PENG YUE, RAHUL RAMACHANDRAN, PETER BAUMANN,
SIRI JODHA S. KHALSA, MEIXIA DENG, AND LIANGCUN JIANG

Recent Activities in Earth Data Science

Recent trends on big Earth-observing (EO) data lead to some questions that the Earth science community needs to address. Are we experiencing a paradigm shift in Earth science research now? How can we better utilize the explosion of technology maturation to create new forms of EO data processing? Can we summarize the existing methodologies and technologies scaling to big EO data as a new field named *earth data science*? Big data technologies are being widely practiced in Earth sciences and remote sensing communities to support EO data access, processing, and knowledge discovery. The data-intensive scientific discovery, named *the fourth paradigm*, leads to data science in the big data era [1]. According to the definition by the U.S. National Institute of Standards and Technology, the data science paradigm is the "extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing" [2]. Earth data science is the art and science of applying the data science paradigm to EO data.

Over the past decade, the EO data managed and processed by information systems have increased from the terabyte level to the petabyte and exabyte levels. The rapid development of sensor and cyberinfrastructure technologies makes EO data, which are generated by global and local sensor systems and networks measuring the state of Earth, an important part of big data. The data are not only bigger than before but they also have increased complexity due to their very special characteristics of volume, variety, velocity, value, veracity, and variability. The big EO data means that capabilities of traditional data systems and computational methods are inadequate to deal with these characteristics. Today, in addition to the analysis of EO data only, Earth scientists are also using social and economic data to complement EO data to gain a better understanding of the social-economic-environmental systems. Infrastructure-based research is being leveraged to enable fast analysis of the data.

Earth data science will encompass various aspects of big EO data, including big data management (i.e., curation, discovery, and access to EO data); web and cloud-based processing of EO data; methods, tools, and best practices for big data analytics; applications of big EO data; and directions and trends of big EO data science. The time is ripe to rethink these aspects for realizing the potential of EO data and better supporting interdisciplinary research in a collaborative environment. The mission of the IEEE Geoscience and Remote Sensing Society (GRSS) Earth Science Informatics (ESI) Technical Committee (TC) is to "advance the application of informatics to the geosciences and remote sensing, to provide a venue for ESI professionals to exchange information and knowledge, and to give technology advice to major national and international ESI initiatives." Many aspects of Earth data science fall into the scope of the ESI TC. This article intends to identify the significant trends in computing, storage, and modeling infrastructures, data life cycle management, and big data analytics, along with the development of relevant standards that enable Earth data science (Figure 1). It then summarizes some ESI TC activities in the past year related to Earth data science and suggests plans for future engagement.

INFRASTRUCTURE

An infrastructure is an integrated information environment that connects distributed hardware and software resources together. It is built on the results of technological developments and institutional efforts, and it encompasses various components from raw data capture to complex Earth system applications. In Earth information infrastructures, Earth data are being collected by

Digital Object Identifier 10.1109/MGRS.2016.2600528
Date of publication: 16 December 2016

distributed sensors, served by distributed geospatial data services, transformed by processing services and workflows, and consumed by smart clients [3]. From the aspect of organizational practices, significant efforts have been made that provide geospatial data and services around the world, including the Infrastructure for Spatial Information in the European Community [15], the Global Earth Observation System of Systems (GEOSS) [16], the Data Observation Network for Earth [17], the Geoprocessing Web [12], [18], the EarthCube [19], and the EarthServer [22]. For example, the EarthCube initiative is a community-governed effort to develop a cyberinfrastructure platform that supports Earth science data access, analysis, sharing, and visualization.

From the technological perspective, the recent development of information and communication technologies has triggered a paradigm shift in computing, storage, and modeling infrastructures (Figure 2). A variety of data models, computing methods, data storage solutions, and integrated modeling strategies have been developed and applied to Earth sciences. This paradigm shift has changed the way geospatial activities are being conducted. For example, Not Only Structured Query Language (NoSQL) database management systems (DBMSs) have been developed to compete with traditional relational DBMSs in managing geospatial data. In addition, array DBMSs have been developed to efficiently store multidimensional raster data. While traditional data models like vector and raster are designed for data storage, the recent MapReduce paradigm requires us to rethink the data model to better support the high performance computation. The parallelization of geocomputation can be supported by hybrid parallel architectures integrating different parallel programming models, such as the Open Multiprocessing (OpenMP), the Message Passing Interface (MPI), and the general-purpose graphics processing unit (GPGPU), and using the best of multicore hardware, graphics cards, and clusters [24]. The modeling infrastructures have moved the focus from observations to data assimilation and simulation. A set of community practices for modeling frameworks have emerged, such as the Earth System Modeling Framework (ESMF), the Open Modeling Interface (OpenMI), the Common Component Architecture (CCA), and the Model Web [21]. The modeling frameworks are leveraging web technologies for publication, discovery, access, and integration of model components and software.

The development of information infrastructures will contribute significantly to geoscientists' capabilities in answering the most daunting questions and enable complex Earth science applications, such as long-term global climate change studies, higher resolution simulations and forecasting of hurricanes, and early warning of tsunamis. The infrastructures provide rich data sources, high performance computing power, and Earth system models, and they allow collaborative research by experts from multiple scientific domains. They facilitate data utilization, information exchange, and knowledge production, and they help scientists and public users understand fundamental scientific questions and dynamic Earth system and environmental processes.

DATA LIFE CYCLE

Traditionally, Earth data products have been produced in siloed environments (e.g., scientific data centers) with predefined processing steps. Before the products were available for public use, they would go through rigorous internal validation at each step. There is now a rapidly growing number of geospatial tasks that need to be addressed immediately

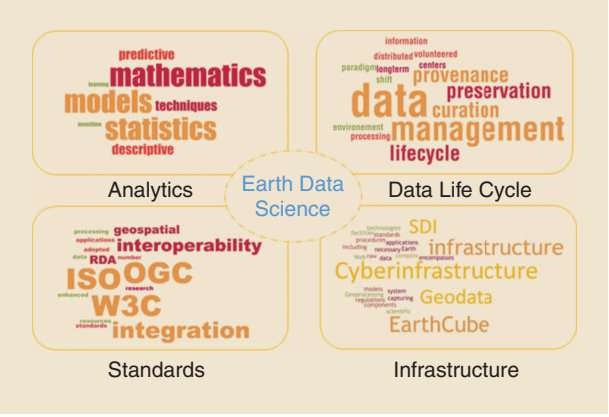

FIGURE 1. The Earth data science aspects.

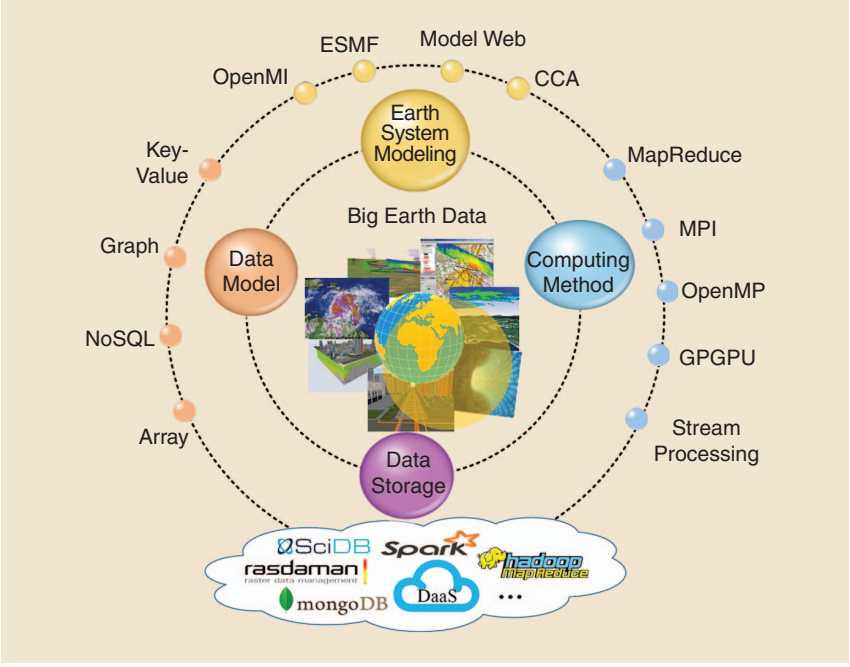

FIGURE 2. A paradigm shift in computing, storage, and modeling infrastructures.

by cooperative efforts from multiple domains and communities, which usually involve a huge volume of EO data from varying sources. These tasks would not be possible using traditional approaches in siloed environments. For example, the Global Forest Watch can look at 143 billion pixels of 30 m each by taking one million central processing unit core hours on 10,000 computers using the Google Earth Engine [25]. As a result, there is an inevitable transition of conducting scientific activities from data centers to the web (i.e., cloud-based computing and web services) and volunteered information and processing environments. Web technologies such as cloud-based computing and web services are used for geospatial data analysis, such as the Google Earth Engine. Some data centers also adopt cloud storage and processing. This transformation brings a paradigm shift in data life cycle management (Figure 3).

The Earth data life cycle refers to several sequential stages involved in the curation and sharing of Earth data for use and reuse. In a traditional context, the data life cycle, which has varying steps according to different organizations, usually includes creation, curation, preservation, transformation, analysis, distribution, access, and provenance. This should be redefined and expanded in a dynamic and distributed environment. Apart from the aforementioned legacy steps, the data life cycle adds data planning, long-term preservation of remote sensed data, and subsetting/disposal. In particular, it highlights the following issues:

- Provenance: In the information infrastructure environment where data are disseminated and processed widely and frequently in an unpredictable way, it is even more important than before to identify original data sources, trace workflows, update or reproduce scientific results, and evaluate the reliability and quality of geospatial data products [4], [23].
- Privacy: Some data and information such as national security-related information are private or sensitive, and they must be handled with caution [5]. The data should be filtered or preprocessed and given restrictive access. There should be a possibility to enforce a fine-grained access control policy to maintain the balance between flexible data access and privacy protection. Tools and procedures are needed to protect confidentiality, intellectual property, and other legitimate requirements for privacy in an Earth data infrastructure.
- Security: Data security, an important component of an information infrastructure, includes data integrity, availability, and veracity. It should be guaranteed in the multiprovider and multitenant environment where Earth data are dynamically created and used and reused.
- Semantics: Semantic annotation of geospatial data across domains is very necessary for improving data interoperability and reusability [20]. Efficient reuse of published data in different discipline contexts needs clear data semantics that may be translated from one community to another in semantic reference systems [6].

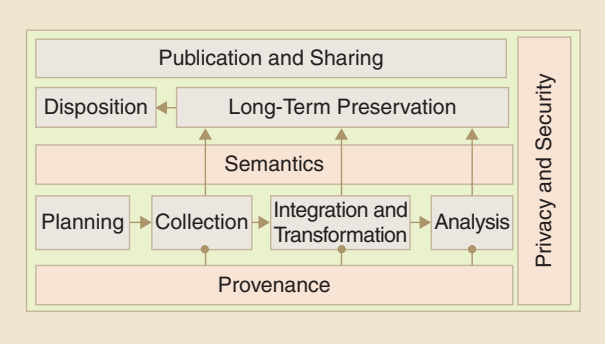

FIGURE 3. A new paradigm in data life cycle.

ANALYTICS

While traditional data analysis is usually designed to solve individual problems, Earth data analytics targets complicated tasks using approaches from informatics like cyberinfrastructure and cloud computing. Earth data analytics relies on simultaneous applications of multiple data sources and computational methods in information infrastructures (Figure 4). Instead of underlining individual analyses or analysis steps (traditionally termed *data analysis*), the term *data analytics* emphasizes the entire methodological process of applying algorithms, methods, technologies, tools, or platforms to transform data into knowledge. Operating on large data sets, Earth science applications are heavily based on data analytics. Emerging advanced analytics techniques are developed including heterogeneous data integration solutions and data-intensive computing methods, and they are employed in EO data management [7], processing [8], [29], and visualization [9].

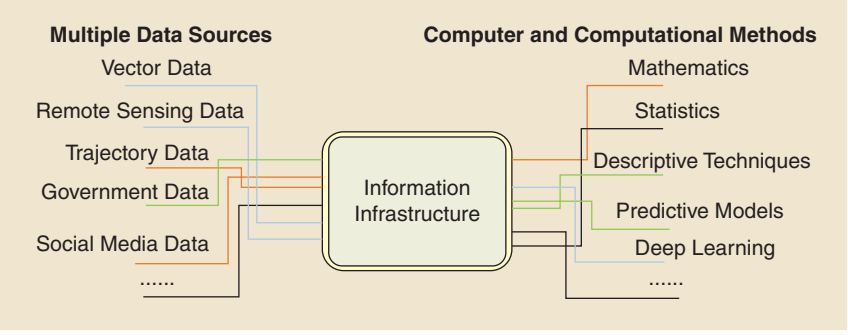

FIGURE 4. Using information infrastructures for big Earth data analytics.

Data and algorithms from a single source are not capable of solving complex problems; therefore, Earth data analytics frequently aggregate diverse data sources and run on parallel computing nodes. Not limited to remote sensing data, data sources also include vector data, global positioning system trajectories, statistical data, and volunteered geographic data. These data sources can complement

each other for enhanced data analysis and scientific discovery. For example, a study on haze detection could use social media data to provide real-time haze information, trigger analysis processing on scientific observation data, or confirm results from scientific analysis [10]. Compared to existing algorithm-based approaches, such data analytics tend to take an infrastructure-based approach, and they run in high performance computing environments to speed up the process. In recent years, an increasing number of Earth data analytic methodologies have been borrowed from transfer learning, mathematics (e.g., fuzzy logic), artificial neural networks, and deep learning [26], [27].

STANDARDS

Interoperability allows proprietary Earth science information systems developed by different vendors to freely exchange information from various sources and work cooperatively to accomplish complex tasks. It has been identified as a fundamental issue when developing Earth science information systems [11] for remote sensing data access, archiving, and analysis. New sources of information for Earth sciences are identified every day. Many of these new sources also bring challenges for making the information suitable for scientific application. New technologies, such as cloud processing, sensor networks, and high velocity big data streaming, bring new opportunities for analysis and understanding of the new data. These technologies can be enhanced through the adoption of geospatial standards thereby providing services for all Earth science information. Standardized open protocols and interfaces allow access to distributed and diverse data and processing functions in a common way [12]. The common data and service environment enabled by standards would greatly lower the complexity of problems caused by the heterogeneity of geospatial data and services [3], [28], [30]. A number of research and applications have adopted geospatial standards for enhanced interoperability and the integration of geospatial data and processing resources.

Earth data science emphasizes a better utilization of existing methodologies and technologies scaling to big EO data. This requires interoperability of data and services more than ever so that efficient information extraction and knowledge discovery from heterogeneous sources can be achieved. This can be achieved through the widespread adoption and implementation of community-developed standards. The standards being developed by the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC), the World Wide Web Consortium, and the IEEE Standards Association (IEEE-SA) are all vital to achieving this goal. The IEEE GRSS ESI TC, along with other organizations such as the Research Data Alliance, play an active role in ISO TC 211, OGC, and IEEE-SA and are coordinating to advance the state of open standards for geosciences and remote sensing. The expertise of the GRSS membership can play an important role in the development and promotion of Earth data standards.

ESI TC ACTIVITIES

The goal of the ESI TC is to advance the research, technology, and applications of informatics to geosciences and remote sensing. A series of productive activities have been conducted in recent years [13], [14]. The ESI TC will continue its role in supporting and promoting the development of Earth data science. The ESI TC sponsored two invited sessions at the IEEE International Geoscience and Remote Sensing Symposium in July 2016 (please see http://www.igarss2016.com/). The first session, titled "Earth Observing Data Science," provided a platform for domain experts from different disciplines to exchange their ideas and report their latest practices on handling EO data. The second session was a joint ESI TC and OGC session titled "Advancing Interoperability for Geoscience Information Systems," which presented the most exciting developments coming from innovative hands-on development processes such as OGC Testbeds and GEOSS Architecture Implementation Pilots. Considering many ESI TC members have rich experiences of applying geoscience and remote sensing technologies to agricultural sustainability, environmental research, and natural resource conservation, the ESI TC also works closely with agriculture sectors, and it is coorganizing the international conference series on Agro-Geoinformatics. For Agro-Geoinformatics 2015, more than one hundred participants joined the conference in Istanbul, Turkey, on 20–24 July 2015. The fifth international conference on Agro-Geoinformatics (Agro-Geoinformatics 2016) was held successfully in Tianjin, China, on 18–20 July 2016 (see http://www.agro-geoinformatics.org/).

The ESI TC also focuses on nurturing young professionals who hold the key to the future development of geoscience and remote sensing. The ESI TC has been supporting the International Geoinformatics Summer School series in the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing at Wuhan University, China. The summer school is held every year to attract young students to advanced studies and training in geoinformatics. It links students, young professionals, and experienced scholars together in lectures, lab exercises, field trips, and social events. In 2015, the ESI TC coorganized successfully the 2015 International Geoinformatics Summer School at Wuhan University (4–14 June 2015). A total of 135 participants from seven countries joined the summer school. In the summer school, the ESI TC developed a course, "Spatial-Temporal Big Data Analytics and Data Science," which provided an introduction to data science and some technologies and approaches for unleashing the power of big data. The ESI TC continued to support the 2016 Geoinformatics Summer School, held at Wuhan University on 19–30 June. (For further information, visit http://www.lmars.whu.edu.cn/geosummerschool/.)

The ESI TC also takes an active part in the geospatial standards development through the ISO TC 211—Geographic Information and the Open Geospatial Consortium. Khalsa and Deng are providing standards expertise

to the International Spaceborne Imaging Spectroscopy TC in OGC and ISO/TC 211/Working Group 6. Baumann is the editor of the OGC Web Coverage Service, and he is the initiator and coeditor of ISO SQL/Multi-Dimensional Arrays. There are ongoing opportunities for GRSS members to contribute their expertise to standards development activities. These include serving as subject matter experts in the review of proposed new standards, lending their opinions and experience to help improve existing standards, or identifying areas of their work that could benefit from standardization.

To learn more or participate in the ESI TC, contact the chairs, Peng Yue, Rahul Ramachandran, and Peter Baumann (pyue@whu.edu.cn, rahul.ramachandran@nasa.gov, p.baumann@jacobs-university.de) and join the IEEE GRSS ESI TC at http://www.grss-ieee.org/join-esi-tc/.

ACKNOWLEDGMENTS

We are grateful to the associate editor and the anonymous reviewers for their constructive comments and suggestions. This work was supported partly by the National Natural Science Foundation of China (91438203 and 41271397), the Hubei Science and Technology Support Program (2014BAA087), and the Program for New Century Excellent Talents in University (NCET-13-0435).

AUTHOR INFORMATION

Peng Yue (pyue@whu.edu.cn) is with the School of Remote Sensing and Information Engineering, Wuhan University, China. He is a Senior Member of the IEEE.

Rahul Ramachandran (rahul.ramachandran@nasa.gov) is with the NASA Marshall Space Flight Center in Huntsville, Alabama. He is a Member of the IEEE.

Peter Baumann (p.baumann@jacobs-university.de) is with the Jacobs University in Bremen, Germany. He is a Member of the IEEE.

Siri Jodha S. Khalsa is with the University of Colorado and the National Snow and Ice Data Center in Boulder, Colorado. He is a Senior Member of the IEEE.

Meixia Deng is with George Mason University in Fairfax, Virginia. She is a Member of the IEEE.

Liangcun Jiang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China.

REFERENCES

[1] T. Hey, S. Tansley, K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1, Redmond, WA: Microsoft Research, 2009.

[2] National Institute of Standards and Technology. (2015, Sept.). NIST big data interoperability framework: Volume 1, Definitions. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[3] P. Yue, P. Baumann, K. Bugbee, and L. Jiang, "Towards intelligent GIServices," *Earth Sci. Informatics*, vol. 8, no. 3, pp. 463–481, Sept. 2015.

[4] L. He, P. Yue, L. Di, M. Zhang, and L. Hu, "Adding geospatial data provenance into SDI-A service-oriented approach," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 2, pp. 926–936, Feb. 2015.

[5] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, "Addressing big data challenges for scientific data infrastructure," in *Proc. 2013 Int. Conf. Collaboration Technologies Systems*, 2013, pp. 48–55.

[6] W. Kuhn, "Semantic reference systems," *Int. J. Geographical Inform. Sci.*, vol. 17, no. 5, pp. 405–409, 2003.

[7] P. Yue, L. C. Jiang, and L. Hu, "Google fusion tables for managing soil moisture sensor observations," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 11, pp. 4414–4421, Nov. 2014.

[8] P. Yue, H. Zhou, J. Gong, and L. Hu, "Geoprocessing in cloud computing platforms: A comparative analysis," *Int. J. Digital Earth*, vol. 6, no. 4, pp. 404–425, 2013.

[9] P. Yue, J. Gong, L. Xiang, and J. Chen, "Analysis-enhanced virtual globe for digital earth," *Sci. China Technol. Sci.*, vol. 53, no. 1, pp. 61–67, May 2010.

[10] P. Yue, C. Zhang, M. Zhang, X. Zhai, and L. Jiang, "An SDI approach for big data analytics: The case on sensor web event detection and geoprocessing workflow," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4720–4728, Oct. 2015.

[11] G. Percivall, "The application of open standards to enhance the interoperability of geoscience information," *Int. J. Digital Earth*, vol. 3, no. S1, pp. 14–30, 2010.

[12] P. Zhao, T. Foerster, and P. Yue, "The geoprocessing web," *Comput. Geosci.*, vol. 47, no. 2012, pp. 3–12, Oct. 2012.

[13] S. J. S. Khalsa and R. Ramachandran, "Earth science informatics comes of age," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 4, pp. 19–21, 2013.

[14] R. Ramachandran and S. J. S. Khalsa, "Moving from data to knowledge: Challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 2, pp. 51–54, 2015.

[15] European Commission. (2016). INSPIRE: Infrastructure for Spatial Information in the European Community. [Online]. Available: http://inspire.jrc.ec.europa.eu

[16] Group on Earth Observations. (2016). GEOSS: Access, Connecting, Users. [Online]. Available: http://www.earthobservations.org/geoss.php

[17] DataONE, Data Observation Network for Earth. (2016). What is DataONE? [Online]. Available: http://www.dataone.org/

[18] P. Yue, J. Gong, L. Di, J. Yuan, L. Sun, Z. Sun, and Q. Wang, "GeoPW: Laying blocks for the geospatial processing web," *Trans. GIS*, vol. 14, no. 6, pp. 755–772. Dec. 2010.

[19] National Science Foundation. (2016). EarthCube. [Online]. Available: http://www.earthcube.org

[20] P. Yue, L. Di, W. Yang, G. Yu, and P. Zhao, "Semantics-based automatic composition of geospatial Web services chains," *Comput. Geosci.*, vol. 33, no. 5, pp. 649–665, 2007.

[21] G.F. Laniak, G. Olchin, J. Goodall, A. Voinov, M. Hill, P. Glynn, G. Whelan, G. Geller, N. Quinn, M. Blind, S. Peackham, S. Reaney, N. Gaber, R. Kennedy, and A. Hughes, "Integrated environmental modeling: A vision and roadmap for the future," *Environ. Modeling Software*, vol. 39, pp. 3–23, Jan. 2013.

[22] P. Baumann, P. Mazzetti, J. Ungar, R. Barbera, D. Barboni, A. Beccati, L. Bigagli, E. Boldrini, R. Bruno, A. Calanducci, P. Campalani, O. Clement, A. Dumitru, M. Grant, P. Herzig, G. Kakaletris, J. Laxton, P. Koltsida, K. Lipskoch, A. R. Mahdiraji, S. Mantovani, V. Merticariu, A. Messina, D. Misev, S. Natali, S. Nativi, J. Oosthoek, J. Passmore, M. Pappalardo, A. P. Rossi, F. Rundo, M. Sen, V. Sorbera, D. Sullivan, M. Torrisi, L. Trovato, M. G. Veratelli, and S. Wagner, "Big data analytics for Earth sciences: The EarthServer approach," *Int. J. Digital Earth*, vol. 9, no. 1, pp. 3–29, 2016.

[23] L. Di, P. Yue, H. K. Ramapriyan, and R. King, "Geoscience data provenance: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 11, pp. 5065–5072, Nov. 2013.

[24] P. Yue and L. Jiang, "BigGIS: How big data can shape next-generation GIS," in Proc. 3rd Int. Conf. Agro-Geoinformatics, Beijing, China, 2014,pp. 1–6.

[25] E. Strickland. (2014, Apr. 16). Google earth engine brings big data to environmental activism. *IEEE Spectrum*. [Online]. Available: http://spectrum.ieee.org/energy/environment/google-earth-engine-brings-big-data-to-environmental-activism

[26] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. (2015). Transfer learning from deep features for remote sensing and poverty mapping. presented at *30th AAAI Conf. Artificial Intelligence*. [Online]. Available: https://arxiv.org/pdf/1510.00098v2.pdf

[27] L. Zhang, G. S. Xia, T. Wu, L. Lin, and X. C. Tai, "Deep learning for remote sensing image understanding," *J. Sensors*, vol. 2015, July 2015.

[28] M. Deng and L. Di, "Building open environments to meet big data challenges in Earth sciences," in *Big Data Techniques and Technologies in Geoinformatics*, H. Karimi, Ed. Boca Raton, FL: CRC Press, 2013, pp. 67–88.

[29] X. Tan, L. Di, M. Deng, A. Chen, F. Huang, C. Peng, M. Gao, Y. Yao, and Z. Sha, . "Cloud-and agent-based geospatial service chain: A case study of submerged crops analysis during flooding of the Yangtze River Basin," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 3, pp. 1359–1370, Mar. 2015.

[30] L. Di, K. Moe, and T. L. van Zyl, , "Earth observation sensor web: An overview" *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 3, no. 4, pp. 415–417, Dec. 2010.

GRS