Hans Journal of Data Mining 数据挖掘, 2026, 16(2), 72-78
Published Online April 2026 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2026.162007

Hans汉斯

气象数据存储技术研究进展与方向

张　平

山东省气象数据中心，山东 济南

收稿日期：2026年3月9日；录用日期：2026年4月2日；发布日期：2026年4月13日

摘　要

随着气象数据规模快速增长，存储技术面临多源异构、高并发、长期保存等多重挑战。文章系统梳理了近年来气象数据存储领域的研究方法、主要创新和不足之处，并展望了未来潜在的热门研究方向。重点分析了数据湖与云原生存储、分布式与混合存储架构、数据智能处理与质量控制、数据安全与传输优化等方面的技术进展，通过性能对比量化分析了不同技术方案的优劣。总结了存储架构、介质优化、智能处理及共享机制等方面的创新成果，指出了标准化不足、系统融合困难、人工智能应用深度有限等现存问题，并从智能分级存储、云边端协同、人工智能赋能、数据要素化支撑及新型存储介质等角度探讨了未来发展趋势。

关键词

气象数据存储，分布式存储架构，云原生技术，人工智能，数据安全管理

Research Progress and Directions in Meteorological Data Storage Technologies

Ping Zhang

Shandong Meteorological Data Center, Jinan Shandong

Received: March 9, 2026; accepted: April 2, 2026; published: April 13, 2026

Abstract

The rapid expansion of meteorological data presents significant challenges for storage technologies, including multi-source heterogeneity, high-concurrency access, and long-term preservation. This paper reviews recent methodologies, innovations, and limitations in meteorological data storage, and prospects potential future research directions. It focuses on analyzing technological advances in data lakes and cloud-native storage, distributed and hybrid storage architectures, intelligent data

文章引用: 张平. 气象数据存储技术研究进展与方向[J]. 数据挖掘, 2026, 16(2): 72-78.
DOI: 10.12677/hjdm.2026.162007

processing and quality control, as well as data security and transmission optimization, and quantitatively compares the advantages and disadvantages of different technical solutions through performance benchmarking. The study summarizes innovative achievements in storage architecture, media optimization, intelligent processing, and sharing mechanisms, identifies existing problems such as insufficient standardization, difficulties in system integration, and limited depth of artificial intelligence applications, and discusses future development trends from the perspectives of intelligent tiered storage, cloud-edge-device collaboration, AI empowerment, data factorization support, and novel storage media.

Keywords

Meteorological Data Storage, Distributed Storage Architecture, Cloud-Native Technology, Artificial Intelligence, Data Security Management

1. 关键技术现状

1.1. 数据湖架构与云原生存储

数据湖架构已成为应对气象数据多样性与大规模增长的主流方案，通过构建统一数据存储池，集中管理多源异构原始数据，有效解决传统数据孤岛问题。中国气象局采用数据湖技术建立了高效的数据存储、共享和服务体系[1]。在实现层面，通过 NAS 存储与虚拟目录一一映射机制，分配不同虚拟目录的权限，实现数据分级管理与安全管控，提升数据共享效率。

云原生存储是气象数据存储技术的前沿方向。传统基于文件的存储方式(GRIB/NetCDF)在面向在线查询时存在显著瓶颈：每次查询需要打开数 GB 的文件并在磁盘中寻址，即使获取单点时间序列数据也需要扫描大量数据块[2]。为解决这一问题，欧洲中期天气预报中心于 2025 年推出分析就绪云优化(ARCO)数据湖，采用 Zarr 格式存储多维数组数据，显著提升了云环境下的数据访问效率[3]。Zarr 将大型多维数组分块存储为独立的对象，配合压缩技术，客户端仅需获取所需的块。

NOAA 研究团队对比了 GRIB 与 Zarr 格式在 HRRR 集合数据上的加载性能。在串行操作中，两者加载时间相当，但 Zarr 内存占用减少 90%，CPU 使用减少 50%；在异步并行加载(多节点并发)场景下，Zarr 加载速度比 GRIB 快 2 倍，同时内存占用减少 90%，CPU 使用减少 50% [4]。这一性能优势源于 Zarr 的分块存储策略：针对单点时间序列查询场景，WeatherPipe 采用优化的 Zarr 分块布局，相比朴素布局提速 30~60 倍。不同存储格式在典型气象数据访问场景下的性能对比见表 1：

Table 1. Storage format performance in meteorological data access
表 1. 不同存储格式在典型气象数据访问场景下的性能对比

存储格式	访问模式	加载时间(相对值)	内存占用	CPU 使用	适用场景
GRIB/NetCDF	串行读取	1.0×	高(基准)	高(基准)	批量处理、归档
Zarr	串行读取	1.0×	−90.0%	−50.0%	单机分析
GRIB/NetCDF	异步并行	2.0×	高	高	多节点并发访问(不推荐)
Zarr	异步并行	1.0×	−90.0%	−50.0%	分布式计算、云原生应用

1.2. 分布式系统与混合存储架构

为应对气象数据量的指数级增长，分布式存储架构成为研究热点。传统的 POSIX 分布式文件系统(如 Lustre)在处理大规模 I/O 负载时面临性能瓶颈[5]。近年来，对象存储系统如 DAOS (Distributed Asynchronous Object Storage)和 Ceph 在数值天气预报领域的应用展现出显著优势。ECMWF 对 DAOS、Ceph 和 Lustre 在相同硬件环境下的 I/O 性能进行了全面基准测试。结果表明，DAOS 和 Ceph 均表现出优异的性能，其中 DAOS 在可扩展性和 I/O 灵活性方面尤为突出，为大规模数值预报应用提供了理想支撑[6]。这一发现对气象数据中心的基础设施选型具有重要指导意义。

混合存储策略根据数据访问频率特性，结合不同存储介质实现性能与成本的最优平衡。某省气象档案馆采用磁光电混合存储技术，融合磁盘、光盘和固态硬盘优势，实施分级存储策略：高频数据存于固态硬盘，温数据存于磁盘阵列，低频归档数据存于光盘。该方案在保证数据可靠性的同时，有效降低了长期存储成本[7]。具体如图 1 所示。

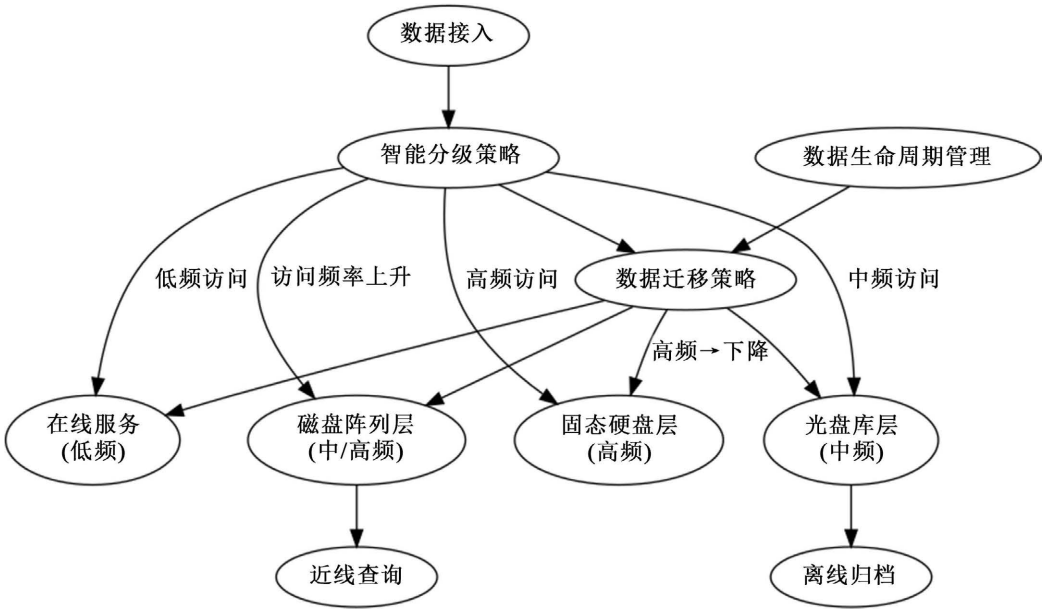

Figure 1. Data flow strategy of magneto-optic-electric hybrid storage
图 1. 磁光电混合存储的数据流转策略

系统基于数据访问频率智能识别数据冷热属性，实现跨存储介质的动态迁移，在保证访问性能的同时优化存储成本。

1.3. 数据智能处理与质量控制

随着人工智能技术的发展，机器学习和深度学习方法在气象数据质量控制与优化领域得到广泛应用。传统质量控制方法基于静态阈值和统计规则，在处理非线性、非高斯分布数据时存在局限[8]。AI 技术的引入显著提升了质量控制的效果。

在可降水量数据处理中，研究者创新性地应用最小协方差行列式(MCD)和隔离森林(Isolation Forest)两种机器学习方法对风云 2E 卫星数据进行质量控制，并将处理后的数据同化到 WRF 模式中评估其对强降水预报的影响。两种方法均显著提升了数据质量，使数据分布更接近高斯分布，并在降水强度、空间分布模拟上取得明显改进。定量分析显示，MCD 方法在预报初期可将均方根误差降低 58%。

在多源数据融合质控方面，基于 XGBoost、LightGBM 等梯度提升模型的算法在强降水质量控制中表现优异。融合自动站观测、雷达、卫星及元数据的模型相比传统方法显著提升质控效果：XGBoost 模型的精确率提高 0.110，召回率提高 0.162，F1 分数提高 0.140。通过 SHAP 可解释性分析，模型的决策逻辑与气象物理原理高度一致，增强了其在业务应用中的可信度[9]。

在纸质历史气象档案数字化中，基于 AI 的手写字符识别技术显著提升了模糊字符的识别准确率与数字化效率，突破了传统 OCR 在气象特殊符号识别方面的局限。

1.4. 数据安全与传输优化

气象数据传输面临网络安全与完整性的双重挑战。研究提出采用加密压缩传输技术，结合 LZ77 算法与数据压缩并行传输，提升传输稳定性与安全性[10]。LZ77 是一种基于字典的无损压缩算法，通过查找输入数据中的重复字符串并用(距离、长度)指针对进行替换，在气象时间序列数据中通常可获得 2:1 至 5:1 的压缩比。实践中，构建双通道线路分散数据流，引入 QoS 服务质量机制，确保关键气象业务数据的优先传输。

全球数据缓存系统是气象数据分发机制的重要创新。北京全球信息系统中心设计的系统采用数据主动发现技术，基于 MQTT 协议构建高性能、低时延的消息收集与传输模型。通过最优路径汇聚与特征信息识别清洗技术，实现数据的快速汇集，为全球用户提供高吞吐、高并发的数据下载服务。截至 2025 年 5 月，系统从全球 82 个 WIS2 节点获取数据，日均缓存约 2 TB，支撑日均下载超 200 万次，体现了在全球气象数据共享中的枢纽作用。

2. 创新点总结

2.1. 存储架构与存取模式创新

传统孤岛式存储架构导致数据分散、共享困难。新一代数据湖架构通过虚拟目录映射技术，实现数据统一管理与分级授权，简化了访问流程，还增强了安全性。欧洲中期天气预报中心的 ARCO 数据湖采用云原生存储理念，将数据预处理为“分析就绪”状态，使用 Zarr 分块格式，显著提升云环境下的数据访问效率，为机器学习和人工智能应用奠定了数据基础。全球数据缓存系统创新性地采用消息驱动的数据主动发现机制，基于 MQTT 协议实现数据主动发现与通知，通过特征值分析实现数据识别与清洗，支撑全球气象数据的实时汇聚与高效分发。

2.2. 存储介质与资源优化创新

磁光电混合存储技术整合不同存储介质的优势，基于数据访问频率实施智能分级存储策略，实现存储性能与成本的最佳平衡，既提升了数据安全性，也为海量历史气象资料的长期保存提供了可行路径。

在分布式系统架构方面，无共享设计理念的应用推动了系统的去中心化与线性扩展。DAOS 对象存储系统相比传统 Lustre 文件系统，在高并发场景下展现出更优的可扩展性和 I/O 灵活性。各节点拥有独立的处理、存储与网络接口，通过动态负载均衡机制优化资源利用率，为气象业务提供高吞吐、可扩展、高稳定的存储解决方案。

2.3. 数据智能处理与质量控制创新

人工智能技术在气象数据质量控制领域取得突破性进展。机器学习与深度学习系统被应用于异常值识别、缺测数据插补等环节，形成了较为完整的技术体系。相比传统方法，AI 质控在效率和准确性上优势显著。如风云 2E 卫星可降水量数据的质控研究中，MCD 方法在预报初期将均方根误差降低 58%；基于梯度提升模型的多源数据质控算法使 F1 分数提升 0.140。在纸质气象档案数字化方面，手写字符智能识别技术结合气象数据特点，设计质量控制标准方法，突破了传统 OCR 在气象特殊符号识别上的局限，

形成了可推广的技术模式。

2.4. 数据安全与共享机制创新

加密压缩并行传输技术通过 LZ77 算法与并行传输机制的有机结合，在保障数据安全的同时提升了传输效率，尤其适用于远程备份与跨地域同步场景。数据要素市场化配置理念的引入，推动了气象数据共享机制的创新。通过构建可信数据空间、数据交易中心与授权运营平台，培育数商生态，促进气象数据在金融、农业等行业的价值释放。

3. 研究结论与局限

3.1. 主要研究结论

数据湖与云原生架构已成为应对气象数据增长的主流方案，能够整合多源异构数据、提升访问效率，特别适用于长时间序列分析场景。Zarr 等云优化格式相比传统 GRIB/NetCDF 格式在并行访问场景下可减少 90%内存占用、降低 50% CPU 使用，加载速度提升 2 倍。分布式系统与混合存储优势明显：DAOS、Ceph 等对象存储系统在高并发访问场景表现突出，磁光电混合存储通过分层策略有效优化存储成本。人工智能技术在气象数据质量控制与处理中作用显著，在质控准确率和预报改进方面展现出自动化处理复杂问题的优势。全球化数据共享架构如全球数据缓存系统，显著提升了气象数据的实时交互与分发效率，有力支持了全球化应用需求。

3.2. 存在不足与局限性

标准化与互操作性不足严重制约数据交换与系统兼容。ARCO 等云优化格式与传统 GRIB/NetCDF 格式之间的转换存在障碍，影响了技术的平滑演进。跨系统数据融合能力有限，多源数据的时空匹配与一致性保障仍依赖经验调整，缺乏坚实的物理基础。

人工智能应用仍处于初级阶段。现有 AI 质控模型的可解释性不足，尽管已有研究尝试 SHAP 等方法进行分析，但模型决策的物理一致性、对训练数据的依赖性等问题亟待解决。当前模型往往针对特定数据类型或区域训练，泛化能力有待验证。

存储系统的资源消耗与经济性问题突出。分布式系统的多副本机制虽然增强了数据可靠性，但也显著增加了存储成本。混合存储中的数据迁移增加了系统复杂度，数据在不同存储层间迁移时的策略优化、迁移时机选择等问题仍需深入研究。

数据安全机制有待加强。现有研究多聚焦于传输加密，对存储本体的防护、隐私保护、跨境传输合规性等方面的研究相对不足。随着气象数据要素化进程加速，如何在保障安全的前提下促进数据流通共享，成为亟待解决的问题。

4. 未来具有潜在价值的热门研究方向

4.1. 智能分级存储与成本优化

未来的研究将聚焦于存储成本与效率的平衡。基于人工智能的数据价值评估与自动分层存储技术有望成为研究热点。通过机器学习模型预测数据的访问频率与价值，实现数据在不同存储介质间的智能迁移，在保证访问性能的同时最小化存储成本。绿色存储技术以及数据生命周期智能管理系统也将受到更多关注，旨在降低数据中心能耗、优化资源使用效率。

4.2. 云边端协同存储与处理架构

随着边缘计算与物联网的发展，云边端协同存储架构将成为重要研究方向。如图 2 所示，该架构涵

盖边缘节点的轻量预处理、断点续传、数据去重以及云端的统一编目与全局索引等功能。研究重点包括边缘与云端的高效同步机制、基于数据价值的边缘缓存策略，以及端到端的时延优化等。边缘层负责实时数据的预处理和质量控制，通过数据去重和断点续传机制确保数据完整性；云端层承担统一存储、全局编目和深度分析职能，形成端–边–云协同的有机整体。

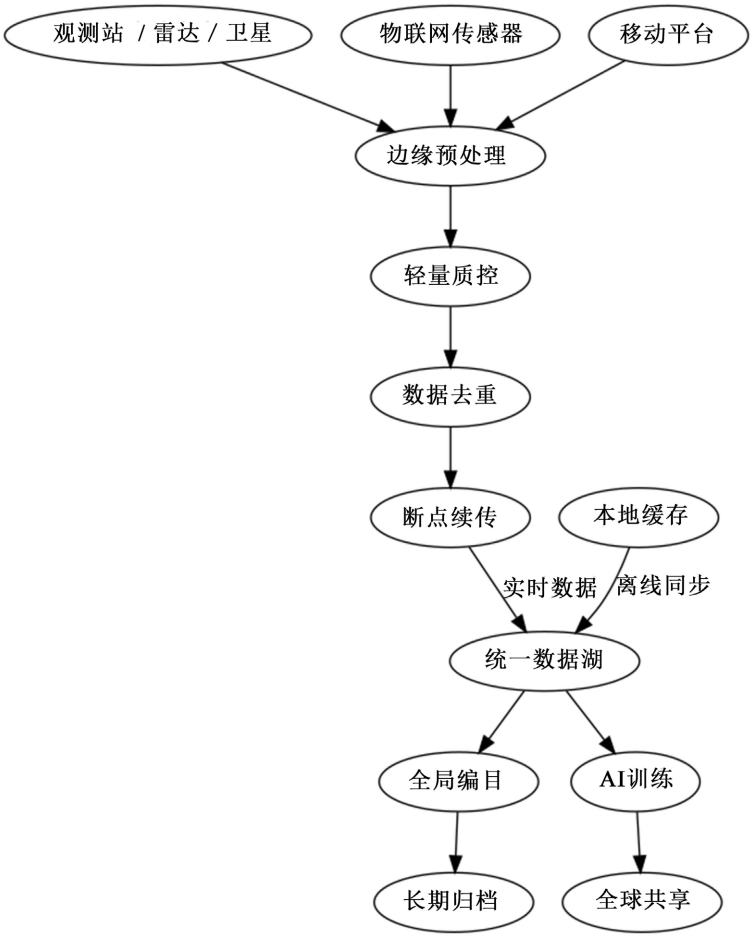

Figure 2. Cloud-edge-device collaboration for data processing and storage
图 2. 云边端协同的数据处理与存储架构

4.3. AI 赋能的气象数据管理

人工智能将实现更深度的融合，形成以 AI 为原生的气象数据管理范式。研究方向包括：基于深度学习的数据质量控制(如利用注意力机制识别异常模式)、利用生成对抗网络进行数据插补与降尺度、运用图神经网络开展多源数据融合、通过强化学习优化存储策略与数据布局等。气象大模型与存储系统的协同优化也将成为重要课题，探索如何将 AI 模型的访问模式与存储布局相互适配，最大化整体效能。

4.4. 面向数据要素化的存储与共享机制

随着数据要素市场化的推进，支持数据使用权控制、安全多方计算、联邦学习等技术的存储方案将受到广泛关注。面向数据要素化的存储架构需要原生支持细粒度的访问控制、可审计的数据使用记录、“数据可用不可见”的隐私保护机制。数据资产编目与估值方法、隐私计算技术在气象数据共享中的应

用也将成为未来研究的重点内容。

4.5. 新型存储介质与架构的探索

从长远来看，DNA 存储、全息存储等前沿技术在超长期气象档案保存中的应用潜力值得深入探索。这些介质具有极高的存储密度和超长的保存寿命，特别适合需要百年保存的历史气象资料归档。存储计算一体化架构、近存储处理与存内计算等技术，可能成为未来高频气象数据分析的重要技术支撑，通过将计算卸载到存储端，显著减少数据移动开销，提升分析效能。

5. 总结与展望

本文系统分析了近年来气象数据存储领域的关键技术现状、创新点与局限性，并通过性能对比量化评估了不同技术方案的优劣，展望了未来潜在的研究方向。技术发展呈现出从分散到统一、从静态归档到智能服务、从成本中心到价值源泉的清晰脉络。

当前，数据湖与云原生架构有效解决了数据孤岛与云适配问题，Zarr 等云优化格式相比传统格式在并行访问场景下可减少 90%内存占用、提升 2 倍加载速度；分布式对象存储(DAOS、Ceph)与混合存储架构应对了规模扩展与成本控制的挑战；人工智能技术显著提升了数据质量与价值密度，机器学习质控方法可将预报均方根误差降低 58%。然而，标准化不足、系统融合困难、AI 应用深度有限、资源消耗大及安全机制薄弱等问题仍是制约发展的主要瓶颈。

未来，随着数据要素化进程加速和全球气候变化应对需求的加剧，气象数据存储技术将向更智能、高效、安全、绿色的方向演进。智能分级存储、云边端协同、AI 原生管理、数据要素化支持及新型存储介质等方向具有重要的科研价值，将为气象事业发展提供坚实的技术基础，增强全球气候监测、极端天气预警与精细化预报的能力。

参考文献

[1] 国家气象信息中心. 气象大数据云平台数据存储架构设计与应用[J]. 气象科技进展, 2022, 12(6): 45-52.

[2] Clima Links (2025) Introducing Weather Pipe: From Raw Forecasts to an API. https://www.climalinks.com/blog/weatherpipe

[3] European Centre for Medium-Range Weather Forecasts (ECMWF) (2025) Work in Progress: Our Data Stores Turn ARCO. https://climate.copernicus.eu/work-progress-our-data-stores-turn-arco

[4] NOAA (2024) Benchmarking Zarr vs GRIB for Weather Data Access in Cloud Environments. NOAA Technical Memorandum.

[5] Manubens Gil, N. (2025) Exploring Novel Data Storage Approaches for Large-Scale Numerical Weather Prediction. The University of Edinburgh. https://arxiv.org/abs/2602.17610

[6] 吴京生, 张驰. 磁光电混合存储在气象档案备份中的应用——以浙江省气象档案馆为例[J]. 浙江档案, 2025(3): 49-51.

[7] Shen, W., Chen, S., Xu, J., *et al*. (2025) Enhancing Extreme Precipitation Forecasts through Machine Learning Quality Control of Precipitable Water Data from Satellite FengYun-2E: A Comparative Study of Minimum Covariance Determinant and Isolation Forest Methods. *EGU General Assembly* 2025, Vienna, 27 April-2 May 2025, EGU25-3931.

[8] Sun, H., Zhou, Q., Shi, L.J., *et al*. (2025) A Machine Learning-Based Quality Control Algorithm for Heavy Rainfall Using Multi-Source Data. https://www.preprints.org/manuscript/202510.1276

[9] Xu, M.X. (2020) A Multidimensional Array Database Engine for Gridded Climate Data and a Precipitation Downscaling Study. George Mason University. https://hdl.handle.net/1920/12461

[10] 王宁, 黄伟, 刘俊宏. 气象数据加密传输系统的设计与实现[J]. 计算机科学与应用, 2024, 14(6): 88-95.