40 zettabytes (10^{21} bytes) in 2020, at a 40% annual increase (Gantz & Reinsel 2012). 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 analysed, which yields a large potential market for data utilization (Burn-Murdoch 2012). The actual data size needed is dependent on the task of data analysis, which further scales down the size of data to be processed. On the other hand, the data storage capacity has increased dramatically in the past decades. In 1956, IBM made the first commercial disk drive with a capacity of 3.75 MB (Oracle 2014). In 1980, the world's first gigabyte-capacity disk drive (2.52 GB), the IBM 3380, was the size of a refrigerator. After 25 years, the first 500 GB desktop hard drive was shipped (Dahl 2005), followed by the 1 TB one in 2007 (Perenson 2007). In 2014, Western Digital shipped the 8 TB hard drive and announced the world's first 10 TB hard drive (Hartin & Watson 2014). The unit cost of data storage will drop down from $2.00 per GB to $0.20 per GB from 2012 to 2020 (Gantz & Reinsel 2012). The storage of data should no longer be a big problem owing to 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.

The storage capacity of the hard disk keeps increasing, nevertheless the I/O speed of the hard disk grows slowly due to the limitation of the hard disk mechanism. Solid state disk (SSD) has a much higher I/O rate and negligible seek time; however, 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 than the single device. The I/O speed of the data storage devices is the bottleneck of extremely large data processing rather than the data storage capacity.

The MapReduce parallel computing

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 data dramatically increased without an appropriate algorithm. The parallel computing and distributed storage were developed to counter 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 is 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 data transfer across the network through sending the code to the local machine and writing the intermediate data to the local disk. The MapReduce system minimizes the impact of slow machines, and can handle machine failures and data loss by redundant execution. The success of the MapReduce programming model relies on several things. First, the model automatically deals with the details of parallelization, fault tolerance, locality optimization and load balancing, which makes it easy for programmers even without experience with parallel and distributed computing. Second, the Map and Reduce functions are capable of a variety of applications, such as sorting, data mining, machine learning, etc. Third, the MapReduce can scale up to large clusters of thousands of commodity machines, which means the computing resources can be utilized for big purposes (Dean & Ghemawat 2008). The Hadoop is an open-source version of the MapReduce framework developed by Apache, freely available to the 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 (Ghemawat *et al.* 2003). 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. (Apache 2015).

Hadoop MapReduce has a weakness during iterative data analysis in 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, this will dramatically slow down the analysis. This happens to most machine learning algorithms, e.g., gradient decent. Apache Spark is the latest programing model in the big data world, featuring its lightning fast data processing speed for iterative jobs (Zaharia *et al.* 2010). The Spark achieved its lightning fast speed by the implementing Resilient Distributed Datasets (RDDs), a distributed memory abstraction that lets the programmer perform in-memory computation (Zaharia *et al.* 2012). The Spark outperforms Hadoop by 20 times in speed by utilizing the RAM instead of hard disk to store the intermittent data.

Modelling big data

There are many data-based computational methods, and they can be classified as 'the seven computational giants of massive data analysis' (Council 2013). Data-based computing is facing challenges due to the expansion of data volume and dimensionality. The first giant is basic statistics including calculating the mean, variance and moments; estimating the number of distinct elements; number counting and frequency analysis; and calculating order statistics such as the median. These tasks typically require O(N) complexity calculations for N data points. The second computational giant is the generalized N-body problem, including nearly any problem involving distances, kernels, or other similarities between pairs or higher-order n-tuples of data points. The computational complexity is typically $O(N^2)$ or $O(N^3)$. N-body problems are involved in range searches, nearest-neighbour search problems and the nearest-neighbour classification problem. They also appear in nonlinear dimension reduction methods, also known as manifold learning methods. N-body problems are related to kernel computation, like kernel estimators – such as kernel density estimation, kernel regression methods, radial basis function neural networks and mean-shift tracking – and modern methods such as support vector machines and kernel principal components analysis (PCA). Other instances include k-means, mixtures of Gaussian clustering, hierarchical clustering, spatial statistics of various kinds, spatial joins, the Hausdorff set distance, etc. Graph-theoretic computation is the third giant, including problems with graph traversing. The graph can be either the data itself or the statistical model in the form of a graph depending on the nature of the problem. Common statistical computations include betweenness, centrality and commute distances; used to identify nodes or communities of interest. Nevertheless, the challenges arise when computing in large-scale, sparse graphs. When the statistical model takes the form of a graph, graph-search algorithms continue to remain important, but there is also a need to compute marginal probabilities and conditional probabilities over graphs; operations generally referred to as 'inference' in the graphical models literature. The fourth computational giant is linear algebraic computations, including linear systems, eigenvalue problems and inverses, deriving a large number of linear models, e.g., linear regression, PCA and many variants. Many of them are suitable for generic linear algebra approaches, but there are two important issues. One is that the optimization in statistical learning problems does not necessarily need to be trained to high accuracy to avoid overfitting. Another important difference is that multivariate statistics has its own matrix form, that of a kernel (or Gram) matrix; while, on the other hand, the computational linear algebra involves techniques specialized to take advantage of certain matrix structures. In kernel methods, such as Gaussian process regression or kernel PCA, the kernel matrix can be too large to be stored in the matrix explicitly, requiring probably matrix-free algorithms. Optimization is the fifth giant in massive data analysis. Linear algebraic computations are the main subroutine of second-order optimization algorithms. Non-trivial optimizations will continue and become increasingly common as methods have become more sophisticated. Linear programming, quadratic programming, and second-order cone programming are involved in support vector machines and recent classifiers, and semidefinite programming appears in manifold learning methods. Other standard types of optimization problems,

e.g., geometric programming, are to be applied in data analysis in the near future. The sixth one is integration of functions, which is required for fully Bayesian inference, and also non-Bayesian settings, most notably random effects models. The integrals that appear in statistics are often expectations. The frontier is the high-dimensional integrals arising in Bayesian models for modern problems. The approaches for this problem include Markov Chain Monte Carlo, or sequential Monte Carlo in some cases, approximate Bayesian computation (ABC) operating on summary data, and population Monte Carlo, a form of adaptive importance sampling. Alignment problems is the seventh giant, consisting of problems involving matchings between two or more data objects or datasets, such as data integration, data fusion. The fundamental alignment problems are usually carried out before performing further data analysis.

Correlation vs. causation

The most significant part of the big data concept is the fundamental and innovative ideas that change how people interact with the world. The enrichment of available data enables people to consider the entire system rather than taking few samples, thereby scientists can discover trends or phenomena that cannot be revealed with small data. The idea of big data always encourages to think bigger, to broaden the horizon to cover a big scope rather than focus on a few small areas. Moreover, the big data analysis focuses on correlation rather than causation, in that the correlations between datasets do not necessarily lead to causation, or that making use of the correlation is sometimes more valuable than exploring the causation behind it (Mayer-Schönberger & Cukier 2013). For simplicity, the associations of two variables can be classified in three types, i.e., causation, common response and confounding. Causation means direct cause-and-effect connection between variables, revealing that they are strongly correlated. Common response means the association between variables is in fact caused by another lurking variable. The change of the observed variables is in response to the changes of the hidden variable, even though the observed variables have no direct causal link. Two variables are confounded when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables (Moore & McCabe 2006). The association between two variables is not that simple, given how complex it is to understand the practical problems with multiple variables. Focusing on correlation makes it much easier for practical data mining application without too much effort on the causation. Machine learning, the artificial intelligence method, is the typical algorithm lying behind the big data analysis, which is a 'black box' model. Users feed inputs to the machine learning algorithm and get outputs from it without knowing what really happens to the data training process. This process is practically useful without necessarily understanding the causation behind it, but the causation is what scientists are always seeking. For academic purposes, detecting the potential of the data correlation is not the ultimate goal. Instead, the big data should help the development of science in a way that the novel association between big datasets can be detected to motivate further research for the causation. From the control theory perspective, the scientific exploration is to open the 'black box' of the objective world iteratively. On the other hand, the scientific model developed from analysing large datasets can then be validated through the correlation of the datasets. Figure 2 gives a clear illustration of the ideas stated above. The major difference is that the science focuses on causation, either derived from correlation, or validated through correlation, while the big data analysis in industry focuses on values of correlation from the data.

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 (Abbott 1991), 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 into 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 (Abbott 1999), are strongly dependent on information of which data, both textual or

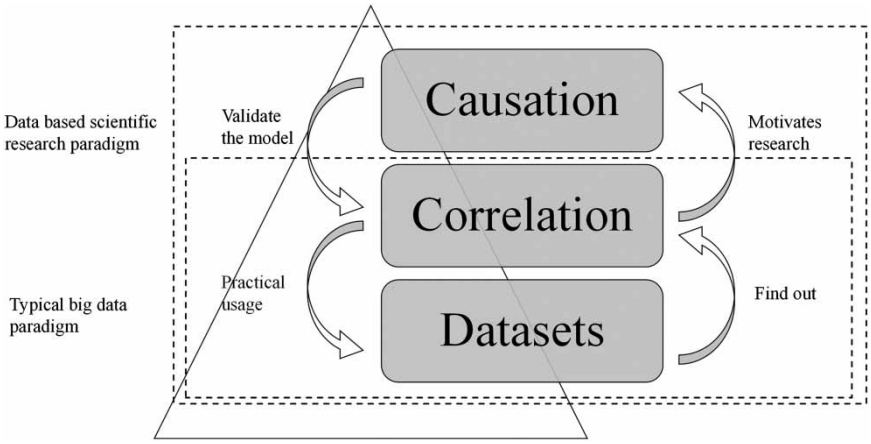

Figure 2 | The relationship between the datasets, correlation and causation.

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 (Pierson 2014). One simple example of big data analysis is called text mining. It has been carried out on the 50th anniversary of *Water Resources Research* to produce word clouds, shown in Figure 3, based on highly cited papers for every 10 years of *Water Resources Research*, which provides a visual representation of the themes emphasized in each decade (Rajaram *et al.* 2015).

Data for hydroinformatics

In general, water-related problems are quite complex due to the interrelationships between water-related environmental, social and business factors. The data being generated and collected relevant to hydroinformatics feature huge volumes and multiple types. For the purpose of simplification, the data sources for hydroinformatics, without loss of generosity, can be classified into three dimensions, i.e., the natural dimension, the social dimension and the business 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 water resource management, flood and drought management. Water-related data include the measurement of precipitation (rainfall, snow and hail), river flow, water quality, soil moisture, soil characteristics, ground water condition, air temperature and humidity, solar flux, etc. The observation methods developed are from local stations for point measurement to remote sensing – radar and satellites, and drone. Earth observation satellites are generating huge volumes of data including weather- and water-related information. ESA launched SMOS for soil moisture observation in 2009, and will launch ADM-Aeolus for Atmospheric Dynamics observation in 2017 (ESA 2016). NASA launched SMAP to map soil moisture and determine the freeze or thaw state in 2015 (SMAP 2015). The GPM mission launched in 2015 aims to provide global rain and snow observations based upon the success of TRMM launched in 1997 (NASA 2011). 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 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 weather forecasting, flood and drought prediction. Unfortunately, although many satellites were launched or are to be launched, the huge amount of available data is rarely used; only 3–5% of data is used on a daily average, while billions of dollars have been invested annually (Selding 2012). Apart from the earth observation data, reanalysis data are another important information source with high data quality. In other words, the

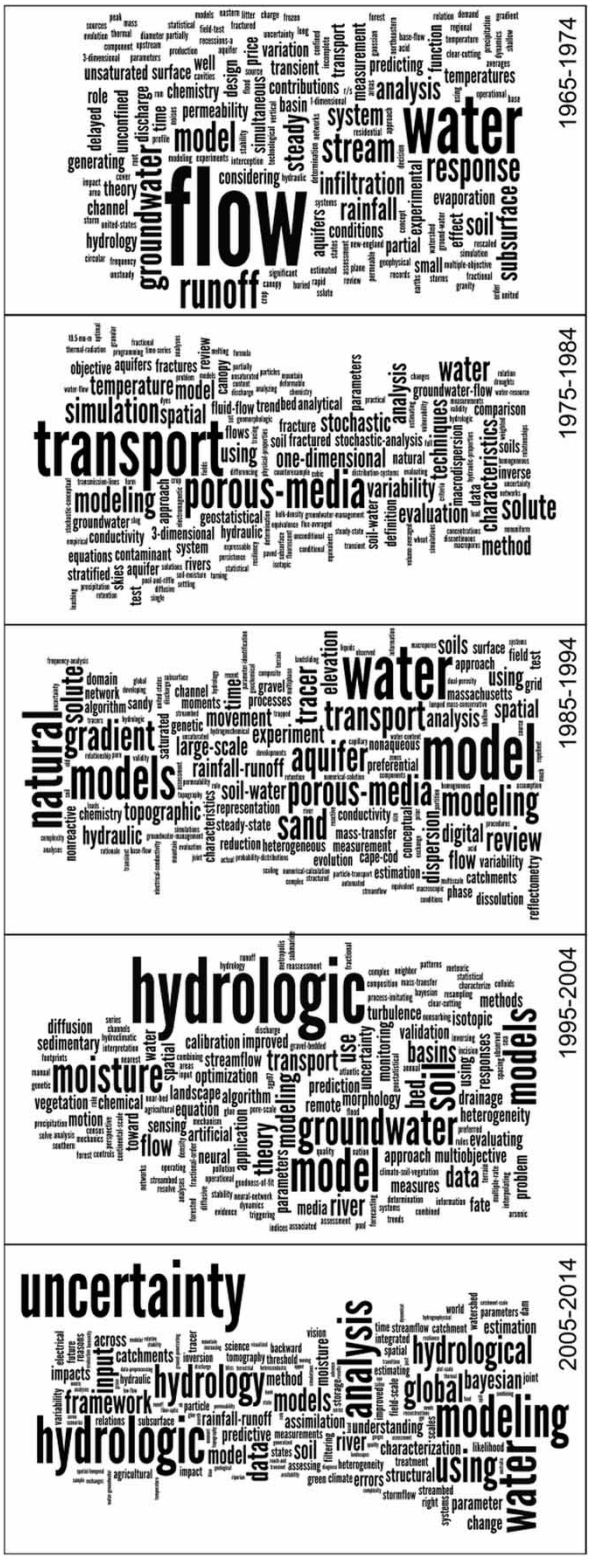

Figure 3 | Word clouds of highly cited papers from *Water Resources Research* in each decade as an example of big data related to water (Rajaram *et al.* 2015).

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 datasets 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 (National Centers for Environmental Prediction 1994). ECMWF has series of ERA projects for global atmospheric reanalysis that can be traced back to 1957 (ECMWF 2015). The Japan Meteorological Agency conducted the JRA-55 project for a high-quality homogeneous climate dataset covering the last half century (Kobayashi *et al.* 2015). The model generated data are four dimensional, three dimensions in space and one in time, and of high spatial and temporal coverage and resolution, resulting in huge volumes of data, which means that 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 amounts 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 an ideal location for hydro-power based on over 30 freely remote sensing and environmental datasets in the UK (Leicester 2015).

The social dimension is about the interaction of water environment and human society. With the digitalization of textual information available online and the explosion of social media, textual mining technologies enable a new research area of the public attitude towards certain issues. For instance, five million scientific articles have been analysed to explore the impact of the Fukushima disaster on the media attitude towards nuclear power (Lansdall-Welfare *et al.* 2014). Similar ideas can be applied 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 to discover real-world situations, demonstrated by the example of GFD mentioned in the previous section. Twitter data is now attracting many researchers to dig out water environment-related research. It was found that Twitter content could infer daily rainfall rates in five UK cities, which revealed that the online textual features in Twitter were strongly related to the topic with significant inference (Lampos & Cristianini 2012). Two Dutch organizations, Deltares and Floodtags, have developed real-time flood-extent maps based on tweets about floods for Jakarta, Indonesia (Eilander 2015). This method gives disaster management a real-time view of the situation with 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 playing the 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 internet users interact with the internet. The data ocean has valuable potential for scientists to discover novel correlations between real-world situations. The fundamental data mining techniques behind the big data application, such as GFD, 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 behaviour 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.

The business dimension covers but is 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 National Oceanic and Atmospheric Administration (NOAA) to explore the business of weather. They built one of the first parallel processing supercomputers for weather modelling in 1995, named Deep Thunder Project. Deep Thunder creates 24- to 48-hour forecasts at 1–2 km resolution with a lead time of 3 hours to 3 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 weather-generated delays by rearranging or combining flights more efficiently (IBM 2015). Another possibility is that, as inspired by the big data application in e-commerce that utilizes accumulated user activity logs for a 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 by the consumer and utility (Stewart *et al.* 2010). 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 ideas 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 ICTs (ITU 2015). Furthermore, the operation data collected by companies in the water industry also have potential values for data mining for optimizing the system and providing more information for decision-making.

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 (Opendefinition 2015). Open data is the further development of free data, i.e. 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 the ability to process 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 full value and accelerate future research only

when being open. Multiple national governments have 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 amounts of environmental, meteorological and climate datasets, 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 (SWITCH-ON 2015). EarthCube is a project launched in 2011 that develops a common cyber infrastructure for the purpose of collecting, accessing, analysing, 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 (EarthCube 2015). The ongoing movement of open data can boost data-based research and data usage by removing the legal restriction on data use. Many data portals are being created for data sharing through web services with much more powerful data search tools where users can find data by location, time and data types, etc.

Issues of data sharing

The trend of open data will motivate data sharing and comprehensive utilization of data by removing the restriction of patents, copyrights, but other issues of data sharing necessitate cooperative effort and innovative ideas. Data format is one of them. As the datasets related to water are collected by different organizations in different countries all around the world, how the data was recorded and expressed has not been identical. A very simple example is that even the expression of dates is different. Chinese use 'yyyy-mm-dd'; Europeans use 'dd-mm-yyyy'; and in the USA people use 'mm-dd-yyyy'. This issue with dates was tackled by ISO 8601, an international agreement to use 'yyyy-mm-dd' for the format of dates. Other issues may include but are not limited to the observation resolution, both temporal and spatial, the expression of missing value, the data processing methods, the units of the data, etc. As the characteristics of different datasets vary, the data should be clearly tagged by the metadata which is essential for the data user to carry out data analysis. The metadata is the information about information, which describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource (Guenther & Radebaugh 2004). The metadata should capture the basic characteristics of a data or information resource including who, what, when, where, why and how about the data resource. In the big data era, ad hoc data analysis for simple tasks may be time-consuming when the data size becomes extremely large. It can be worthwhile for the data provider to process feature extraction offline and incorporate these features into the metadata, such as mean values, extremums, general trend or pattern prior to the data release. Such pre-process of data can make it much easier for data users to find the data they need.

Another challenging issue of integration data usage is the variety of data formats, varying from simple binary or CSV format to advanced self-describing Network Common Data Form (netCDF), Hierarchical Data Format (HDF), GRIddedBinary (GRIB), Extensible Markup Language (XML), waterML, etc. For satellite data, High Rate Information Transmission (HRIT), Low Rate Information Transmission (LRIT), High Rate Picture Transmission (HRPT) and Low Rate Picture Transmission (LRPT) are the CGMS standards agreed upon by satellite operators for the dissemination of digital data to users via direct broadcast. The difference is that HRIT and LRIT transmit data originating from geostationary satellites while the HRPT and LRPT transmit data originating from low earth orbit satellites. Also, their names suggest that they operate at different data broadwidth. The World Meteorological Organization (WMO) has two binary data formats: Binary Universal Form for the Representation of meteorological data (BUFR) to represent any meteorological dataset

employing a continuous binary stream, and GRIB format to transmit large volumes of gridded data to automated centres over high-speed telecommunication lines using modern protocols. The Man computer Interactive Data Access System (McIDAS) goes beyond a simple data format to a set of tools for analysing and displaying meteorological data for research and education. NetCDF is a machine-independent, self-describing, binary data format standard and a set of software libraries for exchanging array-based scientific data; it features self-describing, portable, scalable, appendable, shareable and archivable data. HDF, HDF4 or HDF5 is a library and multi-object file format for the transfer of graphical and numerical data between computers developed by NASA. HDF supports several different data models in a single file, including multidimensional arrays, raster images and tables, which respectively have their specific data type and API. The XML is a general-purpose markup language, primarily used to share information via the Internet (WMO 2015). The WaterML2 is a variation of XML specified for water observation data, and allowing data exchange across information systems (OGC 2015). Standard Hydrologic Exchange Format (SHEF) was created to store and exchange hydrometeorological data in the 1980s, and is readable by both humans and machines (Bissell *et al.* 1984).

The variety of data formats may cost scientists much time dealing with different formats rather than working on scientific problems when utilizing multiple datasets from a variety of sources. To enhance the accessibility of hydrological data, GEOWOW (Global Earth Observation System of Systems (GEOSS) interoperability for Water, Ocean and Water) contributes to international standardization processes within the Hydrology Domain Working Group, a joint working group of the Open Geospatial Consortium (OGC) and the WMO. GEOWOW developed for the first time a common global exchange infrastructure for hydrological data based on standardized formats and services. GEOWOW aims to evolve the GEOSS in the aspect of water, and is part of the GEOSS Conmen Infrastructure (GCI) (GEOWOW 2013). In addition, a middleware that connects the data I/O scripts and the data analysis tools may be a feasible alternate featuring reusability. Middleware is the glue of software, and usually lies between the application layer and the system layer, or connects between different software components. The data-based analysis necessitates such middleware to handle large datasets from different sources in a variety of data formats and many computational models as well as being compatible with multiple programming languages. The open source development has to be implemented to such middleware to enable the whole research community to contribute to and benefit from it.

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 runtimes, 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 more cost-effective than implementing local clusters. Cloud computing is scaleable to suit the job, and does not require extensive knowledge on configuring local clusters. US NOAA has launched its Big Data Project collaborating with Amazon Web Service, Google Cloud Platform, IBM, Microsoft and the Open Cloud Consortium (Department of Commerce 2015). 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 to create new findings. NOAA's Big Data Project indicated a coming trend of combing the tremendous volume of high quality data held by the government and private industry's vast infrastructure and technical capacity of data management and analysis.

BIG DATA FOR PRECIPITATION ESTIMATE

The available precipitation data

Although computer scientists have attempted to use newly emerged social network data to estimate rainfall, as mentioned in the previous section, it is like a 'dessert'; the

main data sources of rainfall measurement are rain gauges, weather radars and satellites, which are the 'main course'. The 'dessert' has some obvious shortages apart from its advantage on data cost and quick response. The use of Twitter data to estimate rainfall or flood situation, as mentioned in the previous section, requires the prevalence of Twitter at a local level, e.g., developed urban area with a large number of users and a wide internet access, which implies the spatial coverage and resolution of the data can be poor in less developed cities and rural areas. The temporal length of the Twitter data is significantly less than the meteorological records, which can be traced back to 1861, while Twitter was launched in 2006. Despite the low cost and quick response of the new data sources foretelling a possible future direction, the existing data sources for rainfall measurement have accumulated a vast quantity of data which can substantially benefit from the big data technology. Table 1 shows information of some widely used datasets containing precipitation data. The features of precipitation data from different sources vary significantly due to the different measuring mechanisms and processing algorithms.

Data fusion

Hydrologists are pursuing fine and accurate estimates of precipitation data in both space and time for drought and flood management. Rain gauge observations are direct

Table 1 | Information of some datasets containing precipitation

Datasets	Data source	Data size	Spatial and temporal coverage and resolution	
GPCC Global Precipitation Climatology Centre monthly precipitation dataset	Gauge based	4.2 GB	Monthly values from 1901/01. Varys, 0.5 °, 1.0 ° and 2.5 ° global grid	Beck *et al.* (2005)
The Next Generation Weather Radar (NEXRAD)	Radar	73.1 TB	Comprising 160 sites throughout the USA. 1 ° grid. 1 hour, 3 hour and total storm accumulated data since 1988	NCEI (2015)
Global Historical Climatology Network Daily Database	Station record	22 GB	Daily since 1861. Contains records from over 80,000 stations in 180 countries and territories	Menne *et al.* (2012)
CPC Global Summary of Day/Month Observations	Station record	13.7 GB	Approx. 8,900 actively reporting stations in global daily data since 1979	Climate Prediction Center National Centers for Environmental Prediction National Weather Service NOAA US Department of Commerce (1987)
GPCP (Daily): Global Precipitation Climatology Project-1DD product	Geostationary infrared satellite	0.78 GB	Daily rainfall accumulation globally on a 1 ° grid in latitude and longitude starting in October 1996	Pendergrass (2015)
The Tropical Rainfall Measuring Mission (TRMM)	Satellite	236 GB	3 hourly from Jan 1st 1998 to mid-2017. 0.25 ° latitude/longitude grid over the domain 50 °S–50 °N	NASA (2013)
The Global Precipitation Measurement Mission (GPM)	Satellite	N/A	Provide half-hourly and monthly precipitation estimates on a 0.1 ° latitude/longitude grid over the domain 60 °N–S	NASA (2011)
NCEP Climate Forecast System Reanalysis	Model reanalysis	67 TB	6 hourly from 1979. 0.1 ° latitude/longitude grid globally	(Saha *et al.* 2010)

measurements of rainfall on the ground, but are often sparse in regions with complex landform, clustered in valleys or populated areas, and of poor temporal consistency. Thus, gauge data may not be able to provide sufficient information about the spatial extent and intensity of precipitation (Verdin *et al.* 2015). Estimating precipitation from satellites provides an alternative method for collecting rainfall data with the inherent advantage of detecting the spatial distribution of the precipitation. They are different in the observation mechanism resulting in a substantial difference in the features of observation results. Satellite-based measurement is intermittent, area-averaged observation, while rain gauge measurement is continuous and point observation (Arkin & Ardanuy 1989). There is a trade-off of accuracy and spatial coverage between each data source. The observations of rain gauges and radar have the best measurement of actual rainfall but with the most limited spatial coverage. Geostationary satellites with infrared sensors are less accurate but the coverage is broad and continuous. Between them are the microwave sensors on low earth orbits which provide more reliable estimates of precipitation but with incomplete temporal sampling and coarse spatial resolution (Gorenburg *et al.* 2001). In the big data era, it is encouraged to make use of the joint data from various sources. It is promising to fuse the existing precipitation data from heterogeneous data sources. As heterogeneous data sources possess different advantages and disadvantages, they can complement each other in an optimal way (Sander & Beyerer 2013).

Verdin *et al.* (2015) used a Bayesian data fusion model with ordinary kriging to blend infrared precipitation data and gauge data on the Central and South American region. The method was applied to pentadal and monthly total precipitation fields during 2009. This blending method significantly improves upon the satellite-derived estimates and is also competitive in its ability. Wang *et al.* (2011) assessed the performance of the multiscale Kalman smoother-based framework in precipitation fusion. They tested the algorithm on 2003 hourly NEXRAD MPE precipitation data of two spatial resolutions, i.e., 1/8 ° and 1/32 °, respectively, covering 152,175 km^2 in the USA. Linear weighted algorithm, multiple linear regression, and artificial neural network were also used to fuse the remote sensing data with the ground data (Srivastava *et al.* 2013). All the previous data fusion studies indicate that data fusion processes can generally improve data quality over the data from a single source. Nevertheless, there is an apparent limitation of the previous studies in that they only proposed the methodology and tested it with limited spatial and temporal coverage; in other words, the amount of data was limited, so they did not concern themselves much about the efficiency of the algorithm which is the key factor in processing big data.

In fact, applying data fusion technique to the existing terabytes of precipitation data is a tough issue for hydrologists as processing the huge amount of data generated every day will be extremely time-consuming. It becomes more problematic when dealing with the accumulated historical data which are equally valuable. Owing to the development of big data, the ability of a cluster of computers to process large amounts of data has been greatly improved primarily by implementing the idea of parallel computing, which is to subdivide the job into small portions and to involve a cluster of computers to work simultaneously. Thus, parallel computing posed a requirement on the fusion algorithm that the data fusion process can be separated into individual parts. Data fusion algorithms, e.g., the Bayesian kriging method proposed by Verdin *et al.* (2015), have the disadvantage of snapshot; in other words, are temporally independent, while the time series of precipitation are available for possible improvements in accuracy. However, this snapshot feature makes the fusion process easy to be separated temporally for parallel computing, which can effectively speed up the processing procedure.

Hadoop MapReduce was designed to handle textual data initially, and how it performs on processing high-volume remote sensing image data has been assessed by only a few papers, of which the results were positive. Almeer (2012) researched the performance of eight pixel-level image processing algorithms using Hadoop, with the result that the method is scalable and efficient in processing multiple large images used mostly for remote sensing applications, and the Hadoop runtime is significantly lower than the runtime of a single PC. Lv *et al.* (2010) developed a parallel model of K-means clustering algorithm based on Hadoop MapReduce to process satellite remote sensing images, of which the results are acceptable, and the runtime drops. It is reasonable to believe that Hadoop MapReduce

on a cluster of machines will work on the data fusion job as the parallel computing is not restricted by the type of data. Thus, the future of fusing precipitation data with the aid of big data techniques can be promising. The reasons are, as mentioned above, that the data fusion of heterogeneous precipitation data sources can offer better results than data from each single source, and the fusion process can be accelerated by parallel computing.

CONCLUSIONS

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 the internet as digitalizing of information among the world becomes much easier and cheaper for future data mining purposes. 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 purposes as well. Hydroinformatics can benefit from the expanding amount of data collected, generated and available 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, analysing climate change, flood and drought management and the global water cycle; the social dimension, focusing on the interaction between water environment and human society; and the business dimension, using data-based decision-making systems for optimizing water resource management systems and future water planning. The data processing tools like parallel computing, distributed storage have been developed to help users handle the large datasets in hundreds of GBs or even TBs in tolerable time to make real-time application possible and interactive human–computer analysis feasible. Cloud computing platforms will make it unnecessary to download the data to a local machine and run the model locally, but provide superior computing efficiency in the future cloud computing era.

The challenges of big data have also been included in this paper. Data sharing is one of them, as water-related datasets have a variety of formats with different observation methods generated from different organizations. Either a general standardized format for data exchange or an open sourced data management tool that glues all relevant scripts for read and write of different data formats can benefit the research community on handling datasets. Many data portals based on web services are being created for data exchange and encouraging data-based research. Contradiction is another challenge of big data in that the correlation between datasets is practically more useful than the causation between datasets, while the causation is the purpose of scientific research. The correlation identified from a vast range of datasets ought to help researchers explore new potential causation between the phenomena for further research, instead of only replacing the logic-based model. The real challenge in the near future is how to make the best use of the available data, as currently there is little being 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.

Precipitation estimation is one possible area to make a start, as data related to precipitation is being collected from multiple sources, such as rain gauges, weather radars and satellites. Global precipitation data collected from NEXRAD and GPM can reach tens of TBs, which is a big data problem. One promising future is to fuse precipitation data from multiple sources – weather radar, satellite remote sensing, rain gauge and model reanalysis data – to generate a rainfall estimation product with a better spatial and temporal resolution and minimized uncertainty. The parallel computing, distributed data storage paradigms and cloud computing platforms developed during the explosion of information are essential to accelerate the data processing procedure. The implementation of big data in precipitation data fusion and the parallel computing model are the tip of the iceberg in the big data era. The utilization of available data is not limited to improving the precipitation estimation. The future should rely on an 'all data revolution', in which innovative analytical ideas, utilizing data from all existing and new sources, and providing a deeper, clearer understanding will significantly shift how we recognize the world (Lazer *et al.* 2014).

REFERENCES

Abbott, M. B. 1991 *Hydroinformatics: Information Technology and the Aquatic Environment*. Avebury Technical, Aldershot, UK.

Abbott, M. 1999 Introducing hydroinformatics. *J. Hydroinform.* **1**, 3–19.

Almeer, M. H. 2012 Hadoop MapReduce for remote sensing image analysis. *Int. J. Emerg. Technol. Adv. Eng.* **2**, 443–451.

Apache. 2015 *What Is Apache Hadoop?* [Online] Apache Hadoop. https://hadoop.apache.org/ (accessed 22 June 2015).

Arkin, P. A. & Ardanuy, P. E. 1989 Estimating climatic-scale precipitation from space: a review. *J. Climate* **2**, 1229–1238.

Beck, C., Grieser, J. & Schneider, U. 2005 Global precipitation analysis products. Global Precipitation Climatology Centre (GPCC). DWD.

Bennett, J. & Lanning, S. 2007 The Netflix prize. In: *Proceedings of KDD Cup and Workshop*, New York. ACM, 35.

Bissell, V. C., Pasteris, P. A. & Bennett, D. G. 1984 Standard hydrologic exchange format (SHEF). *J. Water Resour. Plann. Manage.* **110**, 392–401.

Burn-Murdoch, J. 2012 Study: Less than 1% of the world's data is analysed, over 80% is unprotected [Online]. *The Guardian*. http://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume (accessed 18 June 2015).

Butler, D. 2013 When Google got flu wrong. *Nature* **494**, 155–156.

Climate Prediction Center National Centers for Environmental Prediction National Weather Service NOAA US Department of Commerce 1987 CPC Global Summary of Day/Month Observations, 1979-continuing. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO, USA.

Council, N. R. 2013 *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, DC, USA.

Dahl, E. 2005 *PC Drive Reaches 500GB* [Online]. PCWorld. http://www.pcworld.com/article/120102/article.html (accessed 10 May 2015).

Dean, J. & Ghemawat, S. 2008 MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**, 107–113.

De Mauro, A., Greco, M. & Grimaldi, M. 2014 What is Big Data? A consensual definition and a review of key research topics. In: *4th International Conference on Integrated Information, Madrid, Spain*. doi. 2341.5048.

Department of Commerce 2015 U.S. Secretary of Commerce Penny Pritzker announces new collaboration to unleash the power of NOAA's data [Online]. Department of Commerce. https://www.commerce.gov/news/press-releases/2015/04/us-secretary-commerce-penny-pritzker-announces-new-collaboration-unleash (accessed 2 December 2015).

EarthCube 2015 *About EarthCube* [Online]. Earthcube.org. http://earthcube.org/info/about (accessed 15 July 2015).

ECMWF 2015 *ECMWF Climate Reanalysis* [Online]. ECMWF. http://www.ecmwf.int/en/research/climate-reanalysis (accessed 16 July 2015).

Eilander, D. 2015 *Twitter used to create real-time flood maps* [Online]. https://www.deltares.nl/en/news/twitter-used-to-create-real-time-flood-maps/ (accessed 27 April 2015).

ESA 2016 *Earth Explorers overview* [Online]. European Space Agency. http://www.esa.int/Our_Activities/Observing_the_Earth/Earth_Explorers_an_overview (accessed 12 February 2016).

Gantz, J. & Reinsel, D. 2012 The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future* **2007**, 1–16.

GEOWOW 2013 *GEOWOW Water* [Online]. GEOWOW. http://www.geowow.eu/water.html (accessed 25 July 2015).

Ghemawat, S., Gobioff, H. & Leung, S.-T. 2003 The Google file system. ACM SIGOPS operating systems review. ACM, 29–43.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. 2009 Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014.

Gorenburg, I. P., McLaughlin, D. & Entekhabi, D. 2001 Scale-recursive assimilation of precipitation data. *Adv. Water Resourc.* **24**, 941–953.

Guenther, R. & Radebaugh, J. 2004 *Understanding Metadata*. National Information Standard Organization (NISO) Press, Bethesda, MD, USA.

Hartin, E. & Watson, K. 2014 Announces new innovations that set the standard for performance, reliability, capacity, agility and efficiency for helping companies harness the power of data [Online]. http://www.hgst.com/press-room/press-releases/HGST-unveils-intelligent-dynamic-storage-solutions-to-transform-the-data-center (accessed 10 May 2015).

IBM 2015 *IBM100 – Deep Thunder* [Online]. IBM's 100 Icons of Progress: IBM. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepthunder/ (accessed 14 December 2015).

ITU 2015 *Internet of Things Global Standards Initiative* [Online]. ITU. http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx (accessed 28 July 2015).

Kobayashi, S., Ota, Y. & Harada, Y. 2015 The JRA-55 reanalysis: general specifications and basic characteristics. *J. Meteor. Soc. Japan* **93**, 5–48.

Lampos, V. & Cristianini, N. 2012 Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol. (TIST)* **3**, 72, doi:10.1145/2337542.2337557.

Laney, D. 2001 3D data management: controlling data volume, velocity, and variety [Online]. Gartner Group. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 11 May 2015).

Lansdall-Welfare, T., Sudhahar, S., Veltri, G. & Cristianini, N. 2014 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*. IEEE, pp. 60–66.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. 2014 Big data. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205.

Leicester, U. O. 2015 *Big data technology finds ideal river locations to generate hydro-power* [Online]. ScienceDaily. http://www.sciencedaily.com/releases/2015/04/150413075144.htm (accessed 28 July 2015).

Lv, Z., Hu, Y., Zhong, H., Wu, J., Li, B. & Zhao, H. 2010 *Parallel K-means clustering of remote sensing images based on mapreduce*. Web Information Systems and Mining. Springer-Verlag, Berlin, Heidelberg, Germany.

Mayer-Schonberger, V. & Cukier, K. 2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, MA, USA.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. & Houston, T. G. 2012 An overview of the global historical climatology network-daily database. *J. Atmos. Ocean. Technol.* **29**, 897–910.

Moore, D. & McCabe, G. 2006 *Introduction to the Practice of Statistics*, 5th edn. W. H. Freeman, New York, USA.

NASA 2011 *Global Precipitation Measurement (GPM) Mission Overview* [Online]. Pmm.nasa.gov. http://pmm.nasa.gov/GPM (accessed 11 May 2015).

NASA 2013 *Readme for TRMM Product 3B42 (V7)* [Online]. GES DISC NASA. http://disc.sci.gsfc.nasa.gov/precipitation/documentation/TRMM_README/TRMM_3B42_readme.shtml (accessed 11 May 2015).

National Centers For Environmental Prediction, N. W. S. N. U. S. D. O. C. 1994 NCEP/NCAR Global Reanalysis Products, 1948-continuing. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO, USA.

NCEI 2015 *NEXRAD Products* [Online]. http://www.ncdc.noaa.gov/data-access/radar-data/nexrad-products (accessed 15 May 2015).

OGC 2015 *OGC@WaterML* [Online]. Open Geospatial Consortium. http://www.opengeospatial.org/standards/waterml (accessed 4 July 2015).

OPENDEFINITION 2015 *Defining Open in Open Data, Open Content and Open Knowledge* [Online]. Opendefinition.org. http://opendefinition.org/od/ (accessed 5 July 2015).

ORACLE. 2014 *Time Capsule, 1956 Hard Disk* [Online]. Oracle. http://www.oracle.com/technetwork/issue-archive/2014/14-jul/o44timecapsule-2219543.html (accessed 12 May 2015).

Pendergrass, A. 2015 *The Climate Data Guide: GPCP (Daily): Global Precipitation Climatology Project* [Online]. https://climatedataguide.ucar.edu/climate-data/gpcp-daily-global-precipitation-climatology-project (accessed 15 May 2015).

Perenson, M. 2007 *Hitachi Introduces 1-Terabyte Hard Drive* [Online]. PCWorld. http://www.pcworld.com/article/128400/article.html (accessed 10 May 2015).

Pierson, L. 2014 Civil engineer turned environmental data scientist harnesses big environmental data at UNESCO-IHE [Online]. statisticsviews.com. http://www.statisticsviews.com/details/feature/7136441/Civil-Engineer-Turned-Environmental-Data-Scientist-Harnesses-Big-Environmental-D.html (accessed 15 June 2015).

Rajaram, H., Bahr, J. M., Bloschl, G., Cai, X., Scott Mackay, D., Michalak, A. M., Montanari, A., Sanchez-Villa, X. & Sander, G. 2015 A reflection on the first 50 years of *Water Resources Research*. *Water Resour. Res.* **51**, 7829–7837.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J. & Behringer, D. 2010 The NCEP climate forecast system reanalysis. *Bull. Am. Meteor. Soc.* **91**, 1015–1057.

Sander, J. & Beyerer, J. 2013 Bayesian fusion: modeling and application. In: *Sensor Data Fusion: Trends, Solutions, Applications (SDF), 2013 Workshop on*. IEEE, pp. 1–6.

Selding, P. B. D. 2012 *U.S. Government-leased Satellite Capacity Going Unused* [Online]. SpaceNews.com. http://spacenews.com/32581us-government-leased-satellite-capacity-going-unused/ (accessed 20 June 2015).

SMAP 2015 *SMAP Overview* [Online]. SMAP, JPL. http://smap.jpl.nasa.gov/observatory/overview/ (accessed 5 July 2015).

Snijders, C., Matzat, U. & Reips, U.-D. 2012 Big data: Big gaps of knowledge in the field of internet science. *Int. J. Internet Sci.* **7**, 1–5.

Srivastava, P. K., Han, D., Rico-Ramirez, M. A., Al-Shrafany, D. & Islam, T. 2013 Data fusion techniques for improving soil moisture deficit using SMOS satellite and WRF-NOAH land surface model. *Water Resour. Manage.* **27** (15), 5069–5087.

Stewart, R. A., Willis, R., Giurco, D., Panuwatwanich, K. & Capati, G. 2010 Web-based knowledge management system: linking smart metering to the future of urban water planning. *Australian Planner* **47**, 66–74.

SWITCH-ON 2015 *About SWITCH-ON* [Online]. SWITCH-ON Project. http://www.project.water-switch-on.eu/ (accessed 15 June 2015).

Verdin, A., Rajagopalan, B., Kleiber, W. & Funk, C. 2015 A Bayesian kriging approach for blending satellite and ground precipitation observations. *Water Resour. Res.* **51**, 908–921.

Wang, S., Liang, X. & Nan, Z. 2011 How much improvement can precipitation data fusion achieve with a Multiscale Kalman Smoother-based framework? *Water Resour. Res.* **47** (3), W00H12.

WMO 2015 *Satellite Data Formats and Standards* [Online]. World Meteorological Organization. http://www.wmo.int/pages/prog/sat/formatsandstandards_en.php (accessed 6 July 2015).

Zaharia, A. M., Chowdhury, M., Franklin, M. J., Shenker, S. & Stoica, I. 2010 Spark: cluster computing with working sets. In: *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*. USENIX Association, Berkeley, CA, USA, 10.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S. & Stoica, I. 2012 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*. USENIX Association, Berkeley, CA, USA, pp. 2–2.

First received 21 August 2015; accepted in revised form 22 January 2016. Available online 2 March 2016