Big Data Promotes the Tibetan Plateau and Pan-Third Pole Earth System Science

Xin Li, Xiaoduo Pan, Xuejun Guo, Jun Qin, Baosheng An, Tao Wang, Qinghua Ye, Weimin Wang, Xiaojuan Yang, Xiaolei Niu, Min Feng, Tao Che, Rui Jin, and Jianwen Guo

Abstract The Pan-Third Pole includes the Tibetan Plateau and the northern intra-continental arid region of Asia, extending to the Caucasus Mountains in the west and the western Loess Plateau in the east. This region covers 20 million square kilometers and affects the environment inhabited by three billion people. Two special projects have been implemented to provide important scientific support for eco-environmental refining and sustainable economic and social development of the Pan-Third Pole region, with the Tibetan Plateau as its core: the Second Tibetan Plateau Scientific Expedition Program (a national special project) and the Pan-Third Pole Environmental Change and Construction of the Green Silk Road (hereinafter referred to as the Silk Road and Environment), a strategic pilot science and technology project (Category A) of the Chinese Academy of Science. The Pan-Third Pole big data system is an important data support platform for these two major research programs and has several purposes: the storage, management, analysis, mining and sharing of scientific data for various disciplines, such as resources, the environment, ecology and atmospheric science of the Pan-Third Pole; preparation of key scientific data products of the Pan-Third Pole; the gradual development of functions such as online big data analysis and model application; the construction of a cloud-based platform to integrate data, methods, models; and services for Pan-Third Pole research and promote application of big data technology in scientific research in the region. This

X. Li (✉) · X. Pan · X. Guo · J. Qin · X. Yang · X. Niu · M. Feng
National Tibetan Plateau Data Center, Key Laboratory of Tibetan Environment Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China
e-mail: xinli@itpcas.ac.cn

X. Li · B. An · T. Wang · T. Che · R. Jin
CAS Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, Beijing, China

B. An · T. Wang · Q. Ye · W. Wang
Institute of Tibetan Plateau Research, Chinese Academy of Science, Beijing, China

T. Che · R. Jin · J. Guo
Heihe Remote Sensing Experimental Research Station, Key Laboratory of Remote Sensing of Gansu Province,Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, Gansu, China

China's e-Science Blue Book 2020,
https://doi.org/10.1007/978-981-15-8342-1_7

paper demonstrates in detail various aspects of the Pan-Third Pole big data system, including the system architecture, data resource integration and big data analysis methods. The system improves big data processing capability in geoscience, serves as a new paradigm of geoscience research driven by big data and facilitates scientific research of the Earth system of the Pan-Third Pole.

Keywords Tibetan Plateau · Pan-Third Pole · Data integration · Big data analysis

1 Introduction

The Tibetan Plateau, with the highest elevation in the world, is known as the Third Pole of the Earth. Multiple major rivers in Asia originate from the Third Pole, which is also referred to as Asia's water tower. The Third Pole is considered one of the regions that have been subjected to the highest warming and environmental variation over the last 50 years. This region is a typical region interacting with the global climate system at multiple levels, and it is also a key region sensitive to global climate and environmental change. It plays a vital role in the global circulation of energy and water and has important influences on the global and regional climate [1, 2]. The Third Pole features a unique and yet vulnerable eco-environment, and the inhabitant species and ecosystem are highly sensitive to climate change. Consequently, the Second Tibetan Plateau Scientific Expedition Program (hereinafter shorted as STEP-2) was initiated in 2017. This program investigates variations in glaciers, the environment, lake and hydro meteorology, changes in biology and the ecosystem, the paleoecosystem and the paleoenvironment. The program probes the regularity of these variations, forecasts change scenarios and proposes countermeasure strategies while satisfying requirements for societal development in Tibet and the Belt and Road Initiative. This comprehensive scientific expedition is driven by dual considerations, the needs of the regional development and the demands of the scientific research frontier. Development in the Belt and Road Initiative has resulted in environmental changes that have attracted worldwide attention [3]. This 20-million-square-kilometer fan-shaped region has an extremely vulnerable ecological environment and intensive human activity. Moving westward, with the Third Pole as its center, this region covers the Tibetan Plateau, the Pamirs, the Hindu Kush, the Tianshan Mountains, the Iranian Plateau, the Caucasus and the Carpathian Mountains. Sustaining the resources of the Pan-Third Pole will provide critical scientific and technical support for the Belt and Road Initiative [4, 5].

A big data environmental system is being constructed for the Pan-Third Pole to solve regional environmental problems and enhance scientific research, thereby promoting the Belt and Road Initiative. This system performs strategically necessary functions for regional development: comprehensive summarization, long-term preservation, integrated management and sharing of national-level important scientific and technical data resources, and solution of major issues in economic & social development and national security.

The existing data of Pan third polar cryosphere, lake, ecology, hydrology, atmosphere, solid earth, etc. are characterized by massive but fragmentary, scattered and incomplete space–time coverage. On the one hand, a lot of resources are wasted, and different researchers need to repeatedly collect and preprocess the data when conducting research. On the other hand, they cannot fully and accurately be used to understand the integration of Pan third polar data Physical condition, did not play out the great potential. Until now, there is no scientific research big data platform for the Pan-Third Pole in either China or other countries. Such a platform for the Pan-Third Pole environment is urgently required for in-depth integration and rapid sharing of data to increase data utilization and scientific research efficiency. We develop a new big data platform and corresponding big data analysis methods based the development concept and experience of scientific resource integration and sharing of geoscience-related scientific data centers around the world [6–11]. A novel research paradigm is provided for scientific research of the Pan-Third Pole, and an example of big data-based innovation in geoscience is demonstrated. The platform can serve as a data support platform for STEP-2 and the Pan-Third Pole Environmental Change and Construction of the Green Silk Road project (a stragetic pilot project of the Chinese Academy of Science (category A), hereinafter referred to as the Silk Road and Environment), and the teleconnection analysis of global cryosphere as well.

2 Architecture of Pan-Third Pole Big Data System

A big data platform for geoscience research of the Pan-Third Pole has been constructed by integrating multisource heterogeneous data and combining big data mining and geoscience models. The platform is mainly used in five research fields: glaciers, lakes, ecological simulations, seismic monitoring and sustainable development. Efforts are made to explore and summarize big data driven new approaches for geoscience studies (Fig. 1).

2.1 Data Management and Sharing Platform

The big data system stores, manages, analyzes, mines and releases scientific data on the resources, environment, ecology, atmosphere, etc., of the Pan-Third Pole. Functions such as the online big data analysis and model application are gradually developed and enabled to integrate extensive data, methods, models and services for Pan-Third Pole science (https://data.tpdc.ac.cn/en/) (Fig. 2).

This system incorporates the submission, review and publishing of data resources in various forms, such as data uploaded by noninstitutional users, data from collaborative programs and institutions, journal data and platform data mining products. The system allows for dataset storage, quality control and the sharing of observation

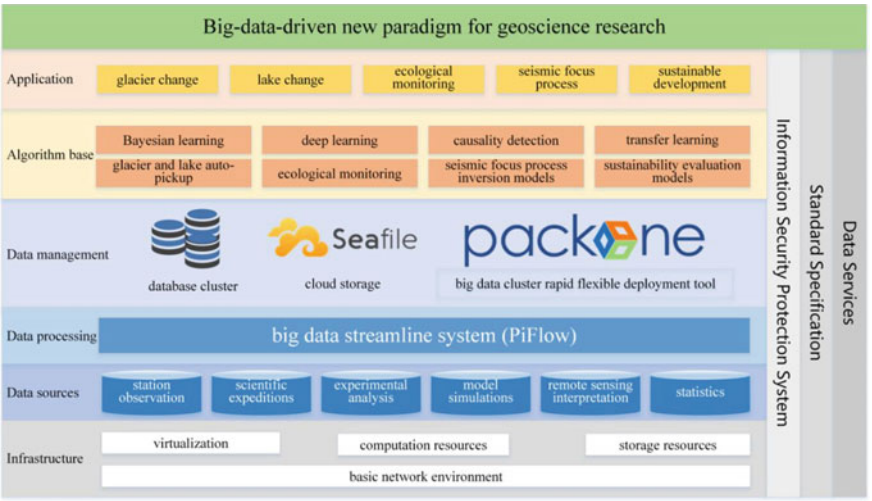

Fig. 1 Overall system architecture

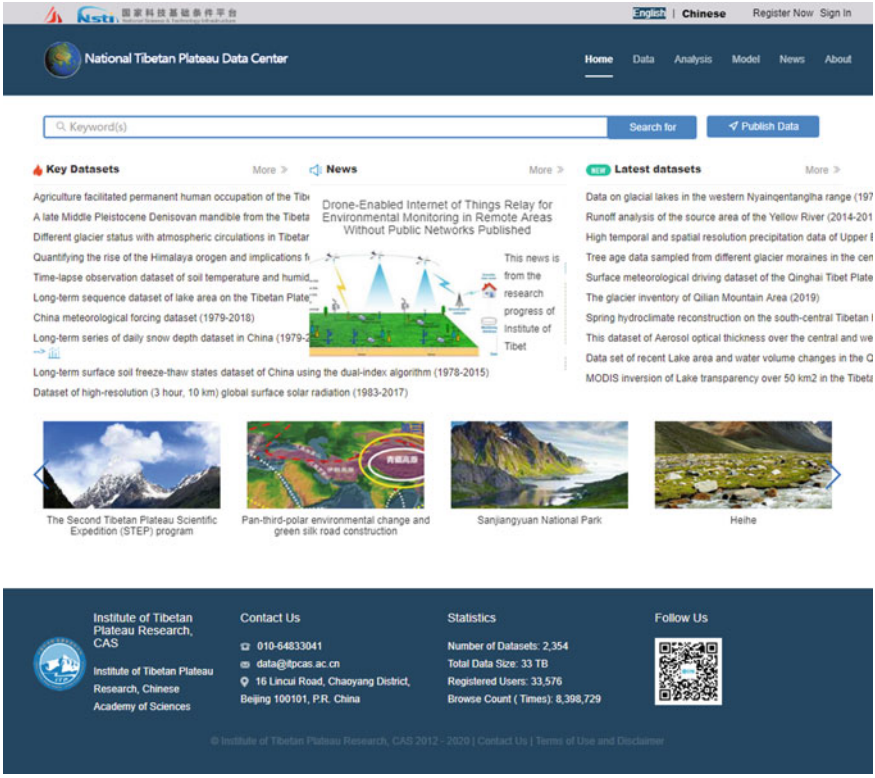

Fig. 2 User interface of Pan-Third Pole big data system

and measurement data from field stations, hierarchical data sharing and online data analysis, mining and visualization.

2.2 Construction of Information Infrastructure

Virtualization technology is used to realize dynamic allocation, flexible distribution and interterritory sharing of hardware resources, which increases the utilization efficiency of IT resources considerably. The Internet, storage and computation resources provided by China Science and Technology Cloud are fully exploited to develop applications such as cloud storage and online archives to facilitate the accumulation and sharing of scientific data. A PB-level data-intensive data storage and processing facility is constructed and embedded with big data processing software and tools (Ambari, PackOne, PiFlow, etc.). An environment is created to demonstrate big data applications for the cryosphere, hydrology and ecology, solid Earth science and regional sustainable development (Fig. 1).

2.3 Standardized System Regulation

The National Tibet Plateau Data Center (hereinafter referred to as the Data Center) is the interface of Pan-Third Pole big data system. The objectives of the Data Center are as follows: to fully implement the *Notice of the General Office of the State Council (of China) on Regulations of Scientific Data Managemen*t (GBF (2018) No. 17) and the "*Notice of the Chinese Academy of Science on Printing and Distributing "Regulations of Scientific Data Management and Public Sharing of the Chinese Academy of Science (Tentative)*" (KFB (2019) No. 11); to further strengthen and regularize scientific data management; to ensure scientific data security; to increase public sharing; and to sustain management of the Data Center. To fulfill these objectives, the Data Center has developed standardized specifications for administration and operation, data management, metadata standards, scientific data management, data-sharing services and IT and data security. Each process for data acquisition, integration and submission, storage, sharing and publishing has been regularized and standardized to ensure data security and protect the rights and interests of the data producer (Fig. 3).

A *Discipline-based Content Standard of the National Tibet Plateau Data Center* is under development. This standard is based on 102 national and sector standards and 35 research papers from 15 disciplines, including glaciers and permafrost, the paleoenvironment, geology, hydrology, soil, the atmosphere, atmospheric chemistry, ecology, remote sensing, bio-diversity and disasters.

Taking the discipline term of glacier as an example, the data content standard includes the following 10 sub-discipline: material balances, glacier terminus changes, a glacier inventory, black carbon content, isotopic oxygen and aerosol concentrations,

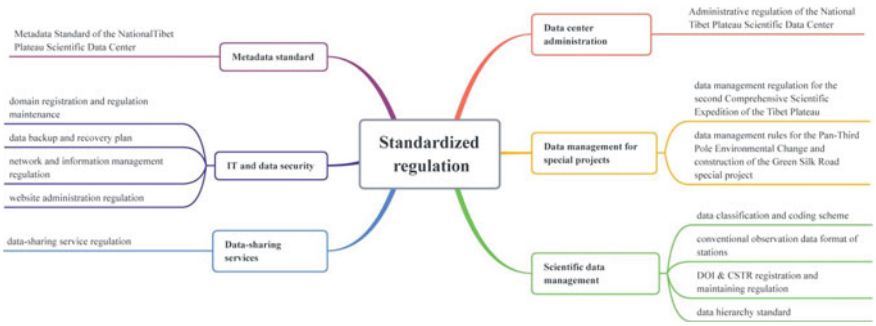

Fig. 3 Standardized system regulation of the National Tibet Plateau Scientific Data Center

glacier surface movement, glacier temperatures, glacier thicknesses, glacier meteorology and glacier streamflow. Each subclass has several indicators. The content and value standard of each indicator are described in detail by Chinese and English names and units, precision, granularity and data value ranges. Specific provisions in these standards provide thorough references for the Data Center for collection, processing, quality assessment and calibration of glacier data.

3 Pan-Third Pole Data Resources

3.1 Keywords of the Data Classification System for the Pan-Third Pole Scientific Data

A new list of keywords is created from the "Keywords of the Data Classification System for the Pan-Third Pole Scientific Data", the Third Pole data catalog and the "First-Level Discipline Orientation Classification and Keywords of Geoscience (Tentative Version, 2012)" of the National Natural Science Foundation of China. This list includes 11 1st-level classes, 62 2nd-level classes and 702 keywords. The 1st-level class contains the cryosphere, hydrology, soil science, the atmosphere, the biosphere, geology, paleoclimate/paleoenvironment, human dimensions/natural resources, disaster, remote sensing and basic geography.

3.2 Combination of Pan-Third Pole Data

The existing 148 scientific datasets of the Data Center are combined into over 2300 scientific datasets: STEP-2, the satellite-aerial-remote sensing-ground station integrated ecological monitoring and data platform of the Three-River-Source National Park, the Cold and Arid Region Scientific Data Center [5] and extensive research

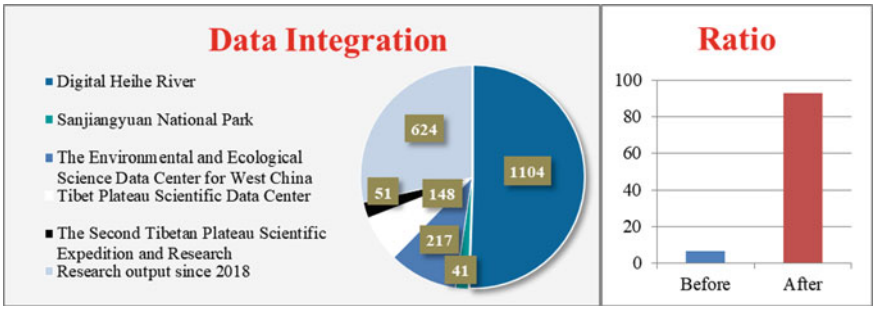

Fig. 4 Sources and quantities of integrated data resources for Pan-Third Pole

results (Fig. 4). The total data volume reaches 33 TB, and the relational data contain over 1.4 billion rows.

3.3 Integration of Pan-Third Pole Data

The metadata of the Tibetan Plateau are managed in accordance with ISO 19115. The cloud-based Pan-Third Pole environment big data system has been constructed to achieve multilevel, multiterminal and multilingual cloud sharing. The datasets originate from various platforms and are often multisource and heterogeneous. Thus, there is an urgent need both for data integration, to eliminate information redundancy among multisource data and to identify interactions between various elements in different layers of the Pan-Third Pole area affected by global warming, and for teleconnections with other areas. The overarching principles of data integration are to input heterogeneous spatial data of various classes into the dataset in a unified manner, implement strict quality control, provide complete metadata and data documents and achieve data sharing. The time sensitivity, completeness, principles and logicality of data are evaluated. In addition to conventional quality control methods such as the manual method, the metadata method and the geographic correlation method, considerable effort has been expended to introduce big data cleansing (quality control). Appropriate techniques such as mathematical statistics, data mining and predefined cleansing criteria are used to clean dirty data. The resulting data are transformed into high-quality data by computer-based automatic extract, transform and load (ETL) processing. Hence, the reliability (accuracy, integrity, consistency, effectiveness and uniqueness) and availability (timeliness, accessibility and satisfaction degree) of the Pan-Third Pole data resources are enhanced.

The quality of the metadata and the data are ensured by an online bilingual data submission system and a semi-intellectual review system. The submission system operates under Chinese-English linking, and drop-down lists for selection are used to the greatest extent possible to prevent manual typing errors. For example, three-level Chinese-English linked drop-down lists are used for discipline keywords,

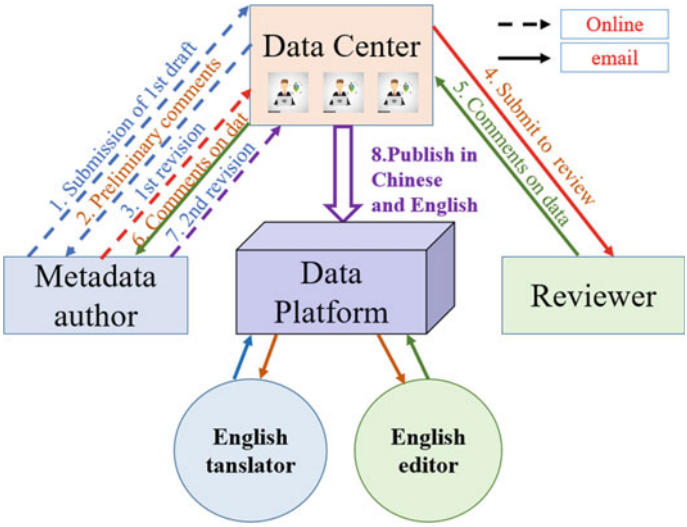

Fig. 5 Workflow of bilingual (Chinese-English editing) and metadata review in the Pan-Third Pole system (Solid lines represent workflows via email, and dashed lines represent workflows that are executed online)

topic keywords and others, and Chinese-English linked drop-down lists are used for temporal and spatial resolution. The semi-intellectual review system of the Pan-Third Pole big data system mainly manifests in the workflow when the system automatically emails a preliminary review notice to data service personnel upon uploading new data. The system invites expert reviewers according to discipline keywords via data review emails: the data are instantaneously open for sharing as soon as the reviewers recommend to publish the data; and if the reviewers require revisions, metadata producers are automatically notified. The workflows for English translation/editing of metadata and the semi-intellectual metadata review are presented in Fig. 5.

3.4 Key Datasets for Pan-Third Pole

The key datasets for Pan-Third Pole are developed from the enormous quantity of multisource and heterogeneous data of various classes and different disciplines through investigating, arranging, reconstructing and merging by using the same dataset system and geographic information system (GIS) platform in accordance with mechanisms related to data collection and submission, management, quality control, sharing and updating. The resulting key data resource of the Pan-Third Pole has considerable significance and high research value (Table 1).

Table 1 Classification of Pan-Third Pole core datasets

English Names of Key Datasets	Principles
Calibration and verification key datasets over Tibetan Plateau	Elements: meteorology, soil temperature and humidity, hydrology and eddy correlation (EC), based on principles of unified time span and time granularity (year, month, day and hour)
Cryospheric key datasets over Tibetan Plateau	Cryosphere elements: glaciers, glacier lakes, permafrost and snow cover over a unified area using a unified projection system
Basic geographic key datasets over Tibetan Plateau	Unified projection system, same spatial resolution and possibly synchronized data preparation
Near-surface atmospheric forcing datasets over Tibetan Plateau	Providing a group of near-surface meteorologically driven datasets with reliable quality and long time span
Scientific discovery key datasets over Tibetan Plateau	Mainly composed of scientific research findings

The permafrost thermal condition distribution map of the Tibetan Plateau (2000–2010) is considered as an example. The geographic-weighing regression model is used to integrate the following data: reconstructed temporal-spatial data, the moderate resolution imaging spectroradiometer (MODIS) surface temperature, the leaf area index, the snow cover ratio and the multimodel soil moisture prediction product of the National Meteorological Information Center, China. The map incorporates precipitation measurements from over 40,000 meteorological stations, the precipitation measurement product of Satellite FY2 and the average atmospheric temperature data of 152 meteorological stations from 2000 to 2010. The abovementioned data are used to simulate the long-term average atmospheric temperature data of the Tibetan Plateau with a spatial resolution of one kilometer. The thermal conditions of the permafrost classification system is used to identify the permafrost as very cold, cold, warm, very warm and likely to thaw. After deducting the areas of the lakes and glaciers, the total area of permafrost in the Tibetan Plateau is approximately 1.0719 million square kilometers, which demonstrates the high accuracy of the developed map and its ability to support planning and design of permafrost projects and environmental management. These results have been published in *The Cryosphere* [12].

An ecological and hydrological wireless sensor network in the midstream Yingke/Damanguan area of the Heihe River Valley is used to collect continuous observations from 50 WATERNET nodes over a 5.5 km * 5.5 km survey matrix. The resulting WATERNET observation and measurement dataset includes the soil moisture, the soil temperature, the conductivity, the complex permittivity, the surface infrared radiation temperature and the atmospheric infrared radiation temperature. The dataset has temporal and spatial continuity and can be used in several research areas: the remote sensing estimation of key water and thermal variables, remote sensing validation for a heterogeneous surface, ecological and hydrological studies

and irrigation optimization. Relevant results have been published in Scientific Data [13–15].

3.5 *Data Sharing: Principle and Mode*

The Pan-Third Pole big data system follows the FAIR (findability, accessibility, interoperability and reusability) principle for data sharing [5]. The system offers user access to bilingual (Chinese-English) data with primary online and secondary offline services (Fig. 6). Both the online and offline sharing approaches guarantee the following data rights and interests: 1) a unique data identification code, that is, a digital object identifier (DOI); 2) data distribution protocol, the default being Creative Commons Attribution 4.0 International (CC BY 4.0), which retains author's copyright; 3) literature citations: one or two publications are suggested to be cited for use of the data; 4) data citations: users are encouraged to cite the shared data via DOI; and 5) feedback on the status of data browsing, downloading and citation is regularly sent to the data producer. The online and offline data-sharing approaches are in different procedure. The online data can be conveniently directly downloaded by users, obviating notification emails to the data producer, whereas the offline data allows data producers to easily specify data rights and interests in addition to those listed above. The offline data service, under the premise of securing the exclusive rights and interests of the data producers, creates an automatic interactive system between data users and producers to increase sharing efficiency. The bilingual service is an improved approach that highlights both Chinese and English contexts and supplies high-quality metadata and data entities in English. The following types of service are provided: regular service, specialized service, typical case service, push notification service and information service (Fig. 7). Moreover, the Pan-Third Pole big data system will develop a mobile platform APP, a website for mobile phones, a

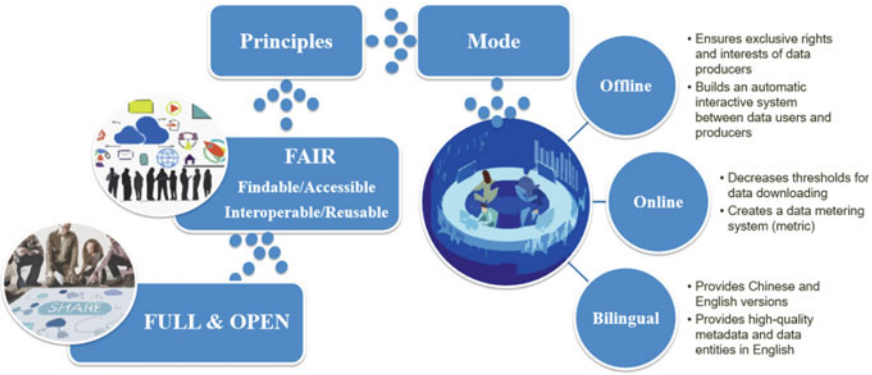

Fig. 6 Schematic showing sharing principles and mode (Pan et al., 2020, in review)

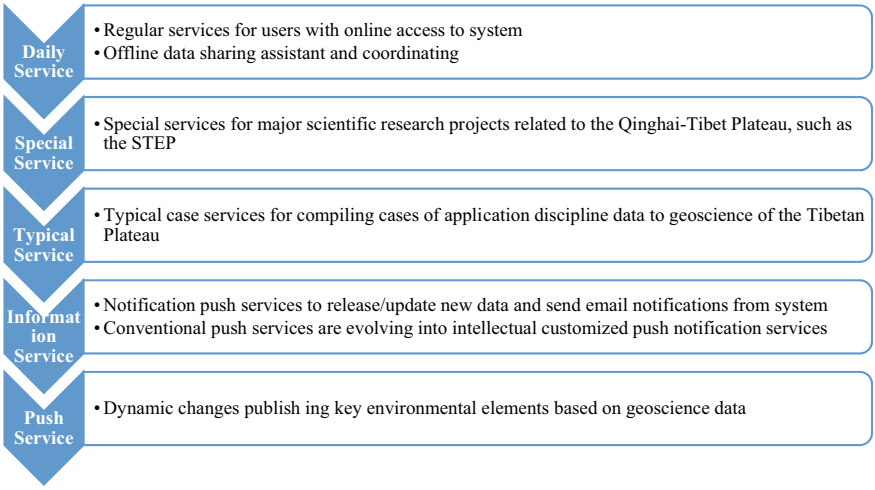

Fig. 7 Service mode of the Pan-Third Pole big data system

WeChat official account and a Weibo account. A newsletter will be regularly issued to promote the influence of the Pan-Third Pole big data system.

4 Intellectual Property and Data Publication

Big data has produced numerous changes in the information age. Increasing attention is being paid to intellectual property protection to provide security and sustainability for the big data industry. Data are products of both physical labor and intellectual activity. Therefore, data sharing requires intellectual property to protect the rights and interests of data producers and make data sharing sustainable. The Pan-Third Pole Data Center adopts multiple methods of securing data intellectual property rights.

4.1 Data DOI

The DOI is a unique system that was developed to address multiple links and the copyright transfer of digital resources over the Internet. The DOI system is used for Pan-Third Pole data resource development. This system can be applied to data sharing and enables tracking, citation, integration and networking. The DOI format of the Pan-Third Pole big data system is shown below:

10.11888/category.tpdc.metadataID.

where 10.11888 is a fixed prefix; 11888 is the registration number of the Research Institute of Tibet Plateau Research, the Chinese Academy of Science; category denotes the data type; tpdc is a fixed term; and metadataID denotes a six-digit data sequence number.

4.2 Data Distribution Protocol

The Pan-Third Pole big data system adopts Creative Commons 4.0 to retain author copyright, while allowing users to reprint, use and deduce data within the limited scope of an agreement and without informing the author. Six options are provided for data producers in the data submission system: CC BY 4.0 (Attribution), the default; CC BY-NC 4.0 (Attribution-NonCommercial); CC BY-ND 4.0 (Attribution-NoDerivatives); CC BY-NC-ND 4.0 (Attribution-NonCommercial-NoDerivatives); CC BY-SA 4.0 (Attribution-ShareAlike); and CC BY-NC-SA 4.0 (Attribution-NonCommercial-ShareAlike).

4.3 Literature Citation of Original Data

The submission module of the Pan-Third Pole big data system contains functions to create a data-related literature information entry and upload literature for reference and citation. Data users are thereby provided with a research background, data preparation, a processing method, quality evaluation and data application. The system provides the following notification to promote a benign academic atmosphere for scientific data sharing: “To use this data, you are suggested to reference the articles listed in the Required Data Citation section”.

4.4 Data Citation

Data citation is a new concept that was proposed in the global publishing industry and data-sharing sector. Data are treated like article references and listed in the references section of a paper. Data citation enables (1) data usage to be tracked, (2) data services to be counted and (3) protects the intellectual property rights of data. The data citation format is typically provided by the Data Center as “Data producer list. Data title. Publishing/release organization of data, date of data publishing/release. Permanent address of data DOI.” in the Pan-Third Pole big data system.

4.5 Data Protection Period

The *Notice of the General Office of the State Council (of China) on Regulations of Scientific Data Managemen*t (GBF (2018) No. 17) requires that all the scientific data in projects of science and technology programs (special projects, funds, etc.) at all government budget funding levels be submitted to the appropriate scientific data center by the leading organization of the project. In the interest of timeliness, it is recommended that data be submitted every year in accordance with the project tasks. Different data protection periods are set in the Pan-Third Pole big data system according to data acquisition method: (1) for data acquired in real time via the Internet of Things, quasi-real-time sharing is implemented; (2) for basic and automatic station data, no data protection period is set, in principle; (3) for incremental data generated in scientific research and field data acquired manually at elevations less than 4,000 m, a data protection period of no more than one year is planned, and direct sharing is encouraged; and (4) for field data acquired manually at elevations no less than 4,000 m, the set protection period should not exceed two years, and direct sharing is also encouraged.

4.6 Data Publication and Data Repository for Research Paper

The Pan-Third Pole big data system collaborates with prestigious journals to promote publication of data for the Tibetan Plateau and would like to increase the number of scientists sharing data. Considerable effort is being expended to ensure that the system will become a data repository certified by major international data publications such as *Scientific Data* and AGU journals and recommended by Earth System Science Data (ESSD). Researchers are being encouraged to publish and share their latest research results and related original data. Thus far, the Pan-Third Pole big data system has satisfied all the criteria for a data repository of major international data publications. A DOI and data-sharing system has been created to facilitate scientific data sharing and peer review of standardized metadata. Once the system becomes a data repository recommended by major international journals, original study data sharing can be promoted, data sources can be expanded, and the analysis of Pan-Third Pole scientific big data will be stimulated.

4.7 Permission of User Authority

The limited publication range and copyright of a dataset requires that user access be controlled under some conditions. After logging in, users are free to access datasets within the scope of authorization (including the authority to browse metadata and download specific data).

A user may be assigned multiple roles, each of which may have a customized data access range (that is, a mapping relation between the dataset and the user role). For example, special project users for the Silk Road environment project and STEP-2 program are authorized differently to control the dataset access range.

Control of user authority creates and maintains a correlation between roles/users and datasets rights and can be classified into role management and membership management. The role management module permits the viewing and downloading of metadata and datasets. The membership module creates membership, edits member information, reviews membership and manages member authority. In member authority management, a member can be assigned a role or provided customized access authority to specific datasets.

5 Big Data Analysis for Pan-Third Pole

5.1 Method Library Framework

Scientific data for the Pan-Third Pole region are characterized by high uncertainty and dimensionality. Various types of data result from multiple sources of observations and models and from the particularity and complexity of the Pan-Third Pole environment. The characteristics of big data are becoming increasingly clear. In geoscience, both big data-based information mining methods and spatial–temporal visualization methods are still in development. Many issues need to be resolved urgently. For instance, it remains unclear how to accurately analyze the overall change trend, detailed variation characteristics and the temporal evolution pattern of the Pan-Third Pole environment using geological big data that are characterized by different structures, decentralized storage and enormous volumes. It is also still difficult to maintain a global perspective on the interactions and synergy among multiple elements of the Pan-Third Pole environment while supporting the joint analysis of different temporal and spatial scales, which in turn makes mining, analysis and cognitive discovery of big data in the Pan-Third Pole environment challenging. The big data analysis system uses incremental integration and independent research and development to construct a method library for the Pan-Third Pole environment. The library comprises big data quality control, automatic modeling and analysis, data mining and interactive visualization. A tool library with high reliability, expandability, efficiency and fault tolerance is established in the meantime. Thus, collaborative analysis methods can be integrated and shared for big Pan-Third Pole environmental data with multisource heterogeneity, multigranularity, multitemporal phases and long time series, thereby facilitating efficient and online big data analysis and processing. The demonstration and application of big data for critical surface processes in the Pan-Third Pole environment opens an overarching technical link for deep data mining. The six major categories of big data analysis in the method library are as follows: advanced

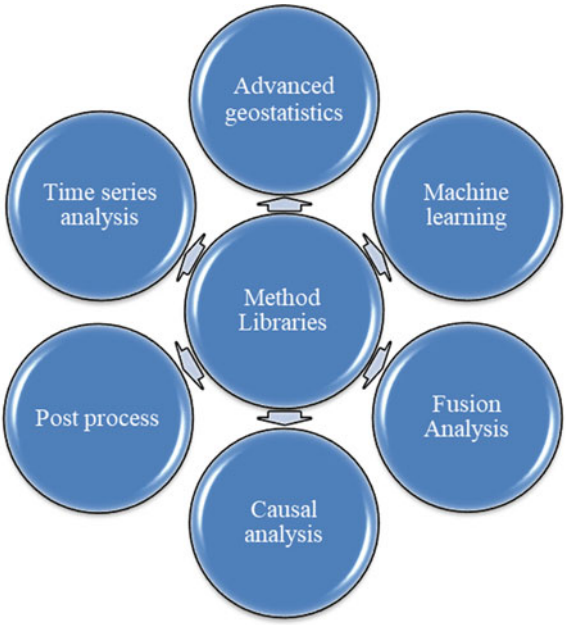

Fig. 8 Method library for big data analysis

geostatistics, time series analysis, traditional machine learning methods, postprocessing, model-observation fusion and causal analysis (Fig. 8). A code-sharing mechanism is established by using meta information to manage and perform intelligent searches/recommendations of these methods. The codes are hosted on GitHub.

5.2 Cases for the Use of Big Data Methods in Geological Research on the Pan-Third Pole Environment

The advent of the big data era has provided new opportunities for understanding and appropriately addressing water, ecology and environmental issues in the Pan-Third Pole. Big data science and technology is the product of a brand new revolution in scientific methods based on empirical, deductive and digital computing and has successfully predicted and analyzed the behavior of complex systems in many case studies. This technology is expected to dramatically change research in Earth system science. Multisource big data (such as from observations, remote sensing and simulations) can be combined using big analysis methods for use in automatic boundary extraction, inheritance of ecological network data and models and big data-driven

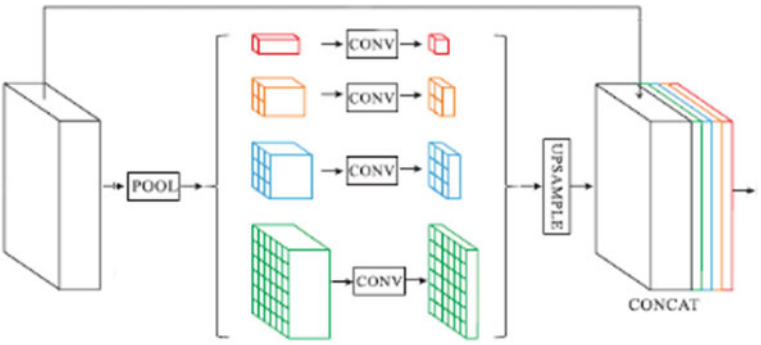

Fig. 9 Multiscale nested deep learning network

imaging of rupturing processes near the seismic source. Earth system science is thereby facilitated in the Pan-Third Pole environment.

(1) Deep learning-based ground object segmentation algorithm

In this study, an image pyramid-based concept is constructed in conjunction with a multiscale nested deep learning network to segment ground objects (Fig. 9).

Consequently, these networks require numerous convolution kernel structures, and in turn, high adaptability, during model optimization and calibration. The large number of required network parameters results in low training efficiency. In this study, the computing efficiency is increased by introducing a deep learning neural network with a variable convolution kernel structure to segment ground objects. A characteristic map pyramid is constructed by down-sampling and convolving the frontal-layer characteristic maps of the network. Up-sampling is then implemented, followed by merging and fusion with the back-layer characteristic maps. The proposed novel network exhibits a significantly improved response to ground objects at different scales than existing semantic segmentation networks.

(2) Algorithm for automatic extraction of glacier and lake vector data

The big data platform can be used to perform multisource data calculations that make full use of existing multisource remote sensing data and comprehensive data information. For instance, glacial data can be used for automatic glacier boundary extraction to accurately estimate glacial retreat on the Tibetan Plateau and surrounding areas under global warming and to predict changing trends. The deep learning neural network with a variable convolution kernel structure is used to segment ground objects based on a large quantity of data for typical lakes in the Tibetan region: Sentinel-1 wide-range synthetic aperture radar (SAR) data, Landsat remote sensing image data, high-resolution digital terrain model remote sensing data, glacier catalog vector and aerial data. Down-sampling and convolution are implemented on the frontal-layer characteristic maps to construct the characteristic map pyramid, followed by up-sampling and fusion with the back-layer characteristic maps. The network response to ground objects at different scales is improved. In addition, terrain

shadows, which are easily confused with glaciers, are removed, and the effects of speckle noise, winds and waves on the lake are prevented. The boundaries of glaciers and lakes can be identified with an accuracy above 90%.

(3) Observation data for ecological observation network and integration of ecological models

The following multiplatform, multispatial-scale and long time series remote sensing data are integrated: top apparent albedo (TOA) data at a Landsat 30-m resolution for the Third Pole region from the past 30 years; MODIS zenith angle albedo products at a 500-m resolution from 2000 to 2017; and MODIS vegetation coverage products and terrain data (SRTM DEM) at a 250-m resolution from 2000 to 2017. Carbon and water flux data from Tibetan Plateau flux stations are employed to optimize the parameters of the ecosystem model. The Morris method is used to perform a sensitivity analysis of the carbon and water cycle processes in the model. The most sensitive parameters are selected for optimization using Bayesian assimilation technology in conjunction with the carbon and water flux data. The parameter optimization significantly improves the simulation accuracy of the model for the net ecosystem carbon exchange (NEE) and reduces the simulation errors of the model for the gross primary productivity (GPP) of the ecosystem and the NEE.

(4) Development of seismic observation big data-driven imaging system for seismic rupturing processes

This study develops a software system that automatically classifies, retrieves, evaluates and picks up far-field P and SH waveform data (Fig. 10). A large quantity of raw data is thus rapidly processed to obtain far-field body wave data for imaging rupturing processes at the seismic source. A program based on the Green function is developed to calculate the static displacement field by the generalized reflection and transmission coefficient matrix method. Velocity structure models of the crust and upper mantle can be automatically extracted along with a static displacement Green function that depends on the depth and epicenter distance. The far-field waveform data are used to rapidly image rupturing processes at the sources of recent earthquakes with magnitudes above 6.0 in the Pan-Third Pole region and above 7.0 throughout the world. The earthquake intensity can then be theoretically evaluated.

6 Conclusions

The advent of the big data era has produced new opportunities and challenges for understanding the mechanisms of environmental problems in the Tibetan Plateau and the Pan-Third Pole and enacting appropriate measures. The Pan-Third Pole big data system is an important data support platform for the STEP-2 program and the Silk Road and Environment project. This system has multiple functions: the storage, management, analysis, mining and sharing of scientific data for various disciplines

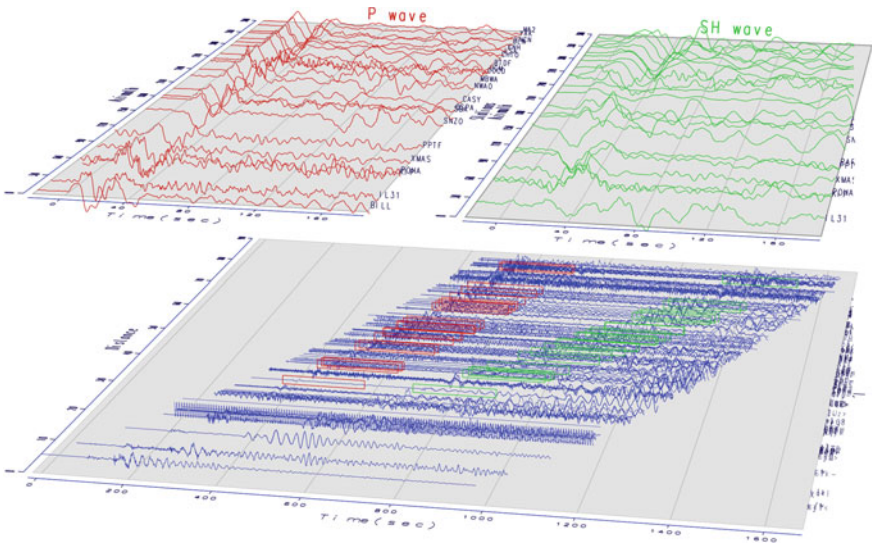

Fig. 10 Automatic selection and imaging analysis of far-field body wave data

of the Pan-Third Pole, such as resources, the environment, ecology, atmospheric science and solid-earth science; the refining of products of key scientific data of the Pan-Third Pole; the gradual development of functions such as online big data analysis and model application; the construction of a cloud-based platform to integrate data, methods, models; and services for Pan-Third Pole research; and the promotion of the application of big data technology to scientific research on the Pan-Third Pole. The system helps protect the eco-environment of the Pan-Third Pole region and facilitates the healthy, environmentally friendly and sustainable development of society and the economy. The system platform and the data resources of the Pan-Third Pole big data system are developed as follows: a standardized system regulation is created; the data catalog for the Pan-Third Pole is reviewed and organized; data resources of various platforms are combined and integrated; and the core dataset is compacted. Comprehensive data intellectual property protection measures for data rights and interests are adopted, and data protection periods are set to secure the rights and interests of data producers. The system follows the FAIR sharing principle and provides primary online and secondary offline services as sharing modes. The barrier for users to download data is lowered. The system has been proactively submitted for qualification as a data storage center for major international journals in an attempt to absorb more original data and energize Pan-Third Pole scientific research. The system offers a bilingual (Chinese-English) interface to provide scientific data resources from the Pan-Third Pole to research institutes and scientists worldwide. The Pan-Third Pole big data system is expected to deepen investigation of the "water-glacier-atmosphere-biology-human activity" multilayer interaction. The process and mechanism for environmental change in the region will be thereby elucidated, along with the impact on

and response regularity to global environmental change. Forecasting, prewarning and mitigation capabilities of regional disasters will also be improved.

After nearly two years of development, the Pan-Third Pole big data system (https://data.tpdc.ac.cn) is in regular operation and is delivering services. Over 2,500 scientific datasets have been published, and primary online services and secondary offline services are adopted as sharing modes. Multiple measures for intellectual property rights of data are enforced. The online big data analysis methodology is gradually being developed. The Pan-Third Pole big data system is dedicated to providing data support for scientific research on the Earth system and regional environmentally friendly development.

References

1. Yao T (2019) Tackling on environmental changes in Tibetan Plateau with focus on water, ecosystem and adaptation [J]. Science Bulletin 64(7):417
2. Yao T, Xue Y, Chen D, Chen F, Thompson L, Cui P, Koike T, Lau WKM, Lettenmaier D, Mosbrugger V, Zhang R, Xu B, Dozier J, Gillespie T, Gu Y, Kang S, Piao S, Sugimoto S, Ueno K, Wang L, Wang W, Zhang F, Sheng Y, Guo W, Alikun YX, Ma Y, Shen SSP, Su Z, Chen F, Liang S, Liu Y, Singh VP, Yang K, Yang D, Zhao X, Qian Y, Zhang Y, Li Q (2019) Recent third Pole's rapid warming accompanies cryospheric melt and water cycle intensification and interactions between monsoon and environment: multidisciplinary approach with observations, modeling, and analysis. Bullet Am Meteorol Soc 100:423–444
3. Yao T, Chen F, Cui P, Ma Y, Xu B, Zhu L, Zhang F, Wang W, Ai L, Yang X (2017) From tibetan plateau to third pole and pan-third pole. Bullet Chin Acad Sci 32(9):924–931
4. Yang X (2017) Scientists from various countries discuss the Pan-Third Pole environment and the Belt and Road Initiative. Sci Technol Daily 12(3)
5. Li X, Niu XL, Pan XD, et al. (2020) National Tibetan plateau data center established to promote third-pole earth system sciences [EB/OL]. https://www.gewex.org/gewex-content/files_mf/1590612006May2020.pdf
6. Li X, Nan ZT, Cheng GD, Ding YJ, Wu LZ, Wang LX, Wang J, Ran YH, Li HX, Pan XD, Zhu ZM (2011) Toward an improved data stewardship and service for environmental and ecological science data in west China. Int J Digital Earth 4(4):347–359. https://doi.org/10.1080/17538947.2011.558123
7. Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, Parsons M, Robinson E, Wyborn L (2019) Make all scientific data FAIR. Nature 570:27–29
8. Wang L, Li X (2019) Challenges and practice of geoscience data sharing: the example of the West China Ecological and Environmental Science Data Center. Science Press, Beijing
9. Wang L, Nan Z, Wu L, Ran Z, Li H, Pan X, Zhu Z, Li X, Ding Y (2010) Environmental and ecological science data center for western china: review and outlook. China Sci Technol Resour Rev 42(5):30–36
10. Pan X, Li X, Nan Z, Wu L, Wang L, Li H (2010) Research on data description documents. China Sci Technol Resour Rev 42(3):30–35
11. Li X, Che T, Li XW et al (2020) CASEarth poles: big data for the three poles. Bull Am Meteor Soc. 101(9):E1475–E1491. https://doi.org/10.1175/BAMS-D-19-0280.1
12. Ran YH, Li X, Cheng GD (2018) Climate warming over the past half century has led to thermal degradation of permafrost on the Qinghai-Tibet Plateau. The Cryosphere 12(2):595–608
13. Li X, Liu SM, Xiao Q, Ma MG, Jin R, Che T, Wang WZ, Hu XL, Xu ZW, Wen JG, Wang LX (2017) A multiscale dataset for understanding complex eco-hydrological processes in a heterogeneous oasis system. Sci Data 4170083. https://doi.org/10.1038/sdata.2017.83.

14. Jin R, Li X, Yan B, Li X, Luo W, Ma M, Guo J, Kang J, Zhu Z (2014) A Nested eco-hydrological wireless sensor network for capturing surface heterogeneity in the middle-reach of Heihe River Basin, China. IEEE Geosci Remote Sens Lett 11(11):2015–2019. https://doi.org/10.1109/LGRS.2014.2319085
15. Kang J, Li X, Jin R, Ge Y, Wang J, Wang J (2014) Hybrid optimal design of the eco-hydrological wireless sensor network in the middle reach of the Heihe River Basin, China. Sensors 14(10):19095–19114

Xin Li is currently a professor at Institute of Tibetan Plateau Research (ITP), Chinese Academy of Sciences (CAS) and the Director of National Tibetan Plateau Data Center at ITP/CAS. His primary research interests include land data assimilation, application of remote sensing and GIS in hydrological and cryospheric sciences, and integrated watershed study. He received the B.S. degree in GIS and Cartography from Nanjing University in 1992 and the Ph.D. degree in Remote Sensing and GIS from CAS in 1998. He was a member of WCRP GEWEX scientific steering committee, World Data Center for Glaciology and Geocryology at Lanzhou, and is a member of the International Science Advisory Panel of Global Water Futures programme. He is in the editorial board of Journal of Hydrology, Science Bulletin, Vadose Zone Journal, Remote Sensing, Big Earth Data and other international journals. He has published over 380 journal articles (SCI > 230) and coauthored 8 books. Total citations to these publications are 15,000+ and h-index is 60+. He is the lead scientist of WATER (Watershed Allied Telemetry Experimental Research, 2007-2010) and HiWATER (Heihe Watershed Allied Telemetry Experimental Research, 2012-2017), which are two comprehensive remote sensing ecohydrology experiments conducted sequentially in recent years in China.