面向人工影响天气的机载多源微物理数据集构建与泄露控制评估

付佳[1,2,3]，陈英英[1,2,3]，李德俊[1,2,3]，陈旭[1]，刘睿[1]
1. 湖北省气象工程技术中心（华中区域人工影响天气技术中心），湖北 武汉 430074；
2. 湖北省气象服务中心，湖北 武汉 430074；
3. 湖北丹江口人工影响天气野外科学观测研究站，湖北 武汉 430074

摘要

针对人工影响天气飞机作业机载数据种类繁多、探头难精确对齐及切分易致数据泄露等问题，构建了面向数据治理与泄露控制评估的人影作业飞机多源微物理基准数据集。通过脏数据剔除、多编码自适应解析、以结冰探头为锚点的1Hz融合及非负物理冰厚生成等方法，将43个架次的机载数据转化为含336570条样本的高质量融合表，并建立了按会话留组的严格嵌套交叉验证协议以阻断事件级泄露。采用XGBoost算法开展基准测试，验证了机载数据结冰起始信号的有效性，并确立了泄露控制下的回归任务基线。该数据集为航空结冰机理剖析与预警算法研发提供了标准化数据底座，为提升风险监控与飞行安全保障提供了严谨的评估参考。

关键词

人工影响天气；机载数据集；数据集构建；数据治理；秒级对齐；泄露控制

中图分类号：TP311.13；P412.1　　文献标志码：A　　doi：10.11959/j.issn.2096–0271.2021xxx

Construction and Leakage-Controlled Evaluation of an Airborne Multi-source Microphysical Dataset for Weather Modification

FU Jia[1,2,3], CHEN Yingying[1,2,3], LI Dejun[1,2,3], CHEN Xu[1], LIU Rui[1]
1. Hubei Meteorological Engineering Technology Center, Wuhan 430074, Hubei, China;
2. Hubei Meteorological Service Center, Wuhan 430074, Hubei, China;
3. Danjiangkou Weather Modification Hubei Field Scientific Observation and Research Station , Wuhan 430074, Hubei, China

Abstract

To address the challenges of diverse airborne data types, difficulties in precise probe alignment, and high data leakage risks during data splitting in weather modification aircraft operations, a multi-source microphysical benchmark dataset oriented toward data governance and leakage-controlled evaluation is constructed. Through methods including dirty data rejection, adaptive parsing of multi-character encodings, 1 Hz fusion anchored by the DXAS probe, and non-negative physical ice thickness generation, airborne data from 43 flights were transformed into a high-quality fused

table containing 336, 570 samples. Furthermore, a strict nested Leave-One-Session-Out (LOSO) cross-validation protocol was established to block event-level leakage. Benchmark testing using the XGBoost algorithm verified the validity of icing onset signals in the airborne data and established a baseline for regression tasks under leakage control. This dataset provides a standardized data foundation for aircraft icing mechanism analysis and early-warning algorithm development, offering a rigorous evaluation reference for enhancing risk monitoring and ensuring flight safety.

Key words
weather modification, airborne dataset, dataset construction, data governance, second-level alignment, leakage control

1 引言

人工影响天气飞机作业通常围绕目标云系开展航线设计、作业高度选择、播撒时机判断和安全风险评估等环节。作业飞机需要在云内多次穿越、盘旋或沿云带飞行；这些区域往往伴随大量过冷水滴、冰晶和复杂粒径谱结构。机体表面一旦发生结冰，会改变翼型气动特性、增加阻力、削弱操纵裕度，并影响作业窗口判断、任务持续能力和影响飞行安全。对于人工影响天气业务而言，飞行人员和业务指挥人员不仅关心“是否存在结冰风险”，还关心结冰是否已经开始、结冰速率情况以及结冰厚度变化。因此，面向结冰风险评估的数据集不能只记录单一探头的数据，而需要把快速云滴谱（fast cloud droplet probe，FCDP）、液态水/总水（liquid water content / total water content，LWC/TWC）含量和结冰探头（dual-channel rime ice detector，DXAS）响应放在同一时间尺度下进行综合预测。。

本文使用的三类机载探头承担不同观测功能：FCDP主要反映云滴粒径谱和粒子数变化，LWC/TWC提供液态水和总水含量相关观测，DXAS结冰探头通过振动频率变化反映探头表面积冰状态。三类探头的采样频率、文件格式、字段命名和有效工作区间并不一致，原始数据不能直接拼接成模型样本。特别是DXAS频率需要结合风洞标定关系和探头失调边界进行处理，连续飞行日志又具有强时间相关性，若简单按行随机切分，会把同一结冰事件的相邻秒样本同时放入训练集和测试集，从而高估模型表现。

围绕结冰风险评估，已有研究多集中于航空结冰安全风险综述、单一探头响应、机器学习模型构建或飞行试验观测分析[1-5]。但要支撑数据驱动的预警与风险评估方法，首先需要解决一个更基础的问题：人工影响天气飞机作业结冰场景下高质量、可复验、受泄露控制约束的机载微物理数据集如何构建。

已有人工影响天气飞机作业结冰相关数据研究大致可分为三类。第一类侧重外场或飞行试验中的云微物理观测，例如大过冷水滴（supercooled large droplet，SLD）条件下的机载原位观测数据集和云滴谱/液态水含量测量工作，为结冰环境识别提供了重要物理基础[3,5-7]。第二类关注探头测量与标定关系，例如液态水/总水含量探头、结冰探头和过冷水响应的风洞或外场评估，为原始信号转化为物理量提供依据[8-11]。第三类则将机器学习用于结冰强度或过冷水条件识别，强调模型预测性能和业务判别能力[4,12]。与这些工作相比，

本文不以单一观测个例、探头机理或新模型性能为主要研究内容，而是把多源机载日志的字段治理、秒级对齐、物理标注边界和泄露控制评估方案作为核心对象，强调数据集构建过程本身的可审计性和可复验性。

实际工程数据集面临三类典型治理难点。第一，FCDP、LWC/TWC和DXAS三类探头采用粒子分析与显示系统（particle analysis and display system，PADS）标准CSV文件输出，原始文件存在中文乱码、字段位置依赖、采样频率不一致和时间字段不统一等问题，难以直接进入统一秒级模型样本。第二，机载结冰过程具有连续时间相关性，相邻秒级记录往往属于同一结冰事件，若按行随机划分，模型评价指标会被同一事件内部的强相关性放大，从而高估泛化能力；类似的数据泄露、模型选择偏差以及时间/空间/层级结构下交叉验证失效问题，在机器学习和生态统计研究中已有系统讨论[13-15]。第三，结冰过程本身包含起始、增长、稳定和探头失调等阶段，缺乏面向连续过程刻画的客观、可审计标注。

为构建一个支持上述需求的标准数据集，，并响应该领域对高质量数据集建设的规范化要求[16]，本文采用可发现、可访问、可互操作、可复用（findable, accessible, interoperable, reusable，FAIR）数据管理原则、数据集说明表和数据卡等数据集文档化思想，强调数据来源、字段语义、质量状态、访问边界和复现实验接口的显式记录[17-19]。具体工作包括：（1）打通从原始PADS文件到1Hz秒级融合表的完整治理链路，涵盖文件角色识别、多字符集编码自适应解析、PADS表头规范化、秒级对齐、以DXAS为锚点的多源融合，以及探头失调与近失调状态的质量标记；（2）基于DXAS风洞标定关系生成冰厚、增长率和结冰阶段标注，并同步处理低频探头失调边界与高于基准频率导致的负反解异常，确保最终公开的物理冰厚满足非负约束与物理真实性；（3）构建严格嵌套留一会话交叉验证（leave-one-session-out，LOSO）评估方案，将特征筛选过程强制约束于训练折内，从切分机制底层阻断了连续日志的事件级泄露风险；（4）依托极限梯度提升（extreme gradient boosting，XGBoost）算法在该方案下开展数据集下游基准测试，证实了结冰起始任务中蕴含有效的稀有事件排序信号，并为持续时间与增长率任务提供了泄露控制条件下的探索性参考基线。

鉴于当前数据规模（43个架次、58个会话单元），采用基于Pandas的单节点脚本已足以支撑全流程的快速复现。若后续应用场景扩展至全国范围人工影响天气作业的海量与实时数据处理，主要算力瓶颈将集中于异构CSV文件的批量解析、会话级聚合重采样、跨探头秒级对齐，以及严格嵌套评估中的高频特征重选。为此，本治理管线在设计上具备向分布式架构平滑迁移的潜力：在离线批处理层面，原始日志可优先转化为按会话或日期分区的Parquet列式存储，并依托Spark、Dask或Polars实现高效的分组聚合与特征派生；在准实时流处理层面，可通过Flink框架按会话键（Session Key）维护状态窗口。同时，DXAS标定参数、失调阈值与插值窗口可作为全局配置统一下发，以确保分布式架构与当前单节点基准规则的严格等效与一致性。

综上所述，本文严格聚焦于多源机载微物理数据集的构建规范、泄露控制治理及评估协议的发布，旨在为后续的结冰状态识别、阶段主导因子挖掘及微物理机制

阐释夯实标准化的数据底座。

2 数据来源与总体框架

2.1 原始数据来源

本研究数据来源于湖北省高性能人工影响天气作业飞机平台搭载的FCDP云滴谱探头、LWC/TWC液态水/总水探头和DXAS结冰探头，以及风洞标定的DXAS探头结冰数据。云滴谱、液态水和总水含量类机载探头是云微物理观测和结冰环境识别中的核心测量手段，其响应、标定和外场适用性已有较多研究积累[6-10]。43个架次原始观测在统一治理后形成58个会话单元（部分架次包含多个独立飞行段，按PADS文件名时间戳作为会话主键）。三类探头原始文件均为PADS标准CSV，文件前部为仪器元信息，真实数据表头位于分隔符之后，并在不同架次中存在中英文混合表头、列名乱码以及位置依赖字段。

表1给出本数据集的信息卡。该表把数据资产的时间范围、传感器构成、样本规模、清洗后可用性和开放边界集中呈现，便于下游使用者判断数据集的覆盖范围与复用条件。

表1 机载结冰微物理数据集信息卡

项目	统计或说明
数据时间跨度	2025-01-04 13:23:17至2026-01-30 00:31:11
飞行地域/平台	湖北区域人工影响天气飞机作业与试验平台，面向云内作业过程中的飞机结冰风险监测
原始架次与会话	43个飞行架次，治理后划分为58个会话单元
异构传感器	FCDP云滴谱探头、LWC/TWC液态水/总水探头、DXAS结冰探头
采样频率特征	DXAS与LWC/TWC主步长约1Hz；FCDP存在重复秒和变步长记录，治理后统一聚合至1Hz
融合前原始测量规模	177个已解析测量文件，共978260行原始测量记录，其中DXAS337424行、FCDP301626行、LWC/TWC339210行
清洗后融合样本	336570条以DXAS为主时间轴的1Hz秒级融合样本，113个融合表字段
候选特征库	336570行、146个可追溯字段，覆盖原始通道、谱统计、FCDP高尾部质量检查字段、跨传感器交互项、对齐偏移校核字段、质量状态变量与标注字段
传感器可用率	FCDP可用209069行(62.12%)，LWC/TWC可用336320行(99.93%)，DXAS锚点覆盖全部融合样本
质量标记分布	DXAS失调无效样本4558行(1.35%)，近探头失调482行(0.14%)
连续标注可用性	h_mm 非空306874行(91.18%)，dh_dt_mm_s 非空306744行(91.14%)；物理冰厚负值0行
结冰事件标注	结冰起始秒级样本398行(0.12%)，31个会话单元至少包含一次结冰起始
阶段标注分布	none 280538行，steady_or_slow 44450行，growth 6626行，invalid_detune 4558行，onset 398行
对外开放字段	脱敏后的会话编号、日期分组键、秒级时间索引、规范化传感器通道、FCDP谱箱统计、LWC/TWC液态水/总水计数、DXAS频率与派生冰厚、对齐偏移、质量标记、标注字段和固定折次编号
受限字段	原始飞行轨迹、未经脱敏的完整原始探头日志、可能关联具体飞行任务的元信息

数据治理流程的整体情况如图1所示。图中补充了编码解析、秒级对齐、短缺测插值、DXAS失调阈值、近失调带宽和留组评估等关键参数。该流程把异构原始日志逐步转化为统一的、可追溯的1Hz秒级融合表，并把会话级LOSO留组方案作为泄露控制评估的入口。

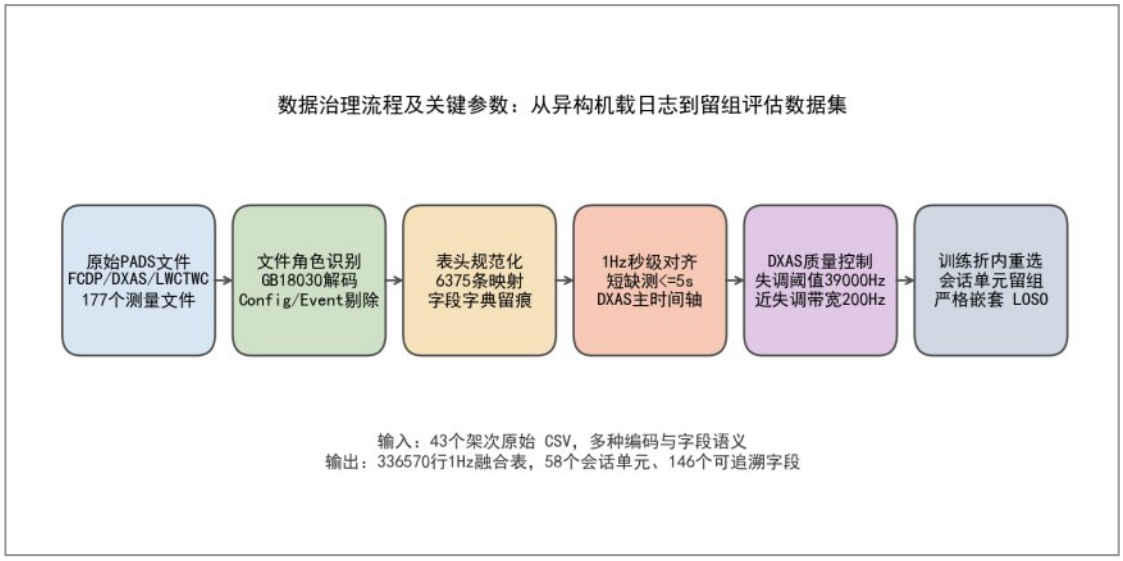

图1　数据治理流程及关键参数

2.2 治理目标与产出

本文数据治理流程的核心目标，是输出具备高度透明性与可复验性的三项标准化数据资产：

（1）原始文件解析审计报告：详尽记录每个原始测量文件的解析状态、自适应采用的字符集编码、时间跨度、表头行号及缺测情况，为底层数据溯源提供完整证据链；

（2）1Hz秒级多源融合表：以DXAS探头的秒级记录为时间锚点，将异构观测数据精确对齐为包含336570条样本、113个字段的高质量融合表，全面覆盖58个会话单元；

（3）候选特征字段库：在上述融合表基础上派生的扩展特征集（包含336570行、146列），涵盖原始通道、谱箱统计量、FCDP高尾部质量校核字段、跨传感器交互项以及时空对齐与可用性诊断变量。本文将其严格定位为标准化的数据底座与下游评估协议的输入接口，故未在此过度展开关于结冰阶段主导因子的物理机理剖析。

此外，该治理管线在运行过程中同步生成了6,75条字段映射记录。其中，针对DXAS、FCDP与LWC/TWC探头分别提炼出53、40与17个规范化字段，由此构建了一套支持跨架次无缝复用的标准化字段字典。

3 数据治理与泄露控制方法

3.1 文件角色识别与基于规则的脏数据剔除

PADS标准格式中除探测文件外，还包含“config”、“event”、“合并数据”等派生文件。如果直接遍历目录，模型样本会被这些非测量文件污染。本文采用基于规则的文件分类策略：

（1）当文件名包含“config”、“event”、“合并数据”或“merged”等特征字符时，将其分类为配置文件或派生文件（config/derived）；

（2）当文件名前缀为“09_FCDP_”、“04_LWC/TWC_”、“08_DXAS_”时，该类CSV文件被分类为测量文件（measurement），并进入秒级处理管线；

（3）其余文件统一标记为其他类别（other），但仍保留在解析摘要中以便后续审计。

在此基础上，对探测数据再施加多层基于规则的数据治理：（1）解析失败、表头缺失或时间字段缺失的文件直接判为跳过；（2）行级别上，所有非时间字段统一转换为数值类型，并把""、“NaN”、“NAN”、“nan”等字符串归一为缺测；（3）基于DXAS风洞标定关系和外场失调方向敏感性测试，把频率进入低频失调边界的样本以invalid_dxas=True标记；（4）对接近失调边界但尚未越界的频率样本以near_detune_flag=True标记，提示这些点处于物理上低可信的过渡带；（5）FCDP数据经秒级聚合后保留原始计数结构，fcdp_total_counts的负值、超过30000counts/s的显示上限外样本、会话内P99高尾部样本、显示限幅值和log1p稳健变换均作为质量检查或派生字段保留，不在主数据集中按固定阈值剔除。所有规则都在脚本中显式参数化，既方便审计，也允许下游用户基于不同业务阈值重新生成数据集。

3.2 多字符集自适应解析与PADS表头规范化

DXAS与FCDP文件中的中文表头在不同采集软件版本下存在UTF-8、GB18030、GBK、CP936等多种编码，部分历史文件甚至无法用单一编码完全解码。直接读取会引发解码错误或产生乱码字段名，进而导致跨架次的字段语义严重割裂。

为此，本文将编码探测显式写入治理流程：依次尝试“utf-8”、“gb18030”、“gbk”、“cp936”、“latin1”，记录最终采用的编码与所有尝试过的候选，并把这些信息保留在CSV文件中，便于复现。在确定编码后，按PADS风格扫描行，定位分隔符所在行，据此找到真实数据表头。表头规范化分三步：

（1）公共字段归一化：依托预设的公共别名表，将诸如End Seconds、Day of Year、Year、Status，以及全球定位系统（global positioning system，GPS）时间字段GPS Time、Date、Time等基础标量，直接映射至统一的系统标识；

（2）多语言混合表头映射：针对FCDP与LWC/TWC探头，构建专门的别名规则映射库。将环境温度、激光传感器温度、采集极/参考极电压电流等中英混杂字段，统一映射为标准化的英文下划线命名；

（3）位置依赖映射：针对DXAS表头出现严重乱码的极端情况，启用基于物理绝对位置的映射机制。按固定列号将第13、19、25、37、38、39、45、46、47、48列，依次强制映射为对应的高度、水含量、环境参数、探头频率及三组故障字变量（如dxas_pressure_altitude_

candidate、dxas_frequency_hz等)。

经过上述系统性的规范化处理，跨架次的异构探头字段在整个数据集层面实现了高度一致的语义对齐。共计6375条原始字段的演变路径均生成了对应的映射审计记录，确保了所有规范化名称具备清晰的数据溯源能力。

3.3 线性插值与1Hz秒级对齐机制

三类探头采样频率不同：FCDP通常以高于1Hz的频率写入并伴随重复秒，DXAS与LWC/TWC主步长接近1s。为获得统一的1Hz主时间轴，本文采用如下治理机制：

(1)时间戳构建：优先用日期和GPS时间组合构造时间戳；若GPS时间缺失，再退化使用日期和普通时间字符串；进一步生成秒级时间索引作为对齐键。

(2)秒级聚合：FCDP以会话编号和秒级时间索引为键，对所有数值字段求均值，并记录聚合行数 fcdp_n_raw 作为采样密度校核字段；LWC/TWC与DXAS则按相同键聚合至1s。

(3)DXAS主时间轴融合：以DXAS秒级记录为主时间轴，FCDP与LWC/TWC通过同秒键左连接写入；同时记录has_fcdp、has_lwctwc、has_dxas三类传感器可用性标志。

(4)线性插值与限制：在每个会话单元内，对短时缺测（默认≤5s间隔）的非标注连续通道可使用线性插值补齐，更长的缺测保持空值，避免引入虚假动态。该窗口主要用于修复连续记录中的短暂通信或写盘空洞；在云内宏观温湿背景和探头状态通常呈平滑渐变的前提下，5s内的线性补齐可降低离散缺测对秒级融合表的破坏，但不用于恢复快速结冰标注或跨越质量异常段。DXAS失调样本只做质量标记和遮蔽，不通过插值跨越失调区恢复为连续物理曲线；冰厚h_mm在差分得到dh_dt_mm_s时限制时间间隔超过5s的差分置空，以阻止跨段错连。

(5)对齐偏移校核：保留dt_fcdp_dxas与dt_lwctwc_dxas，分别给出每个秒级样本到最近FCDP与LWC/TWC记录的有符号偏移。统计结果显示：dt_lwctwc_dxas的绝对偏移中位数与95分位数均为0s，说明LWC/TWC与DXAS已基本同步；dt_fcdp_dxas的中位数为0s、95分位数为14s、最大值为55s，说明FCDP在部分时段存在轻度时延，必须以校核字段的形式显式提供给下游使用者。

经过1Hz秒级对齐治理后，数据集每一秒同时承载多源观测、传感器可用性与对齐质量信息，构成一个具备质量审计能力的融合表。

3.4 线性插值与1Hz秒级对齐机制

图2给出按样本量降序排列的会话样本组成。每根柱代表一个会话，并用S01–S58标注；图中同时按样本量分为不小于10000行、3000–9999行、1000–2999行和小于1000行四组。从下到上依次表示有效非结冰样本、结冰样本、近探头失调样本和失调无效样本。整体上，58个会话的样本量差异较大，从数十秒到三四个小时不等；其中32个会话的有效样本超过1000行，30个会话超过3000行，31个会话至少包含一次结冰起始样本。整张图直接揭示了机载真实采集数据的不均衡性：少数长架次贡献了绝大多数有效样本和结冰样本，但这些会话中的失调占比也明显偏高，因此在评估方案中必须按会话留组而不是按行划分。

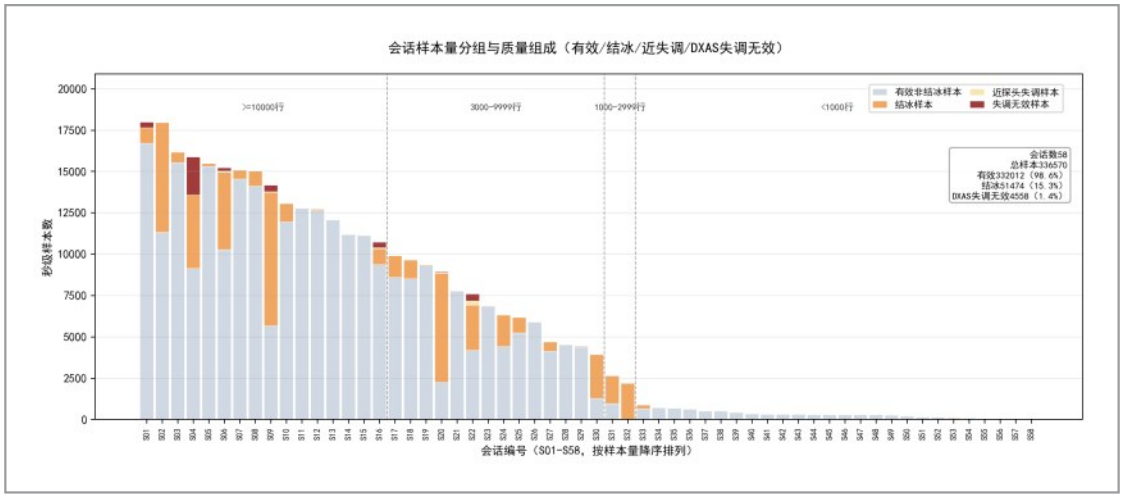

图2 会话样本量分组与质量组成

在质量标记基础上，数据集进一步基于DXAS风洞标定关系生成三类连续过程标注：冰厚（h_mm）、冰厚变化率（dh_dt_mm_s）与结冰阶段（icing_phase）。其中保留频率－冰厚（h_mm_raw）公式的直接反解结果；当外场频率高于基准频率而产生负反解值时，该值仅表示无冰基线偏移或仪器噪声，不解释为负冰厚。公开字段中的h_mm为非负物理冰厚，先在invalid_dxas=False的样本上计算，再将负反解值按物理下限置为0；dh_dt_mm_s基于该非负物理冰厚差分得到。icing_phase把样本进一步划分为无结冰（none）、结冰起始（onset）、显著增长（growth）、稳定或缓慢增长（steady_or_slow）与失调无效（invalid_detune）五类。下列定义清单给出了关键标注与状态变量的最终定义，下游研究可在不动主流程的前提下重新设定阈值生成自定义版本。

上述阈值的选取依托DXAS风洞标定受控实验和外场探头失调方向敏感性测试确定标注可信区间。标定关系给出基准频率f0_hz=42398.05Hz及频率－冰厚映射；39000Hz低频边界对应探头进入失调或机械响应失真区，采用200Hz近失调带宽用于隔离边界附近的低可信过渡样本，二者构成连续冰厚和冰厚变化率标注的低频物理安全边界。与此同时，外场无冰或弱结冰状态下可能出现频率高于f0_hz的零点漂移，此时公式直接反解得到的h_mm_raw<0只用于校核基线偏移，物理冰厚统一置为0，结冰起始也不允许由这一负反解区间内的正差分单独触发。阈值敏感性检查显示，在冰厚变化率阈值固定为0.01mm/s时，冰厚阈值取0.05、0.10、0.20mm的重算结冰起始行数分别为299、382、508行；在冰厚阈值固定为0.10mm时，冰厚变化率阈值取0.005、0.010、0.020mm/s的重算结冰起始行数分别为1138、382、228行，说明冰厚变化率阈值仍是更敏感的标注边界。探头失调方向敏感性结果表明，采用低频失调方向时DXAS失调无效样本为4558行、结冰起始样本为398行，且235025行负原始反解均被置为非负物理冰厚；若错误采用高频方向，即大于等于规则，DXAS失调无效样本会扩大至306879行、结冰起始样本仅剩73行，并显著改变高排序字段组合。这说明探头失调方向和高频负反解处理都不是

表2 物理标注与质量状态变量定义

标注/状态	定义或计算方式	当前参数与处理
delta_f_hz	f0_hz − f(t)	f0_hz = 42398.05Hz
h_mm_raw	DXAS频率 - 冰厚标定公式的直接反解结果	可为负,仅作为基线偏移校核信息,不作为物理冰厚解释
h_mm	非负物理冰厚	仅对非失调DXAS频率计算,并将负反解值按物理下限置为0
dh_dt_raw_mm_s	h_mm_raw 的相邻秒差分	作为校核字段保留,不进入结冰风险识别候选字段集
dh_dt_mm_s	同一会话单元内相邻秒 h_mm 差分除以时间差	时间间隔缺失或 >5s时置空
invalid_dxas	DXAS频率进入探头失调方向的无效区	主结果采用低频失调边界,阈值为39000Hz、方向为小于等于
near_detune_flag	接近探头失调边界但尚未进入无效区	近失调带宽200Hz
icing_onset_flag	从非结冰状态转入结冰状态的起始秒	基于非负 h_mm 的冰厚阈值0.1mm或冰厚变化率阈值0.01mm/s,负原始反解区间不能单独触发结冰起始
time_since_icing_start_s	自最近一次结冰起始起算的持续秒数	非结冰状态重置为空
icing_phase	none、onset、growth、steady_or_slow、invalid_detune	结合DXAS失调状态、结冰起始标志和增长率阈值生成

次要工程细节，而是决定标注可信边界和保留真实过冷水响应的关键物理约束。

按上述规则，融合表DXAS失调无效样本4558行（占1.35%），近探头失调482行（0.14%），结冰起始样本398行（0.12%）；物理冰厚与冰厚变化率非空率分别为91.18%与91.14%，物理冰厚负值行数为0，且所有结冰起始样本均不来自负原始反解区间。若不对负反解值施加非负约束，并直接用原始反解冰厚及其差分生成起始标注，则结冰起始样本会增至2087行，其中1728行来自负原始反解区间内的正差分；采用非负约束后，这部分不具备物理冰厚含义的触发被排除，阶段分布稳定为无结冰280538行、稳定或缓慢增长44450行、显著增长6626行、失调无效4558行以及结冰起始398行。FCDP侧仅有2行总粒子数<0，超过30000 counts/s显示上限的高尾部样本为4850行（占FCDP可用样本2.32%），这些样本被质量标记而非从主数据集中删除。

3.5 清洗前后对比

为直观展示治理效果，本文选取两个代表性窗口分别展示DXAS探头失调标记和FCDP高尾部计数显示治理（图3、图4）。DXAS示例图给出振动频率在39000Hz失调边界附近的低可信段，图中仅保留原始频率、有效频率段和失调无效点三类核心标注；有效频率段在失调区断开，失调样本不参与冰厚反演，也不通过插值跨区恢复为连续物理曲线。FCDP示例图给出总粒子数的独立显示案例：原始通道存在极少数负值和明显右偏的高计数长尾；为避免少数高值拉伸纵轴，图中采用30000 counts/s作为显示上限，并标出超过该上限的高尾部样本。该上限仅用于图示可读性和高尾部提示，不作为FCDP物理失效阈值，也不参与标注生成或主数

据剔除。两张图分别服务于不同探头的质量治理说明，不暗示DXAS失调与FCDP高计数在物理上同步对应。

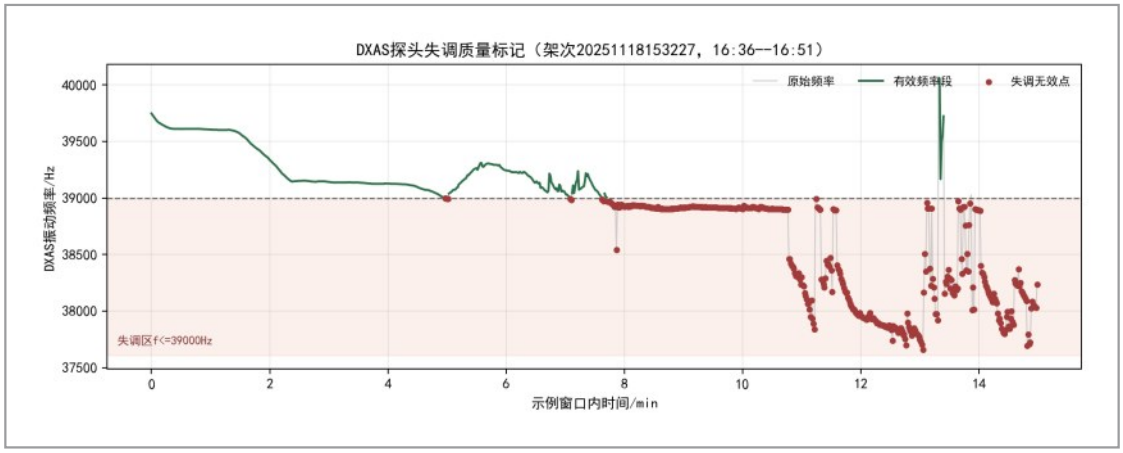

图3 DXAS探头失调质量标记

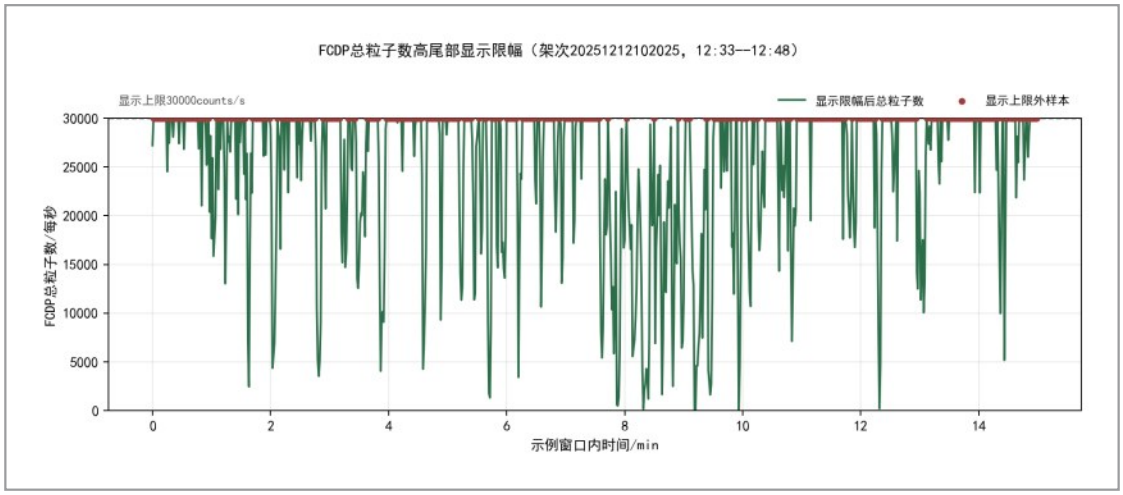

图4 FCDP总粒子数高尾部显示限幅案例

4 泄露控制的LOSO切分协议

人工影响天气飞机作业中探头探测到的结冰数据是典型的事件级时间序列，相邻秒级样本来自同一云层、同一结冰事件，标注和特征都呈现强自相关。直接进行行级交叉验证会让训练集和测试集共享事件特征，从而导致评价偏高。数据泄露研究和结构化交叉验证研究均指出，具有时间、空间、层级或事件结构的数据应采用与预测目标一致的分组验证策略，并避免把特征选择或模型选择过程暴露给测试折[13-15]。本文为该数据集预先确定了一套泄露控制切分方案，并把它作为评估方案的一

部分发布。

（1）切分主键。评估实验以标准化会话单元为最小切分粒度，要求所有LOSO折次中的训练集与测试集必须跨会话相互独立。该机制将连续的飞行日志段整体隔离于单折之内，从根本上避免了同一结冰物理过程被跨界撕裂至训练集与测试集中。此外，鉴于部分同日架次可进一步拆分为多个会话，本文同步提供了以“日期”为分组键的敏感性检查，作为更为严苛的架次代理留组视角，用于审视同日宏观气象背景与仪器状态相关性可能引入的残余乐观偏差。

(2)严格嵌套LOSO。本文将“严格嵌套LOSO”明确界定为：外层循环按会话留出测试评估组，内层环节则强制要求候选字段筛选、缺失值插补以及模型训练过程，仅在剩余的训练折内部独立完成。在每个外层折次中，模型必须依托训练折内动态算出的统计量重新实施Top-K字段选择，随后再对留出的未知测试会话进行推理。该闭环流程有效规避了“全局特征排序后再分折训练”所固有的隐性数据泄露风险，高度契合了模型选择场景下降低误差估计偏差的嵌套验证准则[14]。

（3）物理约束下的缺测补齐与状态保留。在训练折内部处理异常值时，遵循严格的物理约束：对极短时缺测实施受限的线性插值（时间窗口≤5s）或中位数填充；对于落入失调区（invalid_dxas=True）的无效样本，坚决将其排除在连续标签的计算之外；而对于近失调带（near_detune_flag=True）的样本，则将其作为低可信过渡状态显式保留。这一处理确保了探头物理响应失调附近的测量不确定性，能够被下游预警模型透明地感知与审计。

（4）全链路可复现性与对照基线发布。为确保上述协议的可操作性与客观性，本文将评估规则、超参数阈值、动态特征筛选脚本、训练折次编号以及各折入选字段记录，随数据集开源并全套发布，下游研究者仅需替换核心分类器/回归器即可无缝复现完整的LOSO实验。更为关键的是，鉴于目标标签深度依赖DXAS的标定关系派生，本文额外设立了剔除DXAS探头直接相关特征的结冰起始检测基线，旨在严谨界定模型的预测能力究竟源自“DXAS同源字段的一致性拟合”，还是切实捕捉到了“非标签源传感器（如FCDP）的独立预警信号”。

5 协议验证测试

为验证数据集在泄露控制条件下仍可被下游任务调用，本文将经典经验特征组合和XGBoost共同作为评估方案验证。经典经验特征组合包括DXAS过冷水含量、FCDP温度、LWC/TWC液态水/总水计数、FCDP总粒子数以及高粒径/谱尾占比，代表传统单一或少量经验指标驱动的结冰判别思路；开放候选字段和严格嵌套XGBoost则代表在本文数据治理流程上构建的可审计字段库与泄露控制评估方案。引入经验对照主要用于检查三个问题：第一，清洗后的字段是否包含可排序的结冰起始信号；第二，把字段筛选纳入训练折后评估方案是否仍可执行；第三，当标注由DXAS标定派生时，去除DXAS直接相关字段后非标注源传感器是否仍保留可用信号。因此，以下结果被解释为数据集与评估方案的可执行性证据，而不是业务部署精度或算法优越性的证明。

表2给出基准对照结果。对于结冰起始阶段，经典经验特征组合在留组逻辑回归下的平均特征曲线下面积（area under

the curve，AUC）/平均精确率（average precision，AP）为0.3529/0.0092；开放候选字段中前K个高排序字段的平均AUC/AP为0.3886/0.0741，说明经过字段规范化、秒级对齐和跨传感器交互构建后，结冰起始样本存在可排序信号。通过XGBoost算法验证中，将字段筛选纳入训练折后，结冰起始任务AUC为0.6087、AP为0.0129；持续时间任务均方根误差（root mean square error，RMSE）为671.2845s；增长率任务RMSE为0.1107mm/s。回归任务的决定系数R^2明显为负，说明跨会话分布漂移和标注噪声仍然突出，也从反面表明：若不采用会话级留组和严格嵌套方案，连续日志内部相关性很容易产生过度乐观的表观性能。因此这些数值反映了任务难度和评估方案必要性。表2中优先报告按留出组等权平均的AUC/AP，因为每个会话单元或日期代表一个独立验证单元；补充文件同时保留汇总口径指标。由于结冰起始样本较少且不同会话样本量差异显著，汇总口径指标会被少数长会话主导，本文不把汇总口径阈值分类结果作为主结论。

表2 经验对照与泄露控制评估结果

任务	方法/特征设置	验证方式	指标
结冰起始	经典经验特征组合	会话留组逻辑回归	AUC 0.3529 / AP 0.0092
结冰起始	开放候选字段前K项	会话留组逻辑回归	AUC 0.3886 / AP 0.0741
结冰起始	XGBoost,非严格嵌套前K项	会话留组LOSO	AUC 0.5851 / AP 0.0160
结冰起始	XGBoost,严格嵌套前K项	严格嵌套LOSO(训练折内字段重选)	AUC 0.6087 / AP 0.0129
结冰起始	去除DXAS直接相关字段后动态前K项	严格嵌套LOSO(训练折内字段重选)	AUC 0.3877 / AP 0.0776
结冰起始	去除DXAS直接相关字段后动态前K项	日期留组(训练折内字段重选)	AUC 0.2838 / AP 0.0626
结冰起始	最小独立源特征	日期留组逻辑回归	AUC 0.2884 / AP 0.0045
结冰持续时间	经典经验特征组合	会话留组LOSO XGBoost	RMSE 632.0232s
结冰持续时间	开放候选字段前K项	会话留组LOSO XGBoost	RMSE 684.2504s
结冰持续时间	严格嵌套前K项	严格嵌套LOSO(训练折内字段重选)	RMSE 671.2845s
结冰增长率	经典经验特征组合	会话留组LOSO XGBoost	RMSE 0.1103mm/s
结冰增长率	开放候选字段前K项	会话留组LOSO XGBoost	RMSE 0.1051mm/s
结冰增长率	严格嵌套前K项	严格嵌套LOSO(训练折内字段重选)	RMSE 0.1107mm/s

总体上，评估结果说明数据集在修正后的物理标注口径下仍可执行完整留组评估，但结冰起始样本稀有性和跨会话漂移更加突出。逻辑回归与XGBoost结果均显示开放候选字段前K项并非在所有指标上优于经典经验组合；去除DXAS直接相关字段后，动态前K项的AP仍高于专家最小特征，但AUC低于0.5，说明非标注源

字段只保留有限且不稳定的稀有事件排序信息。特征来源审计显示，去除DXAS直接相关字段后的动态前K项全部来自FCDP数据统计；专家最小特征中FCDP谱统计、LWC/TWC水含量通道和环境量的选择占比分别为57.14%、28.57%和14.29%。持续时间和增长率两个回归标注在LOSO条件下的R^2明显为负，说明其更适合作为探索性任务和难度参考；这一结果并不削弱本研究的价值，反而说明机载连续观测若采用随机行切分或全局字段筛选，可能把同一事件内的相关性误当作泛化能力。具体的微物理机制讨论、字段选择稳定性分析和关键因子物理意义等内容，将在后续基于本数据集构建的基础上开展相关研究工作。

6 数据集质量、应用价值与局限

6.1 质量与可审计性

本数据集的全生命周期治理环节均实现了高度的透明化与可溯源。从原始文件的解析状态、多字符集自适应编码、时间跨度、字段语义映射，到多源传感器的可用性、时空对齐偏移及探头失调标记，均在解析审计摘要、字段映射报告及融合表的诊断列中进行了全链路留痕。这一设计严格按照FAIR数据共享原则，符合现代数据集文档化对字段语义透明度与证据链审计的严苛要求。正如在机器学习与深度学习等复杂任务中所强调的那样，严谨的数据治理与客观的数据质量评估体系是保障下游模型可靠性的核心前提[20]。

6.2 应用价值

本数据集立足于实际业务需求，旨在支撑以下三类下游研究与业务应用场景：

（1）赋能机载结冰监控与安全预警系统：高质量、无泄露的微物理数据底座是构建高可靠性预警模型的前提。本数据集为复杂穿云作业中的极端结冰起始秒级检测、积冰强度评估提供了受控基准。这不仅能直接赋能机载实时监控算法的研发，更为提升人工影响天气作业飞机的飞行安全、优化航线防冰决策提供了坚实的数据保障；

（2）治理管线的无缝业务化复用：支持多源机载探头数据治理流程向全国其他区域业务的扩展，包括多字符集自适应解析、PADS表头语义规范化及1Hz秒级时空对齐等核心组件，可直接支撑更大规模人影探测数据的标准化整编；

（3）防泄露方法学参考：提供标准化的会话单元留组、日期留组敏感性测试及严格嵌套 LOSO 评估协议。下游算法开发者可直接套用本文的切分方案，免去繁琐的防漏切分设计，确保预警模型在实际部署前得到最严谨的泛化能力评估。

6.3 可迁移的数据治理经验

尽管本数据集扎根于飞行结冰与飞行安全保障场景，但其系统性的数据治理范式具有广泛的跨领域迁移价值：第一，前置的文件角色识别与自适应编码解析，可有效屏蔽派生文件与乱码对核心样本库的污染；第二，确立物理含义明确的高可信度传感器作为时间锚点，辅以其他传感器的对齐偏移与可用性诊断，是研判多源融合置信度的有效手段；第三，面对低可信边缘样本（如探头失调区），采用“质量标记与边界限定”优于“粗暴删除”，最大程度保留了真实外场探测的不确定性物理结构；第四，针对强自相关时序数据，采用

事件/会话级留组验证并将特征筛选闭环于训练折内，可从根本上规避评价的乐观偏差。上述旨在“排除虚假信号、保障模型在真实物理世界可靠性”的治理范式，均可平滑迁移至车联网、工业传感网络、医疗连续生命体征监测等多源复杂时序数据集的构建任务中。

6.4 数据集局限性与使用边界

（1）物理可信边界：受限于DXAS探头特性，其在低频失调边界附近的物理可信度显著下降，故该区域标签已被显式置空；同时，高于基准频率的负反解值虽作为校核诊断保留，但物理冰厚已严格截断为0。下游预警模型在解释预测结果时须恪守这一底层物理约束或改善该探头的探测精度。

（2）极端类别不平衡：数据集中结冰起始状态的正样本占比仅为0.12%。这种极端的数据不平衡是由穿云探测的客观物理场景决定的，因此相关评估指标的偏低反映了任务的真实难度，不能单纯据此评判下游预警模型的性能上限。

（3）分布漂移与残余评估偏差：在严苛的交叉验证设置下，回归任务在不同折次间波动较大，且以日期留组的评估指标通常低于按会话留组的结果。这表明真实的机载观测在跨会话间存在显著的分布漂移，同日气象过程的强相关性仍可能诱发微弱的性能高估，未来亟待扩充更多跨日、跨平台的独立飞行样本加以验证。

（4）业务化部署与机理探讨：尽管本数据集为算法研发提供了坚实底座，但受限于样本规模，当前更适宜作为防泄露协议与多源特征挖掘的基准测试。

7 结论

本文针对人工影响天气飞机作业中结冰风险监测、预警业务对高质量数据的紧迫需求，构建并发布了一个面向数据治理与泄露控制评估的多源机载微物理基准数据集。通过执行涵盖基于规则的脏数据剔除、多字符集编码自适应解析、PADS表头规范化、1Hz秒级时空对齐、以DXAS为锚点的多源融合、探头失调质量标记及非负物理冰厚标签构建的治理管线，成功将43个架次的异构原始日志转化为包含58个标准化会话单元、336570行融合样本及146个可审计字段的高质量数据资产。

本研究的核心意义在于，通过样本级的探头状态诊断与切分级的严格嵌套LOSO协议，为人工影响天气领域确立了一个具备物理约束且防泄露的数据底座。这不仅极大提升了穿云作业中结冰风险监测的精准度与预警模型的科学性，更为保障作业飞机在复杂微物理环境下的飞行安全提供了关键的标准化支撑。基于XGBoost算法的基准验证表明，在严格的泄露控制协议下，数据集仍有效保留了结冰起始等事件的排序信号；同时，通过特征来源审计与日期留组敏感性检查，清晰揭示了非标签源信号的边界与跨日泛化的真实难度。该数据集及其配套评估方案的发布，将为后续航空结冰的机理解释、阶段控制因子发现及高性能风险预警算法的研发提供标准化的对照实验平台与全链路可复现的数据基准。

参考文献：

[1] Cao Y., Tan W., Wu Z. Aircraft icing: An ongoing threat to aviation safety. [J]. Aerospace Science and Technology, 2018, 75: 353～385.

[2] Luo L., Xue M., Xu X., Deng L., Li J.,

Zhang R. Case study of aircraft icing in-cloud measurements and explicit super-cooled water prediction in Eastern China. [J]. Atmospheric Research, 2026, 334: 108748.

[3] Menekay D., Lucke J., Jurkat-Witschas T., Voigt C., Kirschler S., Bourdon A. An airborne in-situ dataset of cloud microphysical properties in supercooled large droplet icing conditions. [J]. Earth System Science Data Discussions, 2026.

[4] Li S., Qin J., He M., Paoli R. Fast Evaluation of Aircraft Icing Severity Using Machine Learning Based on XGBoost. [J]. Aerospace, 2020, 7(4): 36.

[5] Jensen A. A., Weeks C., Xu M., Landolt S., Korolev A., Wolde M., DiVito S. The Prediction of Supercooled Large Drops by a Microphysics and a Machine Learning Model for the ICICLE Field Campaign. [J]. Weather and Forecasting, 2023, 38(7): 1107～1124.

[6] Baumgardner D., Brenguier J.-L., Bucholtz A., Coe H., DeMott P., Garrett T. J., Gayet J.-F., Hermann M., Heymsfield A., Korolev A., Kramer M., Petzold A., Strapp J. W., Pilewskie P., Taylor J., Twohy C., Wendisch M., Bachalo W., Chuang P. Airborne instruments to measure atmospheric aerosol particles, clouds and radiation: A cook's tour of mature and emerging technology. [J]. Atmospheric Research, 2011, 102(1-2): 10～29.

[7] Lance S., Brock C.A., Rogers D., Gordon J. A. Water droplet calibration of the Cloud Droplet Probe (CDP) and in-flight performance in liquid, ice and mixed-phase clouds during ARCPAC. [J]. Atmospheric Measurement Techniques, 2010, 3(6): 1683～1706.

[8] King W.D., Parkin D.A., Handsworth R.J. A hot-wire liquid water device having fully calculable response characteristics. [J]. Journal of Applied Meteorology, 1978, 17(12): 1809～1813.

[9] King W.D., Dye J.E., Strapp J.W., Baumgardner D., Huffman D. Icing wind tunnel tests on the CSIRO liquid water probe. [J]. Journal of Atmospheric and Oceanic Technology, 1985, 2(3): 340～352.

[10] Lucke J., Jurkat-Witschas T., Heller R., Hahn V., Hamman M., Breitfuss W., Bora V.R., Moser M., Voigt C. Icing wind tunnel measurements of supercooled large droplets using the 12 mm total water content cone of the Nevzorov probe. [J]. Atmospheric Measurement Techniques, 2022, 15(24): 7375～7394.

[11] Renno N.O., Backhus R., Butler T., Cooper C., Hochrein K.A., Madathil R., Marr L., Miller R., Mohan P., Musko S., Ryan T., Saca F., Zewicke J. A new type of aircraft icing detection system. [J]. Scientific Reports, 2026, 16(1).

[12] Aricò M., Piontek D., Bugliaro L., Mayer J., Müller R., Kalinka F., Butter M. A novel machine learning retrieval for the detection of ice crystal icing conditions based on geostationary satellite imagery. [J]. Atmospheric Measurement Techniques, 2025, 18(23): 7129～7152.

[13] Kaufman S., Rosset S., Perlich C., Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. [J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(4): 15:1～15:21.

[14] Varma S., Simon R. Bias in error estimation when using cross-validation for model selection. [J]. BMC Bioinformatics, 2006, 7: 91.

[15] Roberts D.R., Bahn V., Ciuti S., Boyce M. S., Elith J., Guillera-Arroita G., Hauen-

stein S., Lahoz-Monfort J. J., Schroder B., Thuiller W., Warton D. I., Wintle B. A., Hartig F., Dormann C. F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. [J]. Ecography, 2017, 40 (8): 913～929.

[16] 张群洪,李林,杨堃,等.基于可信数据空间架构的高质量数据集建设[J/OL].大数据,1-23 [2026-05-18].
Zhang Q H, Li L, Yang K, et al. High-quality dataset construction based on trusted data space architecture[J/OL]. Big Data Research, 1-23[2026-05-18].

[17] Wilkinson M.D., Dumontier M., Aalbersberg Ij. J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J. -W., Silva Santos L.B. da, Bourne P.E., others. The FAIR Guiding Principles for scientific data management and stewardship. [J]. Scientific Data, 2016, 3: 160018.

[18] Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Daume III H., Crawford K. Datasheets for datasets. [J]. Communications of the ACM, 2021, 64 (12): 86～92.

[19] Pushkarna M., Zaldivar A., Kjartansson O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. [C]. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022: 1776～1826.

[20] 罗春旭,熊海旭,叶雅珍,等.TDQE:一种面向深度学习的文本数据质量评估方法[J].大数据,2025,11(06):95-107.
Luo C X, Xiong H X, Ye Y Z, et al. TDQE: A text data quality evaluation method for deep learning[J]. Big Data Research, 2025, 11(06): 95-107.

作者简介

李德俊（1974– ），男，硕士，湖北省气象工程技术中心（华中区域人工影响天气技术中心），正高。主要研究方向为云与降水物理及人工影响天气技术，专注于雷达和卫星资料在相关业务中的应用。联系邮箱：195318993@qq. com

陈旭（1996– ），女，硕士，湖北省气象工程技术中心（华中区域人工影响天气技术中心），主要研究方向气象数值模拟研究、云微物理参数化方案的改进与开发。联系邮箱：420830134@qq. com

刘睿（1993– ），男，硕士，湖北省气象工程技术中心（华中区域人工影响天气技术中心），主要研究方向为有人机、无人机机载作业数据融合与智能化。联系邮箱：501288582@qq. com

付佳（1981– ），男，硕士，湖北省气象工程技术中心（华中区域人工影响天气技术中心），副高。主持多项省自然科学基金项目，自主开发了多项专业气象服务系统。发表论文多篇，获发明专利多项。主要从事气象大数据治理、人工影响天气技术及跨模态智能计算等交叉应用研究。联系邮箱：2673072@qq.com

陈英英（1982– ），女，硕士，湖北省气象工程技术中心（华中区域人工影响天气技术中心），副高。主要研究方向为大气物理与大气环境，特别是在人工影响天气（云降水物理）和气象水文交叉领域的应用研究。联系邮箱：56912535@qq.com

收稿日期: XXXX-XX-XX

通信作者: 陈英英,56912535@qq.com

基金项目: 湖北省自然科学基金项目(No. 2025AFD414)；中国气象局创新发展专项(No. CXFZ2024J028)；中国气象局创新发展专项(No. CXFZ2025J038)

Foundation Items: Natural Science Foundation of Hubei Province (No.2025AFD414), Innovation and Development Project of China Meteorological Administration (No.CXFZ2024J028), Innovation and Development Project of China Meteorological Administration (No.CXFZ2025J038)