# Space Science Big Data: Opportunities and Challenges Postprint

**Authors:** Zou Ziming, Hu Xiaoyan, Bear Forest

**Date:** 2023-03-19T00:00:00+00:00

## Abstract

With the growing emphasis on space science worldwide and the flourishing development of space exploration technologies, the era of space science big data has arrived. This article elaborates on the principal trends in the current development of space science big data from multiple perspectives, including data scale expansion, evolution of data management philosophies, paradigm shifts in research, advancement of big data technologies and tools, emergence of intelligent applications, and construction of research ecosystems. Furthermore, by integrating future development requirements with national strategic planning and deployment, it analyzes the specific challenges and opportunities confronting space science big data. The article proposes that comprehensive efforts should be directed toward promoting the sharing and collaborative utilization of scientific data, expanding knowledge innovation and scientific-technological output, thereby ushering in a new era of space science development.

## Full Text

## Preamble

**Subject and Field**: ChinaXiv Partner Journal
**Title**: Challenges and Opportunities of Big Data in Space Science
**Authors**: National Space Science Center, Chinese Academy of Sciences

As countries worldwide increasingly prioritize space science and space exploration technologies flourish, the era of big data in space science has arrived. This article elaborates on the major trends in current space science big data development from the perspectives of data volume growth, evolution of data management concepts, transformation of research paradigms, development of big data technologies and tools, emergence of intelligent applications, and construction of research ecosystems. Combining future development needs with

national strategic planning, it analyzes specific challenges and development opportunities facing space science big data. The article proposes that comprehensive efforts should be made to promote scientific data sharing and utilization, expand knowledge innovation and scientific output, and lead a new era of space science development.

**Keywords**: Space Science, Scientific Big Data, Planning Proposal
**DOI**: 10.16418/j.issn.1000-3045.2018.08.017

Space science encompasses both macroscopic and microscopic fields, focusing on frontier scientific themes such as the origin and evolution of the universe and life, and the relationship between the solar system and humanity. It addresses fundamental questions that have long puzzled mankind, including cosmic evolution, dark matter and dark energy, black holes, gravitational waves, solar activity and space weather, global changes on Earth, and the formation and evolution of extraterrestrial life. The discipline generally involves branches such as space astronomy, solar physics, space physics, planetary science, microgravity science, space Earth science, space fundamental physics, and space life science, representing a highly innovation-oriented and frontier interdisciplinary field [1]. Space science is an experimental science based on data, using satellites and other space vehicles as platforms, combined with ground-based observation stations or large ground-based observation networks, to obtain large amounts of scientific detection and experimental data for analysis, physical modeling, correlation mining, and other data-centric scientific work, aiming to break through humanity's understanding of nature, promote fundamental scientific progress, drive disruptive technological innovation, lead strategic emerging industries, and bring strong scientific support to national security.

World powers attach great importance to and strongly support space science, designing and planning a series of space science development programs, such as the United States' *National Space Weather Strategy* [2] and *Global Exploration Roadmap* [3], Europe's *Cosmic Vision 2015—2025* [4], Russia's *Strategy for the Development of Space Activities Until 2030* [5], and China's *Research Report on the Medium- and Long-Term Development Plan for Space Science Projects (2010—2030)* . Under the guidance of these strategic plans, major space science missions such as large scientific facilities, flagship programs, special projects, and international cooperation programs have been effectively advanced and implemented, leading to rapid accumulation of scientific data and continuous emergence of scientific achievements.

*Corresponding author*

**Funding**: Chinese Academy of Sciences Strategic Priority Research Program (XDA19020500)
**Received**: August 7, 2018

Compiled by the Chinese Academy of Sciences Space Science Project Medium- and Long-Term Development Planning Research Group in 2008.

## Space Science Enters the Big Data Era

### Explosive Growth of Data Volume

Driven by space science development planning, space science exploration is forming a new pattern of multi-band, multi-messenger, networked, and integrated space-ground joint detection. On the ground-based side, the East Hemisphere Space Environment Ground-based Comprehensive Monitoring Meridian Chain (Meridian Project) Phase I has been completed and put into operation [6], and Phase II of the Meridian Project has also been launched [7]. Based on the Meridian Project, the International Meridian Circle Program for multi-station, networked, and interdisciplinary collaborative monitoring is also being planned [8]. On the space-based side, the Chinese Academy of Sciences Strategic Priority Research Program (Category A) "Space Science" (Space Science Strategic Priority Program) has successfully launched four scientific satellites: "Wukong" (DAMPE), "SJ-10," "Mozi" (QUESS), and "Insight" (HXMT), with remarkable achievements [9-13]. Phase II of the Space Science Strategic Priority Program (2020—2025) is expected to launch the GECAM, SVOM, SMILE, ASO-S, EP, and EXTP series of scientific satellites.

The direct result of the successful implementation of these major projects is the rapid accumulation of scientific data. Relying on new observation technologies such as large field-of-view, high resolution, and high sensitivity, the data acquisition rate in space science is growing exponentially. The four satellites of Phase I of the Space Science Strategic Priority Program have accumulated more than 200 TB of scientific data, with over 2,000 data types and nearly 2 million data products. It is estimated that by the end of the mission, the total data volume will exceed 420 TB. Compared with Phase I, the effective payload data volume of Phase II satellite missions will increase by multiples, with total daily data production exceeding 8,000 GB, reaching a total data volume of 52.2 PB by the end of Phase II missions.

According to the NASA space science satellite catalog , since 2000, 674 space science satellites have been launched worldwide, with an annual average of more than 35 satellites. In 2016 and 2017, the number doubled year-over-year (Figure 2 [Figure 2: see original paper]). Combining observation data from space- and ground-based platforms both domestically and internationally, the annual data production rate in space science exceeds the EB level. From the perspective of data scale and volume, space science has ushered in the era of big data.

### Data Management and Preservation Receive Attention

The massive scientific data generated by major project plans, as national resources and a repository of human knowledge, should be preserved and managed long-term so that scientists can deeply mine the knowledge behind the data for a long time to come. Both domestic and international entities have established data centers/systems in the field of space science to implement long-term preservation and quality control of data.

The National Aeronautics and Space Administration (NASA) established the National Space Science Data Center (NSSDC) to ensure the permanent safety and long-term availability of archival data from space science satellite missions (including space astronomy, astrophysics, solar and space plasma physics, planetary science, lunar science, and space physics), and to provide open data services to scientists worldwide. As of December 2015, its Planetary Data System (PDS) alone stored and managed more than 947 TB of orbital detection data at various levels for Mars, the Moon, Venus, and Mercury.

The European Space Astronomy Centre (ESAC), as the European Space Agency (ESA) space science data center, also centrally manages and stores archival data from all European space astronomy, solar system exploration, planetary science, and fundamental physics satellite missions.

With support from the Space Science Strategic Priority Program, China has also built a space science data management system (Space Science Data Center) that integrates scientific satellite data collection and distribution, full lifecycle data quality control, and data storage, management, and archiving, effectively supporting the implementation of the "Wukong," "SJ-10," "Mozi," and "Insight" satellite missions, promoting satellite output, and ensuring permanent data safety.

### Research Paradigm Transformation

A typical characteristic of the scientific big data era is the transformation of research paradigms [14]. Unlike traditional research models based on theoretical analysis of small data samples, researchers in the big data era mainly analyze multi-source, multi-element, full-sample space big data, often combined with big data technologies such as neural networks and machine learning, to mine scientific knowledge contained in the data. The research model in the space science field is also shifting toward data-intensive scientific discovery.

As a paradigm of data-driven knowledge discovery, the Dark Matter Particle Explorer (DAMPE, "Wukong") scientific team, through analysis of 2.8 billion high-energy cosmic ray data samples collected over 530 days, found for the first time a break (anomalous fluctuation) in the electron cosmic ray energy spectrum at ~1 TeV, which reflects the typical acceleration capability of high-energy electron radiation sources. The declining behavior at the break plays a key role in explaining whether high-energy electrons originate from dark matter [9].

For the massive data obtained by the Kepler space telescope, NASA scientists used deep learning algorithms to build a machine learning model capable of automatically identifying exoplanets from low signal-to-noise ratio data. The model automatically searched 200,000 target galaxy data in the Kepler database and successfully found two exoplanets, Kepler-80 g and Kepler-90 i [15].

As early as the 1990s, space physics research began using machine learning and other big data technologies to analyze satellite data for space weather research and forecasting [16], such as magnetospheric substorm onset identification [17], solar activity (coronal mass ejections, flares) prediction [18,19], and interplanetary shock forecasting [20]. Among them, solar flare prediction [19] used more than 5.5 TB of solar photosphere and chromosphere image big data from the SDO satellite over four years as training input for the model. Facts have proven that big data analysis technology has important practical significance for research on nonlinear space weather processes and forecasting of highly complex space weather events, and the data-intensive research model is gradually becoming the mainstream model in space physics.

### Big Data Technologies and Tools Flourish

Big data analysis technologies, tools, algorithms, and big data system platforms are key links in big data research and applications, and all have made considerable progress.

**(1) Mature and Available Algorithms.** In addition to the big data technologies and algorithms mentioned above, the academic community has developed many other big data analysis and mining algorithms, such as decision trees, Bayesian methods, neural networks, support vector machines and other classification algorithms; K-means clustering, hierarchical clustering, density-based or network-based clustering algorithms; linear regression, logistic regression and other prediction algorithms; and convolutional neural networks and other machine learning models suitable for image processing. These algorithms have demonstrated enormous potential in fields such as business, urban governance, biology, medicine, and geology, and can also be applied to space science research.

**(2) Tool Suites and Convenience.** The American company Analytical Graphics, Inc. (AGI) developed the “Systems Tool Kit” (STK), which provides engineers and scientists with four-dimensional modeling, simulation, analysis, and operations capabilities. It can perform complex analysis of objects on the ground, in the ocean, in the air, and in space, providing simulation calculations and real-time performance evaluation for systems or payloads operating in these environments (Figure 3 [Figure 3: see original paper]). The Space Environment Information System (SPENVIS) developed by the Belgian Federal Science Policy Office is a comprehensive software for evaluating interactions between the space environment and payloads/astronauts, integrating rich model calculations, simulation models, and supporting data analysis tools. Users can access SPENVIS via the Web to construct models, define parameters, and execute simulations to develop, optimize, track, and detect the performance, working status, and faults of spacecraft and detectors, and to predict the space environment and its effects.

**(3) Integrated and Converged Service Systems.** NASA’s Goddard Space Flight Center (GSFC) established the Coordinated Data Analysis Web (CDA Web), which 汇集了 nearly all detection data from international satellite missions since 1992, as well as some ground-based network observation data, truly achieving one-stop scientific data query, physical quantity extraction, visualization plotting, downloading, and online analysis functions. It is the most widely used data system in the world for space science. NASA's Community Coordinated Modeling Center (CCMC) integrates a large number of computable models, such as atmospheric wind field models, ionospheric electron density models, solar and interplanetary space models, and planetary (Venus, Mars, Jupiter) models. Existing models basically cover the entire pattern requirements of the space physics field, providing users with online model calculation and visual analysis to support space physics research and space weather forecasting. The Solar-Terrestrial space Research Network (STAR-Network) [21] is a domestic space science data application platform integrating IT infrastructure, multi-element space science data, data analysis tool suites, and space weather models. It possesses capabilities to support space science missions and scientific innovation activities.

### Trends in Big Data Intelligent Applications

New scientific concepts oriented toward intelligent applications are emerging. For example, Academician Wei Fengsi proposed the "Digital Space" strategy [22], driven by space science space- and ground-based observation data, based on scientific cognition, and using cloud computing infrastructure and space big data application technology as means. This strategy aims to create a major space infrastructure integrating space science, space technology, space applications, and space services, digitally presenting spatiotemporal element changes in real cosmic space, opening up responses to space weather disasters, enhancing satellite application capabilities, and serving the development of strategic new economic fields such as new space energy, new communications, new transportation, new manufacturing, and new environmental protection. Driven by new scientific concepts, technologies such as high-dimensional spatiotemporal discrete technology [23] and intelligent fusion data mining and application technology [24] have also made important progress.

### Formation of a Space Science Big Data Research and Application Ecosystem

Under the above development trends of space science big data, a big data research and application ecosystem in the space science field is gradually taking shape. Open source communities are blossoming everywhere. For example, scientists in the space astronomy field have organized open source community systems such as Astropy and Openastronomy, and have also formed numerous exchange groups such as Astronomical Circle, AstroIDL, and AstroTex. Academic exchange activities are thriving, with specialized space science big data seminars such as BiDS (Big Data for Space), as well as sessions on data science and other directions at top international academic conferences such as the American Geophysical Union (AGU) meeting. Since 2014, China has organized the Scientific Data Conference, which also provides opportunities for exchange, demonstration, and discussion for big data researchers in various domestic disciplines. The space science big data research network has begun to take shape, with teams from relevant research institutions and universities at home and abroad carrying out extensive work on research hotspots such as data open sharing, data management and preservation, tool development, intelligent applications, and infrastructure construction, and actively responding to initiatives such as scientific data publishing, jointly striving to create a good disciplinary big data ecosystem.

## Challenges and Development Opportunities of Space Science Big Data

Space science big data is showing a thriving development trend, achieving a series of considerable research and construction results, bringing new opportunities for space science development. At the same time, we also need to recognize that due to the lack of top-level planning for domain big data development, insufficient awareness and confidence in big data development within the disciplinary community, and other reasons, the development of space science big data in China also faces many bottlenecks and challenges. Overall, it presents a situation where challenges and opportunities coexist, which will be elaborated from five aspects below.

### Challenges and Opportunities in Data Open Sharing

The flow of domestic space science data between institutions and projects is insufficient, with widespread problems such as inconsistent data standards and uneven data quality, and a lack of good data sharing mechanisms and service platforms, hindering the further development of scientific innovation. Taking advantage of the favorable timing of the state's issuance of the *Administrative Measures for Scientific Data*, we should focus on enhancing the concept of scientific data open sharing, and from the perspectives of data openness, standardization, and security, research appropriate space science data standards and sharing norms to achieve integration and open exchange of space science data and promote the full sharing and use of high-value scientific data.

### Challenges and Opportunities in Big Data Infrastructure

Compared with disciplines such as high-energy physics and biomedicine, the development of big data infrastructure in the domestic space science field, including transmission networks, computing resources, application software, and algorithm tools, is relatively lagging, without long-term planning, lacking top-level design, and with relatively weak and scattered basic capabilities. This results in insufficient national support for the space science field. For example, there are difficulties in deploying specific professional models and algorithms to public supercomputing resources, and transmission network issues (speed, restrictions) limit scientific data access and exchange.

Therefore, attention should be paid to long-term planning of big data infrastructure, strengthening top-level design, rational layout and construction, and targeted enhancement of space science big data infrastructure capabilities. For example, by building necessary international data transmission networks and creating public computing environments oriented toward professional needs in space science, we can match and alleviate the enormous pressure from future major space science missions and innovative research activities. At the same time, through policy guidance and increased virtualization facilities and enhanced virtualization resource scheduling capabilities, as many big data infrastructures as possible can achieve joint sharing.

### Challenges and Opportunities in Long-Term Data Security

Currently, domestic space science data resources are mainly preserved by individual research institutions or project departments. In daily management, more attention is paid to the convenience of data use, while the ability to ensure long-term data security is relatively insufficient. We should fully integrate data activity process domain management with major missions to achieve standardized management of the entire data activity lifecycle. Through comprehensive quality control schemes, we can ensure the scientificity and reliability of data. On this basis, we should make full use of advanced scientific data management and disaster backup technologies to ensure the permanent safety and long-term availability of scientific data.

### Urgent Need for Breakthroughs in Disruptive Domain Big Data Management, Analysis, and Application Technologies

Data-intensive knowledge discovery and intelligent fusion application platform construction have urgent needs for domain big data management, analysis, and application technologies. However, the current situation is that traditional scientific research groups lack attention to how big data can boost scientific output, while investment in domain-specific algorithms and software is relatively small, and independent research and development capabilities are relatively weak, leading to insufficient accumulation of disruptive new technologies. We should increase investment in big data technology research and development and application in major scientific missions and large scientific facilities, and learn from development action plans such as the U.S. EarthCube and ROSES to establish a space science big data technology research fund in China to support the long-term development of advanced space science big data technologies.

### New Opportunities for Research and Application Ecosystem Construction

In view of this, we should actively respond to the national call, seize the good momentum of space science development, make good use of the key opportunities of major space science project implementation, establish a national-level space science data center, form an innovative interdisciplinary talent team, and jointly develop a space science big data open source community, making these communities innovation workshops that continuously provide new ideas for space science big data development. We should develop domain-specific algorithms and application tools to compensate for the insufficient independent innovation capabilities of China's space science big data, and promote the benign development of the space science big data research and application ecosystem.

Benefiting from the high attention of the Party and the state, China's space science industry is developing rapidly, with reasonable long-term development planning and layout, and space science has ushered in the era of big data. Under the big data trend, space science development presents a brand-new situation. We should seize the trends and opportunities of space science big data development, and make comprehensive efforts in all aspects including data open sharing, data storage management, big data infrastructure layout, breakthroughs in big data key technologies, and research and application ecosystem construction, to push China's space science, space technology, and space applications to new heights.

## References

1. Wu Ji. Research Report on Space Science Planning for 2016—2030. Beijing: Science Press, 2016.
2. Space Weather Operations, Research, and Mitigation (SWORM) Task Force Co-Chairs. National Space Weather Strategy. [2018-08-04]. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/final_{{{nationalspaceweathe
3. International Space Exploration Coordination Group. Global exploration roadmap. [2017-01-05]. https://www.nasa.gov/sites/default/files/files/GER-2013_{Small}.pdf.
4. European Space Agency. Space science for Europe 2015-2025. [2017-01-05]. http://www.esa.int/esapub/br/br247/br247.pdf.
5. You Liang, Bai Qingjiang, Sun Lilin, et al. Overview of space science development trends in major space countries worldwide. Bulletin of Chinese Academy of Sciences, 2015, 30(6): 740-750.
6. Wang C. New chains of space weather monitoring stations in China. Bulletin of the Chinese Academy of Sciences, 2013, 27(1): 33-36.
7. Wang Chi, Ren Liwen. Journey of solar-terrestrial space exploration—Latest progress and prospects of space physics detection (Part 2). International Space, 2015(02): 82-86.
8. Fan Quanlin. International Space Weather Meridian Circle Program based on the Meridian Project. China Science Fund, 2008(2): 65-69.
9. Ambrosi G, An Q, Asfandiyarov R, et al. Direct detection of a break in the teraelectronvolt cosmic-ray spectrum of electrons and positrons. Nature, 2017, 552(7683): 63-66.
10. Liao S K, Yong H L, Liu C, et al. Long-distance free-space quantum key

distribution in daylight towards inter-satellite communication. Nature Photonics, 2017, 11(8): 509-518.
11. Yin J, Cao Y, Li Y H, et al. Satellite-based entanglement distribution over 1200 kilometers. Science, 2017, 356(6343): 1140-1144.
12. Li T P, Xiong S L, Zhang S N, et al. Insight-HXMT observations of the first binary neutron star merger GW170817. Science China (Physics, Mechanics & Astronomy), 2018, 61(3): 031011.
13. Hey T. The fourth paradigm-data-intensive scientific discovery. Proceedings of the IEEE, 2011, 99(8): 1334-1337.
14. Shallue C J, Vanderburg A. Identifying exoplanets with deep learning: a five planet resonant chain around Kepler-80 and an eighth planet around Kepler-90. Astronomical Journal, 2017, 155(2): 94.
15. Pulkkinen A, Bernabeu E, Eichner J, et al. Scientific data products and dissemination: the NASA space physics data facility. In: Pulkkinen A, Taktakishvili A, Borth E, eds. Geomagnetically induced currents from the Sun to the power grid. Washington: American Geophysical Union, 2019.
16. Sutcliffe P R. Substorm onset identification using neural networks and Pi2 pulsations. Annales Geophysicae, 1997, 15(10): 1257-1263.
17. Nishizuka N, Sugiura K, Kubo Y, et al. Solar flare prediction using machine learning methods. Astrophysical Journal, 2016, 821(2): 110.
18. Jonas E, Bobra M, Shankar V, et al. Flare prediction using photospheric and coronal image data. Solar Physics, 2018, 293(3): 32.
19. Vandegriff J, Wagstaff K, Ho G, et al. Forecasting space weather: Predicting interplanetary shocks using neural networks. Advances in Space Research, 2005, 36(12): 2323-2327.
20. Zou Ziming, Tong Jizhou, Xiong Senlin, et al. Practice and achievements of scientific informatization in space science in the era of big data. Big Data, 2016, 2(6): 83-96.
21. Wei Fengsi. “Digital Space” is a new strategic highland for space science and technology. Henan Science and Technology, 2016, (19): 7-7.
22. Xiong Senlin, Zou Ziming, Tong Jizhou, et al. High-dimensional spatiotemporal discrete technology for space science big data. Journal of Earth Information Science, 2017, 19(6): 1-9.
23. Tong Jizhou, Zou Ziming, Xiong Senlin, et al. Intelligent fusion data mining and application technology for space science big data. Journal of Earth Information Science, 2018, 20(6): 728-735.

## Author Biographies

**ZOU Ziming** (Corresponding author)
Deputy Director of National Space Science Center, Chinese Academy of Sciences (CAS), Chief Designer of the Ground Support System in CAS Strategic Priority Program on Space Science, Director of Chinese Space Science Data Center in World Data System (WDS), Member of Information Technology Committee in Chinese Geophysical Society, and Member of Science and Technology Committee in CAS Space Environment Prediction Center. His research focuses on

space environment information system integration, space science satellite data processing, space information organization, retrieval, interoperability, and solar-terrestrial information representation and visual analysis. He has published over 30 papers and co-authored one monograph. He has received two First Prizes of the Military Science and Technology Progress Award, the CAS Manned Space Program Important Contribution Award, and the honorary title of "Excellent Worker in CAS Manned Space Program." E-mail: mzou@nssc.ac.cn

*Note: Figure translations are in progress. See original paper for figures.*

*Source: ChinaXiv — Machine translation. Verify with original.*