Chapter 3
An Overview of Big Data Analytics: A State-of-the-Art Platform for Water Resources Management

Nirav Raval and Manish Kumar

3.1 Introduction

Water resources have become stressed with the enhancement of population growth, looming agricultural/industrial production, unprecedented rise in living standard, and uncontrolled climate change (Kumar et al. 2019a; Roshan et al. 2020). Water scarcity refers to the condition when sufficient water resources are not available to meet a particular region's water requirement (Kumar et al. 2019b, 2020). It is estimated that worldwide approximately four thousand million people are not provided with adequate quantity of potable water for at least one month a year. With the population projected to expand to nine thousand million by 2050, the demand for potable water is set to increase dramatically (du Plessis 2019; Patel et al. 2019; Singh et al. 2020). Hence, to meet the demand of blooming population for clean water, monitoring-based management of water resources is essential.

Advancement of the engineered sensors, data monitoring, and communication devices enables continuous monitoring of particular water system. As a result of this, near real-time series data with high frequency can be recorded. Such perpetual measurement produces bulk of data, called as big data (Gandomi and Haider 2015; Mayer-Schönberger and Cukier 2013). The term "Big data" which includes the major processes like, data acquisition, storage, extraction, and cleaning as well as analysis and interpretation were first proposed by Michael Cox and David Ellsworth in 1997. The targeted use of advanced big data analytics is emerging for effective and sustainable management of water resources in the scientific community. Application of the computer models is increasing in the field of water science and engineering because of the urgent necessity for deeper perspicacity into water systems and

N. Raval · M. Kumar (✉)
Discipline of Earth Sciences, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat 382355, India
e-mail: manish.kumar@iitgn.ac.in

M. Kumar et al. (eds.), *Resilience, Response, and Risk in Water Systems*,
Springer Transactions in Civil and Environmental Engineering,
https://doi.org/10.1007/978-981-15-4668-6_3

demand for providing effective solutions toward stressed water resources in a sustainable manner (Bibri and Krogstie 2017; Singh et al. 2020; Mukherjee et al. 2020). However, heterogenous nature of big water data causes difficulty in its storing, handling, and processing. Thus, to manage water resources effectively and sustainably, proper application of advanced water analytics is one of the prime necessities of the current decade.

A key contribution of this chapter is to bring forth the basic overview, characteristics, applications, challenges, and open platform/proposed model supporting water resources-based big data. Also, the future directions related to the integrations of various platforms are provided.

3.2 Big Water Data and Associated Characteristics

Technological advancement facilitates constant acquisition and processing of data at an unprecedented rate which can be further managed with the readily available software and hardware. This capability can be inclusively termed as "big data" (Adamala 2017). To characterize the big data, commonly used parameters include (i) Volume: As name suggests, volume is generally the quantity of data generated, processed, and stored. In the twenty-first century, data generation is constantly increasing as a result of which big data sizes are reported in multiple of terabytes and petabytes. For the storage, handling, and processing of this bulk data, the distributed systems are used instead of traditional database technology (Schroeck et al. 2012). (ii) Velocity: Speed with which generated data can be transferred and processed is known as velocity. At present time, streaming data (collected in real time) is one of the leading edges of big data. The modern applications and computer-based programmes/softwares enable the sorting, transmitting, and processing of generated data at faster rate (David et al. 2014). (iii) Variety: The availability of different types of data represents the variety. Water-related data is highly unstructured. The modern big data technology enables simultaneous collection and usage of structured and unstructured data. In water resource management, there has been more efforts require to integrate all types of water data from across different sections/sectors into one continuous data stream (Zikopoulos and Eaton 2011). (iv) Veracity: The quality or trustworthiness of any water-related data is known as veracity and it is directly associated with the health aspect as water is considered to be one of the primary necessities for the survival of living being. In general, it is a measure of the accuracy of the data. Quality control is one the important parameters to be considered for big data. (v) Value: It refers to the actionable perception gained from generated data. Having access to big water data will not going to complete the work unless and until it's conversion into some value has not been performed. In case of water consumption survey study, the availability of data is not sufficient to reach the decision making untill its conversion into some deliverable value. With the help of the state-of-the-art models/softwares and algorithms, large amounts of data can be converted into deductive information for final decision making (David et al. 2014; Madden 2012).

3.3 Big Data Analytical Methods

Big data analysis can be useful in enlightening the decision-making process in numerous areas such as environmental, natural disaster, and resources management. Numbers of big data analytical methods are used to infer a value/decision from the acquired big data (Chen et al. 2012; Manyika 2011). Most important of them are listed with their characteristics in Table 3.1.

3.4 Big Data and Water Resources Management

Water usage is more than double the rate of the population growth in the last century, which makes water as one of the precious resources of the present decade. This also increased the importance of effective management of water resources via the big data analytics. The major 5 "V" capabilities of big data (shown in Fig. 3.1) can help in proper perception and management of these scarce water resources.

3.4.1 Types of Water Data and Data-Sharing Methodologies

A diverse set of information that addresses the environmental, physical, ecological, social, economic, cultural, and political parameters of water usage, availability, and accessibility is known as water data. Water data can be divided into five categories: (i) water quality: The physical, biological, and chemical characteristics of water are often referred as water quality, an important parameter to determine the potability of water. To identify the water quality, single measurement is not enough but measurement of the number of water characteristics is required. It is a measure of the condition of water usually in reference to the requirements of some ecological process or anthropogenic purpose. (ii) Water quantity: It is often regarded as a rate at which volume of water is moving downstream (Wanielista et al. 1997). (iii) Water use: It includes the human consumptive uses (i.e., per capita), application by various sectors (i.e., agriculture, industry), environmental practice (i.e., evapotranspiration rates), and ecosystem services. (iv) Water extremes: Hazard and natural disaster-related data that include drought/flood monitoring and weather data. (v) Water indicators: Such indicators are generally linked to few common aspects of human or environmental health. Water indicators integrate other water-related data to provide a metric for water sustainability and utilization for human well-being (Sternlieb and Laituri 2010). Water data can be generated as primary and secondary data. The collection of water quality and quantity-related raw data can be defined as the primary data. For the measurement of the primary data, different methods are used depending upon the characteristics of water and availability of resources. Data

Table 3.1 Big data analysis methods and their characteristics

Sr. No.	Methods	Characteristics
1.	A/B or bucket or split testing	For the improvement of given objective variables, determination of the required treatment is done by comparing the control group with a variety of other test groups
2.	Association rule learning or fuzzy learning	This method compromises variety of algorithms to produce and test possible rules for determining interesting relationships
3.	Classification	It consists of supervised and unsupervised learning techniques to recognize the appropriate categories in which new data points fits
4.	Cluster analysis	As training data are not used in this method, it is considered as a type of unsupervised learning
5.	Crowd sourcing	In this method, data has been collected from the large group of people. Open-call technique has been used for this purpose
6.	Data fusion and integration	It is used for the integration and analysis of the data from variable sources in order to develop insights in more accurate and effective manner
7.	Data mining	As the name suggests it is basically the data extraction technique. It includes (i) association rule learning, (ii) classification, (iii) cluster analysis and (iv) regression methods
8.	Ensemble learning	In this type of the supervised learning, multiple predictive models have been used to obtain better predictive performance
9.	Neural networks	It is a series of algorithms that undertake to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Various water quality parameters can be simulated using this computational method
10.	Network analysis	It is basically used to characterize relationships among discrete nodes in a network or graph

(continued)

Table 3.1 (continued)

Sr. No.	Methods	Characteristics
11.	Optimization	In this method, complex systems and processes are redesigned to improve their efficiency according to the specified objectives
12.	Pattern recognition	It is a set of machine-learning methods that assign label (some sort of output value) to instance (given input value) according to a particular algorithm
13.	Predictive modeling	In this method, a mathematical model is created/selected to best predict the probability of an outcome. Water quality predictive models can incorporate both mathematical expressions and expert scientific judgment
14.	Regression	A set of statistical methods to understand how the value of the dependent variable changes when one or more explanatory (independent) variables is modified. Ex.: Water usage can be estimated indirectly by applying multiple regression analysis
15.	Spatial analysis	It is a set of analyses methods used for the identification of geometric, topographic, and geographic informations encoded in a dataset
16.	Statistics	Statistics refers to the science of the data acquisition, organization and interpretation. For example, Response Surface Methodology (RSM) explores the relationships between several explanatory (independent) variables and one or more response variables
17.	Simulation	Modeling the behavior of complex systems, often used for forecasting, predicting, and scenario planning
18.	Time series analysis	Temporal analyses of data points are significant methods to extract significant results from the acquired datasets. For example, water quality and quantity data collected from specific time intervals to represent the real situations

which is derived directly from the sensors or hydraulic measurements are known as secondary data. Primary data can be easily shared as compared to secondary data.

Water resources-based data is highly fragmented as data is generated by number of entities and warehoused in many locations. Due to the fragmented nature, water data sharing is considered as a barrier toward big data capabilities. The data fragmentation

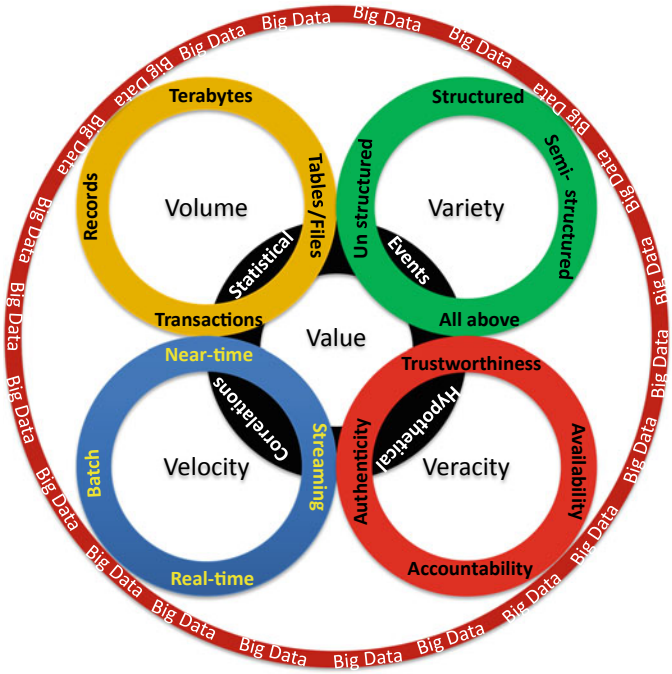

Fig. 3.1 Characteristics of Big data

problem can be overcome by using three common methodologies: (i) one-to-one: As name suggests, the data is generated by one entity and used for a single purpose. The most common example includes the academic research study or a contracted consulting project, (ii) one-to-many: In this case, data are generated by one entity and provided to many users for many purposes, and (iii) many-to-many.

3.4.2 Appositeness of Big Data to Water Resources

Science has been driven by the data but with advancement of technology, the word data has been replaced by big data. Water resources, one of the significant fields of environmental science, comprise a big data issue and flourishes increasingly. Big data helps in identifying the suitable data to resolve the problems, which are difficult to be addressed by traditional data. Some of the major applications of big data are highlighted here:

Irrigation process which requires appropriate amount of water is mainly dependent on the number of climatic factors as well as on crop and soil types. These data can be easily provided with the help of the automated sensors and continuous monitoring systems. By using the big data farming, efficiency can be improved through reducing

water requirements. Variety of automated sensors, continuous monitoring systems, robotics, and computational technology provide useful information related to the water quality which enables to understand the movement of chemicals. In addition, big data can be helpful in monitoring flood, tsunami, and drought conditions as well as the melting of ice and related climatic problems can also be monitored.

In addition to the above-discussed water resources-based applications, big data techniques have been also utilized for many applications such as oceanic (e.g., oil spill pollution detection), agriculture (e.g., food monitoring and security), urban planning, management, and sustainability, climate change (global warming, acid rain), energy assessment, disease problem, ecosystem assessment, land development and use, and so on.

3.4.3 Limitations of the Big Water Data Analytics

Big data analytics help in identifying, analyzing, and interpreting the available data for the proper management of water resources. However, at present, water resources systems in many developing countries are organized with the help of hydrological data. This represents potable water accessibility and availability data from which demand for the current and future generations can be derived. Such type of the conventional datasets mostly leads to ineffective planning, design, and functioning of water management schemes. The following listed limitations need to overcome to acquire complete benefit of the big data analytics.

Because of its large volume, the quality of stored and transmitted database is one of the major concerns in big data. Errors can be introduced from the first stage of data collection to the final deposition. Most of the automated instruments are either battery operated or need some kind of power supply. Sudden failure of which is directly associated with the gap in time series data. For example, data gap usually happens during the measurement of water consumption data using the smart water meters. Water resources quality data are complex to handle, store, and process because of their heterogeneous nature. Hence, modeling is still being done using traditional simulation models supported by GIS data.

3.5 Big Water Data Platform Components and Structure

Water resources management-related conceptual framework of the big water data open platform is shown in Fig. 3.2. It basically consists of nine blocks as discussed below: (i) The first bloc, i.e., decision support tools contain decision support technique to resolve the real-world difficulties. Because of the various available techniques, the first difficulty lies in the selection of the best decision method. (ii) Knowledge-based system deals with collection and storage of water data and ultimately transfers that information to stakeholders, including professionals and experts

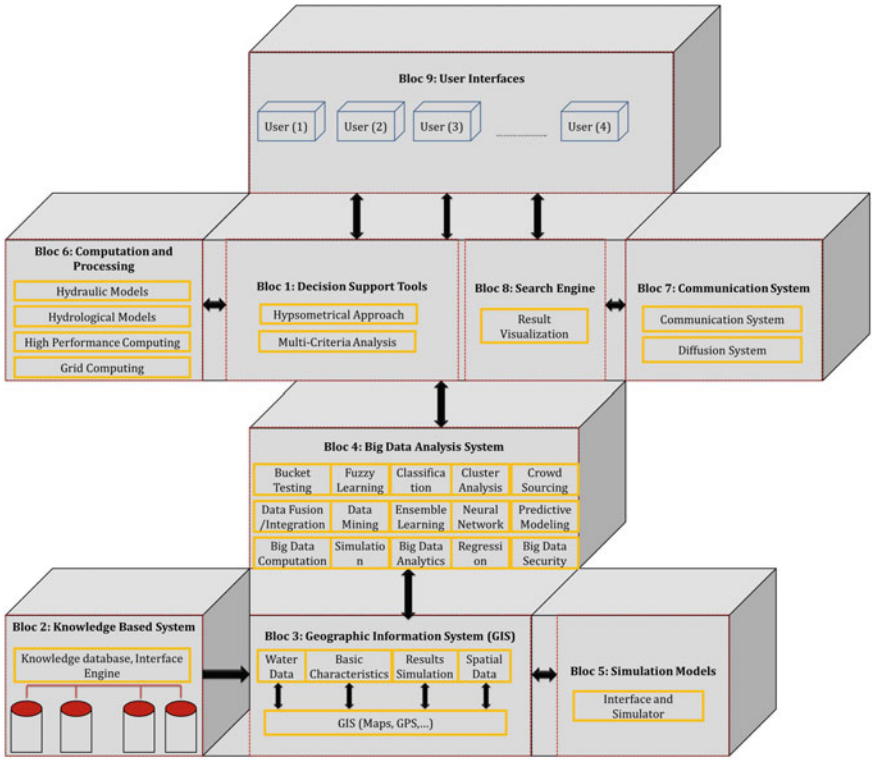

Fig. 3.2 Big data open platform for water resources management

in the field of water research. (iii) The third block geographic information system (GIS) is generally used to capture, store, analyze, and integrate complex hydrological data. ArcGis normally collects maps, applications, data, and allows users to recognize data in order to quickly deduce the best conclusion. (iv) The most important bloc of the big data platform is big data analysis system which consists number of tools to arrange, investigate, envisage, and extract useful water sources regarding information from large quantities and varieties of datasets. It requires suitable technologies (like, big data computing, analytics, mining, and security) to competently process large quantities of data.

(v) With help of the fifth bloc called simulation models, the data acquired from GIS will be linked and tried to simulate the water-associated difficulties using the simulators and interface. (vi) Sixth block of the big data platform, computation and processing, furnishes a receptacle of tools like hydraulic/hydrological models and high performance/grid computing. These tools help for the advancement of water resources prediction. (vii) After acquiring and processing the water data, the next important bloc is the communication system which makes pertinent data and information available to achieve efficiency and effectiveness. (viii) Search engine as

the eighth block enables users to find the suitable information from the big water data warehouses. (ix) User interfaces as the ninth and final bloc help operators to formulate the water resources-based problems by entering related data and portraying the obtained results and graphics.

3.6 Modern Big Data Cycle in the Context of Water Resources

Some meaningful outputs from the collected data can be drawn in order to reach up to the final conclusion. In general, two main processes, i.e., data management and analytics are used for extracting meaningful results from the big water data. The term data management can be defined as the acquisition of data, its temporary storage and final preparation for suitable for analysis. Analytics refer to methods utilized to investigate and get conclusive findings from big data. Both of these processes are normally divided into five stages as shown in Fig. 3.3. Data management is the first process which needs to be performed, after acquiring the big data. From this process, the structured data can be stored and retrieved using some traditional methods such as data marts and data warehouses. Extract–load–transform (ELT) tools are used for extraction, transformation, and loading of data into the final database.

One or more analytical methods discussed in the above section have been used by water engineers and scientists for the modeling and management of water systems. (Shafiee et al. 2018) proposed the framework for the state of the flow of water data as represented in Fig. 3.4. Number of sensors have been installed in the environment for the collection of data. After proper data management, the stored data was embedded into models for analysis and interpretation. In this system (Fig. 3.4), the water data lake collects data during every stage. Analytics handles and further processes raw

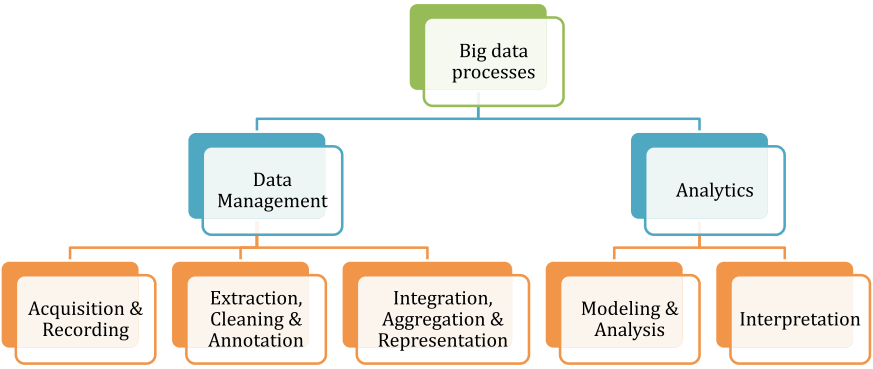

Fig. 3.3 General classification of the big data processes

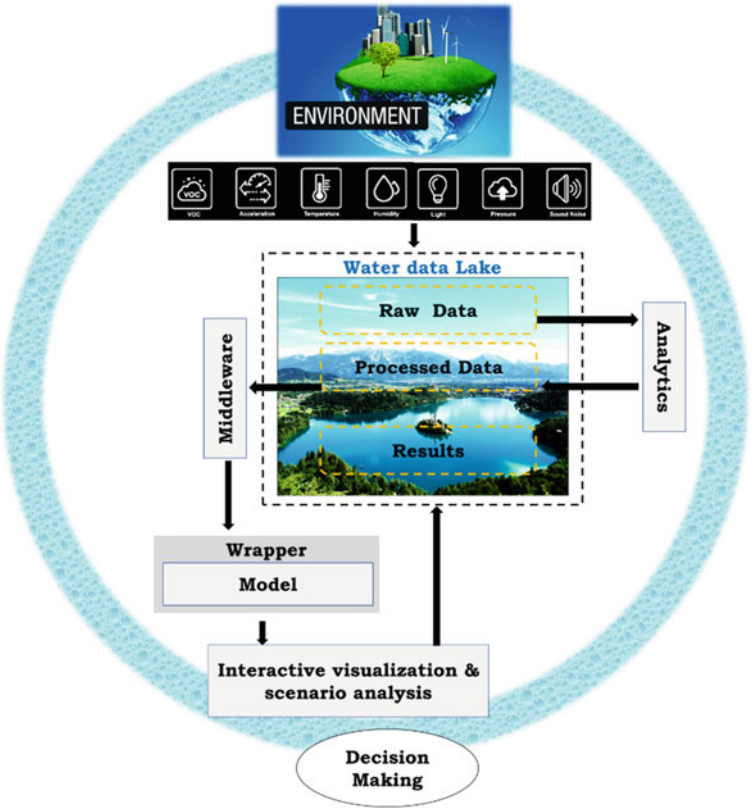

Fig. 3.4 Typical water data lake

data and finally, returns cleaned/forecasted data. Middleware pulls, aggregates, and formats data for a model. A wrapper provides communication capabilities to a model.

3.7 Future Perspectives of Big Data for Water Resources Management

At present, a number of big data platforms are available related to the water resources. Table 3.2 displays some of the common big data platforms pertaining to the water resources with their major objectives, significance, and limitations. Based on the associated limitations, the prospective applications of big data in water resources management are highlighted below.

As it has been mentioned, big data techniques have demonstrated wide applications in the decision-making process by predicting the outcomes. However, despite having access to a broad range of data sources and technical resources, the water

Table 3.2 Common big data platforms pertaining to the water resources with their major objectives, significance, and limitations

Sr. No.	Model	Objective	Importance	Limitation	References
1.	Big data platform	To solve water resources problems using big data analytics	It is important in providing effective tools to solve complex water resources systems, water modeling issues and helps in decision making	The heterogeneity of water data coming from various sources causes problem in architecture of the big data platform	Chalh et al. (2015)
2.	Cloud GIS platform	To develop the water resources and hydropower cloud GIS Platform	It has powerful ability of data mining and analysis. Big data-based cloud GIS platform of water resources and hydropower can provide the suitable decision support for the design, development, operation, and maintenance	The security of the data is the key element which restricts the development of the cloud GIS platform.	Wang and Sun (2013)
3.	Big data water resource management standards	To apply big data method in the development of water management standards	It can assemble and organize the basic data of water resources and assists the setting of water resource management standards	Supervision and data standardization are the major associated problems	Bai et al. (2017)

utility sector appears to make partial use of it for the enhancement of water quality and source distribution. With the high-density survey of big data, the risk of the error will also increase mainly due to the lack of the availability of instant processing techniques. Hence, the future perspective of big data research is not to obtain more and more data but it should mainly focus on the development of the new generation of smaller, cheaper, and accurate sensors to produce real-time data. The integration techniques can be helpful in improved decision making and management of the water resources. For example, machine learning, one of the analysis techniques, is able to extract accurate patterns and relationships from the data. At present, a number of the models, methodologies, and techniques are accessible for the planning and management of the water resources. However, none of them provides a convenient solution.

3.8 Conclusion

With advancement of the computer science and Web technology, the data generation has become increasing in day-to-day life. These large datasest ultimately pose challenges in its storage, handling, analysis, and interpretation. Water is one of the prime requirements for the survival of life and is progressively becoming a precious resource due to its inflated usage. Increased population and economic/industrial growth cause stress on available water resources. Similarly, climate change also significantly affects the water resources due to its direct effects on important hydrological processes, i.e., precipitation and evaporation. With the help of big data, each and every component of environment, such as water resources, can be managed.

The aim of the current chapter is to present an overview of big water data, associated characteristics, applications, and limitations. It also gives a summary related to the open big data platforms/proposed models supporting water resources. The authors can get the specific idea about the available models by referring to this chapter. It also highlights the future perspective required for the proper utilization of big data technique for the water resources management. Despite the increasing importance of modeling in water resources management and planning, no single methodology/tool provides an acceptable solution. Hence, more research is required in development of single but comprehensive methodology/tool. The basic available models are generally restricted to local/regional-level strategies, while the challenges are transdisciplinary and encompass knowledge from various sciences and engineering backgrounds.

References

Adamala S (2017) An overview of Big Data applications in water resources engineering. Machine Learning Res 2(1):10–18

Bai Y, Bai X, Lin L, Huang J, Fang H, Cai K (2017) Big Data technology in establishment and amendment of water management standard. Appl Ecol Environ Res 15(3):263–272

Bibri SE, Krogstie J (2017) On the social shaping dimensions of smart sustainable cities: a study in science, technology, and society. Sustain Cities Soc 29:219–246

Chalh R, Bakkoury Z, Ouazar D, Hasnaoui MD (2015) Big data open platform for water resources management. In: 2015 International Conference on Cloud Technologies and Applications (CloudTech), pp 1–8

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from Big Data to big impact. MIS Q 36(4)

Cox M, Ellsworth D (1997) Application-controlled demand paging for out-of-core visualization. In: Proceedings. Visualization '97 (Cat. No. 97CB36155), pp 235–244

David H, Branko K, Amin R, Avi O, Barbara M, Katherine BM (2014) Sensing and Cyberinfrastructure for smarter water management: the promise and challenge of ubiquity. J Water Res Plan Manage 140(7):01814002

du Plessis A (2019) Current and future water scarcity and stress. In: du Plessis A (ed) Water as an inescapable risk: current global water availability, quality and risks with a specific focus on South Africa, pp 13–25. Springer International Publishing

Gandomi A, Haider M (2015) Beyond the Hype: Big data concepts, methods, and analytics. Int J Inf Manage 35(2):137–144

Kumar M (2019a) Micro-components quantification of end uses of water consumption in low income settings. Interim Project Report. 1–44

Kumar M, Chaminda T, Honda R, Furumai H (2019b) Vulnerability of urban waters to emerging contaminants in India and Sri Lanka: resilience framework and strategy. APN Science Bulletin

Kumar M, Deka JP, Kumari O (2020) Development of water resilience strategies in the context of climate change, and rapid urbanization: a discussion on vulnerability mitigation. Groundwater Sustain Dev 10:100308

Madden S (2012) From databases to Big Data. IEEE Internet Comput 16(3):4–6

Manyika J (2011) Big data: the next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

Mayer-Schönberger V, Cukier K (2013) Big Data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt

Mukherjee S, Patel AKR, Kumar M (2020) Water scarcity and land degradation nexus in the era of anthropocene: some reformations to encounter the environmental challenges for advanced water management systems meeting the sustainable development. In: Kumar M, Snow D, Honda R (eds) Emerging issues in the water environment during Anthropocene: A South East Asian Perspective. Springer Nature. ISBN 978-93-81891-41-4

Patel AK, Das N, Kumar M (2019) Multilayer arsenic mobilization and multimetal co-enrichment in the alluvium (Brahmaputra) plains of India: a tale of redox domination along the depth. Chemosphere 224:140–150

Roshan A, Kumar M (2020) Water end-use estimation can support the urban water crisis management: a critical review. J Environ Manage Ms. Ref. No.: JEMA-D-20-00036R1. https://doi.org/10.1016/j.jenvman.2020.110663

Schroeck M, Shockley R, Smart J, Romero-Morales D, Tufano P (2012) Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report. In: IBM Institute for Business Value and Said Business School at the University of Oxford

Shafiee ME, Barker Z, Rasekh A (2018) Enhancing water system models by integrating big data. Sustainable cities and society 37:485–491

Singh A, Patel AK, Kumar M (2020) Mitigating the risk of Arsenic and Fluoride contamination of groundwater through a multi-model framework of statistical assessment and natural remediation techniques. In: Kumar M, Snow D, Honda R (eds) Emerging issues in the water environment during Anthropocene: a South East Asian perspective. Springer Nature. ISBN 978-93-81891-41-4

Sternlieb FR, Laituri M (2010) Water, sanitation, and hygiene (WASH) indicators: measuring hydrophilanthropic quality. J Contemp Water Res Educ 145(1):51–60

Wang X, Sun Z (2013) The design of water resources and hydropower cloud GIS platform based on big data 313–322

Wanielista M, Kersten R, Eaglin R (1997) Hydrology: water quantity and quality control. In: Hydrology: water quantity and quality control

Zikopoulos P, Eaton C (2011) Understanding Big Data: analytics for enterprise class Hadoop and streaming data, 1st edn. McGraw-Hill Osborne Media