




Review

Big Data Analytics and Its Role to Support Groundwater Management in the Southern African Development Community

Zaheed Gaffoor ^{1,2}, Kevin Pietersen ³, Nebo Jovanovic ^{1,4}, Antoine Bagula ⁵
and Thokozani Kanyerere ^{1,*}

¹ Department of Earth Science, University of the Western Cape, Cape Town 7535, South Africa; zaheed.gaff@gmail.com (Z.G.); njovanovic@uwc.ac.za (N.J.)

² IBM Research, Africa Labs, Johannesburg 2000, South Africa

³ Institute for Water Studies, University of the Western Cape, Cape Town 7535, South Africa; kpietersen@mweb.co.za

⁴ Council for Scientific and Industrial Research, Stellenbosch 7600, South Africa

⁵ Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa; abagula@uwc.ac.za

* Correspondence: tkanyerere@uwc.ac.za; Tel.: +27-21-959-9292

Received: 13 July 2020; Accepted: 25 September 2020; Published: 9 October 2020



Abstract: Big data analytics (BDA) is a novel concept focusing on leveraging large volumes of heterogeneous data through advanced analytics to drive information discovery. This paper aims to highlight the potential role BDA can play to improve groundwater management in the Southern African Development Community (SADC) region in Africa. Through a review of the literature, this paper defines the concepts of big data, big data sources in groundwater, big data analytics, big data platforms and framework and how they can be used to support groundwater management in the SADC region. BDA may support groundwater management in SADC region by filling in data gaps and transforming these data into useful information. In recent times, machine learning and artificial intelligence have stood out as a novel tool for data-driven modeling. Managing big data from collection to information delivery requires critical application of selected tools, techniques and methods. Hence, in this paper we present a conceptual framework that can be used to manage the implementation of BDA in a groundwater management context. Then, we highlight challenges limiting the application of BDA which included technological constraints and institutional barriers. In conclusion, the paper shows that sufficient big data exist in groundwater domain and that BDA exists to be used in groundwater sciences thereby providing the basis to further explore data-driven sciences in groundwater management.

Keywords: transboundary aquifers; data-mining; Internet of things; machine learning; remote sensing

1. Introduction

Big data analytics is a revolutionary new buzz-word describing the use of advanced and traditional analytical techniques to leverage vast quantities of heterogeneous data, in-order to provide valuable insights that can be used to propel optimization, development and knowledge discovery [1,2]. To date, the surge of data from online social media activities, internet activities, business transactions, scientific missions, digitization and sensor technologies, among many others, benefit many industries in understanding their operational environment. Collectively these data are referred to as big data. For instance, some healthcare institutes now readily utilize data from electronic patient records, physician notes, medical equipment, social media, to predict the outcome of treatments, the onset

of diseases and the spread of infectious diseases [3]. This has helped reduce treatment costs and improved overall patient care. In the astronomy field, the recent development of sky survey missions has created an avalanche of astronomy big data ranging in the Petabytes. The consequence is that astronomers now must rely on more advanced data collection, storage, analysis and dissemination tools and techniques, to be able to draw value from data [4]. This has moved the astronomy discipline into a data-driven science.

The earth sciences discipline, like many other scientific disciplines, has in itself been driven into the big data era with the advancement of sensor technologies, such as remote sensing, that continually collect new data [5]. This has paved the way for the introduction of data-driven approaches in the earth science discipline. It is not a surprise that in recent times the potential for big data to support knowledge discovery in the hydrogeological discipline has become apparent [1]. For example, [6] showcased the use of the big data open platform to support water resource management in the Fouta Djallon watershed, Morocco. The platform utilizes a number of tools such as stochastic models, simulations, hydraulic and hydrological models, high performance computing, grid computing, decision support tools, big data analysis systems, communication and diffusion systems, database management, geographic information system (GIS) and knowledge-based expert systems to extract information from a variety of heterogeneous datasets. Through decision support tools such as hypsometrical approach, users can understand the impacts of various future management scenarios. Ref [7] demonstrated the potential of big data analytics to mapping groundwater potential in Goyang-si, South Korea, by combining data from borehole-pumping activities and satellite-based earth observation data. In fact, recent interest in big data analytics has spurred a special section in Water Resource Research focusing entirely on the application of big data analytics in hydrological research [8]. Nonetheless, applications of big data are still very incipient in the discipline of hydrogeology. As such, the range of applicability of big data in the hydrogeological field has not been fully explored, hence the motivation for this review paper.

One area of application where big data may be useful is in the support of sustainable groundwater management. Groundwater well field data can be fully digitized, collecting real-time information from sensors-equipped monitoring systems and other relevant sources, fed into advanced analytical algorithms to provide well field managers with useful insight to support them during decision-making scenarios. This concept is already applied successfully in the shale gas industry, where data from operational equipment, written notes, user inputs are analyzed on-the-fly to support drilling and production operations [9]. Perhaps this vision is far-off with the current state of the art in groundwater management. However, it indicates the potential for big data to support operational decision-making in groundwater management. By combining data from well field operations, drilling reports, groundwater models, remote sensing and field monitoring programs (and other unconventional sources), we can extend limited groundwater datasets. Through big data analytics, we can transform these data into information for groundwater managers to exploit. IBM is leading the way in this charge, trying to develop digital aquifers that rely on Internet of things (IoT) equipped wells, smartphone data from humans, weather data, and paper records to model operations of the aquifer in the cloud [10]. The goal is to support sustainable management of groundwater.

The beneficial uses of groundwater in the Southern African Development Community (SADC) region have been documented [11,12] and the groundwater challenges facing Member States have been elucidated in a number of reports [11–13]. These challenges include issues such as over-abstraction, institutional mismanagement, pollution and climate change and variability among others. Efforts to address some of the challenges are often curtailed by the lack of relevant data, especially at a local scale. This is largely as a result of decision-making processes being often ill-informed due to the lack of relevant data and information. This lack of data is generally due to inadequate monitoring efforts in the region, the disparate storing of the data, and the ineffective transformation of data into information to support the decision-making process [12]. These challenges are further exacerbated when dealing with transboundary aquifers, where data collection and management of water resources across state boundaries is inconsistent.

These issues hamper the sustainable management of groundwater in the SADC region. In this case, big data may provide a useful tool that can be used to fill data gaps, by exploring, collecting and integrating various sources of groundwater big data (both conventional and unconventional). As well as provide the analytics to transform these data into valuable groundwater information to support sustainable groundwater development in SADC region through big data analytics methods. Big data analytics may also provide the opportunity to address scale issues, where methods can be employed to downscale regional groundwater data to local conditions in support of localized groundwater management (e.g., individual boreholes, wellfields). In many groundwater management scenarios, a local scale analysis is more desired [14]. For SADC groundwater, investing in big data may enable effective harnessing of data from a multitude of new sources, to improve monitoring by using new sensor technologies, to centralize data storage and management and to apply new advanced analytics that can uncover new patterns and trends to drive knowledge discovery.

The aim of the paper is to highlight the potential role big data analytics and big data can play in supporting groundwater management in SADC. Therefore, in this paper we present the current state of the art of big data and big data analytics in the groundwater discipline and explain how it can be applied to groundwater management in the SADC region. The vast spectrum of data sources, analytics tools, and technologies are challenging to navigate while trying to ensure data integrity and accuracy. For this purpose, specialized big data analytics frameworks are employed to facilitate the management and application of big data analytics. Therefore, by drawing on the findings of the review, a novel conceptual big data analytics framework is proposed, that is uniquely designed to address challenges of groundwater management in the SADC region. This paper provides a foundation for the application of big data analytics in the groundwater discipline, in particular for problem-solving applications to the SADC region, paving the way for further work into data-driven sciences.

2. Research Approach

The information presented in this paper is based on a review of relevant literature. By summarizing and extracting critical information from key research in the big data, big data analytics, water resources and groundwater discipline, we establish the current state of knowledge on big data in the groundwater discipline. For example, we first define what big data are (Section 3.1), then describe where big data comes from in the groundwater sciences in general (Section 3.2) and then describe the various big data analytical methods that are relevant (Section 4). Thereafter we present a brief review of the big data analytical framework (Section 5). Where appropriate, a relation is made to a SADC setting, in order to contextualize the review. The findings of the review are then used to facilitate the development of a proposed conceptual framework for SADC (Section 6). Finally, we end the paper with a discussion of the expected challenges facing the application of big data analytics in a SADC context. Figure 1 illustrates a road map of the paper and how the findings of each section relate to the framework.

The relevant literature was sourced through key word and phrase searchers in popular web-based search engine, such as Google and Google Scholar, SpringerLink, Scopus and Mendeley. Key words and phrases used to search for relevant literature include, “big data”, “big data analytics”, “big data and groundwater”, “big data analytics and groundwater”, “big data and water”, “big data analytics and water”, but not limited too. A total of 135 papers are cited based on the review process.

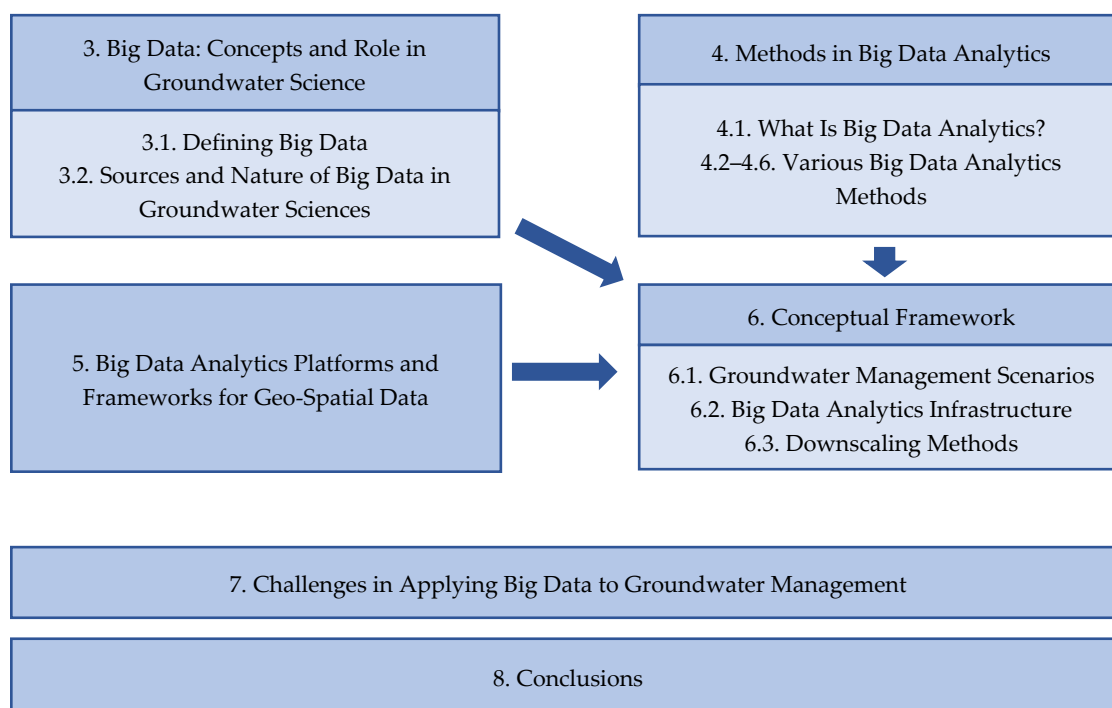


Figure 1. Road map of the research.

3. Big Data: Concepts and Role in Groundwater Science

In this section, we present the landscape of big data in the groundwater discipline. Drawing from the literature, we define what big data are and we introduce the various sources of big data relevant to the groundwater discipline including what these big datasets are composed of and how they relate to support groundwater management in SADC.

3.1. Defining Big Data

Big data are referred to as collections of very huge datasets with a great diversity of types that makes it difficult to be collected, stored and analyzed by conventional tools and techniques [15,16]. Big data have a few characteristics that separate them from generally large datasets. These characteristics are recognized as the Vs of big data [17]: *volume*—big data consist of enormous quantities of data, generally beyond a threshold of one terabyte, however this changes with time, sector, data types and use case; *velocity*—big data are generated at an exceptionally high rate, such that the volume of big data increases rapidly over time; *variety*—big data are composed of a variety of different data types from a variety of sources [17].

The three Vs (volume, velocity and variety) are the commonly defined features of big data, which were first coined by [18]. Since then, industry experts have added additional Vs to define big data. For example, IBM added *veracity*—which describes the inherent inaccuracy and uncertainty present in most large datasets and complex datasets [19]. SAS introduced *variability & complexity*—which describe the ever changing nature of big data over time with respect to velocity and variety [17,20]. Oracle introduced *value* as an additional V—which stipulates that big data must contain new knowledge or improve operational efficiency for them to have any meaning in terms of financial investment [17,20]. This value is usually achieved through the use of analytics which transforms the raw data into useful information. For SADC groundwater to realize the value of big data, thought must be given to understanding the Vs in the context of groundwater big data in Southern Africa, as well as the analytics required to turn these data into useful information for groundwater management.

Big data types play a role in how big data are managed from data to information. They can be broadly categorized into structured and unstructured data [17]. Structured data are any type of data

that can easily be stored, categorized and referenced in tabular form. The main tool to store, access and query this type of data is through relational databases, making them easily readable by machines [20]. For example, conventional hydrological data generated through in situ monitoring commonly constitute point information that can easily be captured in relational databases and conventional spreadsheets. This is typical of structured data.

On the other hand, text, video, audio and images are examples of unstructured data. These lack higher structural organization and are not easily stored in relational databases [20]. For example, videos of a flooding events or social media posts related to various aspects of water and groundwater, constitute unstructured data relevant to groundwater. In addition, remote-sensing images constitute unstructured data, but the meta-data attached to the image is structured [5,21]. Unstructured data are particularly difficult for machine programs to extract information from, at least with traditional techniques. Semi-structured data have some form of structure; however, these tend to be very irregular and often heterogeneous, which makes categorization challenging. Emails and XML files fall into the semi-structured data type [17,20,22].

3.2. Sources and Nature of Big Data in Groundwater Sciences

The previous section introduced the characteristics that define big data in general. Intrinsically, these are also the characteristics that make big data difficult to leverage with traditional information systems alone [23]. In this section, we try and define sources and nature of data in the groundwater domain, within a big data context.

A common awareness among data scientists is that not all big data are the same and that the structure and nature of big data and how we analyze them depend on the domain [5]. For example, geospatial data differ from text data (such as from social media posts) and the techniques and tools used to collect, store and analyze each of these types of data will be different [15]. The result is that one needs to fully understand the specificities of the relevant data sources and what information is required from these data before appropriate big data tools, techniques and analytics can be applied.

Data in the groundwater domain has not been static. Over the years, groundwater scientists have explored various sources to collect groundwater data. Table 1 illustrates these sources of data relevant to groundwater. Table 1 includes the traditional sources of groundwater data such as in situ observations or hydrogeological maps, as well as modern data sources such as remote sensing, social media or Internet of things (IoT). Individually, some of these sources may not have the characteristics of big data, but when harnessed together they provide some substantial opportunities for knowledge discovery. Large scale data assimilation models are one example of such systems that incorporate data from different sources, such as field activities, remote sensing and computer simulations. However, at the moment they do not ingest data from unconventional big data sources, such as social media [24].

3.2.1. Field Activities and Historic Sources of Data in Groundwater

In the groundwater sciences, one of the primary sources of data are observations collected during field operations. These activities include drilling operations, pumping operations and monitoring operations. Drilling operations collect data on geological and hydrogeological properties of the aquifer, such as lithology and water strikes. Pumping operations collect data on hydraulic properties of the aquifers, such as yield. Field-based hydrological monitoring operations typically involve the selection of sampling sites (in a hydrogeological context these are mostly boreholes, piezometers and springs) and the collection of in situ point data through the use of various techniques and instrumentation [25]. These data are considered direct observations and are thus typically robust in terms of accuracy. In addition, these data represent local conditions within an aquifer, and are thus preferred for groundwater management. Drilling and pumping operations tend to be occasional activities, while field monitoring data collection is generally carried out on a quarterly basis but may even be less frequent. In modern times, the use of sensors equipped at sampling sites have increased the frequency at which observations are recorded at sampling sites. In some cases, these sensors are

connected to remote monitoring centers, allowing off-site data collection. However, this is not the norm across SADC region member states.

Table 1. Sources of data in the groundwater domain from a big data context.

Source	Description	Characteristics
Field activities	Data generated from field activities such as monitoring, drilling and pumping activities	Structured data format Limited coverage (spatially and temporally) Local
Historical	Legacy reports, maps and documents	Unstructured Local or regional Text or images
Remote sensing	Satellite, airborne or ground-based earth observation	Unstructured and structured Multidimensional Voluminous Regional
Computer simulation	Data generated through computer-based models	Unstructured and structured Multidimensional Voluminous Regional
Social media and the web	Data available on webpages and social media post	Unstructured Textual, images, videos or audio Multidimensional Heterogeneous Voluminous Local
Internet of Things	Data available from connected devices	Unstructured and structured Heterogeneous Multidimensional Local

In the SADC region, there are generally two challenges affecting the impact of these data to support groundwater management. The first challenge is that collecting data from field activities is generally sporadic. For example, field monitoring data collection in SADC has been curtailed by the number and distribution of sampling sites having generally decreased over the years [26]. The results are limited networks of sampling sites that are actively monitoring on a regular basis. This has manifested into a generally sparsely populated (both temporally and spatially) data record across SADC. Second, data storage is disparate and in various formats. For example, some countries store data in centralized databases, while others only store data on spreadsheets or in hardcopy form [26]. This challenge ultimately affects data retrieval and sharing.

Nowadays, data from this source may be stored in databases, digital spreadsheets or in GIS files. However, in the past, the results of field activities were recorded in reports and on physical maps. These historic data exist either in hardcopy form or scanned documents. Many times, these sources of data idle in archives, as digital forms are more favorable. However, through a process called optical character recognition (OCR), written text can be converted into machine readable characters [27]. Similarly, computer vision applications combined with deep neural networks have shown potential to transform raster maps (images) into vector data [28]. Digitizing and transforming these sources of information into machine readable data can create a new stream of big data [29].

3.2.2. Remote-Sensing Big Data

Field monitoring hydrological data do not necessarily constitute big data, in the ontological sense of the word [2]. These data are easily managed and analyzed by standard information systems. When looking for big data sources for groundwater, remote earth observation systems or remote-sensing data are the obvious candidates. Remote-sensing data truly are big data, constituting highly dimensional, highly heterogeneous and increasingly voluminous datasets [5]. Remote-sensing

data constitute all data collected from ground, airborne or spaceborne earth observation instruments. Remote sensing for earth observation started in the late 1950s with the launch of the Sputnik 1 satellite [30]. Since then, hundreds of earth observation satellites have been launched, some specifically to collect data on Earth's hydrological systems, such as Landsat or gravity recovery and climate experiment (GRACE) [31]. Some of these remote-sensing missions, such as Landsat, have been active since the early 1970s [30]. Over the years new remote-sensing missions have been undertaken and advanced, which has contributed to an ever-increasing big dataset. For example, NASA's SMAP missions collect 458 GB of soil moisture data every day [32]. Table 2 illustrates some of the remote-sensing products that are relevant to groundwater. For a more detailed description of the missions, refer to [32,33]. These remote-sensing missions generally provide global coverages of gridded data products, including the SADC region. In SADC, where local in situ monitoring data are scarce, remote-sensing data can fill the gap, providing a better temporal and spatial coverage.

Table 2. Key remote-sensing missions for collection of hydrological data.

Mission/Sensor	Hydrological Component	Spatial Resolution	Temporal Resolution	Launch—End Year
Gravity recovery and climate experiment (GRACE)	Terrestrial water storage	110–330 km	monthly	2002–2017
Gravity recovery and climate Experiment-follow on (GRACE-FO)	Terrestrial water storage	110–330 km	monthly	2018—ongoing
Soil moisture active and passive (SMAP)	Soil moisture	3–36 km	1–7 days	2015—ongoing
Soil moisture and ocean salinity (SMOS)	Soil moisture	35–50 km	1–3 days	2009—ongoing
Global precipitation measurement (GPM)	Precipitation	5–15 km	30 min—monthly	2014—ongoing
Tropical rainfall measuring mission (TRMM)	Precipitation	5–550 km	3 hours—monthly	1997–2015
Terra/MODIS	Evapotranspiration, LST, NDVI	0.5 km	8 day—annual	2000—ongoing
Sentinel 3 and 3B	LST, NDVI, GVI	various	various	2016—ongoing

N.B. The ranges expressed in the spatial and temporal resolution reflect the specification of various science data products emanating from the missions. LST—land surface temperature; NDVI—normalized difference vegetation index; GVI—global vegetation index.

Remote-sensing big data also have the potential to provide spatial and temporal coverage needed to close terrestrial water budgets [34–37], although uncertainty in sensor estimates and over-simplification of water budget models has often resulted in erroneous results.

On the other hand, one challenging aspect of remote-sensing data for groundwater management is the coarse spatial resolution of the data. Hydrological investigations using remote-sensing data generally have been carried out at regional or global scales. This is because much of the remote-sensing data are at a spatial scale that does not support local or site-specific analysis (Table 2) [33]. This is especially true for GRACE data, which has a spatial resolution of 110 km. At this scale, many of the smaller transboundary aquifers would be contained in one or overlap only a few GRACE pixels. This hinders their applicability to local scale use. In fact, most of the studies done using GRACE data have focused on regional scale investigations [38–42]. In order to be applicable to local scale groundwater management, the resolution of GRACE data must be refined. Big data analytics has the potential to apply methods for downscaling remote-sensing data to support local groundwater management.

Ground-based and airborne geophysical surveys also contribute to remote-sensing data. Geophysical surveys are a broad category of observational techniques, which can be used to collect data on aquifers and groundwater properties. Active geophysical methods rely on generating some type of

artificial energy fields, such as electro-magnetic field and recording the interaction with the water or rock interfaces. Passive geophysical methods rely on measuring natural fields of the Earth at various location, such as the magnetic field and inferring rock or water properties from these observations [43]. Geophysical methods are numerous and include ground penetrating radar, electric resistivity and seismic reflection/refraction, among many others [44]. Geophysical survey methods allow data collection at greater spatial scales than in situ point observations, but smaller than satellite-based remote sensing. However, they are expensive, and they are generally only performed during groundwater exploration exercises. Thus, these types of data are not encountered frequently in SADC, but are available for some transboundary aquifers in SADC, such as the Zeerust/Lobatse/Ramotswa dolomite aquifer [45].

3.2.3. Simulated Hydrological Data

In this section, we discuss hydrological data generated through computer models or through reanalysis applications. In essence, these datasets represent synthesized data, generated through numeric methods and data assimilation techniques. The data available through these sources are comprehensive, providing detailed spatiotemporal data on numerous hydrological variables.

In this category of big data, one source stands out as being extensive—that is the results of atmospheric models. This is a broad category of numeric weather and climate models that are used to predict future weather and climate patterns in the short and long-terms and at the regional or global scale [46]. It includes models such as global circulation models (GCM), regional climate models (RCM) and numeric weather prediction (NWP) models. Some of the most advanced atmospheric models, such as GCMs, are often coupled with land-surface models, sea ice component and ocean circulation models and are only capable of being run on powerful supercomputers [47]. Hence, the amount of data processed and generated by these models is enormous. These data are often made available to the general public for free or through various paid license agreements. For example, European Centre for Medium-Range Weather Forecasts (ECMWF, Reading, UK) disseminates much of their data via their website (<https://www.ecmwf.int/en/forecasts/datasets>). Of particular relevance to hydrological sciences is the forecast data for precipitation, which may be useful when understanding the future trends in groundwater resources in the SADC region.

Land-surface modeling applies complex mathematical equations to integrate hydrological, biologic and radiation-based energy exchange processes at the land-surface, between the land surface and the atmosphere and within the soil-column [48]. These models assimilate an extensive array of both in situ and remote-sensing-based observational data to derive natural fluxes at the earth surface [49]. For example, the land data assimilation system from NASA provides numerous datasets on various hydrological variables, such as evapotranspiration, soil moisture and run-off, on a global scale (visit <https://earthdata.nasa.gov/> for data retrieval). Datasets from these systems are particularly useful for hydrological applications, providing data for an integrated systems analysis.

Lastly, reanalysis datasets provide an additional trove of historical data, which are useful to understand past trends in natural earth systems. Reanalysis data refer to original in situ observational datasets that have been reanalyzed and amended using data assimilation techniques and are generally the by-products of land-surface models and atmospheric models [50]. Examples of re-analysis datasets are ERA5 from ECMWF, NCEP/NCAR Reanalysis I from the National Center for Environmental Prediction and National Center for Atmospheric Research (NCEP, College Park, MD, USA; NCAR, Boulder, CO, USA), and the Japan Meteorological Agency's JRA-55 (JMA, Tokyo, Japan), which are easily retrievable and widely used in hydrological applications [50]. Although these datasets are primarily geared towards atmospheric sciences and land-surface states, many of the parameters included in the datasets are correlated to groundwater processes (e.g., stream discharge, soil-moisture), making them valuable data sources.

3.2.4. Social Media and the Web Data

With the advent of the Internet, a new channel for communication and transfer of information was created. Today, almost all industries and individuals rely on the Internet in some way. It is no surprise then that the volume of data being generated and transmitted over the Internet is both enormous and complex. Of concern to groundwater is all the hydrological information being transmitted over the Internet, which is not already stored in specialized data repositories. This means information present on webpages and social media threads, among others.

Social media provide an unconventional new big data source in groundwater sciences. It may be hard to visualize how unstructured social media data may be useful in a conventional sense, but these data types make up majority of big datasets [17]. With modern advanced analytical techniques such as natural language processing or video analytics, valuable information can be extracted from these data sources [51,52]. For example, [53] demonstrated a framework to infer actual levels of rainfall from the contents of Twitter feeds. This study used certain words/phrases combination commonly appearing in tweets, as a rainfall magnitude reference. Statistical learning was then applied to model and forecast the magnitude of a rainfall event based on the wording in twitter posts and the actual rainfall amount. In addition, [54] showed the production of real-time flood extent maps from live twitter feeds in Jakarta, Indonesia. In this case, twitter posts that contained geo-located information on water depth and extent were used to infer near real-time flood extent maps by combining the data with digital elevation models using a flood-fill algorithm. Not only do these data represent local conditions, they are also streaming in real time and they have real world applications in supporting disaster relief and risk management efforts [1,55]. These examples clearly demonstrate the potential value of unstructured data, specifically from social media, in hydrology-related applications.

However, from a groundwater perspective in the SADC region, data from social media platforms or other similar data conduits, may have limited value. These types of applications work well in developed areas, with a large number of users and sufficient Internet access. In the less developed urban areas and rural settings of the SADC region, the spatial coverage of this type of data may be limited [56]. In addition, it is very difficult to visualize groundwater from the surface, as it is hidden below layers of soil and rock. Thus, it remains to be seen whether social media related groundwater data are prevalent and quantifiable in countries within the SADC region.

3.2.5. Internet of Things Data

According to [20], an estimated 20.8 billion connected devices will exist in 2020. Connected devices are electronic equipment that can connect with each other and various digital systems over the Internet [57]. These devices include objects such as smartphones, sensor equipment or even house-hold appliances. Some of these objects are continually streaming environmental data. For example, [58] demonstrated the use of atmospheric pressure and temperature data collected through a smartphone application to improve near real-time weather predictions. Similarly, [59] showed the advantages of using smartphones and connected personal weather stations to monitoring weather patterns in Amsterdam. The real-time spatial and temporal distributions of data from these sources allow a level of data insight that was not possible before. This is the realm of Internet of things (IoT).

The application of IoT systems to groundwater science can generate large amounts of data on local groundwater conditions, faster than conventional or manual data collection, providing improved management of groundwater resources [60]. For example, real-time IoT groundwater monitoring and data management systems have been piloted in various regions, such as California and India, to improve sustainable groundwater management [61,62]. Sensor equipment is continually decreasing in cost and increasing in accuracy and may certainly improve the data collection capabilities of the groundwater domain in Southern Africa.

Additionally, citizen-science missions have shown promise in collecting environmental data such as groundwater levels [63]. In fact, virtual citizen science missions have shown to outperform conventional data collection methods, collecting data in a few days that would normally take months [64]. However,

the quality of citizen science data is not always of a high standard and proper quality assurance measures must be in place to ensure robust results. By incorporating these technologies and data collection tools into various groundwater-related initiatives, new local scale data can be generated, in some cases in real time. These data can be fed into big data analytical platforms (collections of software and hardware utilities for management of big data), integrating them with other datasets, turning them into useful information to support groundwater management [65].

4. Methods in Big Data Analytics

The value of big data is truly realized when it is transformed into useful information. Big data analytics covers a comprehensive package of advanced analytical, statistical, mathematical and graphic methods that can be used to transform the data into useful information [51]. In this chapter, we discuss big data analytics in more detail, focusing on specificities that are important for transforming groundwater big data into useful information.

4.1. What Is Big Data Analytics?

According to [51], big data analytics is advanced analytics operating on big data. Many of the tools and techniques employed in big data analytics, such as machine learning, have been available for many years [66]. It is only recently, with the surge in big data, that the value of these advanced analytical techniques has been realized. Compared to traditional analytics approaches, advanced analytical techniques perform well when dealing with very large, heterogeneous datasets, requiring less data pre-processing, as shown in Table 3 [67]. For example, machine learning can work on both structured and unstructured data, while traditional analytics works well only on structured data. One of the major differences between traditional analytics and big data analytics is the processing platforms required. Big data generally requires parallel processing methods to effectively analyze these large datasets. Big data analytics methods are designed to operate over multiple distributed processors, whereas traditional analytics methods are generally designed to operate on single machines [67]. Traditional analytical methods are only efficient when significant sampling and dimensional reduction methods (e.g., principal component analysis, genetic algorithm) are used to reduce data size. In addition, traditional analytics is not suited for parallel processing frameworks. Big data analytics together with traditional analytics may allow us to leverage various sources and types of groundwater big data, turning them into useful information for a groundwater manager to use.

Table 3. Traditional analytics vs big data analytics (adapted from [67,68]).

	Traditional Analytics	Big Data Analytics
Focus	Descriptive analytics and diagnosis analytics	Predictive analysis and prescriptive analytics
Datasets	Limited datasets with structured data. Adoption of simple data models	Large scale datasets with more types of data. Adoption of complex data models
Analysis	Looks to what happened and why?	Provides new insights and forecasts
Processing	Generally capable of being run on a single machine (centralized processing)	Requires parallel processing across multiple machines (distributed processing)

Generally, big data analytical techniques include traditional analytics such as data mining, statistical analysis, SQL queries (Structured Query Language queries) and data visualization, which work well on structured data. Advanced analytical techniques such as natural language processing, text analytics, video analytics, audio analytics, artificial intelligence and machine learning work well with heterogeneous unstructured data [17,51]. An assemblage of these techniques is usually used to turn raw big data into information. For example, in shale analytics, a combination of data mining, machine learning, artificial intelligence, correlation analysis and pattern recognition is used to extract information from text reports, sensor data and geophysical surveys from thousands of existing well operations. This information is then used to predict the success of new well operations [9]. In this case,

the combination of analytics is uniquely designed to extract value from the types of data present in shale gas operations. In order to leverage big data in groundwater in SADC, a similar set of unique analytical operations is needed to extract information from the types of data expected. It is also important to note that the type of analytics required should address the problem being investigated.

The spectrum of big data analytical techniques is vast and an explanation of all these techniques is beyond the scope of this study. However, understanding the role various big data analytics play in deriving information from data are key to derive the knowledge required to improve decision-making. For example, Table 4 presents a summary of common big data analytical techniques and the typical methods they include. These techniques can be used for a myriad of tasks such as extracting information from text data (text analytics), video files (video analytics) and audio data (audio analytics) and even geospatial data [17]. Hence, data collected from citizen science initiatives, remote-sensing data, social media data and conventional hydrological data can be turned into useful information for advancing understanding in groundwater management.

Generally, the role of big data analytics is to understand historical events or observations (descriptive analytics), what will occur based on historical observation (predictive analytics) and what is the best solution under uncertainty (prescriptive analytics) [69]. Translating this to a groundwater context allows us to understand what the fundamental interrelation and operation of various hydrogeological processes are based on current data (descriptive analytics), using this knowledge to predict future groundwater scenarios (predictive analytics) and then understanding what the best actions are going forward (prescriptive analytics). This is where the paradigm shifts towards emphasis on data-driven solutions, allowing our analysis to be prescribed by trends in the data rather than theory.

Table 4. Summary of big data analytical techniques [15,17,70,71].

Techniques	Description	Examples of Computational Methods
Statistics	Collection, organization and interpretation of data	Descriptive statistics, regression, correlation, factor analysis, clustering, hypothesis testing, probabilistic statistics
Data-mining	The process of extracting new information, such as patterns, from large datasets	SQL queries, machine-learning, statistics, feature selection
Artificial intelligence (AI)	The role of developing computer systems that imitate, amplify and automate intelligent behavior of human beings	Statistical learning, optimization methods, deep learning
Machine learning	Subset of AI, concerned with using self-learning computer algorithms to recognize features in empirical data	Artificial neural networks, support vector machine, random forest, k-means clustering, natural language processing
Uncertainty analysis	Techniques used to quantify and handle uncertainty in big data	Data cleaning, probability theory, Bayesian theory, Shannon's entropy, rough set theory, fuzzy set theory
Visualization	The use of graphic means to represent large datasets	Tables, graphs, images, feature extraction, geometric modeling

4.2. Statistical Methods

Statistical methods in this case relate to conventional data analysis techniques that have been at the forefront of traditional empirical analysis [72]. These methods are rooted in statistical and mathematical sciences [69]. They are designed to perform functions of association among data points, the segmentation and clustering of data, the categorization of data, anomaly detection, regression and prediction analysis within structured datasets [72]. For example, a multivariate regression analysis can be used to quantify the causal relationship between a series of variables, which can then be used to predict the outcome of a set of dependent variables. These techniques are still widely employed today in extracting information on groundwater data as well as modeling of groundwater processes

(e.g., geostatistics). For example, statistical techniques are used to design groundwater monitoring systems, assess groundwater quality and simulate groundwater flow. However, they suffer from certain drawbacks. Statistical methods are not fully optimized to handle large streams of heterogeneous, highly dimensional and noisy data [17]. Standard statistical techniques are more suited to operate on samples of population statistics, which are then used to infer across the entire population based on the statistical significance of the results [17]. Contrary, big data analytics operates on the majority, if not all, of the data in the population. Hence, the idea of statistical significance is no longer relevant. Furthermore, these standard statistical techniques are difficult to implement in parallel-processing environments, which is often necessary when dealing with big data [15]. However, incorporating these methods into big data analytics applications may still prove useful in handling of traditional structured data on groundwater.

4.3. Data Mining

Data mining is a term used to describe the use of big data analytical techniques to extract new information, such as patterns in data, relationships among variables, groupings of closely related data points or prediction of outcomes, from very large datasets [15,69,70]. Data mining involves the use of many statistical and machine-learning methods. Data mining is not restricted to very large datasets, and has been in use since before the advent of big data [72]. Only now, some of the traditional analytical methods have been extended to cope with processing big data. For example, traditional clustering algorithms such as K-means have been extended by partitioning large datasets into samples that can be processed across multiple machines. Results of the samples are combined to represent the overall dataset [15]. Typical algorithms for this approach include clustering large applications (CLARA) algorithm and clustering large applications based upon randomized search (CLARANS) [15]. Data mining is also not restricted to structured data and can be applied in text, image, video and audio analytics, etc. [17,52,73]. Data mining is the cumulative task of transforming the data into useful information and is thus an important step in any big data analytics application.

4.4. Artificial Intelligence and Machine Learning

One of the most common big data analytical techniques employed in the literature is machine learning, which is a branch of artificial intelligence. Machine learning consists of self-learning algorithms which form the backbone of most artificial intelligence programs [69]. Machine-learning methods provide a robust avenue to analyze large, highly dimensional, highly complex nonlinear systems [74]. For example, the complex interactions between various components of hydrological systems in nature are often nonlinear. We model these systems using conventional statistical analysis results in simplified and inaccurate outputs. In this instance, machine learning methods are better suited.

Machine learning can generally be classified into three broad learning categories: supervised, unsupervised and reinforcement machine-learning methods [75]. Supervised machine learning requires so-called labeled data that can be used as validation during the training process [76]. Labeled data are data points that have been tagged with known properties for that class of data, for algorithms to learn from. The model calculates expected outputs through a reiterative back-propagation training/learning process. The algorithm/or rules are defined that best predict labeled output based on input data. In unsupervised machine learning, the model operates on unlabeled data, hence there are no outputs for which to train the algorithm. In this case, the algorithm finds hidden patterns and groups in the data to perform clustering. Reinforcement learning algorithms are trained on labeled data that are intermediate between supervised and unsupervised. Instead of using labeled data that provide a correct answer for rules, the labeled data only provide an indication whether an action is correct or not. Indications of correct rules or actions or incorrect rules or actions are received through reward or punishment signals, respectively [75]. Through this process an algorithm is trained.

Generally, machine learning is used to perform four basic tasks, which include regression, classification, clustering and association [70]. Supervised machine-learning algorithms primarily

perform either regression (e.g., linear regression) or classification (e.g., k-nearest neighbor) [76]. Unsupervised machine learning primarily performs the tasks of clustering (e.g., k-means clustering) or association (e.g., A priori algorithm) [77]. Reinforcement learning is best suited for determining the best possible actions within an environment based on maximizing the reward. Machine-learning approaches are all data-driven requiring a large number of data points to achieve realistic accuracy. The benefit is that these models rely on real world data, without making any a priori assumptions about the system. When coupled with logical understanding of processes, these techniques allow new unforeseen relationships to be uncovered.

Traditionally, physics-based numeric models and conventional statistics have been the pervasive tools to simulate groundwater processes. These models require advanced a priori knowledge of the aquifer as well as complex data on a multitude of aquifer parameters to develop realistic simulations of groundwater processes [78]. This makes numeric models particularly complex to develop. A comparative study among machine learning techniques and numeric models for modeling groundwater dynamics in the Heihe River Basin, North Western China, showed generally favorable results for machine learning compared to numeric models [79]. Hence, machine learning-based groundwater models provide an alternative tool to physics-based process models.

Although the use of machine learning in groundwater is fairly nascent, there are a few case applications which allow us to illustrate its use. For example, groundwater level modeling and forecasting have been accomplished using various machine-learning methods [80–85]. Groundwater level forecasting is a particularly useful application area for machine learning, and it provides predictions of future groundwater levels to aid groundwater management. Similarly, groundwater quality mapping has been demonstrated by [86], using multivariate cluster analysis. [87] demonstrated the use of a boosted regression tree framework to model and predict nitrate concentrations in Central Valley aquifer, California, USA. [88] explored the use of various machine learning algorithms to predict groundwater recharge. Machine-learning algorithms have even been used to map surface water bodies from Landsat images [89], as an example of image analytics in hydrological sciences. The benefit of machine learning and artificial intelligence is in its ability to describe and predict real world scenarios, as well as to prescribe the best actions for a desired outcome. This feature may be key in developing data-driven solutions to support groundwater management.

4.5. Uncertainty Analysis

One of the most important concepts when collecting and analyzing big data are dealing with the uncertainty [71]. In big data analytics, this uncertainty is generally a result of large, highly heterogeneous, multidimensional datasets. These features of big data introduce many unstructured, inconsistent, incomplete and noisy data to the big data analytics process. The collection of data from heterogeneous sources in a variety of formats creates complexity in assuring the quality of data. For example, data from social media posts are not generated through rigorous scientific processes, and should thus be subjected to enhanced data quality measures [90,91]. Failure to address data quality and uncertainty early in the analysis process can create compounding effects across the big data value chain and can ultimately reduce the accuracy of outputs [71].

Additionally, the lack of training and understanding on the perceived nuances within various big data analytical algorithms may lead to erroneous applications [92]. This emphasizes the need to select proper techniques when dealing with big data, which is a statistical skill. Traditionally, mechanisms to deal with uncertainty involve tasks such as outlier detection, removal of duplicates, missing data detection and handling and unifying datasets [92]. However, even with data preparation taking place, there will still be inherent errors in big data that are difficult to detect. [71,93] discussed several strategies in mitigating errors during statistical learning in big data analytics. This includes incorporating techniques such as probability theory, Bayesian theory, Shannon's entropy, rough set theory, fuzzy set theory into the big data analytics process.

From a groundwater data perspective, the inclusion of remote sensing and other sensor-based measurements should be done with caution. Previous studies have shown these sources to contain uncertainty associated with various remote-sensing errors [94,95]. Additionally, it is also common for local scale variables such as groundwater levels to be uncertain. This is largely a result of missing data, poor data capturing and measurement errors. Efforts must be put in place to ensure reduction in the uncertainty of collected data, and that methods applied are relevant to the type of data being explored.

4.6. Visualization Tools

Visualization tools are techniques used to intuitively investigate big data using graphic means [15]. This typically involves the use of graphs, tables, images, diagrams and other ways to display data. These data visualization tools allow for an intuitive view of data, allowing patterns to be discerned based on expert judgment, instead of sophisticated quantitative analysis. For example, this is applicable when dealing with geospatial data, whose properties are reliant on neighboring data points [96]. However, one of the issues with representing big data in graphic way is that they are too large and contain too many dimensions to represent fully in graphs and tables. Data scientists must condense data, through feature extraction or geometric modeling to properly display them [15].

5. Big Data Analytics Platforms and Frameworks for Geo-Spatial Data

In a sea of big data tools, techniques and methods, groundwater scientists looking to leverage big data to support groundwater management can become overwhelmed. Big data platforms are enterprise scale solutions used to facilitate the use of big data to meet a specific industry need. They are generally a collection of hardware and software layers, built upon a specific big data processing framework. The function of modern big data platforms is to leverage big data. This is achieved through a process of data acquisition, data storage and preprocessing, data transformation through analytics and information dissemination [97]. Figure 2 illustrates a general reference framework for big data, which includes the typical features or components required for any big data platform.

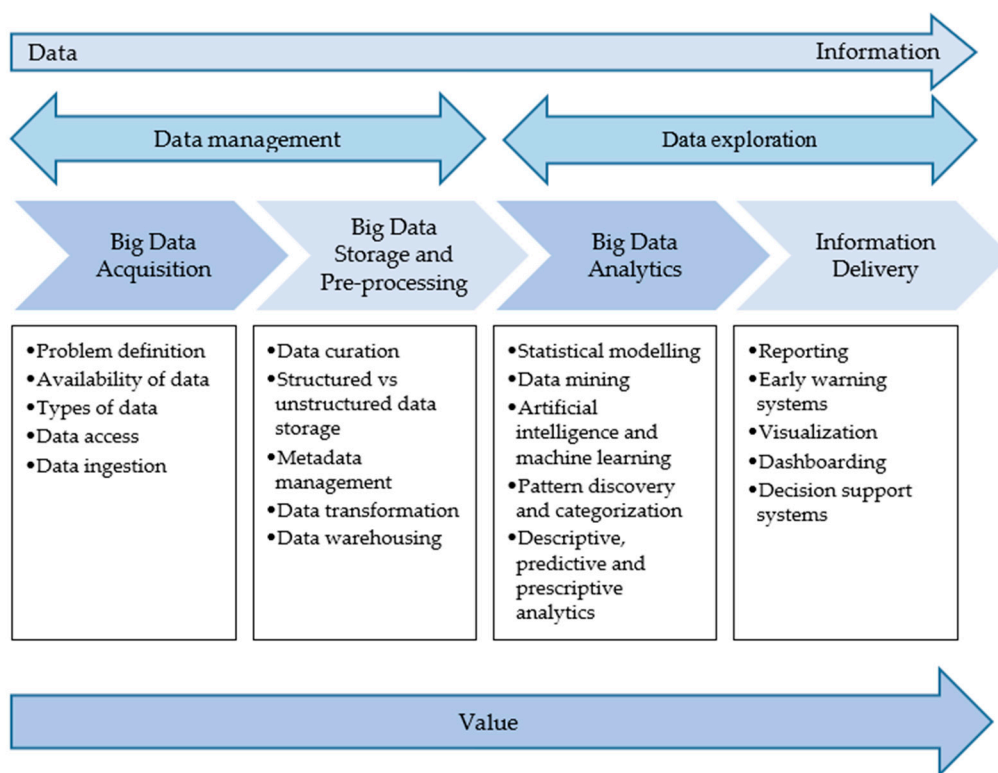


Figure 2. Big data analytics value chain (adapted from [98,99]).

Data acquisition revolves around connecting to relevant data sources, determine individual data products and ingestion mechanisms. Here, one must consider the type of data being collected (e.g., structured versus unstructured), access and usage protocols for the various sources, the volumes of data required (which influences how data will be transmitted from the source to the processing location) and meta-data generation [100]. For example, the size of some data products makes it impractical to retrieve data from the sources repeatedly for analytical queries. In this case, it may be more advantages to ingest entire datasets and store on local systems. The complexities associated with data collection make the data itself an important component of any big data platform.

Data pre-processing focuses on addressing the quality and uncertainty in the data, as well as the conversion of unstructured data to structured data. The purpose of this component is to create analysis-ready datasets. In this step, one must consider the type of data required for analytical operations, data cleaning protocols that are necessary, the uncertainty of the data and the post-processing algorithms that can be applied to improve accuracy in the raw data. The caveats (i.e., limitations and inaccuracies) of individual datasets will be important in this step [101]. Once the data have been preprocessed, then data storage can take place. This requires knowledge on how data are to be curated, the type of data being stored (i.e., structured or unstructured), the processing environment required, meta-data and the indexing paradigm. For example, in the Earth Science domain data will most certainly be geospatial in nature, indexing the data along temporal and spatial dimension would support faster and more versatile analytical operations [102].

Figure 2 also illustrates how the value of big data increases across the value chain. Big data analytics plays an important role in the value chain, leveraging big data in driving the knowledge discovery process, as we move from raw data to useful information. In this component, many of the analytical methods described in Section 4, will be useful. However, developing data-driven modeling through machine learning and artificial intelligence is perhaps the current status quo in terms of extracting value from the data. Descriptive, predictive and prescriptive analytical models, if feasible, can provide additional tools to support groundwater management. For example, descriptive and predictive models may allow simulation of current and future conditions groundwater conditions, while prescriptive models may allow determination of the impact of various management decisions. Finally, usable information must be disseminated in the form of maps, figures and tables (etc.). This information can be usable as it is or it can be incorporated into decision support systems, early warning systems or dashboards to facilitate decisions (Figure 2).

Addressing some of the challenges facing groundwater management in SADC may require a holistic solution such as a big data platform. For example, the disparate nature of groundwater big data could be centralized, the application of analytics could be simplified with built-in methods and functions, and the information could easily be accessed through web-based services. Hence, big data frameworks and platforms that can be used to implement a big data approach in support of sustainable groundwater management in SADC are reviewed below.

Many of the data sources described in Section 3 house their data in large data warehouses or centers, which are distributed across the globe. These data centers can be accessed through various web-based platforms, such as Earth Explorer (<https://earthexplorer.usgs.gov/>), EarthData (<https://earthdata.nasa.gov/>), ESA Earth Online (<https://earth.esa.int/web/guest/data-access>). For example, most data generated by NASA missions get stored in distributed active archive centers across the United States, which can be accessed through various web-based platforms and software [103]. However, navigating, extracting and processing vast amounts of remote-sensing data from various data sources to apply to a specific objective, such as to support groundwater management in SADC region, can be technically challenging [33]. Often, specialist skills and tools are required to properly integrate and use the vast volumes of groundwater big data available.

In order to address some of these challenges, many agencies have developed special platforms that can be used to leverage these big data. The Australian Geoscience Data Cube (AGDC) is an example of a purpose-built big data platform that focuses on leveraging remote-sensing big data, particularly

Landsat, for Australian geoscience applications [104]. Hence, the platforms, data collection, storage and analysis features are tailored toward managing geo-spatial remote-sensing data. For example, data ingestion and preprocessing components focus largely on refining incoming raw data into analysis-ready products before data storage, using standard techniques. Data storage follows a multidimensional data array format with geospatial indexing (Data Cube). The architecture for this system is supported by the National Computational Infrastructure (NCI) Facility and their high-performance computing framework.

EarthServer is a geospatial big data platform that is more generalized and interoperable, by focusing development on open geospatial data standards, such as those provided by the Open Geospatial Consortium (OGC) [105]. The platform is supported by the Rasdaman framework, which is an array-based, fully implemented parallel storage and processing platform. The platform allows various front-end applications to be attached for specific use cases.

IBM's physical analytics integrated data repository and services (PAIRS) is another geospatial big data platform [106,107]. Its focus is largely on facilitating and simplifying the collection, integration, preprocessing, storage, retrieval and analysis of heterogenous spatial data. Data are collected and preprocessed into analysis-ready products, indexed and stored along a common geo-spatial grid. Frameworks such as Hadoop and HBase support the storage and processing. Unlike the other platforms that focus on raster data, PAIRS provides facility for unstructured data types such as from IoT and social media. The unstructured data are transformed and stored alongside the raster data.

The Earth System Grid Federation (ESGF) is a multi-agency, international collaboration focusing on the sharing of climate related data [108]. The design of the ESGF is based on geographical independent data nodes that are built on common infrastructure. The nodes adopt common federation protocols and API's (Application Programming Interfaces) that facilitate peer-peer communication and transfer of data. At the moment the ESGF is not an analytics platform, instead focusing on data indexing and data access.

Besides the aforementioned big data platforms, there are a number of big data geospatial frameworks that can be implemented as geospatial big data processing solutions. These include ST-Hadoop [102], SpatialHadoop [109], Hadoop-GIS [110], GeoWave [111] and GeoSpark [112], among others. These frameworks facilitate the distributed or parallel processing of geospatial big data.

6. Conceptual Framework

Although many of the platforms and frameworks described here are suited to management of the geospatial data, the unique groundwater management challenges faced by SADC warrant additional features. For example, a SADC framework must include some features that allow local scale data gaps to be filled and allow nonconnected disparate data sources to be easily ingested. Based on the findings of the review work in this paper, a conceptual big data analytics framework is proposed (Figure 3). The framework illustrates the required features of a big data analytics framework that can support groundwater management in SADC. This includes the typical features such as data collection, processing, storage, analytics and information delivery. However, key features that are unique in the context of this paper include the groundwater management scenarios and downscaling. By including these two features, we position the framework to address the specific groundwater management challenges in SADC. These unique features are discussed in more detail in the following sections. The features of the framework are grouped together under the big data analytical infrastructure, which represents the hardware and software stack that needs to be developed to implement the framework. This framework also focuses on the groundwater big data sources, which in itself is an important feature of any big data framework. The framework is not intended to be a schematic architecture for a big data platform, but rather a reference framework that can be used to develop a big data solution for SADC groundwater management.

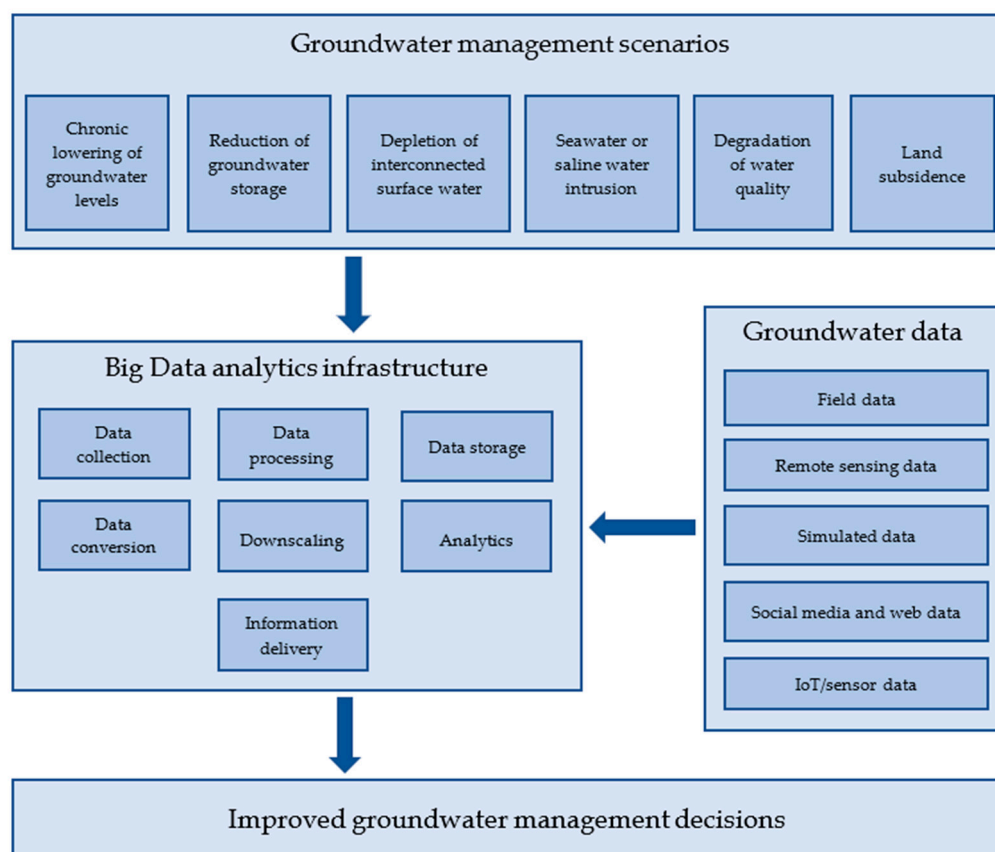


Figure 3. Conceptual framework for SADC groundwater management using big data analytics.

6.1. Groundwater Management Scenarios

According to [113], one critical task that is often overlooked during the application of big data analytics is the establishment of a sound problem definition. Without the problem definition being clearly defined, which ultimately informs the type of information required, the process is open to ambiguity. A good problem definition must be easily translatable into a quantifiable feature that can be statistically modeled [114]. Therefore, the framework begins by defining the problem. In our case, we define various groundwater management scenarios as the problems that need to be addressed. These scenarios are adopted from the California Department of Water Resources Best Management Practices for the Sustainable Management of Groundwater [115]. They represent the typical issues facing groundwater managers and can easily be assessed through quantifiable criteria, such as thresholds indicating undesirable conditions. These groundwater management scenarios ultimately dictate the type of data required, the scale of the data, the individual datasets required, the analytics needed and the information output. For example, during groundwater drought, it is important to monitor groundwater storage, in order to avoid issues of reduction in groundwater storage. In this scenario, it may be required to acquire groundwater level data, GRACE and other hydrogeological data. However, the scale issue associated with GRACE data limits its application to the local scale. This means that downscaling may be necessary before any valuable information can be generated. In this case valuable information may be a series of high-resolution groundwater storage maps over time, which may allow addressing the impacts on interconnected surface water. Other possible problems may be saline water intrusions in coastal aquifers or degradation of water quality in urban, industrial and agricultural areas. Land subsidence is another problem that must be considered especially in karst aquifers. Table 5 presents possible big data and analytical solutions to the groundwater management scenarios.

Table 5. Groundwater management scenarios and possible big data requirements and analytics solutions.

Groundwater Management Scenario	Possible Data Requirements	Possible Big Data Analytical Approaches
Chronic lowering of groundwater levels	<ul style="list-style-type: none"> • In situ groundwater levels • Groundwater abstraction volumes • GRACE TWS 	Predictive modeling of groundwater levels according to various abstraction and climate change scenarios [81,85]
Reduction of groundwater storage	<ul style="list-style-type: none"> • Aquifer properties • GRACE TWS • In situ groundwater levels 	Current storage conditions need to be established, with future groundwater storage modeled through data-driven solutions [116–118]
Depletion of interconnected surface water	<ul style="list-style-type: none"> • In situ groundwater levels • Landsat • Surface hydrology levels 	Correlation and prediction of groundwater levels with surface water levels [82]
Seawater or saline water intrusion		Associated with chronic lowering of groundwater levels
Degradation of water quality	<ul style="list-style-type: none"> • In situ groundwater quality • Geological properties • Land use/cover 	Correlating groundwater quality with land use and geology, following by predictive mapping [86,87]
Land subsidence	<ul style="list-style-type: none"> • GRACE TWS • Elevation (e.g., LiDAR) 	Changes in groundwater storage can be compared to changes in ground elevation

As a specific example, Figure 4 depicts a compartmentalized dolomitic aquifer underlying parts of Botswana and South Africa. In this particular aquifer groundwater over-abstraction has resulted in significant reduction in groundwater levels and has further reduced groundwater storage [119]. In order to address the groundwater management challenges in this aquifer using big data solutions, one could bring together a number of datasets, such as groundwater level observations (shown by the blue circles overlying the aquifer in Figure 4), GRACE data, precipitation data from remote-sensing sources, abstraction data and other complimentary datasets. Together these data can be used to develop data-driven models of spatial and temporal patterns in groundwater storage changes, as well as predict future changes under current conditions [81]. This information can then be used to better inform intervention strategies to reduce excessive degradation of the groundwater resources in the aquifer.

The framework developed in this study is intended to be a conceptual framework that can be used to support groundwater management in SADC region using big data analytics. Thus, it does not include technical details on the individual techniques and methods required for each component to function. For example, how best to integrate and connect the various disparate sources of groundwater data in the SADC region, what methods or models are the best for transforming the data into useful information are questions that still need further research. However, the framework provides a step-wise guidance for the application of big data analytics to different aquifer problems in the SADC.

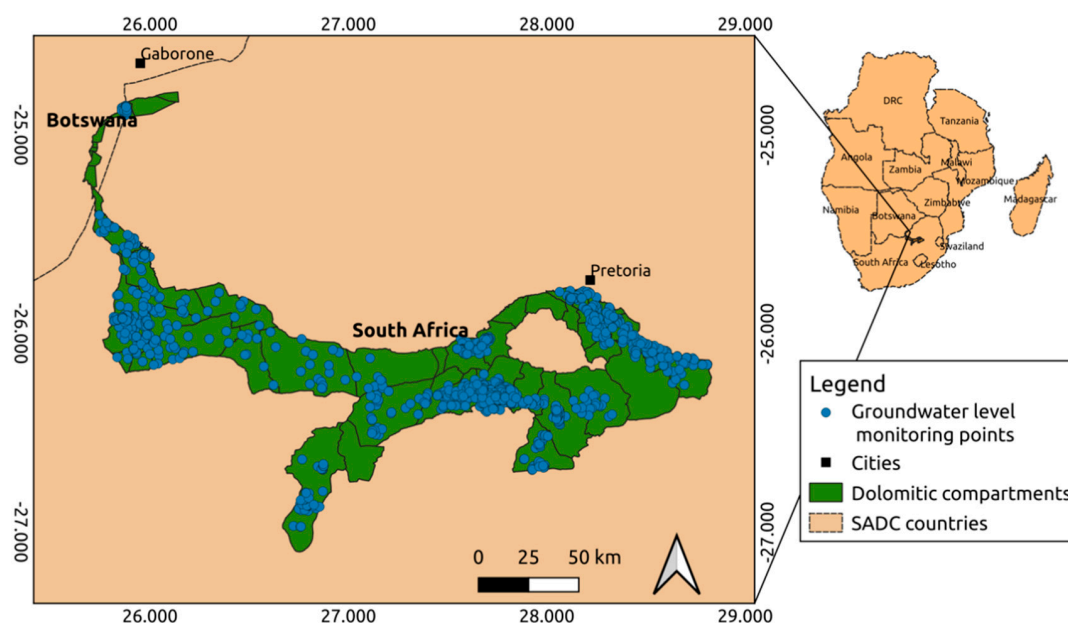


Figure 4. Dolomitic aquifers of North-Western South Africa and Southern Botswana.

6.2. Big Data Analytics Infrastructure

This component of the framework is the main analytics engine that drives the collection, preprocessing, storage, analysis and delivery of groundwater big data and information. For example, consider the data collected to address the challenges in the dolomitic aquifers of Figure 4. Here, the groundwater level observations, GRACE data, precipitation data, abstraction data and other complimentary datasets, are not expected to be in an analysis ready form. Data cleaning, reduction of uncertainty and uniform indexing of the data will need to be conducted before the data can be stored for later use. Once the data are in a quality assured form and stored, it can be transformed through various analytics, such as downscaling or machine learning models into useful information. This information can then be disseminated to relevant stakeholders through information portals, dashboards, reports, maps and other figures.

In a big data context this type of work would be carried out within a big data analytics platform, such as those mentioned in Section 5 or within a purpose built platform designed for leveraging of groundwater big data in SADC.

6.3. Downscaling Methods

Downscaling methods are of particular interest within the context of the proposed framework (Figure 3). The use of remote sensing, atmospheric models, and land surface model provide a useful avenue to explore new insights into the characteristics and processes occurring in aquifers. However, these big data sources, in many cases, offer only regional scale aspects due to the coarse resolution. In order to improve localized groundwater management (e.g., individual boreholes, wellfields), fine resolution information is essential. Big data analytics techniques can address the mismatch between the regional scale data and the local scale information through the process of downscaling. Downscaling is the process of refining the resolution of coarse scale data to a finer resolution for local scale groundwater management.

Generally, there are two approaches to perform downscaling: dynamical downscaling and statistical downscaling [120]. Dynamical approaches rely on numerical/physics-based models to simulate regional or local variables from global scale models [121]. Statistical approaches model the empirical relationships between large-scale variables (predictors) and small-scale covariates (predictants) [122]. Each of these approaches has their own merits and constraints (Table 6). For example,

dynamical downscaling approaches require heavy computational resources and complex data, but can produce physically consistent downscaling results [123–125]. Statistical downscaling approaches require low computational resources, are easy to implement and require generally less complex data (i.e., fewer variables) from multiple sources [125]. However, statistical approaches are less suitable to model nonlinear relationships between predictors and predictants. Nonetheless, the low computational resources and low data requirements meant that researchers have favored statistical approaches over dynamical approaches [126,127].

Machine-learning methods in particular have been applied to downscale remote-sensing data. Of the many machine learning methods available, artificial neural networks, support vector machine, least square support vector machine, relevance vector machine, generalized linear model, random forest, genetic programming, multi-point geostatistical approach, correlative relation methods and boosted regression tree have been applied to downscale various remote sensing and GCM data [74,81,121,123,125–133].

Table 6. Merits and constraints of downscaling approaches.

Characteristics	Dynamical Downscaling	Statistical Downscaling
Execution difficulty	Difficult to execute, requiring heavy computational resources [131]	Easy to execute, requiring less computational resources [128]
Data requirements	Requires complex data from multiple sources [125]	Data requirements are generally lower than for dynamical approaches [125]
Downscaling consistency	Physically consistent downscaling of climate variables [124]	Can downscale to finer resolutions, however nonlinear relationships are hard to model [133]
Hydrogeological model inputs	Requires extensive a priori knowledge of hydrological processes	Requires limited a priori knowledge of hydrological processes [81]
Uncertainties	Uncertainties introduced through model approximation, assumptions and parameterizations [123]	Uncertainties introduced through non-stationarity and high spatial variability between predictors and predictants [123]

7. Challenges in Applying Big Data to Groundwater Management

According to [134], there are numerous challenges that are faced by experts when trying to implement big data analytics, but these can be divided into three broad categories: (1) Data challenges relate to the nature of big data itself (e.g., volume, velocity and variety, etc.); (2) Process challenges relate to how to capture, integrate and transform data, how to select the right model for analysis and how to provide the results; (3) and management challenges cover issues such as privacy, governance, institutionalization, security, among others. These challenges are further exacerbated by the technological limitations of current information systems [23]. In this section (Section 7), we discuss some of these challenges in the context of groundwater big data in the SADC region in Africa as well as how it would affect the implementation of the framework.

Like all other domains, big data in groundwater within SADC region are expected to have considerable volume, velocity and variety. For example, the data for a $10^\circ \times 10^\circ$ tile from MODIS Evapotranspiration dataset for the SADC region can be as large as 20 GB. Multiplying by additional variables and additional tiles needed to model a groundwater management scenario across the entire SADC region would result in the dataset growing rapidly. The technological requirements to store and process such large heterogeneous volumes of data often require dedicated systems beyond the capabilities of conventional desktop systems [23]. In this instance, technologies such as parallel processing infrastructure and clustered computing systems have come to the fore [23]. However,

the computational capabilities of many SADC member states may not be advanced enough to facilitate big data approaches. Furthermore, an obvious bottleneck when ingesting huge volumes of data are the high network speed required to move and process big data [135]. This requirement is often lacking in less developed African regions and may even be non-existent in rural regions.

Lastly, big data management challenges are experienced within a SADC context, especially when dealing with transboundary aquifers. The transparency of data sharing across international boundaries is not always welcomed by individual states. Data ownership and data access is often restricted to certain individual or institutions and come with many caveats for their use [12]. This is certainly the case when security issues are present with sharing or use of data. The institutional barriers may become a roadblock. Furthermore, management practices employed by member states are not always aligned with each other [12]. The consequence is that the decisions taken based on the data may be contradicting within transboundary aquifers, ultimately affecting the sustainable management of groundwater.

8. Conclusions

Groundwater science is generating increasing amounts of data from scientific experiments, sensor arrays, monitoring programs, remote sensing—even social media. Increasing attention is being paid to leveraging these vast volumes of data for new knowledge discovery in groundwater. Improving sustainable groundwater management in SADC is one use case where big data and big data analytics may be useful. Big data analytic's contribution to groundwater management can be two-fold. Firstly, big data analytics can address issues of data scarcity by consolidating data available from different sources, both traditional and unconventional. Secondly, big data analytics can transform data into usable information that can support groundwater management, especially at a local scale. The general consensus in the literature is that big data analytics techniques and methods provide benefits beyond traditional analytics, when dealing with large heterogeneous datasets and are particularly useful when performing data-driven modeling. Advanced analytics such as machine learning have shown a promising insight when modeling groundwater processes. However, the choice of data and the choice of analytical techniques to achieve the analysis goal is critical to ensure data integrity and accuracy along the life cycle of the data. Proper management of data and analytical processes is imperative in this case. A conceptual framework was presented that can be used to facilitate the application of big data analytics in the context SADC groundwater management. This framework considers the required elements in the value chain based on the literature and local experiences in the SADC. Specific big data techniques and methods (e.g., data acquisitions, storage, data mining, machine-learning algorithms) can be used to execute the framework and transform data into usable information. However, it is also clear that some challenges will hinder the progression of big data analytics in the SADC. These challenges include a lack of computing infrastructure (e.g., data storage, network speed) and institutional barriers. Nonetheless, it is clear from this research that there are sufficient data and big data analytics techniques developed well enough to explore its operational use in the SADC region. Future work should focus on highlighting solutions to these challenges and experimenting with specific use cases (such as various aquifer settings) with big data analytics in order to continue developing data-driven sciences in groundwater.

Author Contributions: Conceptualization, Z.G., K.P., N.J. and A.B.; methodology, Z.G., N.J.; formal analysis, Z.G.; investigation, Z.G.; writing—original draft preparation, Z.G., N.J.; writing—review and editing, N.J., K.P., A.B., T.K.; visualization, Z.G.; supervision, K.P.; project administration, K.P., T.K.; funding acquisition, K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, a multi-agency coalition with the Department of Science and Innovation South Africa, the USAID, the Southern African Development Community-Groundwater Management Institute (SADC-GMI), the South African Water Research Commission (WRC), the Department of Science and Innovation South Africa, the Sustainable Water Partnership and the United States Geological Survey (USGS).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

References

1. Adamala, S. An Overview of Big Data Applications in Water Resources Engineering. *Mach. Learn. Res.* **2017**, *2*, 10–18. [[CrossRef](#)]
2. Kitchin, R.; McArdle, G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc.* **2016**, *3*. [[CrossRef](#)]
3. Roy, A.K. Advances and Scope in Big Data Analytics in Healthcare. *Curr. Trends Biomed. Eng. Biosci.* **2017**, *9*, 1–7. [[CrossRef](#)]
4. Zhang, Y.; Zhao, Y. Astronomy in the Big Data Era. *Data Sci. J.* **2015**, *14*, 1–9. [[CrossRef](#)]
5. Guo, H. Big Earth data: A new frontier in Earth and information sciences. *Big Earth Data* **2017**, *1*, 4–20. [[CrossRef](#)]
6. Chalh, R.; Bakkoury, Z.; Ouazar, D.; Hasnaoui, M.D. Big Data Open Platform for Water Resources Management. In Proceedings of the 2015 International Conference on Cloud Technologies and Applications (CloudTech), Marrakesh, Morocco, 2–4 June 2015.
7. Lee, S.; Hyun, Y.; Lee, M.J. Groundwater potential mapping using data mining models of big data analysis in Goyang-si, South Korea. *Sustainability* **2019**, *11*, 1678. [[CrossRef](#)]
8. Water Resources Research. Big Data and Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science. Available online: [https://agupubs.onlinelibrary.wiley.com/doi/toc/10.1002/\(ISSN\)1944-7973.MACHINELEARN](https://agupubs.onlinelibrary.wiley.com/doi/toc/10.1002/(ISSN)1944-7973.MACHINELEARN) (accessed on 22 May 2020).
9. Mohaghegh, S.D.; Gaskari, R.; Maysami, M. Shale Analytics: Making Production and Operational Decisions Based on Facts: A Case Study in Marcellus Shale. In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, The Woodlands, TX, USA, 24–26 January 2017.
10. Fleming, K. The Internet of Things: Creating Water Stability with Streaming Data. Available online: <https://www.ibmbigdatahub.com/blog/internet-things-creating-water-stability-streaming-data> (accessed on 16 May 2020).
11. Pietersen, K.; Kellgren, N.; Roos, M.; Chevallier, L. *Explanatory Brochure for the South African Development Community (SADC) Hydrogeological Map & Atlas*; Southern African Development Community: Gaborone, Botswana, 2010.
12. Pietersen, K.; Beekman, H. *Groundwater Management in the Southern African Development Community*; Southern African Development Community Groundwater Management Institute: Bloemfontein, South Africa, 2016.
13. Farr, J.L.; Gumirehete, R.; Davies, J.; Robins, N.S. *Southern African Development Community Regional Situation Analysis*; British Geological Survey: Nottingham, UK, 2005.
14. Bates, B.C.; Kundzewicz, Z.W.; Wu, S.; Palutikof, J.P. *Climate Change and Water: IPCC Technical Paper VI*; Intergovernmental Panel on Climate Change: Geneva, Switzerland, 2008; ISBN 9789291691234.
15. Chen, P.C.L.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [[CrossRef](#)]
16. Ylijoki, O.; Porras, J. Perspectives to Definition of Big Data: A Mapping Study and Discussion. *J. Innov. Manag.* **2016**, *4*, 69–91. [[CrossRef](#)]
17. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [[CrossRef](#)]
18. Laney, D. *3D Data Management: Controlling Data Volume, Velocity, and Variety*; Stamford: Lincolnshire, UK, 2011.
19. Zikopoulos, P.C.; DeRoos, D.; Parasuraman, K.; Deutsch, T.; Corrigan, D.; Giles, J. *Harness the Power of Big Data*; McGraw-Hill: New York, NY, USA, 2013; ISBN 9780071808187.
20. Lee, I. Big data: Dimensions, evolution, impacts, and challenges. *Bus. Horiz.* **2017**, *60*, 293–303. [[CrossRef](#)]
21. Wang, S.; Li, G.; Yao, X.; Zeng, Y.; Pang, L.; Zhang, L. A distributed storage and access approach for massive remote sensing data in MongoDB. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 533. [[CrossRef](#)]
22. Lin, Y.; Jun, Z.; Hongyan, M.; Zhongwei, Z.; Zhanfang, F. A method of extracting the semi-structured data implication rules. *Procedia Comput. Sci.* **2018**, *131*, 706–716. [[CrossRef](#)]

23. Fan, J.; Han, F.; Liu, H. Challenges of Big Data analysis. *Natl. Sci. Rev.* **2014**, *1*, 293–314. [[CrossRef](#)] [[PubMed](#)]
24. Zhang, Z.; Moore, J.C. Data Assimilation. In *Mathematical and Physical Fundamentals of Climate Change*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 291–311.
25. Loaiciga, H.A.; Charbeneau, R.J.; Everett, L.G.; Fogg, G.E.; Hobbs, B.F.; Rouhani, S. Review of ground-water quality monitoring network design. *J. Hydraul. Eng.* **1992**, *11*, 11–37. [[CrossRef](#)]
26. IGRAC; IGS. *State of Groundwater Data Collection and Data Management in SADC Member States*; Southern African Development Community—Groundwater Management Institute: Bloemfontein, South Africa, 2019.
27. Sankar, K.P.; Ambati, V.; Pratha, L.; Jawahar, C.V. Digitizing a million books: Challenges for document analysis. In *Document Analysis Systems VII*; Bunke, H., Spitz, A.L., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2006; Volume 3872, pp. 425–436.
28. Ignjatić, J.; Nikolić, B.; Rikalović, A.; Čulibrk, D. Deep learning for historical cadastral maps digitization: Overview, challenges and potential. *Comput. Sci. Res. Notes* **2018**, *2803*, 42–47. [[CrossRef](#)]
29. Kaplan, F.; di Lenardo, I. Big Data of the Past. *Front. Digit. Humanit.* **2017**, *4*, 1–12. [[CrossRef](#)]
30. Tatem, A.J.; Goetz, S.J.; Hay, S.I. Fifty Years of Earth Observation Satellites. *Am. Sci.* **2008**, *96*, 390–398. [[CrossRef](#)]
31. Jensen, J.R. *Introductory Digital Image Processing: A Remote Sensing Perspective*, 2nd ed.; Clarke, K.C., Ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1996.
32. Chen, L.; Wang, L. Recent advance in earth observation big data for hydrology. *Big Earth Data* **2018**, *2*, 86–107. [[CrossRef](#)]
33. Cui, Y.; Chen, X.; Gao, J.; Yan, B.; Tang, G.; Hong, Y. Global water cycle and remote sensing big data: Overview, challenge, and opportunities. *Big Earth Data* **2018**, *2*, 282–297. [[CrossRef](#)]
34. Sheffield, J.; Ferguson, C.R.; Troy, T.J.; Wood, E.F.; McCabe, M.F. Closing the terrestrial water budget from satellite remote sensing. *Geophys. Res. Lett.* **2009**, *36*. [[CrossRef](#)]
35. Lv, M.; Ma, Z.; Yuan, X.; Lv, M.; Li, M.; Zheng, Z. Water budget closure based on GRACE measurements and reconstructed evapotranspiration using GLDAS and water use data for two large densely-populated mid-latitude basins. *J. Hydrol.* **2017**, *547*, 585–599. [[CrossRef](#)]
36. Tang, Q.; Gao, H.; Lu, H.; Lettenmaier, D.P. Remote sensing: Hydrology. *Prog. Phys. Geogr.* **2009**, *33*, 490–509. [[CrossRef](#)]
37. Syed, T.H.; Famiglietti, J.S.; Rodell, M.; Chen, J.; Wilson, C.R. Analysis of terrestrial water storage changes from GRACE and GLDAS. *Water Resour. Res.* **2008**, *44*. [[CrossRef](#)]
38. Rodell, M.; Chen, J.; Kato, H. Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeol. J.* **2007**, *15*, 159–166. [[CrossRef](#)]
39. Rodell, M.; Velicogna, I.; Famiglietti, J.S. Satellite-based estimates of groundwater depletion in India. *Nature* **2009**, *460*, 999–1002. [[CrossRef](#)] [[PubMed](#)]
40. Long, D.; Chen, X.; Scanlon, B.R.; Wada, Y.; Hong, Y.; Singh, V.P.; Chen, Y.; Wang, C.; Han, Z.; Yang, W. Have GRACE satellites overestimated groundwater depletion in the Northwest India Aquifer? *Nat. Publ. Group* **2016**, 1–11. [[CrossRef](#)] [[PubMed](#)]
41. Seyoum, W.M.; Milewski, A.M. Improved methods for estimating local terrestrial water dynamics from GRACE in the Northern High Plains. *Adv. Water Resour.* **2017**, *110*, 279–290. [[CrossRef](#)]
42. Bhanja, S.N.; Mukherjee, A.; Saha, D.; Velicogna, I.; Famiglietti, J.S. Validation of GRACE based groundwater storage anomaly using in-situ groundwater level measurements in India. *J. Hydrol.* **2016**, *543*, 729–738. [[CrossRef](#)]
43. Soupios, P.; Kokinou, E. Environmental Geophysics: Techniques, Advantages and Limitations. In *Geophysics: Principles, Applications and Emerging Technologies*; Aiello, G., Ed.; Nova Science Publishers, Inc.: New York, NY, USA, 2016; ISBN 9781634848312.
44. Day-lewis, F.D.; Slater, L.D.; Robinson, J.; Johnson, C.D.; Terry, N.; Werkema, D. An overview of geophysical technologies appropriate for characterization and monitoring at fractured-rock sites. *J. Environ. Manag.* **2017**, *204*, 709–720. [[CrossRef](#)]
45. Ebrahim, G.Y.; Magombeyi, M.; Villholth, K.G.; Lautze, J.; Nijsten, G.-J.; Keodumetse, K.; Kenabatho, P.; Sivashni, N.; Makoba, P.; Mndaweni, S.; et al. *Hydrogeological Modelling for Ramotswa Transboundary Aquifer Area*; IWMI: Colombo, Sri Lanka, 2018.

46. Collins, S.N.; James, R.S.; Ray, P.; Chen, K.; Lassman, A.; Brownlee, J. Grids in Numerical Weather and Climate Models. In *Climate Change and Regional/Local Responses*; Ray, P., Zhang, Y., Eds.; IntechOpen: London, UK, 2013.
47. Li, J.; Liu, K.; Huang, Q. Utilizing Cloud Computing to Support Scalable Atmospheric Modeling: A Case study of Cloud-Enabled Model E. In *Cloud Computing in Ocean and Atmospheric Sciences*; Vance, T.C., Merati, N., Yang, C., Yuan, M., Eds.; Academic Press: Cambridge, MA, USA, 2016; pp. 347–364. ISBN 9780128031933.
48. Sato, H.; Ito, A.; Ito, A.; Ise, T.; Kato, E. Current status and future of land surface models. *Soil Sci. Plant Nutr.* **2015**, *61*, 34–47. [[CrossRef](#)]
49. NASA Land Data Assimilation System. Available online: <https://ldas.gsfc.nasa.gov/> (accessed on 21 April 2020).
50. Parker, W.S. Reanalyses and observations: What's the Difference? *Bull. Am. Meteorol. Soc.* **2016**, *97*, 1565–1572. [[CrossRef](#)]
51. Russom, P. *Big Data Analytics; Transforming Data with Intelligence*: Renton, WA, USA, 2011.
52. Subudhi, B.N.; Rout, D.K.; Ghosh, A. Big data analytics for video surveillance. *Multimed. Tools Appl.* **2019**, *78*, 26129–26162. [[CrossRef](#)]
53. Lampos, V.; Cristianini, N. Nowcasting Events from the Social Web with Statistical Learning. *ACM Trans. Intell. Syst. Technol.* **2012**, *3*, 1–22. [[CrossRef](#)]
54. Eilander, D.; Trambauer, P.; Wagemaker, J.; Van Loenen, A. Harvesting Social Media for Generation of Near Real-time Flood Maps. *Procedia Eng.* **2016**, *154*, 176–183. [[CrossRef](#)]
55. Kryvasheyev, Y.; Chen, H.; Obradovich, N.; Moro, E.; Van Hentenryck, P.; Fowler, J.; Cebrian, M. Rapid assessment of disaster damage using social media activity. *Sci. Adv.* **2016**, *2*, 1–12. [[CrossRef](#)]
56. Chen, Y.; Han, D. On Big Data and Hydroinformatics. *Procedia Eng.* **2016**, *154*, 184–191. [[CrossRef](#)]
57. Macaulay, T. Connected Devices. In *RIoT Control: Understanding and Managing Risks and the Internet of Things*; Macaulay, T., Ed.; Elsevier: Amsterdam, The Netherlands, 2016.
58. McNicholas, C.; Mass, C.F. Impacts of assimilating smartphone pressure observations on forecast skill during two case studies in the Pacific Northwest. *Weather Forecast.* **2018**, *33*, 1375–1396. [[CrossRef](#)]
59. de Vos, L.W.; Droste, A.M.; Zander, M.J.; Overeem, A.; Leijnse, H.; Heusinkveld, B.G.; Steeneveld, G.J.; Uijlenhoet, R. Hydrometeorological monitoring using opportunistic sensing networks in the Amsterdam metropolitan area. *Bull. Am. Meteorol. Soc.* **2019**, 167–185. [[CrossRef](#)]
60. Cecchini, C.; Jimenez, M.; Mosser, S.; Riveill, M. An Architecture to Support the Collection of Big Data in the Internet of Things. In Proceedings of the International Workshop on Ubiquitous Mobile Cloud, Anchorage, AK, USA, 27 June–2 July 2014.
61. Malche, T.; Maheshwary, P. Internet of Things (IoT) Based Water Level Monitoring System for Smart Village. In Proceedings of the International Conference on Communication and Networks, Silicon Valley, CA, USA, 26–29 January 2017.
62. Wolfson, R. IBM Pilots Blockchain and IoT Sensor Solution to Track Sustainable Groundwater Usage in California. Available online: <https://www.forbes.com/sites/rachelwolfson/2019/02/08/ibm-pilots-blockchain-and-iot-sensor-solution-to-track-sustainable-groundwater-usage-in-california/#1b0901713edb> (accessed on 28 November 2019).
63. Little, K.E.; Hayashi, M.; Liang, S. Community-Based Groundwater Monitoring Network Using a Citizen-Science Approach. *Groundwater* **2016**, *54*, 317–324. [[CrossRef](#)]
64. Reeves, N.; Simperl, E. Efficient, but effective? Volunteer engagement in short-term virtual citizen science projects. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–35. [[CrossRef](#)]
65. Senočetnik, M.; Herga, Z.; Šubic, T.; Bradeško, L.; Kenda, K.; Klemen, K.; Pergar, P.; Mladenčić, D. IoT Middleware for Water Management. *Proceedings* **2018**, *2*, 696. [[CrossRef](#)]
66. Watson, H.J. Tutorial: Big data analytics: Concepts, technologies, and applications. *Commun. Assoc. Inf. Syst.* **2014**, *34*, 1247–1268. [[CrossRef](#)]
67. Tsai, C.W.; Lai, C.F.; Chao, H.C.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**, *2*, 1–32. [[CrossRef](#)]
68. Almeida, F. Big Data: Concept, Potentialities and Vulnerabilities. *Emerg. Sci. J.* **2018**, *2*, 1–10. [[CrossRef](#)]

69. Sun, Z.; Huo, Y. The Spectrum of Big Data Analytics. *J. Comput. Inf. Syst.* **2019**. [[CrossRef](#)]
70. Ali, A.; Qadir, J.; ur Rasool, R.; Sathiaseelan, A.; Zwitter, A.; Crowcroft, J. Big data for development: Applications and techniques. *Big Data Anal.* **2016**, *1*, 2. [[CrossRef](#)]
71. Hariri, R.H.; Fredericks, E.M.; Bowers, K.M. Uncertainty in big data analytics: Survey, opportunities, and challenges. *J. Big Data* **2019**, *6*, 44. [[CrossRef](#)]
72. Chen, H.; Chiang, R.H.L.; Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *Bus. Intell. Res.* **2012**, *36*, 1165–1188.
73. Lin, Y.T.; Der Yang, M.; Han, J.Y.; Su, Y.F.; Jang, J.H. Quantifying flood water levels using image-based volunteered geographic information. *Remote Sens.* **2020**, *12*, 706. [[CrossRef](#)]
74. Srivastava, P.K.; Han, D.; Ramirez, M.R.; Islam, T. Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application. *Water Resour. Manag.* **2013**, *27*, 3127–3144. [[CrossRef](#)]
75. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [[CrossRef](#)] [[PubMed](#)]
76. Schrider, D.R.; Kern, A.D. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet.* **2018**, *34*, 301–312. [[CrossRef](#)] [[PubMed](#)]
77. Alashwal, H.; El Halaby, M.; Crouse, J.J.; Abdalla, A.; Moustafa, A.A. The application of unsupervised clustering methods to Alzheimer’s disease. *Front. Comput. Neurosci.* **2019**, *13*, 1–9. [[CrossRef](#)]
78. Kenda, K.; Čerin, M.; Bogataj, M.; Senožetnik, M.; Klemen, K.; Pergar, P.; Laspidou, C.; Mladenčić, D. Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. *Proceedings* **2018**, *2*, 697. [[CrossRef](#)]
79. Chen, C.; He, W.; Zhou, H.; Xue, Y.; Zhu, M. A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Sci. Rep.* **2020**, *10*, 1–13. [[CrossRef](#)]
80. Sahoo, S.; Russo, T.A.; Elliott, J.; Foster, I. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resour. Res.* **2017**, *53*, 3878–3895. [[CrossRef](#)]
81. Sun, A.Y. Predicting groundwater level changes using GRACE data. *Water Resour. Res.* **2013**, *49*, 5900–5912. [[CrossRef](#)]
82. Gong, Y.; Zhang, Y.; Lan, S.; Wang, H. A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida. *Water Resour. Manag.* **2016**, *30*, 375–391. [[CrossRef](#)]
83. Gong, Y.; Wang, Z.; Xu, G.; Zhang, Z. A comparative study of groundwater level forecasting using data-driven models based on ensemble empirical mode decomposition. *Water* **2018**, *10*, 730. [[CrossRef](#)]
84. Zhou, T.; Wang, F.; Yang, Z. Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. *Water* **2017**, *9*, 781. [[CrossRef](#)]
85. Nayak, P.C.; Satyaji Rao, Y.R.; Sudheer, K.P. Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resour. Manag.* **2006**, *20*, 77–90. [[CrossRef](#)]
86. Alahmadi, F.S. Groundwater quality categorization by unsupervised machine learning in Groundwater quality categorization by unsupervised machine learning in Madinah, Western Kingdom of Saudi Arabia. In Proceedings of the International Geoinformatics Conference, Riyadh, Saudi, 10–14 February 2019.
87. Ransom, K.M.; Nolan, B.T.; Traum, J.A.; Faunt, C.C.; Bell, A.M.; Ann, J.; Gronberg, M.; Wheeler, D.C.; Rosecrans, C.Z.; Jurgens, B.; et al. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Sci. Total Environ.* **2017**, *601–602*, 1160–1172. [[CrossRef](#)] [[PubMed](#)]
88. Huang, X.; Gao, L.; Crosbie, R.S.; Zhang, N.; Fu, G.; Doble, R. Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning. *Water* **2019**, *11*, 1879. [[CrossRef](#)]
89. Isikdogan, F.; Bovik, A.C.; Passalacqua, P. Surface water mapping by deep learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 4909–4918. [[CrossRef](#)]
90. Conrado, S.P.; Neville, K.; Woodworth, S.; O’Riordan, S. Managing social media uncertainty to support the decision making process during Emergencies. *J. Decis. Syst.* **2016**, *25*, 171–181. [[CrossRef](#)]

91. Bendler, J.; Wagner, S.; Brandt, T.; Neumann, D. Taming uncertainty in big data: Evidence from social media in urban areas. *Bus. Inf. Syst. Eng.* **2014**, *6*, 279–288. [[CrossRef](#)]
92. Chung, Y.; Servan-Schreiber, S.; Zraggen, E.; Kraska, T. Towards Quantifying Uncertainty in Data Analysis & Exploration. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* **2018**, *41*, 15–27.
93. Wang, X.; He, Y. Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies. *IEEE Syst. Man Cybern. Mag.* **2016**, *2*, 26–31. [[CrossRef](#)]
94. Loew, A.; Bell, W.; Brocca, L.; Bulgin, C.E.; Burdanowitz, J.; Calbet, X.; Donner, R.V.; Ghent, D.; Gruber, A.; Kaminski, T.; et al. Validation practices for satellite-based Earth observation data across communities. *Rev. Geophys.* **2017**, *55*, 779–817. [[CrossRef](#)]
95. Demaria, E.M.C.; Serrat-Capdevila, A. Challenges of Remote Sensing Validation. In *Earth Observation for Water Resources Management: Current Use and Future Opportunities for the Water Sector*; The World Bank: Washington, DC, USA, 2016; pp. 167–171. ISBN 978-1-4648-0475-5.
96. Keim, D.A.; North, S.C.; Sips, M. Visual data mining in large geospatial point sets. *IEEE Comput. Graph. Appl.* **2004**, *5*, 36–44. [[CrossRef](#)]
97. Cavanillas, J.M.; Curry, E.; Wahlster, W. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer: New York, NY, USA, 2016; ISBN 9781420037050.
98. Faroukhi, A.Z.; El Alaoui, I.; Gahi, Y.; Amine, A. Big data monetization throughout Big Data Value Chain: A comprehensive review. *J. Big Data* **2020**, *7*, 1–22. [[CrossRef](#)]
99. Jony, R.I.; Rony, R.I.; Rahat, A.; Rahman, M. Big Data Characteristics, Value Chain and Challenges. In Proceedings of the 1st International Conference on Advanced Information and Communication Technology, Chittagong, Bangladesh, 16–17 May 2016.
100. Nasser, T.; Tariq, R.S. Big Data Challenges. *J. Comput. Eng. Inf. Technol.* **2015**, *4*, 4–11.
101. Padgavankar, M.; Gupta, S. Big data storage and challenges. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 2218–2223.
102. Alarabi, L.; Mokbel, M.F.; Musleh, M. ST-Hadoop: A MapReduce framework for spatio-temporal data. *Geoinformatica* **2018**, *22*, 785–813. [[CrossRef](#)]
103. Blumenfeld, J. EOSDIS DAACs Celebrate Milestones of Service to Global Data Users. Available online: <https://earthdata.nasa.gov/learn/articles/daac-overview-and-milestones> (accessed on 26 November 2019).
104. Lewis, A.; Oliver, S.; Lymburner, L.; Evans, B.; Wyborn, L.; Mueller, N.; Raevksi, G.; Hooke, J.; Woodcock, R.; Sixsmith, J.; et al. The Australian Geoscience Data Cube—Foundations and lessons learned. *Remote Sens. Environ.* **2017**, *202*, 276–292. [[CrossRef](#)]
105. Baumann, P.; Mazzetti, P.; Ungar, J.; Barbera, R.; Barboni, D.; Beccati, A.; Bigagli, L.; Boldrini, E.; Bruno, R.; Calanducci, A.; et al. Big Data Analytics for Earth Sciences: The EarthServer approach. *Int. J. Digit. Earth* **2016**, *9*, 3–29. [[CrossRef](#)]
106. Klein, L.J.; Marianno, F.J.; Albrecht, C.M.; Freitag, M.; Lu, S.; Hinds, N.; Shao, X.; Bermudez Rodriguez, S.; Hamann, H.F. PAIRS: A scalable geo-spatial data analytics platform. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 1290–1298.
107. Lu, S.; Shao, X.; Freitag, M.; Klein, L.J.; Renwick, J.; Marianno, F.J.; Albrecht, C.; Hamann, H.F. IBM PAIRS curated big data service for accelerated geospatial data analytics and discovery. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 2672–2675.
108. Cinquini, L.; Crichton, D.; Mattmann, C.; Harney, J.; Shipman, G.; Wang, F.; Ananthakrishnan, R.; Miller, N.; Denvil, S.; Morgan, M.; et al. The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. *Future Gener. Comput. Syst.* **2014**, *36*, 400–417. [[CrossRef](#)]
109. Eldawy, A.; Mokbel, M.F. SpatialHadoop: A MapReduce framework for spatial data. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, South Korea, 13–17 April 2015; pp. 1352–1363.
110. Aji, A.; Wang, F.; Vo, H.; Lee, R.; Liu, Q.; Zhang, X.; Saltz, J. Hadoop GIS: A high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow.* **2013**, *6*, 1009–1020. [[CrossRef](#)]

111. Whitby, M.A.; Fecher, R.; Bennight, C. GeoWave: Utilizing Distributed Key-Value Stores for Multidimensional Data. In *Advances in Spatial and Temporal Databases*; Gertz, M., Renz, M., Zhou, X., Hoel, E., Ku, W.-S., Voisard, A., Zhang, C., Chen, H., Tang, L., Huang, Y., et al., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10411, pp. 105–122. ISBN 978-3-319-64366-3.
112. Yu, J.; Wu, J.; Sarwat, M. GeoSpark: A cluster computing framework for processing large-scale spatial data. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; pp. 1–4.
113. Escandón-Quintanilla, M.-L.; Gardoni, M.; Cohendet, P. Big Data Analytics as Input for Problem Definition and Idea Generation in Technological Design. *IFIP Int. Conf. Prod. Lifecycle Manag.* **2016**, *492*, 423–432. [[CrossRef](#)]
114. Passi, S.; Barocas, S. Problem formulation and fairness. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 39–48.
115. CDWR. *Best Management Practices for the Sustainable Management of Groundwater: Sustainable Management Criteria*; California Department of Water Resources: Sacramento, CA, USA, 2017.
116. Humphrey, V.; Gudmundsson, L.; Seneviratne, S.I. Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes. *Surv. Geophys.* **2016**, *37*, 357–395. [[CrossRef](#)]
117. Rodell, M.; Famiglietti, J.S.; Wiese, D.N.; Reager, J.T.; Beaulieu, H.K.; Landerer, F.W.; Lo, M.-H. Emerging trends in global freshwater availability. *Nature* **2018**, *557*, 651–659. [[CrossRef](#)] [[PubMed](#)]
118. Seyoum, W.M. Characterizing water storage trends and regional climate influence using GRACE observation and satellite altimetry data in the Upper Blue Nile River Basin. *J. Hydrol.* **2018**, *566*, 274–284. [[CrossRef](#)]
119. Cobbing, J.; Eales, K.; Rossouw, T. *The Path to Successful Water User Associations in the North West Dolomite Aquifers: Report to the Water Research Commission*; Water Research Commission: Pretoria, South Africa, 2016.
120. Fistikoglu, O.; Gunduz, O.; Simsek, C. The Correlation between Statistically Downscaled Precipitation Data and Groundwater Level Records in North-Western Turkey. *Water Resour. Manag.* **2016**, *30*, 5625–5635. [[CrossRef](#)]
121. Sachindra, D.A.; Ahmed, K.; Rashid, M.M.; Shahid, S.; Perera, B.J.C. Statistical downscaling of precipitation using machine learning techniques. *Atmos. Res.* **2018**, *212*, 240–258. [[CrossRef](#)]
122. Jha, S.K.; Mariethoz, G.; Evans, J.P.; McCabe, M.F. Demonstration of a geostatistical approach to physically consistent downscaling of climate modeling simulations. *Water Resour. Res.* **2013**, *49*, 245–259. [[CrossRef](#)]
123. Pour, S.H.; Shahid, S.; Chung, E.S.; Wang, X.J. Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh. *Atmos. Res.* **2018**, *213*, 149–162. [[CrossRef](#)]
124. Moghim, S.; Bras, R.L. Bias correction of climate modeled temperature and precipitation using artificial neural networks. *J. Hydrometeorol.* **2017**, *18*, 1867–1884. [[CrossRef](#)]
125. Yin, W.; Hu, L.; Zhang, M.; Wang, J.; Han, S.C. Statistical Downscaling of GRACE-Derived Groundwater Storage Using ET Data in the North China Plain. *J. Geophys. Res. Atmos.* **2018**, *123*, 5973–5987. [[CrossRef](#)]
126. Duhan, D.; Pandey, A. Statistical downscaling of temperature using three techniques in the Tons River basin in Central India. *Theor. Appl. Climatol.* **2015**, *121*, 605–622. [[CrossRef](#)]
127. Pang, B.; Yue, J.; Zhao, G.; Xu, Z. Statistical Downscaling of Temperature with the Random Forest Model. *Adv. Meteorol.* **2017**, 2017. [[CrossRef](#)]
128. Beecham, S.; Rashid, M.; Chowdhury, R.K. Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalized Linear Model. *Int. J. Climatol.* **2014**, *34*, 3654–3670. [[CrossRef](#)]
129. He, X.; Chaney, N.W.; Schleiss, M.; Sheffield, J. Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.* **2016**, *52*, 8217–8237. [[CrossRef](#)]
130. Seyoum, W.M.; Kwon, D.; Milewski, A.M. Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system. *Remote Sens.* **2019**, *11*, 824. [[CrossRef](#)]
131. Ghosh, S.; Mujumdar, P.P. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Resour.* **2008**, *31*, 132–146. [[CrossRef](#)]
132. Hashmi, M.Z.; Shamseldin, A.Y.; Melville, B.W. Statistical downscaling of precipitation: State-of-the-art and application of bayesian multi-model approach for uncertainty assessment. *Hydrol. Earth Syst. Sci. Discuss.* **2009**, *6*, 6535–6579. [[CrossRef](#)]

133. Liu, Z.; Xu, Z.; Charles, S.P.; Fu, G.; Liu, L. Evaluation of two statistical downscaling models for daily precipitation over an arid basin in China. *Int. J. Climatol.* **2011**, *31*, 2006–2020. [[CrossRef](#)]
134. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286. [[CrossRef](#)]
135. Bonner, S.; Kureshi, I.; Brennan, J.; Theodoropoulos, G. Exploring the Evolution of Big Data Technologies. In *Software Architecture for Big Data and the Cloud*; Mistrik, I., Bahsoon, R., Ali, N., Heisel, M., Maxim, B., Eds.; Elsevier Inc.: Burlington, MA, USA, 2017; pp. 253–283. ISBN 9780128054673.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).