面向AI-Ready数据制备的冰冻圈野外观测数据接引质量治理方法

张智星[1]，康建芳[2,3]，路长发[4]，赵彦博[2,3]
1. 深圳技术大学中德智能制造学院，广东 深圳 518118；
2. 国家冰川冻土沙漠科学数据中心，甘肃 兰州 730000；
3. 中国科学院西北生态环境资源研究院，甘肃 兰州 730000；
4. 中国科学院计算机网络信息中心，北京 100083

摘要
冰冻圈野外观测数据常存在重复、缺失和异常等多种质量问题，影响AI-Ready数据制备和建模应用。针对原始观测数据从“可读取”到“可建模”过程中质量状态难以显式表达、处理过程难以追溯的问题，构建了面向AI-Ready数据制备的数据接引阶段的质控流程。该流程通过规则质控和统计检测相结合的方法，对确定性质量问题和疑似异常进行分层识别、质量标记与输出。基于典型数据集的实证结果表明，该流程能够识别重复、缺失、超范围和局部异常候选等多类质量问题，并通过标准化质量语义字段输出，为AI-Ready数据集制备和可追溯建模提供数据质量治理基础。

关键词
冰冻圈野外观测数据；AI-Ready数据；数据接引质控；质量标记；数据质量治理

中图分类号：TP391.1 文献标志码：A doi：10.11959/j.issn.2096-0271.BDR26245

Quality governance method for ingesting cryospheric field observation data for AI-Ready data preparation

Zhang Zhixing[1], Kang Jianfang[2,3], Lu Changfa[4], Zhao Yanbo[2,3]
1. Sino-German College of Intelligent Manufacturing, Shenzhen Technology University, Shenzhen 518118, China
2. National Cryosphere Desert Data Center, Lanzhou 730000, China
3. Northwest Institute for Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China
4. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China

Abstract
Cryosphere field observation data often contains duplicate, missing, and anomalous records, which affect AI-Ready data preparation and modeling applications. To address the difficulty of explicitly representing quality states and tracing processing actions when raw observations move from readable files to model-ready data, a quality control workflow for the data ingestion stage of AI-Ready data preparation was constructed. By combining rule-based quality control with

statistical detection, the workflow could hierarchically identify, label, and output deterministic quality issues and suspected anomaly candidates. Empirical results on representative datasets showed that the workflow could identify multiple types of quality issues, including duplicates, missing values, out-of-range records, and local anomaly candidates. These standardized quality-semantic fields provide a data ingestion quality governance basis for AI-Ready dataset organization and traceable modeling.

Key words
cryosphere field observation data, AI-Ready data, data ingestion quality control, quality flag, data quality governance

0 引言

冰冻圈野外观测数据是揭示寒区水热过程、评估冰冻圈变化趋势、支撑气候模式验证和开展生态环境风险评估的重要基础[1-3]。随着冰川、积雪、冻土自动气象站和多源环境监测设备的持续部署，冰冻圈研究积累了大量时间分辨率较高、变量类型较多的原位观测数据。这类数据不仅支撑传统过程分析，也逐渐成为通过人工智能方法开展参数反演、时序预测和多源融合建模的重要基础。

然而，野外自动观测数据通常直接来自长期无人值守设备，受高寒低温、冻融循环、供电波动、传感器漂移、通信中断和现场维护条件限制等因素影响，原始记录中常出现缺测、重复、错误码、超范围值、连续平台和局部跳点等问题。若这些问题未经系统治理便进入样本构建和模型训练流程，模型可能将采集故障或处理残留误学习为真实过程特征，进而影响预测的稳定性和科学解释的可靠性。

现有质量控制方法大体可分为规则方法、统计方法和数据驱动方法。规则方法在重复检测、范围检查和缺失识别方面具有计算开销小、可解释性强和便于在线运行等优势[4-5]，但对局部跳变、上下文异常和统计异常的识别能力有限；统计方法通过分布假设检验、滑动窗口和局部稳健估计识别偏离正常模式的观测值[6]，能够在不标注数据的情况下检测局部异常，但阈值设定依赖经验且对缓变趋势不够敏感；数据驱动方法能够利用时序上下文和多变量信息提高复杂异常识别能力[7-8]，但往往依赖标签数据，而且难以脱离基础质控独立发挥作用。对于冰冻圈野外观测场景，异常往往同时具有统计特征和物理背景，单一阈值或单一模型均难以覆盖全部质量问题。

本文面向AI-Ready数据制备前数据接引阶段的质量控制需求，以5个冰冻圈代表性较强的野外观测站数据为例，以其TOA5（一种数据采集器ASCII输出格式）等原始观测文件为处理对象，围绕原始数据解析、字段识别、去重入库、规则质控、疑似异常候选识别和质量标记输出展开。本文不以提出新的异常检测算法为目标，而是关注不同质量问题在数据接引治理流程中的处理：重复、缺失、错误码和超范围值等确定性问题应由规则直接标记，局部跳变、连续平台和趋势变化等问题作为候选事件保留识别依据，并进入后续筛选或复核环节。

1 数据接引质量问题

1.1 AI-Ready数据内涵

近年来，随着数据驱动方法和大规模

预训练模型在地学中的快速应用，“高质量数据”和“AI-Ready数据”逐渐成为科学数据治理的重要议题。已有研究指出，AI系统在实际场景中的表现很大程度上取决于训练数据的可用性、可解释性与可追溯性，而不仅取决于模型本身的复杂度[9-10]。在通用机器学习语境中，AI-Ready通常强调数据能够进入训练管线，包括样本对齐、特征规整、标签明确和数据文档完整等要求[10-11]；在科学数据语境中，相关讨论则更多与FAIR原则[12-13]、数据质量维度[14-15]和地理空间数据质量标准[16]相联系。

参考FAIR原则、ISO 19157数据质量标准[16]、数据质量维度研究[15]、高质量数据集产品形态研究[17]以及可信AI数据治理讨论[9-11,18]，本文将面向冰冻圈野外观测的AI-Ready数据界定为：在原始观测数据基础上，经统一解析、质量识别和元数据保留后，能够稳定支撑下游样本筛选、模型训练与可重复性验证基本需求的数据集合。本文界定的AI-Ready仅关注原始观测数据进入下游标准化治理之前的接引治理阶段，并不涵盖缺失补全、时间对齐、统一封装和样本窗口构建等终态环节。

1.2 典型代表性数据及质量问题

本文选取5个覆盖冰冻圈环境的不同站点、设备类型和采样频率的观测数据集作为研究对象，分别记为D1～D5。5个观测数据集的基本信息与规模统计见**表1**。各数据集的时间分辨率包括12 h、24 h和30 min，变量类型涵盖设备状态、气象变量、积雪变量和土壤变量等。D1为多通道热敏电阻观测数据，可用于分析多通道记录中的重复问题；D2为CR800日值气象观测数据，包含雪深、风速、气温、湿度、气压和辐射等要素，主要用于无效值和超范围值检查；D3为CR1000高频观测数据，记录规模较大，适合分析高频记录中的重复和缺失问题；D4为土壤剖面观测数据，包含多层土壤含水量、电导率和土壤温度等变量，可用于字段缺失和连续缺测识别；D5为综合塔站观测数据，覆盖气象、辐射、雪深及多层土壤温湿度等要素，本轮列出的规则检查项未触发明显确定性问题。上述数据集并非用于证明方法的统计普适性，而是用于覆盖数据接引治理中常见的数据结构和质量问题类型。

冰冻圈野外观测变量在物理语义上可分为3类，各类数据在数据接引过程中呈现不同的质量问题模式：连续气象要素（温度、湿度、辐射等）要求具有时间连续性，缺测、超限和局部跳变是主要问题；累积量（降水量、蒸发量等）具有单调非

表1 5个观测数据集的基本信息与规模统计

观测数据集	数据类型	采样频率	时间范围	记录数	主要问题
D1	Test1_04多通道热敏电阻观测数据	12 h	2025-12-12至2026-03-07	339	重复记录明显
D2	CR800日值气象观测数据	24 h	2025-11-01至2026-03-07	127	无效值与超范围风速
D3	CR1000高频观测数据	30 min	2025-11-01至2026-03-05	11 944	重复与无效值并存
D4	土壤剖面观测数据	30 min	2025-11-01至2026-03-07	6 089	字段缺失与连续缺测
D5	综合塔站观测数据	30 min	2025-11-21至2026-03-07	5 103	本轮规则检查未触发明显确定性问题

减特征，单调性破坏（如负增量）是典型的确定性质量问题；状态类指标（土壤冻融状态、积雪覆盖等）取值离散，异常高频跳变是重要的识别对象。上述语义差异为不同类型变量选择适用的质量治理规则提供了依据，也有助于将本文方法推广至其他野外观测场景。

1.3 原始数据格式

本文处理的原始观测数据以Campbell采集器输出的TOA5格式为主，该类文件通常同时包含站点、采集器型号、程序名、字段名、单位、统计方式以及按时间顺序排列的观测记录。不同台站和设备配置会带来字段数量、变量命名和单位标注差异，部分文件还存在非标准表格结构。对这类数据而言，人工打开文件并不等同于机器可读，也不足以支撑批量质控和后续建模。数据接引解析的任务是在不改变原始观测值的前提下，提取时间字段、变量字段、单位信息和记录结构，将原始文件转换为可批量读取、字段含义明确且能够保留来源信息的结构化记录。本文的字段解析过程对照冰冻圈野外观测规范语义，将TOA5文件中的原始字段名映射为规范命名，确保入库字段含义与观测规范保持一致。

1.4 质量问题类型

本文聚焦原始观测数据进入AI-Ready制备流程前常见的典型质量问题。根据判定确定性差异，这些问题可分为两类：一类为可由规则明确识别的问题，包括重复记录、缺失值、错误码、字段缺失和物理范围超限等，可通过规则质控直接标记和处理；另一类为需要进一步分析的疑似异常，包括局部跳点、短时跌落、连续平台、低方差段等，通常难以仅依靠固定规则判定，需要结合局部统计特征、连续段检测和受控异常实验进行辅助识别。需要说明的是，本文并不试图覆盖冰冻圈观测数据中的全部异常类型，而是着眼于数据接引质量治理流程中常见且会直接影响后续样本构建的问题，并区分两类处理方式：确定性问题可通过规则识别直接标记，疑似异常则更适合作为候选事件输出并保留复核信息。

2 数据接引质量治理流程

2.1 总体流程

数据接引质量治理流程首先将原始观测文件解析为结构化记录，再通过去重入库保留数据来源、字段元数据和处理状态。在此基础上，规则质控用于识别重复记录、缺失值、错误码和物理范围超限等确定性问题；局部统计和连续段检测用于生成疑似异常候选。最终输出不直接覆盖原始观测值，而是以质量标记记录问题类型、触发依据、处理建议和事件编号。数据接引质量治理流程如**图1**所示，其核心作用是将原始文件转化为后续可筛选、可复核的数据对象，而不是在数据接引阶段完成全部AI-Ready数据产品封装。

2.2 原始数据解析与字段识别

原始数据解析主要面向.dat/TOA5文件和部分非标准表格文件，处理过程中提取字段名、单位和统计方式，并对时间字段、数值字段和错误码进行统一转换，无法解析的单元、空值和设备错误码被映射为缺失或无效标记，以便后续规则检查能够在统一的数据结构上运行。该步骤解决

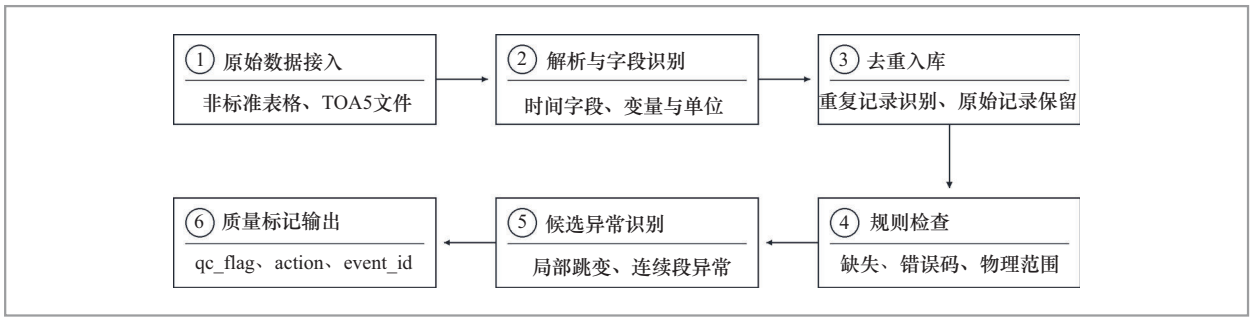


图1 数据接引质量治理流程

的是野外观测文件从“人工可查看”到“机器可批量检查”的转换问题，这也是质量标记和事件追踪能够建立的前提。

2.3 去重入库与记录保留

字段识别完成后，标准TOA5数据将时间戳和记录号的组合作为联合主键进行重复识别；部分非标准结构文件缺少稳定记录号时，则依据时间戳进行重复判定。去重后的数据增量导入SQLite数据库，入库过程保留原始记录、字段元数据、来源文件和质控标记字段。这样处理的目的并非提前修正所有异常值，而是为后续规则检查、统计分析和复核提供可追溯的数据基础。

2.4 规则质控与统计输出

规则质控覆盖完整性检查和有效性检查两个方面。完整性检查包括重复记录统计、缺失值统计、字段缺失识别和时间断点检查；有效性检查依据变量物理阈值识别风速、气温、雪深、土壤含水量和设备状态字段中的超范围问题。规则检查的输出包括分数据集质量统计、典型问题样例和异常事件清单。这些结果用于确定哪些问题可以在数据接引阶段直接标记，哪些问题需要进入后续候选识别或人工复核。

2.5 质量标记输出设计

为使数据接引质控结果能够被后续样本筛选、建模和复核流程使用，本文将质量识别结果表示为结构化质量标记。该标记不直接覆盖原始观测值，而是在保留原始记录的基础上，对质量状态、识别依据和处理建议进行显式记录。质量标记输出字段设计示例见**表2**。

表2 质量标记输出字段设计示例

字段	含义	示例
timestamp	观测时间	2026-01-20 12:00:00
variable	变量名	RH_1m_Avg
original_value	原始观测值	68.42
qc_flag	质量标记	MISSING/OUT_OF_RANGE/SUSPECTED_ANOMALY
qc_detail	识别依据或参数说明	rule=range; limit=[0,100]
action	建议处理动作	keep/mask/review
event_id	异常事件编号	D5_EVENT_0007

表2中，qc_flag用于描述质量问题类型，qc_detail用于记录触发规则、统计窗口或阈值等辅助信息，action表示后续处理建议。对于连续缺测、连续平台或局部异常区段，event_id可将多个记录关联为同一异常事件。重复记录、缺失值、错误码和物理范围超限等确定性问题可由规则质控直接标记；局部跳点、短时跌落和连续平台等疑似异常则不在数据接引阶段直接判定故障原因，而是作为候选事件输出，并保留识别依据供后续筛选或复核使用。

3 质控方法与受控异常基准

3.1 质控组合方法

本文的异常识别方法选择遵循数据接引质控场景下的可解释、低成本和易复核原则：对于重复记录、缺失值、错误码和物理范围超限等确定性问题，采用固定规则进行直接识别和标记，这与气象观测资料质量控制中常用的规则检查思路一致[4-5]；对于局部跳点、短时跌落和连续段异常等难以由固定阈值直接判定的问题，采用3σ规则、Hampel滤波和局部稳健统计方法进行候选识别。上述方法能够与原始观测数据解析、质量标记和后续样本筛选过程直接衔接，便于保留触发依据和处理记录。极端梯度提升（extreme gradient boosting，XGBoost）、孤立森林（isolation forest，iForest）、局部离群因子（local outlier factor，LOF）和长短期记忆异常检测（long short-term memory anomaly detection，LSTM-AD）等模型方法在异常检测研究中已有广泛应用[19-23]，本文仅将其作为复杂多变量背景下异常候选生成的相关方法参照，不开展模型性能排序。

具体到受控注入实验，跳点和跌落点采用以21点滚动中位数和中位数绝对偏差（median absolute deviation，MAD）为基础的Hampel判别，候选阈值$|z|>4.5$；连续平台、高值平台段和低值平台段采用连续段检测，近零连续段采用接近零值的持续时间规则，连续事件至少包含3个观测点。上述参数用于形成可复核的候选标记，并不作为面向所有变量的通用最优阈值。

3.2 受控异常基准构建

真实业务数据中可明确定位的标注异常样本数量有限，因此，本文采用受控注入方式构建异常测试序列，用于评价数据接引质控方法对值级异常和连续段异常的识别能力。具体做法是在真实观测序列的基础上，保持原有时间顺序和记录行数不变，仅修改局部时间段内的观测值，模拟缺测、跳点、跌落、连续平台、近零段和高/低值平台段等异常形态，并同步记录异常位置与类型作为评价参照。本轮实验共选取10组真实观测序列，构造67组含异常测试序列，形成536个异常事件，真值标签由原始序列与注入序列逐行比较生成。重复记录、格式错误等结构性问题会改变行结构，不适合纳入逐行对齐的受控注入实验，仍在真实数据规则检查中分析。

注入幅度和持续时间基于原始序列的局部统计特征确定。跳点和跌落点以局部窗口中位数为基准，分别向上或向下偏移；偏移幅度取5.5倍MAD换算标准差与局部有效值极差35%中的较大值。连续段类异常在4～24个观测点的局部窗口内构造，包括连续缺测、连续平台、低幅值持

续段以及高/低值平台段。其中，连续缺测以空值表示，连续平台设定为局部中位数附近的常值，低幅值持续段设定为接近0的低值，高值和低值平台段分别对应原序列95%和5%分位数附近的常值。为保证注入事件具有可评价性，本文仅保留实际变化点数不少于注入窗口长度80%且至少包含3个观测点的连续段事件，以避免原始值本身接近注入值时形成难以区分的样本。

这种构造方式的作用不在于模拟所有可能出现的真实故障，也不以优化某一类质控算法为目标，而是为数据接引质量治理流程提供位置明确、类型可控的参照样本。连续缺测、连续平台和高/低值平台段具有持续时间属性，可用于检验质量标记能否以事件形式覆盖异常区段；跳点和跌落点表现为短时局部偏离，可用于检验局部统计方法能否生成候选标记。

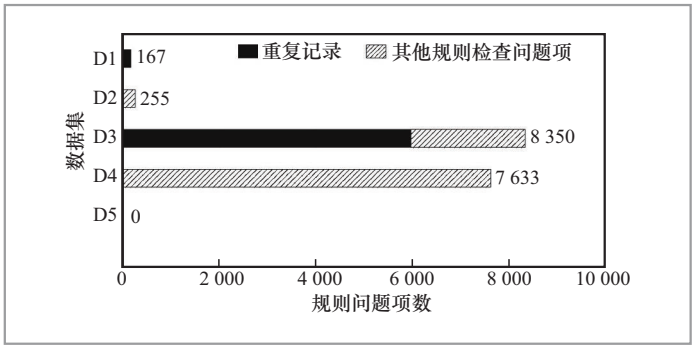


图2 数据集规则检查结果

表3 数据集规则质控结果统计情况

数据集	重复记录数	无效值/错误码数	超范围点数
D1	167	0	0
D2	0	254	1
D3	5 972	2 378	0
D4	0	7 633	0
D5	0	0	0

3.3 评价指标

受控注入数据的特点是异常发生的位置和类型已知，因此，可用于检查不同方法是否能够识别这些人为设置的问题。连续缺测、连续平台和高/低值平台段通常表现为一段持续异常，评价时更关注该区段能否作为一个质量事件被识别。跳点和跌落点多表现为单点或短时偏离，评价时则更关注被改动的观测位置是否被标记。基于这一思路，本文统计事件检出情况和注入位置覆盖情况，用于分析不同方法对典型异常形态的识别能力。由于原始观测序列中未被人工改动的位置并不一定完全正常，本文不将这些位置上的候选标记直接解释为误报，而是统计其数量并折算为每万点候选数，用于估计后续复核或样本筛选需要处理的候选规模。

4 实验结果与分析

4.1 数据集规模与规则检查结果

5个样例数据集触发的规则检查项差异明显，数据集规则检查结果如**图2**所示。数据集规则质控结果统计情况见**表3**。D1和D3主要表现为重复记录，其中D3同时包含较多无效值或错误码；D2记录数较少，但存在254个无效值或错误码以及1个超范围点；D4未触发重复和超范围检查，却出现7 633个无效值或错误码；D5在本轮列出的重复、无效值/错误码和超范围检查项中均未触发标记。上述结果说明，数据接引质控首先需要稳定处理具有明确判定依据的质量问题，并将其与后续疑似异常候选区分开。从AI-Ready数据制备的

角度看，确定性质量问题（如D1/D3的重复记录、D4的无效值）的质量标记（qc_flag）可直接用于下游的样本筛选，避免将已知存在质量问题的数据纳入训练集。

4.2 规则检查识别的典型质量问题

规则检查能够直接定位部分具有明确判定依据的质量问题。以样例数据为例，D2中ws_Max出现1 000 m/s的超范围值远高于设定物理阈值，可由范围规则直接标记；D4中，部分字段缺测率较高，说明质量问题可能发生在字段层面，而不仅是单条记录异常；连续缺测区段则表明缺失可能形成影响样本构建的时间窗口。这类问题具有明确的触发依据，适合在数据接引质控阶段直接记录变量、时间、触发规则和事件信息。规则检查识别的典型质量问题示例如**图3**所示。从数据语义角度看，D2的ws_Max超范围问题属于气象要素的物理合理性问题，1 000 m/s的风速值远超设备量程和物理上限，规则可直接判定；D4的字段级缺测（VWC1、VWC2字段缺测率远高于其他字段）说明质量问题发生在传感器通道层面，而非整行记录，字段级标记可以保留其他同时刻的有效字段；连续缺测区段（D4中VWC3字段出现的连续缺失窗口）则影响样本构建的时间连续性，适合以event_id关联为事件整体进行处理。

4.3 受控注入异常识别结果

规则检查能够处理缺失、错误码和超

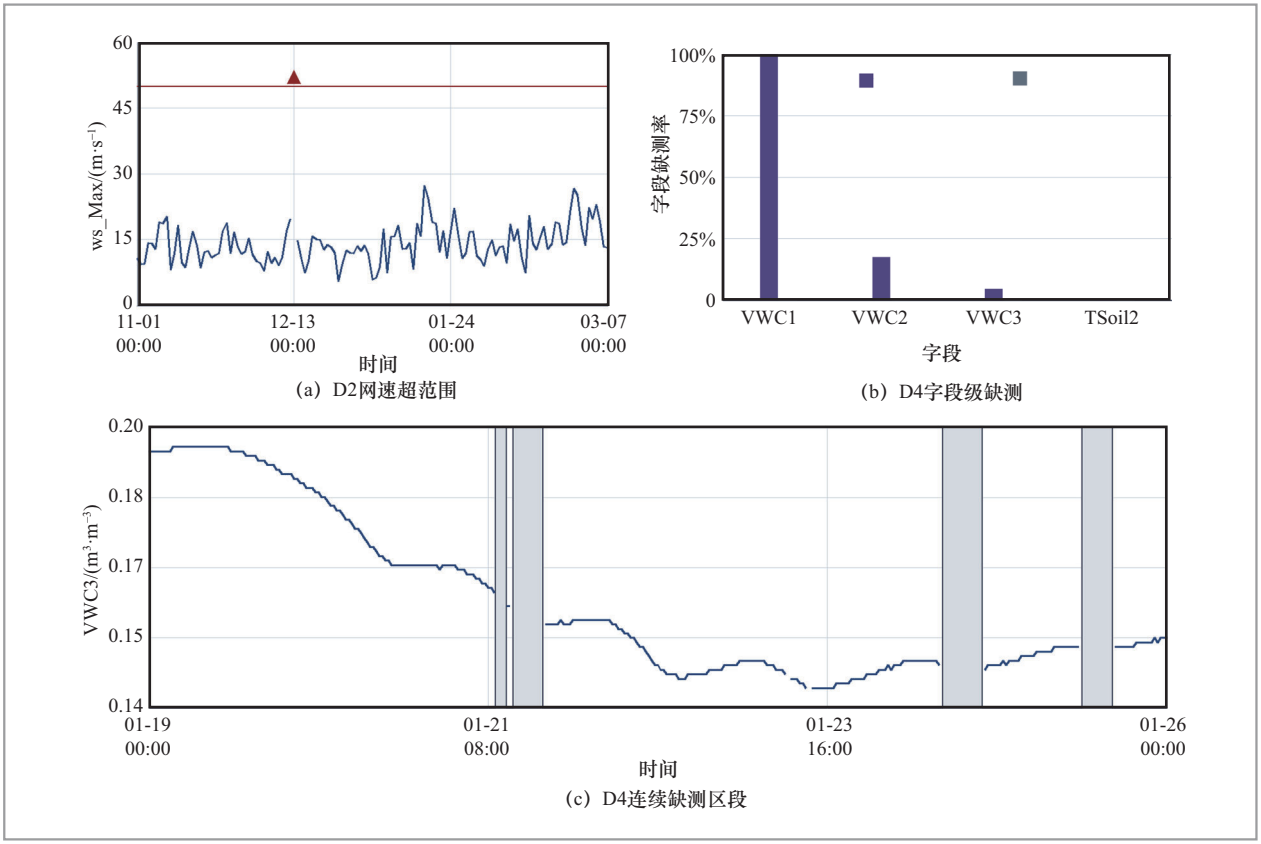


图3 规则检查识别的典型质量问题示例

范围值等具有明确判定依据的问题，但局部尖峰、短时跌落和连续平台等疑似异常在原始观测中通常缺少人工标注位置，难以直接评价不同方法的识别效果。为获得位置和类型明确的参照样本，本文在原始观测序列上进行受控注入，构造局部尖峰、短时跌落、连续平台和缺测区段等典型形态。局部尖峰和短时跌落表现为相对邻近背景的短时偏离，适合由 Hampel 滤波或局部稳健统计方法生成候选标记；连续平台和缺测区段具有持续时间属性，更适合按事件整体标记。受控注入异常示例如**图 4** 所示。图 4 以局部窗口对照方式展示了不同形态的注入效果。**图 4（a）**单点向上偏离局部背景；**图 4（b）**短时向下偏离并持续数个观测点；**图 4（c）**一段时间内取值趋近局部中位数的常值；**图 4（d）**连续观测位置取值为空。每个子图均以原始序列（连线）和改造后序列（散点）对比展示，阴影标注异常或缺测窗口。

受控注入结果进一步显示，异常形态的持续时间属性会明显影响候选识别效果。受控注入异常识别结果如**图 5** 所示，受控注入异常识别结果统计情况见**表 4**。**图 5** 展示了 7 类异常形态的事件检出率和注入点覆盖率。连续缺测、连续平台、近零连续段、高值平台段和低值平台段的事件检出率均为 1.0，说明具有持续时间特征的异常形态更容易被规则或连续段检测稳定识别。其中，连续缺测由连续空值规则直接识别，非注入观测点候选数为 0，这一结果只说明缺测位置能够被准确标记，并不意味着

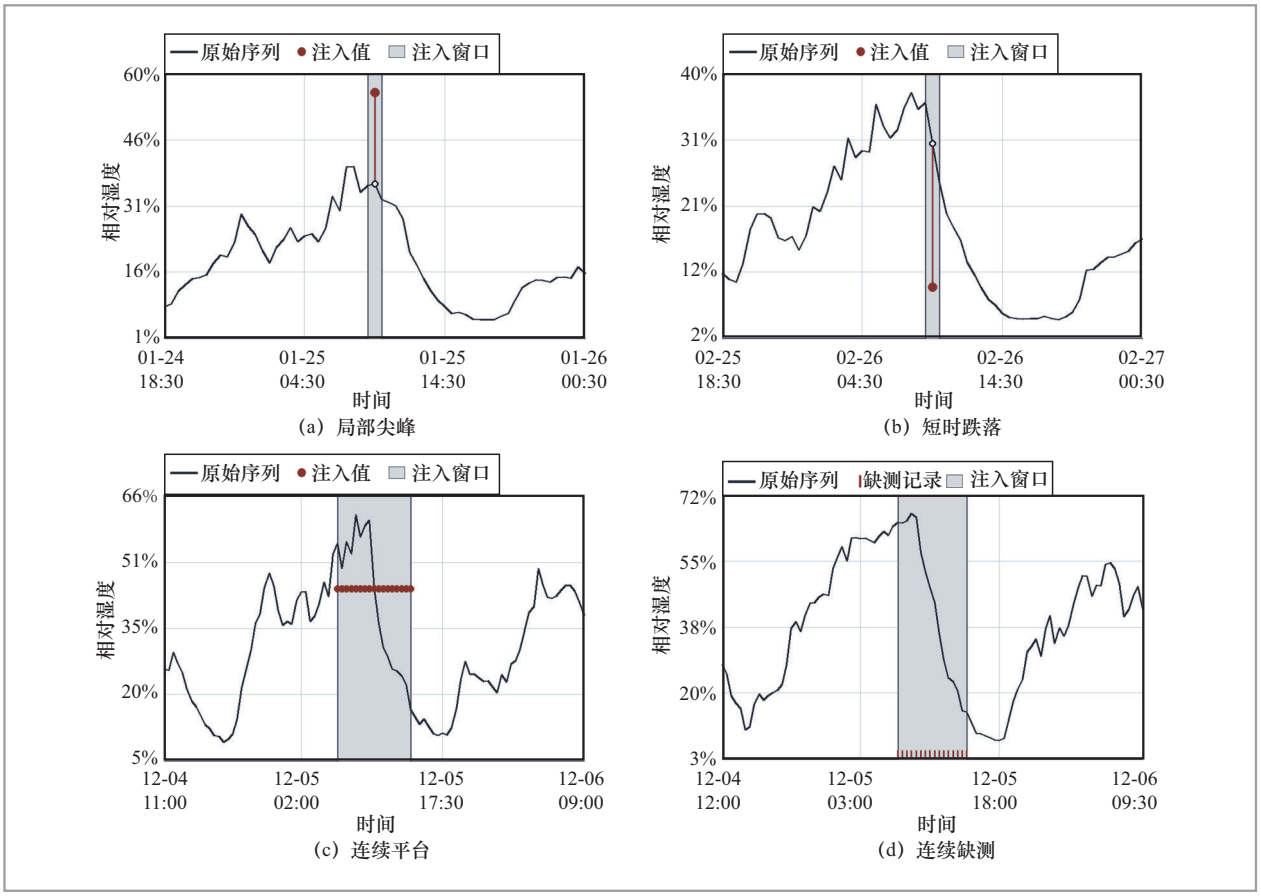


图4 受控注入异常示例

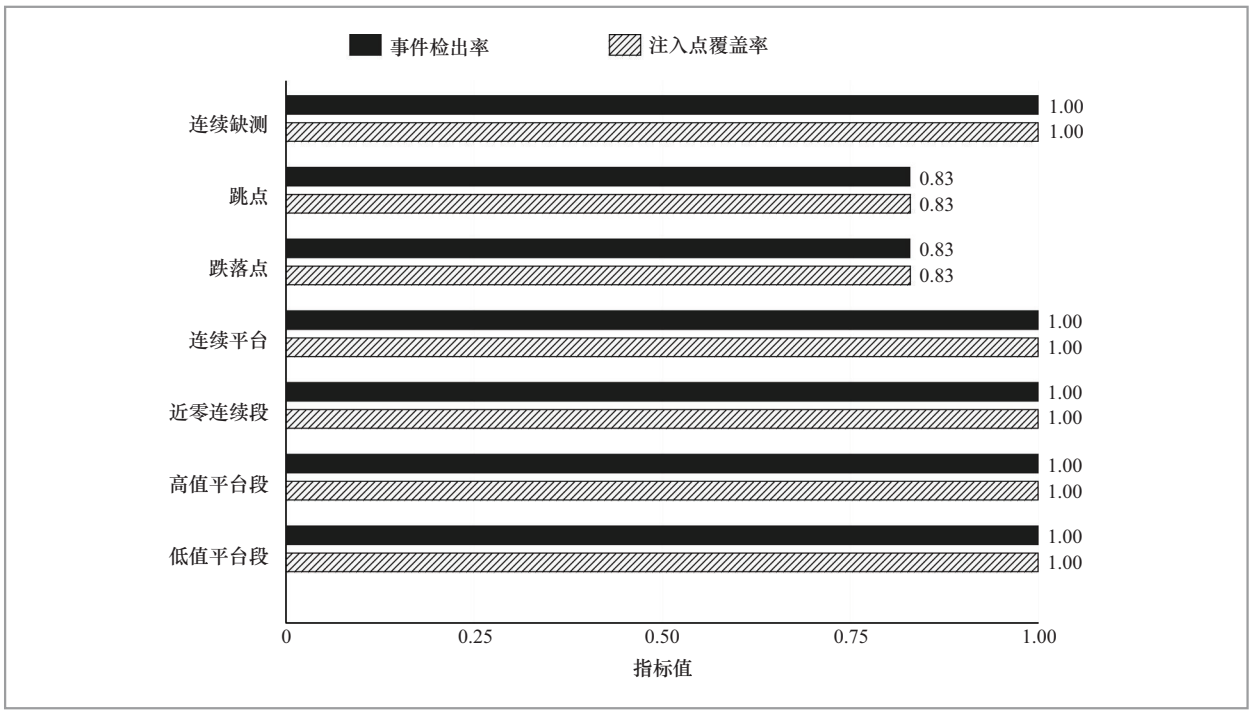


图5 受控注入异常识别结果

表4 受控注入异常识别结果统计情况

异常类型	主要识别方法	注入事件数	注入点数	事件检出率	注入点覆盖率	非注入观测点候选数/(点·万点$^{-1}$)
连续缺测	连续空值规则	70	1 479	1.00	1.00	0
跳点	Hampel/局部MAD	106	106	0.83	0.83	641
跌落点	Hampel/局部MAD	106	106	0.83	0.83	200
连续平台	连续段检测	70	1 479	1.00	1.00	352
近零连续段	近零持续规则	44	933	1.00	1.00	688
高值平台段	平台段检测	70	1 479	1.00	1.00	658
低值平台段	平台段检测	70	1 479	1.00	1.00	477

缺测值可以在数据接引质控阶段被恢复。

连续平台、近零连续段以及高/低值平台段的事件检出率和注入点覆盖率均接近或达到1.00，说明这类具有持续时间属性的问题较容易被连续段规则识别。然而，部分连续段规则在非注入观测点上也会产生较多候选标记，其中近零连续段、高值平台段和低值平台段分别约为688点/万点、658点/万点和477点/万点，这说明相关规则对原始序列中的低值持续、平稳片段或阶段性稳定状态较为敏感。对于野外观测数据而言，这些片段未必都对应仪器故障，因此更适合作为疑似异常事件输出，并保留触发规则、持续时间和变量信息，供后续复核。

跳点和跌落点的事件检出率均为0.83，低于连续段类异常。这两类异常主要表现为单点或短时局部偏离，识别结果更容易受到原始序列波动、统计窗口和阈值设置影响。跳点和跌落点在非注入观测点上的候选标记数分别约为641点/万点和200点/万点，表明跳点类规则更容易在真实观测波动中产生额外候选。该结果支持将局部跳变类问题定位为疑似异常候选，而不是数据接引阶段可直接自动清洗的确定性错误。

从7类异常形态的识别结果来看，持续段类（连续缺测、连续平台、近零连续段、高值平台段、低值平台段）的事件检出率均为1.00，注入点覆盖率均为1.00，说明这类形态在数据接引阶段可稳定识别并完整标记。其中，连续缺测的非注入观测点候选数为0点/万点，标记精准；连续平台段约352点/万点，近零连续段约688点/万点，高值平台段约658点/万点，低值平台段约477点/万点，这些候选在野外观测中可能对应真实的低温冻结、稳定晴空或仪器平稳运行状态，因此更适合作为疑似异常事件输出，保留复核依据而非直接清洗。跳点和跌落点的事件检出率均为0.83，非注入观测点候选数分别约为641点/万点和200点/万点，表明点级异常的识别结果更依赖局部统计窗口和阈值设置，识别结果具有更大的不确定性，定位为候选事件而非确定性错误在操作上更为稳妥。

5 讨论

5.1 面向AI-Ready制备的质量标记

AI-Ready数据集不仅需要保存原始观测值，还需要说明观测值的质量状态、处理依据和来源信息，以便后续进行样本筛选、人工复核和可重复性验证[10-11]。本文建立的数据接引质量治理流程能够在原始台站数据进入人工整理之前识别重复记录、缺测、错误码和超范围值等确定性问题，减少后续排查成本。对于局部跳变、跌落点、连续平台和近零连续段等不宜直接判定故障原因的问题，则以候选事件形式保留触发依据和复核状态，由此形成的字段级质量统计、事件编号和质量语义字段，可作为后续标准化整理、样本筛选和建模准备的结构化输入。具体而言，qc_flag字段可用于训练集筛选，快速排除确定性质量问题记录；event_id可作为异常事件分组标识，支持将连续段异常作为整体排除或保留；qc_detail保留了候选标记的触发规则和参数，便于数据使用方根据自身建模需求决定是否接受特定类型的疑似异常数据。

5.2 接引数据质控的作用范围

本文的AI-Ready主要体现在数据接引治理层面，其作用是为后续数据产品化提供质量依据，而不是直接生成可用于模型训练的终态数据集。数据接引治理能够回答的是原始记录是否可读、字段含义是否明确、质量问题是否被标记、处理过程是否可追溯等问题，对于时间对齐、单位统一、缺失补全、变量语义标准化、多变量一致性验证和统一封装，则仍属于后续数据产品化环节。

因此，本文工作的边界不在于完成所有AI-Ready数据产品构建步骤，而在于将原始观测文件进入后续整理之前的质量状态表达清楚。哪些问题可以由规则直接标记，哪些问题只能作为异常候选进入复核，哪些记录不应被直接覆盖，都是数据接引治理需要明确的内容。

6 结束语

本文围绕冰冻圈野外观测数据从原始文件进入样本构建前的质量控制需求开展研究，得到以下3点结论。

（1）在概念层面，本文将面向冰冻圈野外观测的AI-Ready数据集界定为经统一解析、质量识别和元数据保留后，能够支撑后续样本筛选、建模准备和可重复性验证的数据集合，并将其数据接引要求概括为机器可读、字段明确、质量可识别和记录可追溯。本文将质量治理界定为数据进入AI-Ready制备链之前的准入准备阶段，即对野外观测数据进行前期质量管控。

（2）在方法层面，5个典型数据集的规则检查结果表明，冰冻圈野外观测的质量问题分布在记录级、字段级和事件级3个层次，单一粒度的质控标记不足以支撑后续样本筛选。本文构建的数据接引质量治理流程，通过qc_flag、qc_detail和event_id这3个字段的组合设计，使质量状态在3个层次上均可显式表达和追溯。该质控流程不依赖特定数据集或变量类型，可为其他野外长期观测场景下的数据接引治理提供参考。

（3）在受控注入分析中，连续缺测、连续平台、近零连续段和高/低值平台段等持续性异常的事件检出率均达到1.00，而跳点和跌落点的检出率为0.83，且在非注入点上产生更多候选标记。这一结果表明，不同形态的异常在数据接引质控中的适用处理层次存在根本差异：持续段类问题可由规则稳定标记，而点级跳变更适合作为候选事件保留复核依据，而非在数据接引阶段直接清洗。该区分对于构建可追溯的AI训练样本具有实际意义——过早地在数据接引阶段自动清洗跳变类数据，可能导致真实的极端气象事件被误删。

然而，本文仍存在一定局限：受控注入数据不能完全替代人工标注的真实异常样本，删行、乱序和重复行等结构性异常仍主要依赖真实数据上的规则检查；不同变量的物理背景差异也会影响阈值规则的适用性，部分平稳片段可能被输出为候选事件。后续工作将扩展更多站点和更长时间序列，引入人工复核标签和多变量一致性约束，并在实际建模任务中评估质量标记对样本筛选和模型结果的影响。

参考文献：

[1] Qin D H, Liu S Y, Li P J. Snow cover distribution, variability, and response to climate change in western China[J]. Journal of Climate, 2006, 19(9): 1820-1833.

[2] 效存德, 苏勃, 王晓明, 等. 冰冻圈功能及其服务衰退的级联风险[J]. 科学通报, 2019, 64(19): 1975-1984.
Xiao C D, Su B, Wang X M, et al. Cascading risks to the deterioration in cryospheric functions and services[J]. Chinese Science Bulletin, 2019, 64(19): 1975-1984.

[3] Han C T, Chen R S, Liu Z W, et al. Cryospheric hydrometeorology observation in the Hulu Catchment (CHOICE), Qilian Mountains, China[J]. Vadose Zone Journal, 2018, 17(1): 1-18.

[4] QX/T 118—2020 气象观测资料质量控制 地面[S]. 2020.
QX/T 118-2020 Quality control of meteorological observation data: surface[S]. 2020.

[5] 任芝花, 张志富, 孙超, 等. 全国自动气象站实时观测资料三级质量控制系统研制[J]. 气

象, 2015, 41(10): 1268-1277.
Ren Z H, Zhang Z F, Sun C, et al. Development of three-step quality control system of real-time observation data from AWS in China[J]. Meteorological Monthly, 2015, 41(10): 1268-1277.

[6] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[7] Hill D J, Minsker B S. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach[J]. Environmental Modelling & Software, 2010, 25(9): 1014-1022.

[8] Schmidt J Q, Kerkez B. Machine learning-assisted, process-based quality control for detecting compromised environmental sensors[J]. Environmental Science & Technology, 2023, 57(46): 18058-18066.

[9] Sambasivan N, Kapania S, Highfill H, et al. "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2021: 1-15.

[10] Liang W X, Tadesse G A, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI[J]. Nature Machine Intelligence, 2022, 4(8): 669-677.

[11] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets[J]. Communications of the ACM, 2021, 64(12): 86-92.

[12] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3: 160018.

[13] Bavay M, Fiddes J, Godøy Ø. Automatic data standardization for the global cryosphere watch data portal[J]. Data Science Journal, 2020, 19: 6.

[14] Peng G, Downs R R, Lacagnina C, et al. Call to action for global access to and harmonization of quality information of individual earth science datasets[J]. Data Science Journal, 2021, 20: 19.

[15] Pipino L L, Lee Y W, Wang R Y. Data quality assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.

[16] ISO 19157-1: 2023 Geographic information—Data quality—Part 1: General requirements[S].

[17] 杨琳, 朱扬勇. 高质量数据集产品的形态和生产流程研究[J]. 大数据, 2026, 12(2): 64-74.
Yang L, Zhu Y Y. Research on the form and production process of high-quality dataset products[J]. Big Data Research, 2026, 12(2): 64-74.

[18] 李继峰, 张成龙, 刘鑫, 等. 面向人工智能的数据治理框架[J]. 大数据, 2025, 11(1): 3-20.
Li J F, Zhang C L, Liu X, et al. Data governance framework for artificial intelligence[J]. Big Data Research, 2025, 11(1): 3-20.

[19] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[20] Liu F T, Ting K M, Zhou Z H. Isolation forest[C]//Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2008: 413-422.

[21] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 93-104.

[22] Malhotra P, Vig L, Shroff G, et al.

Long short-term memory networks for anomaly detection in time series[C]// Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges: ESANN, 2015: 89-94.

[23] Blázquez-García A, Conde A, Mori U, et al. A review on outlier/anomaly detection in time series data[J]. ACM Computing Surveys, 2022, 54(3): 1-33.

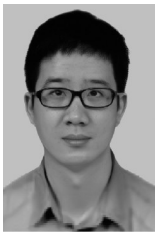
作者简介



张智星（1990–），男，博士，深圳技术大学中德智能制造学院助理教授，主要研究方向为数据驱动的人工智能辅助逆向设计和智能制造。



康建芳（1981–），女，中国科学院西北生态环境资源研究院高级工程师，主要研究方向为数据科学、防灾减灾数据工程研发等。



路长发（1986–），男，博士，中国科学院计算机网络信息中心高级工程师，主要研究方向为大数据管理技术与科学数据软件研发等。



赵彦博（1981–），男，中国科学院西北生态环境资源研究院高级工程师，主要研究方向为冰冻圈科学模型集成和智能化计算等。

录用日期: 2026-10-26

通信作者: 张智星,zhangzhixing@sztu.edu.cn

基金项目: 国家重点研发计划项目(No.2024YFF0729101)

Foundation Item: The National Key Research and Development Program of China (No.2024YFF0729101)