第2卷 第2期 2020年4月
Vol.2 No.2 Apr.2020

ISSN 2096-742X
CN 10-1649/TP

文献DOI：
10.11871/jfdc.issn.
2096-742X.2020.
02.007

文献PID：
21.86101.2/jfdc.
2096-742X.2020.
02.007

页码：91-100

开放科学标识码
（OSID）

# 地学大数据处理架构与关键技术研究

张耀南[1,2,3*]，艾鸣浩[1,2]，康建芳[1,2,3]，敏玉芳[1,2]

1. 中国科学院西北生态环境资源研究院，甘肃 兰州 730000
2. 国家冰川冻土沙漠科学数据中心，甘肃 兰州 730000
3.甘肃资源环境科学数据工程技术研究中心，甘肃 兰州 730000

**摘　要：**【目的】大数据以其独特的数据科学思维为地学研究知识发现带来重大机遇，但地学数据独特的多源异构、时空关联、多尺度和不确定性等特征亦给地学大数据处理带来一系列挑战。【方法】本文在分析地学数据特点基础上，结合数据关联、中间件系统、微服务及容器等技术手段，提出一种面向地学大数据的处理框架，重点解决地学领域多源数据汇集融合、异构数据综合集成处理问题，并将地学模型引入框架，增强数据处理的地学专业性。【结果】框架及其关键技术已在国家冰川冻土科学数据中心建设、高寒环境联合观测研究云及中巴走廊灾害数据集制备中应用实施。【结论】地学大数据平台处理框架拓宽数据处理维度，可为多主题、多尺度地学研究分析和知识发现提供支撑，未来框架将适应互联网、社交网络、平面媒体等更广泛来源的地学数据处理，进一步融合人工智能技术，提供更智能更迅捷的地学数据处理结果。

**关键词：**地学大数据；地学数据处理方法；汇聚融合；异构集成

# Research on Geoscience Big Data Processing Framework and Key Techniques

**Zhang Yaonan[1,2,3*], Ai Minghao[1,2], Kang Jianfang[1,2,3] , Min Yufang[1,2]**

*1. Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou, Gansu 730000, China*
*2. National cryosphere Desert Scientific Data Center, Lanzhou, Gansu 730000, China*
*3.Gansu Data Engineering ang Technology Research Center for Resource and Environment, Lanzhou, Gansu 73000, China*

**Abstract:** [**Objective**] As one of the special data science methods, big data brings great opportunities for geological research. Meanwhile, the characteristics of geological data such as multi-source heterogeneity, spatial-temporal correlation, multi-scale and uncertainty bring great challenges for the data processing. [**Methods**]On the basis of a detailed analysis of aforementioned characteristics, this study proposes a geological data processing framework to solve problems of multi-source data integration and heterogeneous data synthesis in geoscience field combined with a variety of big data technologies like data association, middleware systems, micro services and container technique. Besides, geological models are embedded in this framework in order to improve the expertness of data process. [**Results**] The framework and its key technologies have been applied in the construction of the National Glacier and Frozen Soil Scientific Data Center, the disaster datasets for the China-Pakistan Corridor as well as the High and Cold Environment United

基金项目：中国科学院信息化专项（XXH13506）；国家科技基础条件平台建设（Y719H71）
*通讯作者：张耀南（E-mail: yaonan@lzb.ac.cn）

Observation Cloud. **[Conclusions]** This study is expected to broaden the data processing dimension and support multi-theme, multi-scale research and knowledge discovery in geoscience. In future, it will be adapted to the processing of geological data from a wider range of sources such as the internet, social networks, and printed media. The integration of artificial intelligence technologies will enable the framework to provide smarter and faster geological data processing results.

**Keywords:** earth scientific big data; geological data processing methods; data convergence; heterogeneous data integration

## 引言

更好地传感技术、更强大的计算平台和更敏捷的云服务，使我们以前所未有的速度获得了海量数据的收集、存储、分析和应用的能力，大数据时代对我们感知、认知、预知和决策方式已经产生了深远影响。数据量的激增，数据存储管理、数据处理分析、领域知识链接、人工智能及可视化等方面研究已发生了颠覆性变革，数据科学应运而生[1-3]。除改变互联网、零售、广告等商业行业外，大数据提供了一个在虚拟信息世界中了解和掌握现实世界前所未有的机会，在推动科学发现方面发挥越来越重要的作用[4]。从科学史上来看，科学的发展首先生成假设或理论，然后收集数据以确认或反驳这些假设。但在大数据时代，在不考虑特定理论或假设的情况下，通过对不断收集的大量数据进行挖掘分析，引入人工智能方法技术，为发现新知识新现象提供了一个崭新的机遇。特别是当数据集的维数很高或过于复杂以至于很难用传统的统计方法处理时，以人工智能和机器学习为代表的数据科学处理法尤为具有吸引力[5]。事实上，大数据技术在科学学科发展中的作用已经开始从提供简单的分析工具逐渐转变为提供成熟的知识发现框架[6]。

地学是通过对自然现象的观察，发现蕴含在观测数据之中的自然规律，从而研究地球系统多尺度下的各种过程、变化及相互作用规律。从地学经验范式到数据密集型研究范式，所需的数据数量、复杂性和多样性方面都急剧增加[7]。在全球气候变化背景下，单一学科、单一尺度、单一区域的研究已越来越不适应地学研究发展的需要，地球科学研究态势已出现综合性、跨学科性、跨区域性和协同性等显著特点。数据科学基于将不同来源、不同区域、多尺度数据汇集融合和集成分析，使得地学研究中开展大尺度、广视角、多系统联合和多过程耦合研究成为可能。深化地学大数据与地球系统知识发现研究，也将成为地球关键带过程与功能、全球环境变化与地球圈层相互作用、人类活动对环境影响、重大灾害形成机理研究等研究的重要支撑[8]。

近年来，国际上先后部署一系列地学大数据相关重大计划和研究项目，美国的“地球立方体”项目，欧盟的“地球模拟器”项目，我国的“地球大数据科学工程”都是旨在以整体视角审视地球系统，利用地球大数据驱动跨学科、跨尺度宏观科学发现[9]。“全球气候服务框架（GFCS）”将实施优先领域定在构建气候服务信息系统，定期收集、存储及处理各类地学数据，开发并分发一系列数据产品和服务，为农业、健康、灾害等各种决策提供支持[10]。长期生态学研究网络（LTER）、英国环境变化监测网络（ECN）及中国生态系统研究网络（CERN）等提供多尺度生态信息，使得获取海量、大尺度、多源生态数据成为可能[11-13]；NASA 地球交换平台（NASA Earth Exchange Platform）将超级计算、数据可视化、海量在线数据、模型和算法、社交网络和协同技术集成在一起，形成用于地学研究和知识发现的大数据平台[14]。NOAA 的大数据项目（NOAA-BDP）遍及 68000 个数据集管理与共享，引入多种云平台增强数据集的可发现性和可访问性[15]。

地球科学在数据量、速度和多样性方面已成为数据最丰富的领域之一，但与生物学、天文学、管理学等其他科研领域的成功相比，大数据在地学领

域的应用进展较为缓慢。当前地学大数据研究主要涵盖在地学数据管理方法、汇集与共享方式以及机器学习方法在地球科学的深度运用，为数不多的地学数据处理研究多集中在遥感影像和特定模型的数据前处理。本文针对地学领域多源异质数据综合集成展开研究，在分析地学大数据特点基础上提出一种地学大数据处理框架，打通研究要素聚合时空壁垒，疏通海量地学数据与地学知识发现之间的数据通道，为生态、环境、资源领域的长期监测、机理认识和精准预测提供方法技术支撑。

## 1 地学大数据特点

地学数据是一种与地球参考空间（二维或三维）位置有关的、表达与地理客观世界中各种实体和过程状态属性的数据[16]。地学数据来源于野外调查、卫星遥感、定位观测、仪器测试分析、模拟计算结果、调查统计普查及地图文献资料，涵盖地球从内到外的各个圈层，涉及地球系统多种地学因子，涉及大气、生态、水文、土壤、海洋、地质等诸多学科，还与物理学、化学和信息科学息息相关。地学大数据具备数据量大（volume）、类型繁多（variety）、速度快实效高（velocity）及价值密度低（value）等传统大数据所具备的“4V”共性[17]。同时地学研究对象发展演化时空范围庞大，相互作用影响因素众多，以及数据获取手段和数据处理方式的差异，使得地学数据在内容上具有“参数信息不完全、结构信息不完全、关系信息不完全和演化信息不完全”的特征，在形态上具有显著的多类、多维、多标签、多尺度和多主题特征[18]。这与其他领域学科所产生的大数据存在很大差别，可归纳为四方面。

### 1.1 时空相关特征

地学数据是以地理特征和地学过程为对象，基于统一时空基准，与位置相关联的地学要素的定量体现。地学数据高度时空相关特性体现在三方面：（1）地学数据具有时间、空间和属性三种基本特征，其属性值紧密依赖于时间与空间，这导致地学数据存在抽象意义的时空相关。（2）数据之间的关系与数据的空间位置、空间拓扑关系和时间关系相关联。由于地学过程的时空连续性，在时间或空间上接近的数据呈现高度相关。（3）时空基准不统一。地学数据根据其研究需要和观测实际，时间粒度可从分秒横跨至数十万年，加之描述其空间位置的坐标系、投影参数不同，造成不同数据之间时空基准存在差异。

### 1.2 多源异构特征

地学数据来自于气象、水利、国土、高校、科研院所等诸多部门。这些数据的获取与制备往往面向特定研究或业务背景，针对不同的地学单元，产生于不同的采集方式，再加之数据生产者专业背景和数据理解各异，致使不同来源数据具有不同的数据管理形式。除常规气象数据、遥感数据、基础地理数据等少数几种以外，多数地学数据组织呈现多源异构特点。一是结构化与非结构化数据并存，关系与非关系数据库并存，文本、表格、图像等多种格式数据无序堆叠，数据间关联关系混乱，数据组织方式随意。二是由于缺乏统一标准，对同一地学本体的命名方式、描述方式、采集标准、数据单位多种多样。

### 1.3 多尺度特征

地球作为一个复杂的巨系统，由多种复杂的子系统构成。子系统间相互作用从微米级颗粒和气溶胶到陆表面大规模变化，作用过程从持续数小时或数天到持续数年数十年。由此可见，地学数据具有明显的尺度依赖性和多重表达性，其对同一种地学要素，同一种地学过程，在不同时空尺度描述下取值和呈现的趋势不尽相同。尺度的变化也会影响地学数据处理、分析及表达的方式，在不了解尺度意

义下改变数据尺度会使研究对象的过程和形态得不到预想的结果。地学基础研究趋势是实现综合尺度下的地学过程相互耦合，从整体上解决地学问题复杂性。地学数据的多尺度互动和联结既蕴含巨大机遇，其跨尺度所带来的数据矛盾也成为重大挑战。

### 1.4 不确定性特征

野外自动观测数据受限于仪器本身质量和人工维护质量，数据错误、数据缺失时有发生，勘探调查数据多带有人的主观因素，复杂的数据测试分析链条也会不可避免引入人为误差和系统误差，同时由于地学过程本身具有高度的复杂性，人类对许多地学规律的认识尚存争议，对各种计算模拟和定位观测数据的一致性理解也不相同，故而地学数据体现不确定性特征。此外，地学数据对描述复杂对象的定量化程度有限，采集形成全球样本标准数据集到现阶段还无法实现，基于地学数据的分析结果也多带有模糊和不确定性。

## 2 地学大数据处理框架

地学大数据处理框架旨在搭建数据来源与数据应用之间的桥梁。流程由数据来源开始，多重来源数据通过基础硬件环境进入处理框架中，首先经过汇集融合，按照不同数据特点进入不同异构自治的数据源中，并以元数据集描述数据，以数据字典描述数据源。当应用层发起获取数据请求时，各数据经过统一集成后响应数据请求，实现包括过程研究、知识发现、数据挖掘等方面的应用。地学大数据处理架构核心是解决不同来源的多源数据汇集融合处理，解决跨地域、跨时空、跨学科数据抽取的异构数据综合集成处理，以及解决基于地学专业的“数据 - 模型”一体化应用处理。整体架构如图 1 所示。

### 2.1 多源数据汇集融合处理

多源数据汇集融合处理，主要应对不同来源地学数据在采集和管理过程中出现的体系松散、结构混乱、缺乏组织的现状，主要解决海量地学数据规

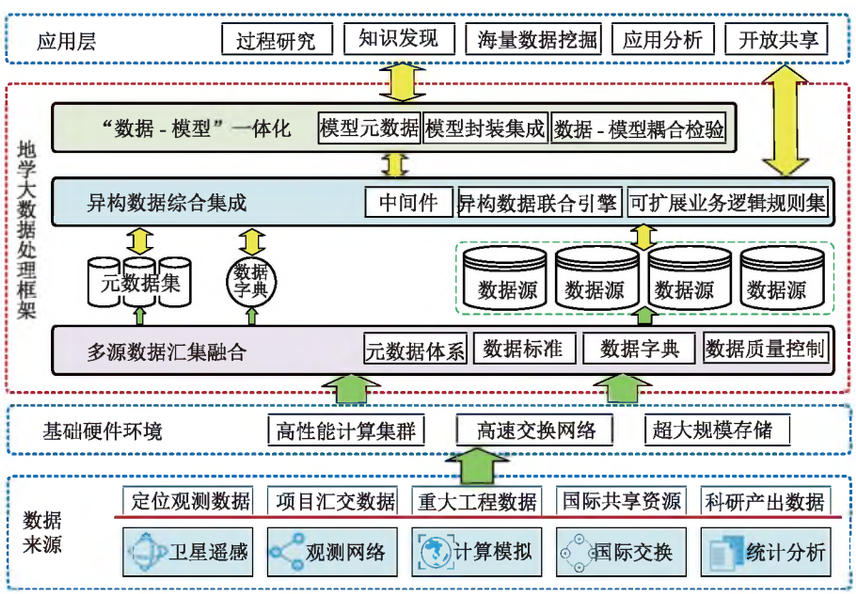


**图 1 地学大数据处理框架示意图**

**Fig.1 Schematic diagram of geoscience big data processing framework**

律挖掘和地学多过程机理研究中相关数据难以有效汇集的问题。地学数据融合不只是多源数据抽取、数据格式转换、结构化 / 非结构化存储等信息技术问题，更多需要专业视角下构建的数据关联方法、数据标准化方法及数据质量控制方法。

### 2.1.1 元数据体系

针对地学大数据特点，梳理地学研究对象包括气候、水资源、资源灾害、人地相互作用及重大工程建设等研究领域中各类数据观测过程、数据生产方式和数据应用需求，厘清各地学过程研究所需环境要素的种类、精度、尺度和制备方式。从数据内容和时空特征两方面入手，建立栅格数据、矢量数据、时间序列野外观测数据、仪器分析模拟数据、模型模拟数据等多源地学数据关联模型。数据内容方面以科研需求、学科分类和研究主题三种维度构建数据之间语义化链接，如图 2 所示。时空特征方面，将地学元数据中表达时间、空间特征的名词映射到数据实体建立时空语义关联，以数据间空间位置距离关系为规则建立空间结构关联。所有概念属性、关联关系和关联规则都在元数据中进行描述。

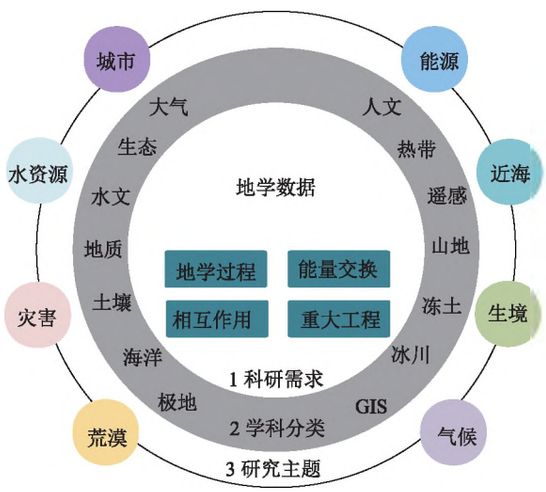


**图 2 地学大数据内容关联示意图**

**Fig.2 Schematic diagram of geoscience big data content correlation**

### 2.1.2 数据标准

针对地学科研数据一般规律，分析野外定位持续观测数据、重大项目产出数据、国际相关数据源及历史数据等四大类数据特点，构建数据标准的设计可以划分为基础性标准与应用性标准两类。基础性标准主要用于在不同系统间，形成信息的一致理解和统一的坐标参照系统，是信息汇集、交换以及应用的基础，包括数据分类与编码、数据字典、数字地图标准；应用型标准则是为数据平台功能发挥所涉及的各个环节，提供一定的标准规范，以保证信息的高效汇集和交换，包括元数据标准、数据交换技术规范、数据传输协议、数据质量标准等。

### 2.1.3 数据字典

针对多数据源内容、命名、数据单位等异构问题构建数据字典，对数据源进行描述。记录每一个数据源的变量命名、变量描述、类型、数据类型及纲量等信息，为每一个数据源建立“异构混乱 - 标准统一”的映射关系，使整体处理框架对每一种异构数据源的存储、组织和命名方式“了如指掌”。

### 2.1.4 数据质量控制

针对数据存在异常值、结构性错误、记录重复和数据缺失等问题，设计循环质量评估流程，如图 3 所示，并依据数据应用方式和使用尺度制定数据处理方法。对长时间序列观测数据建立数据插值、异常点监测及时空滤波等方法；对遥感数据，建立不同监测要素数据的时间序列重建方法、空间插值重建方法，实现异构数据的自动 - 半自动化质量控制过程。

## 2.2 异构数据综合集成处理

异构数据综合集成主要解决两类问题：一类是内容相同但时空属性不同的地学数据集成；另一类是数据资源在存储管理上互异自治，存储在不同操作系统及不同的数据库管理系统和文件系统中。中

间件系统（middleware）因其能够屏蔽底层数据源的平台、环境、数据模型和语义异构性，另有快速部署、管理方便、利于复用的优势，成为大数据领域常用的解决异构数据综合集成的方案之一，其“分而治之”的异构数据融合策略能够应对地学数据多源异构的现状。中间件通过全局数据模型隐藏底层数据细节，保持数据依旧存放于异构自治的数据源中，通过各数据源适配“包装器（Wrapper）”将数据通过映射到全局数据模型上；对于应用层的数据服务请求，则采用“中介器（Mediator）”将其解析、分析和拆分为一个或多个针对相应数据源的子查询，然后将查询结果按照相应逻辑和业务规则综合集成反馈。为适应地学大数据处理需求，打通“异构数据—分析应用”之间技术屏障，一方面中间件全局数据模型需与多源地学数据模型融合，另一方面具备数据联合引擎和中介器逻辑规则扩展集成两种专门面向地学数据处理的能力，异构数据综合集成架构如图 3 所示。

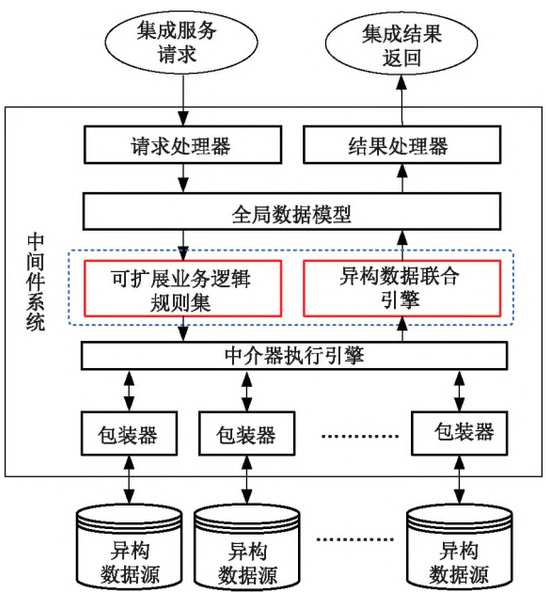


**图 3 异构数据综合集成架构图**

**Fig.3 Heterogeneous data integration structure diagram**

### 2.2.1 异构数据联合引擎

异构融合将相互关联的数据集成到一起供应用层服务访问，然而不同的数据源有不同的时空基准和命名规范，例如应用层需要集成两个地理位置的近地表 2 米气温逐小时数据做分析，这些数据在两个数据源中保存，每个数据源都具有自己的自治标准。一个命名为“2 米气温”，以地理经纬度坐标表示地理位置，采集间隔为小时，单位摄氏度；一个命名为“temperature_2m”，以墨卡托投影表示地理位置，采集间隔为分钟，单位开尔文温度，需要在中间件中提供由异构数据联合引擎采用数据字典将字段相互关联命名统一，利用各种对应转换关系统一时间维、空间维和数据单位。

### 2.2.2 可扩展业务逻辑规则集

地学数据具有空间性、时间性、尺度性等多种独特性质，而研究人员由于专业背景不同、研究领域不同、研究尺度不同，即便面对同样的一条地学数据，理解和分析的角度也不尽相同，因此很难有一套通用固定的方法进行数据的异构融合。构建中介器内数据业务逻辑规则集开放框架，包括架构组件和交互通道，支持不同语言、不同环境的逻辑规则和处理方法集成与组合，使规则逻辑集能够根据需求改变而灵活扩展，每一种规则算法能够即插即用，即删即无。

## 2.3 “数据 - 模型”一体化处理

当前数据处理中常用方法难以迎合地学前沿所需长时间序列、高时空分辨率、大空间范围数据处理需求。大数据领域数据清洗、数据插补等方法多是基于数值方法、统计方法或机器学习，地学数据在这样的数据处理链条中容易发生地学意义和地学规律上的误差，且误差会随数据生命周期进行演化，最终使地学数据驱动的研究分析和知识发现结果发生畸变。因此地学数据处理框架除集成一般数据处理方法外，还需集成具备地学背景的地学模型。与一般处理方法不同，地学模型处理中存在模型异构性和复杂性等问题，且尺度精细化的地学数据处理

常伴随超大规模计算。需要对模型进行封装和管理，构建“数据 - 模型”间的数据互通接口，令地学数据与模型耦合起来形成数据处理链；通过组件技术和容器技术，解决地学模型与超级计算关键集成问题。关键技术包括模型元数据设计、模型集成及“数据 - 模型”耦合有效性检验。

### 2.3.1 模型元数据设计

模型元数据描述模型物理意义、适用范围和模型输入输出数据，从而支持模型与数据的耦合，异构地学数据的传递是数据提取、模型装配和模型耦合的重点。模型元数据包括三方面：(1) 模型标识信息，包括模型的用途描述、时空尺度、建模原理和适用范围；(2) 模型运行信息，包括模型的操作系统、运行环境、所需库文件和编程语言；(3) 模型数据信息，包括输入输出数据的数据名称、变量名称、数据时空尺度、数据存储格式及数据类型等。

### 2.3.2 模型封装集成

模块融入地学数据处理框架需要三个步骤：组件化封装、微服务化集成和容器化部署。首先，根据模型元数据定义接口，将构建在不同平台上、用不同编程语言编码的地学模型封装成即插即用的组件。其次使用微服务框架将每个模型组件作为轻量级 Web 服务发布，并通过服务链接实现模型集成服务。这些服务既可以作为一个整体模型独立运行，也可以通过服务链接组成一个模型链运行。最后将微服务以容易形式打包，不仅打包模型和微服务本身，还将模型所有依赖、附着操作系统一同打包，部署于并行化运行环境中。

### 2.3.3 数据 - 模型耦合校验

复杂数据和模型的集成可能会在整个模型链中传播不确定性，需要对数据 - 计算耦合有效性进行数据兼容性校验。以模型元数据为依据，校验内容包括模型接口、输入输出变量名称、时空分辨率、时空一致性及语义相似度。

## 3 示范与应用

本文以“高寒环境联合观测研究云”中巴走廊冻土分布的地学处理为例，介绍地学大数据处理框架的实际应用。“高寒环境联合观测研究云”（简称“高寒云”）是中国科学院部署，横跨“十二五”、“十三五”的综合性信息化项目，旨在通过高寒环境下模型研究资源的虚拟集成，构建地学大数据处理平台，整体提升高寒区研究水平。中巴经济走廊是我国“一带一路”的重要组成部分，其成功建设和安全运营具有重要战略意义。以中巴经济走廊沿线高寒区灾害为专题，基于地学处理框架开展冰川、冻土、洪水、滑坡及泥石流等灾害数据产品、计算模型和决策工具的研究支撑是“高寒云”的重要示范之一。

“高寒云”中采用 TTOP 模型计算中巴走廊冻土分布，需要高时空分辨率地表温度数据、土地覆被数据和土壤类型数据为输入，其中高时空分辨率地表温度数据则源于对 Landsat 遥感影像的反演和野外定位观测数据验证，上述两种方法模型为不同科研人员提供。“高寒云”建设了包括存放算法模型的模型资源池，存放多源数据的数据资源池，以及由计算集群构成的计算资源池；基于资源池，构建了记录异构数据与全局数据模型映射关系的数据字典，描述“数据 - 模型”业务流程的规则集，以及包含多种格式数据空间变换、时序插值、字段重组等常见地学数据处理工具集。所有资源以服务形式供“高寒云”中间件系统访问使用。

以冻土分布计算数据为例，地学大数据处理过程如下所述。

(1) 封装。“高寒云”将地表温度反演方法和冻土分布模型输入输出数据的数据名称、数据格式、时间尺度、空间分辨率以及模型的初始化状态、预处理方法和运行环境等在模型元数据记录和描述，再将模型元数据、模型运行环境和模型预处理方法标准化封装，存放于模型资源池并“暴露”接口。

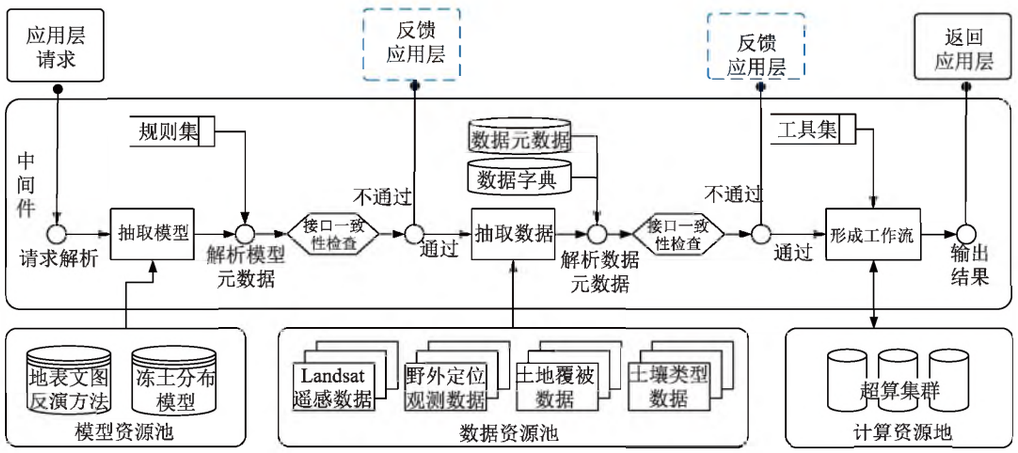


**图 4 以冻土分布计算为例的数据处理流图**

**Fig.4 Data processing workflow diagram based on frozen soil distribution calculation**

（2）模型抽取与检查。中间件系统接受应用层处理请求后，从模型资源池内抽取模型并对地表反演方法和冻土分布模型的输入输出进行一致性检查，主要利用业务逻辑规则集检查地表反演方法的输出数据与冻土分布模型输入是否匹配。

（3）数据抽取与检查。根据应用层请求的时空范围从数据资源池中抽取数据及其元数据，通过数据元数据和数据字典检查数据的字段名称、字段类型、单位纲量及存储格式是否与模型输入匹配。

（4）“数据 - 模型”一体化计算。将数据、方法模型以及数据镶嵌、裁剪及时空插值等常用地学处理工具组合形成工作流进入超算集群计算，向应用层返回计算结果。

处理流程如图 4 所示。

目前“高寒云”已实现包括冻土、积雪、冰湖、荒漠化、滑坡、泥石流和洪水等多种高寒环境自然灾害相关环境因子的提取、分析和数据再生产。形成 1 套在线平台、9 篇数据文章和 14 套中巴走廊灾害数据集，如图 5 所示。利用地学大数据框架有效地解决当前中巴走廊自然灾害综合研究地学数据处理时间成本高、重复工作多、方法不能共用、结果难以集成等问题，为多维度综合研究中巴经济走廊自然环境的时空演变特征及规律提供高效数据处理支撑。


（a）


（b）

**图 5“高寒云”基于地学大数据处理框架应用成果：(a) 数据处理云平台；(b) 中巴走廊灾害研究成果论文集**

**Fig.5 “Alpine & Cold Region Research Cloud” application achievements of geoscience big data processing framework: (a) data processing cloud platform;(b) the study collection of China-Pakistan economic corridor disaster**

## 4 总结与展望

目前地学领域数据挖掘和知识发现的能力远远落后于数据的获取能力，地学数据处理的复杂性和专业性成为重要原因之一。本文在分析地学大数据特点基础上给出一种地学大数据处理框架。针对地学大数据多源异构、时空相关、多尺度和不确定性四个特征提出多源数据汇集融合、异构数据综合集成和“数据 - 模型”一体化三种处理方法，并介绍框架关键技术高寒环境联合观测研究云在中巴走廊灾害研究中的应用。地学处理框架以异构数据联合引擎和可扩展逻辑规则集，适用于来自应用层多尺度、多视角的数据抽取聚集需求；将地学大数据与地学模型组装在一起，拓展了地学数据处理的广度和深度，支撑更为复杂和专业的地学大数据分析与应用。

将大数据技术与地学研究深度融合是一个值得继续探索的问题。本文在地学数据关联的研究和应用还十分浅薄，仅将其作为检查数据与模型一致性的规则，下一步应深入研究地学数据关联特征，构建适用于地学“数据 - 计算”一体化的关联模型。本文所提大数据包含在互联网、社交网络及平面媒体中带有时空属性的数据当前也已进入地学大数据的范畴。本文所提出的地学数据处理框架仅面向野外观测、仪器分析、卫星遥感等传统意义上的地学数据，未来会探索针对上述“新兴”地学大数据的处理问题。此外，地学大数据处理框架将会进一步加深与人工智能的融合，提供更智能更迅捷的地学数据处理结果。

## 利益冲突声明

所有作者声明不存在利益关系。

## 参考文献


[1] Marx V. Biology: The big challenges of big data[J]. Nature, 2013, 498(7453):255-260.

[2] Lake B, Salakhutdinov R, Tenenbaum J. Human-level concept learning through probabilistic program induction[J]. Science, 2015, 350(6266):1332-1338.

[3] Waller M and Fawcett S. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management[J]. Journal of Business Logistics, 2013, 34(2):77-84.

[4] Li G, Cheng X. Research status and scientific thinking of big data[J]. Bulletin of the Chinese Academy of Sciences, 2012, 27(6):647-657.

[5] Karpatne A, Atluri G, Faghmous J. Theory-guided data science: a new paradigm for scientific discovery from data[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(10):2318-2331.

[6] James F, Vipin K. A big data guide to understanding climate change: the case for theory-guided data science. Big data, 2014, 2(3):155-163.

[7] 宋长青.地理学研究范式的思考[J]. 地理科学进展，2016, 35(01):1-3.

[8] 中华人民共和国国土资源部.国土资源“十三五”科技创新发展规划[Z].国土资发(2016)100号.

[9] 郭华东.地球大数据科学工程[J].中国科学院院刊，2018,33(08):818-824.

[10] World Meteorological Organization. Step by step guidelines for establishing a national framework for climate services. https://library.wmo.int/index.php?lvl=notice_display& id=20216#.Xh0jU0czaUk.

[11] Aronova E, Baker K S, Oreskes N. Big science and big data inbiology: From the international geophysical year through the international biological program to the long term ecological research (lter) network, 1957-present[J]. Hist Stud Nat Sci, 2010, 40:183-224.

[12] Hampton S E, Strasser C A, Tewksbury J J, et al. .Big data andthe future of ecology[J]. Frontiers in Ecology and the Environment, 2013, 11(3):156–162.

[13] 于贵瑞, 何洪林, 周玉科.大数据背景下的生态系统观测与研究[J].中国科学院院刊, 2018, 33(08):832-837.

[14] NASA Earth Exchange (NEX). https://nex.nasa.gov/nex/.
[15] NOAA Big Data Project. https://www.noaa.gov/big-data-project.
[16] 李军，周成虎.地学数据特征分析[J].地理科学，1999(02):63-67.
[17] 翟明国，杨树锋，陈宁华，陈汉林.大数据时代：地质学的挑战与机遇[J].中国科学院院刊，2018,33(08):825-831.
[18] 吴冲龙，刘刚，张夏林，何珍文，张志庭.地质科学大数据及其利用的若干问题探讨[J].科学通报，2016,61(16):1797-1807.

收稿日期：2020 年 1 月 16 日

**张耀南**，现任中国科学院西北生态环境资源研究院大数据中心主任，国家冰川冻土沙漠科学数据中心主任。博士，研究员，博士生导师。主要研究方向为环境科学数据工程、基于高性能计算环境的地学模型模拟、遥感图像处理及多源数据融合。
本文主要承担地学处理框架设计及整体项目应用实施。
Zhang Yaonan, PH.D, is a professor and the dean of Big Data Center of Northwest Institute of Eco-Environment and Resources. His current research interests include integrated modeling environment, remote sensing image processing and multi-source heterogeneous data fusion.
In this work, he is mainly responsible for the overall framework design and project implementation of earth data process.
E-mail: yaonan@lzb.ac.cn

**艾鸣浩**，中国科学院西北生态环境资源研究院，在读中国科学院大学博士，工程师。主要研究方向为多源数据集成、遥感数据处理及人工智能应用等工作。
本文主要承担多源异构数据综合集成架构设计，论文编写。
Ai Minghao, is an engineer and also a PH.D candidate in University of Chinese Academy of Sciences. His research interest include multi-source data integration, remote sensing data processing and the application of artificial intelligence.
In this work, he is responsible for the design of multi-source heterogeneous data integration and paper writing.
E-mail: aimh@lzb.ac.cn

**康建芳**，中国科学院西北生态环境资源研究院，工程师。负责地学大数据管理与分析等工作。
本文主要承担数据关联方法设计与系统集成。
Kang JanFang, an engineer in Big Data Center of Northwest Institute of Eco-Environment and Resources, is working on analysis and management of geoscience data.
In this work, she is responsible for designing the geoscience data association method and system integration.
E-mail: kjf@lzb.ac.cn

**敏玉芳**，中国科学院西北生态环境资源研究院，在读中国科学院大学博士，工程师。专注于地学大数据管理与分析研究及地学模型集成耦合等工作。
本文主要承担数据 - 模型一体化架构设计与应用。
Min Yufang, is an engineer and currently pursuing a Ph.D degree at Big Data Center of Northwest Institute of Eco-Environment and Resources, University of Chinese Academy of Sciences. Her main research interests include geoscience data management and geoscience model coupling method.
In this work, she is responsible for the design of data & model integration.
E-mail: myf@lzb.ac.cn

引文格式：张耀南,艾鸣浩,康建芳,等.地学大数据处理架构与关键技术研究[J]. 数据与计算发展前沿,2020,2(2):91-100.DOI:10.11871/jfdc.issn.2096-742X.2020.02.007.PID:21.86101.2/jfdc.2096-742X.2020.02.007.
Zhang Yaonan, Ai Minghao, Kang Jianfang, et al..Research on Geoscience Big Data Processing Framework and Key Techniques[J].Frontiers of Data & Coputing,2020,2(2): 91-100.DOI:10.11871/jfdc.issn.2096-742X.2020.02.007.PID:21.86101.2/jfdc.2096-742X.2020.02.007.