Big Earth data: A new frontier in Earth and information sciences

Huadong Guo

Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Big data is a revolutionary innovation that has allowed the development of many new methods in scientific research. This new way of thinking has encouraged the pursuit of new discoveries. Big data occupies the strategic high ground in the era of knowledge economies and also constitutes a new national and global strategic resource. “Big Earth data”, derived from, but not limited to, Earth observation has macro-level capabilities that enable rapid and accurate monitoring of the Earth, and is becoming a new frontier contributing to the advancement of Earth science and significant scientific discoveries. Within the context of the development of big data, this paper analyzes the characteristics of scientific big data and recognizes its great potential for development, particularly with regard to the role that big Earth data can play in promoting the development of Earth science. On this basis, the paper outlines the Big Earth Data Science Engineering Project (CASEarth) of the Chinese Academy of Sciences Strategic Priority Research Program. Big data is at the forefront of the integration of geoscience, information science, and space science and technology, and it is expected that big Earth data will provide new prospects for the development of Earth science.

1. Big data for science

With data volumes expanding beyond the petabyte and exabyte levels across many scientific disciplines, the role of big data in scientific research is becoming increasingly apparent. According to an article by the International Data Corporation (IDC), the total amount of digital data created, replicated, and consumed more than doubles every two years. There were about 4.4 zettabytes (ZB) of data created, replicated, and consumed worldwide in 2013. The IDC estimates that by the year 2025, Figure 1 will reach 163 ZB (Reinsel, Gantz, & Rydning, 2017). In the current big data age, national competitiveness will be reflected in the size, quality, and applicability of a country’s data. Big data has become a manifestation of informational sovereignty; it will be the next topic of international debate and will play a significant role in border, coastal, and air defense (Fang, 2013; Guo, Wang, & Liang, 2016).
From 2016 to 2025, the digital universe will grow by a factor of 10 – from 16ZB to 163ZB. It more than doubles every two years.

Figure 1. Global growth of data volume from 2016 to 2025.
Big data has begun to significantly influence global production, circulation, distribution, and consumption patterns. It is changing humankind’s production methods, lifestyles, mechanisms of economic operation, and country governance models. Big data occupies the strategic high ground in the era of knowledge-driven economies and it is a new strategic resource for all nations.

Big data’s contributions to scientific discoveries have started to be recognized. In one case, CERN scientists analyzed the records of 800 trillion particle collisions (big data) in the process of trying to find the Higgs particle. Unprecedentedly large data-sets generated, sensed, and harvested from experiments, observations, and simulations have brought about great opportunities for making scientific progress for two reasons. (1) Huge data-sets can serve as important inputs and support the adjustment and validation of current theories that relate to important scientific problems, thus leading to new findings. A good example is the new paradigm of “big data meets big models” that is important in large inverse problems. (2) Massive data-sets themselves are able to provide endless sources of new knowledge without the need to model scientific phenomena. This has been called the fourth paradigm – data-intensive scientific discovery. There is, therefore, no doubt that big data will significantly change the way scientific discoveries are made in the future. Scientists must be prepared to welcome a new age in which big data will play an important role and might dominate the methodologies used in scientific research.

Big data research is different from traditional logical research. It uses analytical induction applied to a vast amount of data to statistically search, compare, cluster, and classify. It involves correlation analysis and implies that there may be certain a regularity in the relation between the values of two or more variables; it also aims to uncover hidden correlated networks within data-sets (Li & Cheng, 2012). Thus, it can be seen that the substantive characteristics of big-data computing comprise a paradigm shift from model-driven science to data-driven science, as well as the establishment of a data-intensive scientific approach. Scientific research has employed observation-based science from the very beginning, including the experimental science that began thousands of years ago, the theoretical science that emerged in the seventeenth century, and the computing paradigm that arose in the twentieth century. In today’s big data era, a new paradigm of data-intensive scientific discovery has emerged that is less dependent on models and a priori knowledge (Guo, Wang, Chen, & Liang, 2014). By seeking relationships within large amounts of data, new models, new knowledge, and new laws can be discovered and explored.

In an initiative of the author, the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU) has worked with other international science organizations and initiatives to explore the value of big data in scientific research and to reinforce the crucial role of science in the development of big data. After the June 2014 “International Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities”, sponsored by CODATA in Beijing and co-sponsored by the ICSU World Data System, Future Earth, Integrated Research on Disaster Risk, the Research Data Alliance, the Group on Earth Observations, the International Society for Digital Earth, and the Chinese Academy of Sciences Institute of Remote Sensing and Digital Earth, CODATA, and others developed a joint statement of recommendations and actions (Guo, 2014). This statement emphasized providing a better understanding of big data for scientific research, and strengthening international science for the benefit of society by developing research, policies, and frameworks related to big data.
Although only a starting point at that time, this statement was a practical step toward focusing attention on the potential of big data, recognizing that big data presents particularly significant challenges and notable opportunities for transdisciplinary international research programs as well as for scientific data services and infrastructure providers. The major points in the statement are: (1) responding to the importance of big data for international scientific programs; (2) exploiting the benefits of big data for society; (3) improving the understanding of big data through international collaboration; (4) promoting universal access to big data through global research infrastructure; (5) exploring and addressing the challenges of big data stewardship; (6) encouraging capacity building and skills development in big data science; and (7) fostering the development of policies to maximize the exploitation of big data (Workshop on Big Data for International Scientific Programmes: Challenges & Opportunities, 2014).

Since then, a series of meetings on big data for science have been either organized or co-organized by our research team. These have included the “Xiangshan Science Conference on Frontiers of Scientific Big Data”, “The Academic Divisions of the Chinese Academy of Sciences Forum on Frontiers of Science and Technology for Big Earth Data from Space”, and the “Exploratory Round Table Conference on Big Data in Natural Sciences, Humanities and Social Sciences”. It is our opinion that scientific big data will play a key role in promoting scientific development.

2. Scientific big data

As a branch of big data, scientific big data is a typical representative of data-intensive science. Scientific big data has a number of characteristics, including complexity, comprehensiveness, and global coverage, as well as high degree of integration with information and communication technology. The approaches used in science in general are also being transformed – from single-discipline to multidisciplinary and interdisciplinary approaches, from natural science to the integration of natural and social sciences, and from work being carried out by individuals or small research groups to projects being run by international scientific organizations. In addition to scientists being able to solve hard or previously insoluble problems through the real-time dynamic monitoring and analysis of various related data, data itself can become the object and tool of research: scientists can conceive, design, and implement their research based on the data (Hey, Tansley, & Tolle, 2009).

Although scientific big data has become important to research, and the paradigm of data-intensive scientific discovery has been widely recognized, the associated theories, methodologies, and models have yet to be widely used. At present, the concept and application of big data have been accepted and developed in the fields of network sciences and economics. However, there has been relatively little theoretical study of big data and the practice of it is generally weak (Guo et al., 2014). This is because big data has its own specific scientific connotations and characteristics.

In order to better understand and study scientific big data, we scientists must sort out what distinguishes the essential attributes and characteristics of scientific big data from other types of big data. On the whole, scientific big data has the following external features: (1) From the perspective of data content, scientific big data generally represent objective natural objects and processes. (2) From the perspective of data volume, scientific big data differ greatly between different disciplines. (3) From the perspective of the data update rate,
there is again great variation between disciplines – e.g. Earth observation data have a fast update rate while that dealing with geological processes have a slow update rate. (4) In terms of data acquisition approaches, scientific big data is generally acquired from observation, experimental records, and subsequent processing. (5) From the perspective of data analysis methods, scientific big data is generally combined with a scientific principle or model to form a method of scientific discovery. However, it is rare to rely entirely on data analysis and to put aside general scientific principles.

By summarizing the external features of scientific big data, the internal features also become fairly clear and can be summarized as follows. (1) Non-repeatability: in general, observations of natural and physical objects and processes cannot be repeated. (2) High uncertainty: big data involves different approaches to observation and recording, as well as to indirect observation and sampling (Kennedy & O'Hagan, 2001). (3) High dimensionality: a wide range of data sources and difficult mathematical analysis methods lead to the curse of dimensionality. (4) High degree of computational complexity: this results from the high degree of uncertainty, high dimensionality, and complexity in the main methods used for data analysis. It can be said that scientific big data has different characteristics from normal big data, and that its internal mechanisms and how to apply it to research are, therefore, worth further study (Guo, 2014).

“Observing system”, “Data management”, and “Earth system modeling” are three of the Future Earth initiative’s eight cross-cutting issues in its research themes and these have a close relationship to big data. A large number of observational data-sets provided by Earth observation networks, including satellite, airborne and ground sensors, have large volumes. The Earth observation data system not only needs to obtain a large amount of data quickly, but also needs to carry out real-time processing and analysis. Besides this, using metadata management and by applying an appropriate data policy, the data system should be able to reduce the amount of uncertainty regarding the quality of the data (Guo, 2014). Earth system modeling involves social models, scientific models, Earth observation, and economic and other data. The data-sets involved can be very rich.

In 2013, the author raised the concept of “scientific big data”. Following this, the Chinese Government approved and supported our proposal to carry out research into scientific big data. In the “Action Plan for the Promotion of Big Data Development” issued by the State Council of the People’s Republic of China in 2015, “Development of Scientific Big Data” was proposed to

Actively promote public welfare research supported by national public finance activities. The purpose is to obtain and generate data to be gradually opened and shared in order to build scientific big data of the country’s major infrastructure with authoritative collection, long-term preservation, integrated management and comprehensive sharing. For economic and social development needs, it is essential to develop scientific big data applications service centers to support the resolution of economic and social development and major national security issues.

3. Big Earth data

3.1. Characteristics of big Earth data

Earth science research, including that into the atmosphere, land and ocean, has produced huge data-sets derived from satellite observations, ground sensor networks, and other sources. This is collectively called big Earth data. Big Earth data has features in common with
scientific big data, but also has its own particular characteristics. Big Earth data is characterized as being massive, multi-source, heterogeneous, multi-temporal, multi-scale, high-dimensional, highly complex, nonstationary, and unstructured. It provides support for data-intensive research in the Earth sciences.

Taking global change research as an example, this demands the systematization of the Earth and the making of comprehensive observations, and has thus led to the rapid development of ground observation technology. Modern Earth science requires globally established, quasi real-time, all-weather Earth data acquisition capabilities, and has developed an integrated space-air-ground observation system with high spatial, temporal, and spectral resolutions. Global change research focuses on global sustainable development and must deal with key multidisciplinary challenges, including global change process monitoring, simulation analysis, and response strategies. These studies rely on big Earth data, such as long-term, multi-spatio-temporal Earth observation data; accurate, continuous ground station observation data; and experimental data based on theoretical speculation and estimated data. Therefore, big Earth data can provide a new approach to the development of global change research (Guo, Chen, et al., 2016). As a tool in cross-disciplinary research, big Earth data has the potential to provide a virtual Earth that can be used not only in the Earth sciences but that also has a close relation to information science, space science, the humanities, and the social sciences. Overall, big Earth data includes the main features of big data.

Big Earth data is big data as used in the Earth sciences. It possesses the general properties of big data, but it also has strong temporal, spatial, and physical correlations. As far as being “massive”, it has a high-resolution, is highly dynamic and consists of multiple bands. It is also characterized by a high data acquisition rate and short cycle. It is “multi-source” in that its data sources and acquisition methods are diverse because imaging mechanisms and models vary widely. It can be considered “multi-temporal” because of the short sampling intervals and high data acquisition frequency. Big Earth data is also “high value” because of the importance of research into the ecological environment, land resources, natural disasters, and other geoscience.

Big Earth data are characterized by multiple spatial and temporal scales. This arises due to the multi-grade subsystems used in Earth observation. Each subsystem has its own spatio-temporal scale and so the acquired Earth observation data have different rules and characteristics at the different scales. Monitoring of the Earth’s systems is complex because it spans local, regional, and global scales, and the temporal scales range from seconds to millennia. Integrating all these different types of data with a single system or platform is, in itself, a tough task, even without any subsequent processing and analysis.

Despite the new impetus behind Earth science, big Earth data poses great technical challenges with respect to transmission, storage, processing, analysis, management, sharing, and scientific discovery. Scientists are engaged in research into, as well as the development of, big Earth data-oriented computing platforms, algorithms, and software systems including high-performance systems, mass storage techniques, fully automated processing techniques, efficient computing techniques, and data-sharing and service systems. Although these techniques have delivered some innovations, a set of key technical problems remain to be solved. These problems include mass multivariate data integration and mining; the multi-level hybrid parallel computing of mass concurrent tasks, data, and algorithms; and multi-source dynamic collaborative data processing. These problems persist due to the characteristics of early big data technology used in Earth science, and the marked differences between big Earth data
and the types of big data used in other fields. Another concern is big Earth data-intensive scientific discoveries. Discoveries made possible through the use of big Earth data not only involve the extraction of information but also the mining of hidden, implicit patterns, and laws. As a result of the large-scale and high dimensionality but low information density of big Earth data, scientists are exploring appropriate ways to simplify the size and dimension of the data by means of artificial intelligence, and of reducing the scale of the data before carrying out subsequent research. In addition, the very large data amounts make it possible to gradually change the scientific discovery from “model-driven” to “data-driven”. However, there is an urgent need to develop innovative theories and approaches, such as cognitive models and data mining, for big Earth data-oriented scientific discovery because the efficient mining of the knowledge contained in big Earth data remains at an early stage (Guo, Chen, et al., 2016).

With the development of technology in the field of Earth science, a huge amount of scientific big data have been generated through various Earth observation, geo-technology surveying, and ground sensor networks. These data-sets contain rich information that is heterogeneous, multi-source, multi-temporal, multi-scalar, highly dimensional, highly complex, and unstructured (Guo et al., 2017). There is great potential to promote the in-depth development of Earth science research, and big Earth data can be seen as a new key to understanding the world.

3.2. Big Earth data technologies

Technologies related to big Earth data include Earth observation, communication technology, computing, and networks, among others. Along with the recognition that big Earth data has deepened humankind’s capability for understanding the Earth, it has also become necessary to meet the challenges brought by the transmission, storage, processing, analysis, management, and sharing of big Earth data. The huge amount of Earth observation data combined with real-time or near real-time acquisition rates, as well as the multiple scales involved, are a particular challenge for existing technology. More numerous, longer term, and cheaper sensors and more time-critical requirements for sharing data have increased the complexity of storage, processing, and computing. Storage environments are built using clustering technology that maps thousands of physical storage devices into a large storage platform through virtualization and these devices work together cooperatively and productively. With distributed and cloud storage, users can store and manage geospatial data on platforms anytime and anywhere to provide service-on-demand support for geospatial data processing and analysis. Among the subsets of big Earth data, Earth observation big data is one of the most important (Guo et al., 2017).

After nearly half a century of development, Earth observation technology is providing a new vision and new methods for Earth science research. It has further deepened humanity’s understanding of the Earth, especially in terms of macro-knowledge. Big Earth data can exert a profound influence on the development of the Earth sciences.

With the help of Earth observation platforms, humans can observe the Earth without interruption. We can rapidly reproduce and objectively reflect on the status, phenomena, processes, spatial distribution, and locations within the epigeosphere through information processing in the service of economic construction and social development. Earth observation technology, which forms the core of geospatial information science and technology,
has become a comprehensive national embodiment of the capacity for scientific and technological achievement, economic strength, and national security. The demand for Earth observation applications and the development of satellite and sensor technologies has grown, substantially increasing the number of Earth observation satellites as well as performance indexes. The volume of Earth observation data has doubled and redoubled. According to statistics from the global satellite mission generated by the Committee on Earth Observation Satellites (CEOS), in the last half century, globally, more than 514 Earth observation satellites were launched for comprehensive observation of the Earth’s systems, including the atmosphere, ocean, and land.

Since China successfully launched its first meteorological satellite in 1988 after about 30 years of development, the nation has gradually developed an Earth observation satellite system, including a series of resource, environmental, meteorological and ocean satellites; high-resolution Earth observing satellites; and the BeiDou navigation satellites (see Figure 2).

With the 1999 launch of the first Earth Observing System (EOS) satellite, Terra, NASA introduced the first satellite-based observation system to offer integrated measurements of the Earth’s processes. NASA has also built the largest scientific data system in the world, the Earth Observing System Data and Information System, which currently collects environmental measurements from more than 30 satellites (http://eospso.gsfc.nasa.gov/). ESA (European Space Agency) is developing a new family of missions called Sentinels (Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-4, Sentinel-5, Sentinel-5 Precursor, and Sentinel-6), specifically for the operational needs of the Copernicus program (Guo, Chen, et al., 2016).

The volume of environmental data that has so far been collected by different countries, regions, and organizations is truly amazing. This is a manifestation of the high-resolution Earth observation age, where the volume of Earth observation data is growing (He et al., 2015). The era of big Earth data has arrived.

### 3.3. Big Earth data drives scientific discovery methods

The analysis and data mining of big Earth data is an important way of demonstrating the value of big data in general and of effectively using it; however, due to its nature, it is more complex and difficult to mine big Earth data than other types of big data. Hence, an innovative theory and method should be developed for use with big Earth data.

It is difficult to use traditional methods of data mining or scientific discovery with big Earth data. As for the traditional methods of dealing with big data in remote sensing, these rely on divide-and-conquer and scale-change strategies and include such methods as traditional classification methods and machine learning. As well as presenting serious problems for data-intensive computing because of the need to search a huge and complex attribute space, using these methods, it is also easy to find meaningless modes because of the highly complex correlation and high noise within the data (Nagarajan et al., 2009). From the descriptions of the spatial autoregressive model, the Markov Random Field (MRF) classification method, Gaussian processes for machine learning, and other computationally complex traditional spatio-temporal data mining methods, it is easy to see that, in big data scientific discovery, there are high computational requirements and frequent I/O operations. The traditional spatio-temporal data mining methods must, therefore, be modified for big data applications. To effectively improve data mining and scientific discovery for big Earth data,
there is an urgent need for the development of automated methods of data mining based on intelligent reasoning theory (Guo, Chen, et al., 2016; Vatsavai et al., 2012).

### 3.4. Big Earth data drives development of Earth sciences

Big Earth data is the new power behind the development of the Earth sciences (Graham & Shelton, 2013; Guo et al., 2010). Using the example of global change, the following discussion describes the significant advances that have been supported by big Earth data.

Since the middle of the twentieth century, human society has faced various challenges related to sustainable development, such as global change, natural resource depletion, food and water insecurity, energy shortages, environmental degradation, natural disaster response, and population growth. Global change science has developed in response to these challenges (Xu et al., 2013).

All of the resource and environmental issues caused by global change are essentially the result of the interactions between various layers of the Earth. An approach to solving ecological-environmental issues through research into the complex Earth system and the interactions between all its layers and subsystems is thus required. This requires a wide range of integrated data from the different elements of the Earth system – atmosphere, hydrosphere, biosphere, cryosphere, and lithosphere – to obtain continuous, accurate, repetitive global data over the long term. The observation of these layers can help reveal the implicit spatio-temporal structure of the processes occurring at the Earth’s surface and the related laws. It also helps in the exploration and understanding of the mechanisms that are changing the Earth system under the influence of human activities to research comprehensive strategies for adapting to global change and to reduce the effects of global change by providing a scientific basis for ecological-environmental protection and sustainable development. Therefore, global change science is a major effort that includes the monitoring of global change processes, the simulation and analysis of global change, and the development of response strategies. All of this research has a close relationship with big Earth data and involves long-term data sequences, multiple temporal and spatial observation scales, accurate, continuous ground-station observations and experimental data, the theoretical speculation and estimated data based on existing scientific basis. Big Earth data is flourishing, while the number of scientific satellites used in global change research is increasing (Guo et al., 2014) and these satellites will all provide scientific data, information, and knowledge for use in global change science.

In 2014, the UN established an award called the “Big Data Climate Challenge”. A project entitled “Big Earth Observation Data for Climate Change Research” by the author’s research team at the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, was one of the winners. This project showed how big data can drive climate change research.

In 2015, the UN initiated 17 Goals for Sustainable Development that aimed to end poverty, inequality, and climate change by 2030. Of these 17 goals, Big Earth Data could help with the realization of at least 8 of them in different ways. Specifically, these 8 goals are clean water, affordable energy, sustainable cities, climate change, life below water, life on land, good health, and peace.
**Figure 2. Development of Earth observation satellites in China.**

| Satellite category | 1988 | 1990 | 1997 | 1999 | 2000 | 2002 | 2003 | 2004 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | Satellites to be launched before 2020 |
|--------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------------------------|
| FY-1 polar orbit satellite | FY2A | FY3B | FY2C | FY2D | FY2A | FY2B | FY2C | FY2D | FY2E | FY2F | FY2G | FY2H | FY2J | FY2K | FY2L | FY2M | FY2N | 8 service satellites and 2 experimental satellites |
| FY-3 polar orbit satellite | FY3A | FY3B | FY3C | FY3D | FY3A | FY3B | FY3C | FY3D | FY3E | FY3F | FY3G | FY3H | FY3I | FY3J | FY3K | FY3L | FY3M | 8 satellites |
| HY-3A ocean color satellite | HY4A | HY1B | HY2A | HY3A | HY4A | HY1B | HY2A | HY3A | HY4A | HY1B | HY2A | HY3A | HY4A | HY1B | HY2A | HY3A | HY4A | 5 satellites |
| ZY-1 satellite | CBERS-01 | ZY-1/2 | CBERS-04 | ZY-1/2 | CBERS-01 | ZY-1/2 | CBERS-04 | ZY-1/2 | CBERS-01 | ZY-1/2 | CBERS-04 | ZY-1/2 | CBERS-01 | ZY-1/2 | CBERS-04 | ZY-1/2 | CBERS-01 | ZY-1/2 | CBERS-04 | 8 satellites |
| ZY-2 satellite | CBERS-02 | ZY-2 | CBERS-04 | ZY-2 | CBERS-02 | ZY-2 | CBERS-04 | ZY-2 | CBERS-02 | ZY-2 | CBERS-04 | ZY-2 | CBERS-02 | ZY-2 | CBERS-04 | ZY-2 | CBERS-02 | ZY-2 | CBERS-04 | ZY-2 | CBERS-02 | 4 satellites |
| ZY-3 satellite | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | ZY-3 | 1 Global coverage |

**Note:** The table above illustrates the development of Earth observation satellites in China from 1988 to 2015. The data includes the launch years, satellite names, and the number of satellites. The table also indicates the satellites to be launched before 2020. The satellites are categorized into Meteorological satellites, Ocean satellites, Resource satellites, and Baidu navigation and positioning satellites.
4. CASEarth – a big Earth data project

The Chinese Academy of Sciences (CAS) has been fully aware of the importance of big Earth data and decided to launch a project entitled “Big Earth Data Science Engineering (CASEarth)” in the Strategic Priority Research Program (SPRP) of the CAS in order to carry out structured research on big Earth data. SPRP is a major scientific and technological project approved by the Chinese Government that is organized and implemented by the CAS. It is oriented toward resolving major scientific and technological problems concerning overall and long-term development. It is a strategic action plan that integrates technical problem-solving with team- and platform-building. The plan aims to make innovative breakthroughs and foster cluster advantages. If big Earth data is the engine that propels discovery and innovation in Earth science, then the purpose of CASEarth is to provide a new impetus for interdisciplinary, cross-scale, macro-scientific discoveries using big Earth data. The project will also investigate a series of critical scientific problems using a structured, holistic view and major breakthroughs in the scientific understanding of the Earth’s system and new strides with respect to decision support are expected.

The overall objective of CASEarth is to establish an International Center for Big Earth Data Science, which will have three main aims.

1. Building the world’s leading Big Earth Data infrastructure. To overcome the bottleneck of data access and sharing, CASEarth will develop a multidisciplinary Big Earth Data and cloud service platform, which will form part of the key national scientific and technological big data infrastructure that supports national macro-level decisions and major scientific discoveries.

2. Developing a world-class big Earth data platform to drive the discipline. CASEarth will explore the new big data-driven, multidisciplinary, globally collaborative paradigm of scientific discovery, and showcase as well as spur major breakthroughs in Earth system sciences, life sciences, and associated disciplines.

3. Constructing a decision support system. The system should serve high-level government authorities and be applied to multiple issues using multi-angle and panoramic visual analysis, simulation, and deduction. The system should support a number of national projects, including the “Belt and Road” initiative, “Beautiful China”, and the national globalization strategy, Human Destiny Community.

CASEarth will showcase its features and demonstrate its important output in the following three ways (see Figure 3).

1. Scientific discoveries: CASEarth will develop new approaches and paradigms for big data-driven scientific discoveries. The Big Earth Data System will reveal the complex coupling interactions and correlations between different elements at global and regional scales, reveal details at different resolutions and different scales, and herald the implementation of scientific research approaches for the Earth System by reproducing the spatial distribution and temporal dynamics of the land, oceans, atmosphere, and human elements.

2. Technological innovations: CASEarth will construct a high-precision Big Earth Data cloud service platform and the world’s leading new Digital Earth system; integrate and display the numerous data and information products generated by CASEarth through precise geographical association and physical association; establish a big
Earth data cloud service that features a “transparent national deployment service combining centralization and decentralization”; and develop an evolving multidisciplinary, serviceable big Earth data critical infrastructure.

(3) Serve government decision support. CASEarth will create a big data-driven, visualized, interactive, dynamically evolving decision support environment, allowing for the integrated digital reproduction of multi-source spatial information and the assessment of multiple elements. It will provide macro-level, real-time big Earth data decision support, and support the China-led UN initiative, “The 2030 Agenda for Sustainable Development”.

CASEarth consists of the following eight research components to help it achieve technological breakthroughs and obtain innovative results, paying special attention to data sharing and encouraging interested scholars to rely on the platform to carry out research.

(1) CASEarth Small Satellites. CASEarth Small Satellites will be developed to provide continuous satellite data for the project, along with a CASEarth satellite operation management and evaluation system to complete the CASEarth satellite data reception and product service. Research into the overall design of observation missions, including highly integrated, miniaturized payload technology, infrared and multi-spectral payload technology, and mass data compression storage and transmission technology will also be carried out. A complete system, from satellite requirements to data products, will be developed through research into the overall design of and payloads carried on Earth observation satellites, the development of satellite engineering and management, the evaluation of satellite operations, the reception of satellite data, and research into product services.

(2) Big Data and Cloud Service Platform. This initiative will involve building a big Earth data cloud service platform with integrated service capabilities to provide unified computing and storage. The program calls for research into and the implementation of multi-source, heterogeneous mass data access, aggregation of data and storage management, as well as unified access through standard specifications, protocols, tools, and systems. The platform will integrate massive multi-source scientific data resources and big Earth data collected from integrated space-air-ground observations. The goal is to build the world’s leading global data repository and achieve breakthroughs in new methods of distributed computing resources, unified scheduling and aggregation services, grid data computing, and big data processing and analysis in order to achieve big data-driven scientific discovery and decision support.

(3) Digital Belt and Road. Following China’s Silk Road Economic Belt and the twenty-first Century Maritime Silk Road initiative, in collaboration with the Digital Belt and Road international science program (DBAR), this initiative will integrate big Earth data, including spatio-temporal data acquired in 49 major categories over the past 50 years, to build a big Earth data integration and technology evaluation system and science database. The system will carry out scientific analysis of big Earth data for the Belt and Road to better understand the spatial distribution of regional resources and environments, development potential, and change trends. The program will also compile a regional spatial assessment index for the United Nations 2030 Sustainable Development Goals (UN SDGs), including 8 related objectives and 33 indicators, to scientifically monitor the key environmental indicators of sustainable development.
Figure 3. Framework of the big Earth data project (CASEarth).
goals along the area covered by the Belt and Road. DBAR will also establish an international big Earth data analysis and decision-making system covering the same area.

4) Beautiful China. This initiative will take Earth system science and human–Earth relationship theory as a guide to develop a decision support and evaluation system using a multi-angle, multi-dimensional, multi-link, multi-factor, multi-level perspective. Beautiful China will carry out research and development based on big data related to resources, the environment, and their patterns of evolution; clean air and environmental health; the construction of an ecological civilization; regional development and smart cities. The system will be able to evaluate and forecast the status quo and future scenarios for Beautiful China and provide policy recommendations for the project and UN SDGs.

5) Biodiversity and Ecological Security. This initiative will carry out studies into data integration and sharing standards, along with the organic integration of biological, ecological, environmental, meteorological, economic, and other data to create a complete data layer. It will use analysis models and visualization technology for the data mining and utilization of biological diversity resources; construct a common interface for processing big data on open and exoteric biological diversity and ecological security; and establish a comprehensive big data platform with biodiversity and ecological security information as the core. The initiative aims to achieve different forms of personalized data services and decision support at different levels.

6) Three-Dimensional Information Ocean. This initiative will form a “two-point and one-side” ocean information resource pool. “One side” refers to the global scale of the marine information resource pool and data products, and the establishment of a basic global ocean data service system. “Two points” refers to focusing on the directional and regional advantages and the carrying out of information integration and scientific research on the two strategic points of “China offshore” and “two oceans and one sea”. In the key area of “two oceans and one sea”, scientists will study multi-source data and build a change database and structural model of the reefs in the South China Sea, a deep-sea biogeographic information system of the Western Pacific, a multi-dimensional demonstration system of extreme deep-sea habitats, marine disaster assimilation data products, and a marine disaster prediction and early warning system for the Indian Ocean and its key port areas. Finally, the Three-Dimensional Information Ocean will form a system with real-time marine displays, dynamic simulations, and scenario analysis visualization.

7) Spatio-temporal Three-Pole Environment. This initiative aims to take the lead in providing decision support for polar governance and Arctic development. It focuses on the topic of Earth’s “three poles”, i.e. the north and south poles as well as the high mountains of Central Asia, due to their importance to climate change research. The Spatio-temporal Three-Pole Environment initiative involves big data sharing and integration, comparative remote sensing studies, data analysis methods and multi-layer interactive models; research into ecological and spatio-temporal dynamics, forecasting, water environments and water security, climate change and its impact on China; as well as Arctic Channel monitoring and refined forecasting, changes in the cryosphere and their effects, permafrost, and other special studies.
(8) Digital Earth Science Platform (DESP). DESP is the integrated platform of CASEarth and focuses on comprehensively displaying big Earth data and the construction of a decision support system and network information service system. It mainly provides data, services, computing and other resources for visual analysis for multidisciplinary integration, big data-driven scientific discovery, and technological innovation. It also supports research into resources, the environment, biology, ecology, and other fields of scientific communication and public service. DESP is flexible, scalable, and multimodal to ensure the safe and reliable operation of the system.

CASEarth will break through the bottleneck of open data and data sharing, realize the comprehensive integration of data, models, and services in the fields of resources, the environment, biology, and ecology; promote multidisciplinary integration; and build a big Earth data and cloud service platform as well as a big data-driven Digital Earth Science Platform with global influence. It will provide a comprehensive display and dynamic simulation for sustainable development processes and ecological conditions along the Belt and Road and provide accurate evaluation and decision support for Beautiful China’s sustainable development. CASEarth will explore a new paradigm of scientific discovery involving big data-driven multidisciplinary integration and global collaboration, and it will constitute a major breakthrough in Earth systems science, life sciences, and related disciplines. CASEarth can become an international center for exoteric big data, comprehensively enhancing national technological innovation, scientific discovery, macro-decision-making and public knowledge dissemination, and other significant outputs.

5. Conclusion

Big data constitutes a strategic highland in the era of knowledge economies and is a new strategic resource for countries. Big data is changing human life and providing a deeper understanding of the world. It relies less on cause and effect and more on correlation to find new modes of developing knowledge, and has become a typical representative of the data-intensive scientific paradigm that is following on from the earlier paradigms of empirical, theoretical, and computational science, bringing innovative methodologies to scientific research and driving the development of disciplines.

Big Earth data, as a subset of big data, provides a new methodology for the Earth sciences, and is becoming a new key for understanding the Earth and the new engine for conducting Earth science. Hence, big Earth data could potentially revolutionize Earth science research. Using all types of big Earth data combined with Earth system models to create theories and methods for discovering and developing knowledge is a major scientific challenge that should be addressed in the Earth sciences. Research into Big Earth data as a discipline should continue and attention should be paid to the fields of Earth science, information science, space science, and technology in a cross-disciplinarily way in order to broaden the research direction of big Earth data and achieve new heights in Earth system science research. The ongoing CAS Strategic Priority Research Program is an open science program that has the goal of building an International Center for Big Earth Data Science. This forward-looking project is involving domestic and international scientists in research, producing discoveries in the Earth sciences, and supporting government decisions.
Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Project title: CASEarth (XDA19000000) and Digital Belt and Road (XDA19030000).

ORCID

Huadong Guo http://orcid.org/0000-0003-0337-1862

References

Fang, J. (2013). Network science and engineering faced with a new challenge and developing opportunity under the wave impact of big data. *Chinese Journal of Nature, 35*(5), 345–354. Retrieved from [http://nature.shu.edu.cn/EN/abstract/abstract13704.shtml](http://nature.shu.edu.cn/EN/abstract/abstract13704.shtml)

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Social Science Electronic Publishing, 3*(3), 255–261.

Guo, H. (2014). Big data, big science, big discovery – Review of CODATA workshop on big data for international scientific programmes. *Bulletin of Chinese Academy of Sciences, 29*(4), 500–506.

Guo, H., Chen, R., Zhiwei, X., Sun, J., Bi, J., Wang, L., … Lengauer, T. (2016). Big data in natural sciences, humanities and social sciences – Review of the 6th exploratory round table conference. *Bulletin of Chinese Academy of Sciences, 31*(6), 707–716.

Guo, H., Fu, W., Li, X., Pei, C., Liu, G., Zhen, L., … Bai, L. (2014). Research on global change scientific satellites. *Science China Earth Sciences, 57*(2), 204–215.

Guo, H., Liu, L., Lei, L., Wu, Y., Li, L., Zhang, B., … Li, Z. (2010). Dynamic analysis of the Wenchuan Earthquake disaster and reconstruction with 3-year remote sensing data. *International Journal of Digital Earth, 3*(4), 355–364.

Guo, H., Liu, Z., Jiang, H., Wang, C., Liu, J., & Liang, D. (2017). Big earth data: A new challenge and opportunity for digital earth's development. *International Journal of Digital Earth, 10*(1), 1–12.

Guo, H., Wang, L., Chen, F., & Liang, D. (2014). Scientific big data and digital earth. *Chinese Science Bulletin, 59*(35), 5066–5073.

Guo, H., Wang, L., & Liang, D. (2016). Big earth data from space: A new engine for earth science. *Science Bulletin, 61*(7), 505–513.

He, G., Wang, L., Ma, Y., Zhang, Z., Wang, G., Peng, Y., … Zhang, X. (2015). Processing of earth observation big data: Challenges and countermeasures. *Chinese Science Bulletin, 60*(5–6), 470.

Hey, T., Tansley, S., & Tolle, K. (2009). *The forth paradigm: Data intensive scientific discovery*. Washington: Microsoft Research.

Kennedy, M. C., & O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, 63*(3), 425–464.

Li, G., & Cheng, X. (2012). Research status and scientific thinking of big data. *Bulletin of Chinese Academy of Sciences, 27*(6), 647–657.

Nagarajan, M., Gomadam, K., Sheth, A. P., Ranabahu, A., Mutharaju, R., & Jadhav, A. (2009). Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. *Paper presented at the International Conference on Web Information Systems Engineering, 539–554.*
Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data age 2025: The evolution of data to life-critical don't focus on big data*. Framingham: IDC Analyze the Future.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). *Spatiotemporal data mining in the era of big spatial data: algorithms and applications*. Paper presented at the ACM Sigspatial International Workshop on Analytics for Big Geospatial Data.

Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities. (2014). *Big data for international scientific programmes: Challenges and opportunities a statement of recommendations and actions*. Retrieved June 9, 2014 from http://codata.org/blog/wp-content/uploads/2014/06/CODATA-Big-Data-Workshop-STATEMENT-v07-FINAL.pdf

Xu, G., Ge, Q., Gong, P., Fang, X., Cheng, B., He, B., … Bin, X. (2013). Societal response to challenges of global change and human sustainable development. *Chinese Science Bulletin, 58*(25), 3161–3168.