面向气候韧性的地理空间基础模型与深度学习：关于迁移、不确定性和决策准备度的综述

Parastoo Farajpoor[a]、MohammadrezaNarimani[a,*]

[a]加州大学戴维斯分校生物与农业工程系，美国加利福尼亚州戴维斯市，邮编 *95616*

文章信息

关键词：地理空间人工智能 地理空间基础模型 地球观测 深度学习 气候极端事件 领域偏移 不确定性量化 气候适应

摘要

地理空间基础模型能够复用从大型地球观测档案中学到的信息，但这些信息能否迁移到陌生事件、区域和观测条件，决定了它们对气候韧性的价值。本综述将深度学习的进展与灾害监测和早期预警所需的证据联系起来。我们对137项全文灾害应用研究进行了描述性审查，并结合了另一个包含37条模型与基准记录的独立语料库。98项应用研究报告了分布外实验，而在标题所示的评估中，31项研究纳入了留出事件，25项纳入了留出区域，仅有两项纳入了留出传感器。纳入地理空间基础模型微调编码的10项研究中，没有一项将传感器留出作为主要测试。不确定性相关方法见于16项研究，其中两项明确采用了经校准的概率输出。这些相互重叠的计数描述的是编码语料中的评估实践，而非对模型家族的排名。我们区分地球观测表征学习与大气预测，并探讨标签、可观测性、预报提前期和决策要求如何制约它们在洪水、野火、干旱、高温、冰冻圈及级联灾害应用中的作用。随后，我们提出一个决策反推评估框架：明确行动及采取行动时可获得的信息，复现相关分布偏移，保留匹配的特定任务基线，并报告漏报、虚警、可靠性和端到端延迟。该框架将迁移能力这一宽泛的模型主张转化为气候适应和风险管理中可检验的要求。

arXiv:2611.30198v1 [eess.IV] 23 Nov 2026

1. 引言

淹没发生后提供的洪水地图和峰值流量出现前发布的预警，都可能是准确的，但其用途却截然不同。前者支持巡查与恢复；后者则必须足够早地到达，以便改变行动。它们的实用性不仅取决于模型的预测，还取决于哪些信息是可观测的、信息何时变得可用，以及哪些错误对用户而言至关重要。这一区分对于将地理空间人工智能（GeoAI）应用于气候韧性至关重要。能够识别熟悉景观的模型，并不一定能在陌生的极端事件期间保持可靠。

深度学习扩大了卫星影像、环境时间序列和物理模型输出在地球系统分析中的应用。地理空间基础模型通过在大规模异构档案上进行预训练，并将所得表征调整到下游任务中，进一步推动了这一发展。与此同时，机器学习天气模型预测大气场的演变。这些发展为气候监测和适应创造了机遇，但对其进行评估时，必须将计算能力与环境过程和实际决策联系起来 [1–3]。

迁移对于灾害尤其重要，因为下一次事件并不是来自不变分布的随机新图像。它可能发生在不同的流域、生物群系或建成环境中；可能与持续云层同时出现；也可能需要开发期间不存在的传感器和处理链。参考标签也各不相同：地表水、新近淹没土地、活动火灾、烧毁区域和结构损伤是不同的目标。因此，泛化具有多个维度，在一种数据划分下取得高分，并不能证明在所有这些维度下的性能。时空结构化验证为区分这些情况提供了方法论基础 [4, 5]。

现有综述阐述了人工智能在极端事件中的应用、卫星灾害监测方法的发展，以及综合预警系统的机遇。基准研究也表明，看似可比的基础模型得分为何会受到评估设置的影响。我们的贡献与此互为补充：我们考察灾害应用中采用的评估设计，并将其与部署必须支持的决策联系起来。表 1 展示了这一贡献的定位，同时不将架构创新、更广泛的预训练或某个特定模型家族视为业务就绪度的证据 [6–8]。

本综述探讨三个问题。哪些变化在经过编码的灾害文献中，哪些内容得到了实际测试？表征学习有望缓解哪些限制，哪些限制则需要更优质的观测数据、过程信息或决策设计？哪些受控实验能够证明预训练具有实际价值？我们通过固定的描述性审计、对相关模型与应用研究的综合分析，以及一种从决策出发的协议来回答这些问题。本文重点关注气候适应与风险管理：评估改进时，必须以预期采取行动者能够获得的信息和资源为依据。

*通讯作者mnarimani@ucdavis.edu (M. Narimani)

表 1本综述对相关综合分析的补充

综合分析	主要贡献	与本综述的关联
洪水制图综述（Bentivoglio 等，2022年）	深度学习应用、数据与开放问题：洪水制图。	将观测和制图任务与其验证要求联系起来。
极端事件人工智能（Camps-Valls 等，2025年）	检测、预测、理解及其对极端事件的影响。	为应用层审查提供更广泛的地球系统背景。应用层审查。
卫星灾害监测（Illarionova 等，2026年）	基于卫星的人工智能方法的发展及其在环境应用中的应用。	将较新的表征与成熟的灾害监测方法并置。
综合预警（Reichstein 等，2025年）	关联气象、地理空间与决策系统能力。	强调对耦合系统进行发布时验证天气影响系统。
基础模型基准审查（Corley 等人，2026年）	考察……的一致性与可复现性模型评估报告。	补充了当前对灾害特定标题迁移测试的关注。

这些贡献相辅相成。本综述针对固定的137项研究应用语料库补充了描述性计数，并提出了决策反向测试框架；但这并不意味着本综述认为所引述的综合研究排除了所有这些主题。

2. 综述方法与解读

2.1. 证据来源与语料库边界

应用语料通过 35 次 Ope-nAlex 标题与摘要查询构建，涵盖 2015年至2026年的出版物（检索日期：2026年11月1日）；受匿名每日 API 额度限制，每次查询最多检索 200 条相关性排序记录。查询集涵盖基础模型和自监督遥感术语，以及洪水、野火、干旱、高温、冰冻圈、复合型灾害、损毁制图、预测和评估等类别。确切的查询字符串、报告的命中数和检索数均保存在项目检索日志中，可在合理要求下向通讯作者索取。检索共获得 6,803 条原始记录；去重后得到 6,092 个唯一 DOI/标题键；移除近重复的同标题记录后，剩余 5,867 条记录。经过自动标题预筛选和人工标题—摘要筛选，最终保留 141 篇主要候选文献，涵盖洪水、野火、干旱、高温、冰冻圈、损害评估，以及复合型或级联灾害。评估设计编码采用了全文可用的 137 条记录。本文属于结构化批判性综述，并包含描述性证据审计，而非针对预测准确度开展汇总荟萃分析或 PRISMA 系统综述。

另有一组包含37条模型与基准记录的独立资料，用于提供架构背景：18个地球观测基础模型、9个天气或气候模型、5项评估基准和5个基准数据集。这些记录不计入137项应用研究的分母。为更新叙述并核实参考文献而查阅的文献（截至2026年11月22日）为综合分析提供了依据，但不会追溯性地改变审计计数。这样的区分既能保持描述性结果稳定，也有助于讨论近期进展。

2.2. 编码与分析单位

应用研究根据可用全文，按照固定的编码方案进行编码。每条记录均针对灾害、任务、模态、模型家族、预训练、研究地理位置（指地点，而非隶属机构）、主要验证设计、其他分布外测试、基线、不确定性相关方法、可解释性以及代码或权重可用性进行编码。分析单位是研究记录，而非图像切片、模型运行或文中提及的国家。一条记录可以包含多个灾害、模型和验证标签。编码人员仅依据研究文本进行编码：验证描述含糊时，采用较保守的编码；不使用仅摘要证据来推断评估设计。137 条可用应用记录（共核实 138 条）的研究层级编码表，以及平行提取的 37 条基础模型和基准记录，可向通讯作者合理申请获取。

主要验证编码描述了支持研究中主要报告评估的测试。同一篇论文的其他部分可能还包含额外的分布外实验。因此，“任何分布外测试”这一指标与具体主要留出集的计数回答的是不同问题。代码可用性指编码论文中明确提供的可用性声明或链接，并不代表代码库会一直保持可访问。与不确定性相关的编码记录的是报告的方法；它本身并不能证明该方法经过校准、具有覆盖率或适用于决策。

表2定义了综合分析所采用的主要对比项。对于传感器变化，区分未见传感器、缺失模态和针对新输入配置的适应尤为重要。这些设置在实际应用中都可能有意义，但它们并不代表同一种迁移形式。同样，下游测试集分离并不能证明某个地点或采集时段未包含在基础模型的预训练存档中。

2.3. 描述性综合

除非另有说明，百分比均按已编码研究数量除以 137 后乘以 100 计算。模型组比较包括 10 项基于基础模型微调的研究和 127 项其他研究。灾害相关数据使用相应的灾害子集。交叉制表统计同时带有两个标签的记录；因此，按灾害或模型家族汇总的数量并非互不重复的论文数。本文未使用显著性检验来宣称某一模型组的评估文化优于另一组，也未计算跨论文准确率平均值。

表2评估设置及其可支持的论断

设置	所需的数据隔离或控制	解读
随机划分	在抽样数据中划分样本 分布。	评估分布内性能；地理- 图形或事件独立性并不意味着如此。
空间区块	连续空间单元被留出。	测试所选区块尺度下的空间分离； 距离和残差依赖很重要。
留出时间	目标时间不参与拟合、调优和 校准。	当所有预测变量 在所述发布时点均可用时，支持时间性主张。
留出事件	测试事件中的所有观测数据均不纳入 下游开发。	测试新事件；多个独立留出的 事件可确立广度。
留出区域	一个指定的地理域不参与 下游开发。	测试该尺度上的地理迁移；环境 相似性仍然重要。
传感器迁移	使用新的传感器或输入配置进行测试， 明确的适应预算。	区分未见传感器评估、缺失 模态测试与目标传感器适应。
预训练独立性	记录测试地点、时间或传感器是否 出现在预训练存档中。	仅凭下游留出集并不能证明 其独立于预训练。

这些设置可能存在重叠。描述　针对的是某种设计所支持的主张，而非对测试难度的普遍排序。一般方法学基础　s: [4, 5]。

研究所在国家的覆盖情况依据报告的地点，而非作者所属机构。137 条应用记录中，有 110 条具有符合条件的国家代码；编码涵盖 86 个 ISO 3166-1 三位字母代码所代表的地区。一项涉及多个国家的研究会分别计入每个已编码地区，训练区域和评估区域均可能纳入统计。因此，所得地图反映的是证据的地理分布，而非灾害发生率或经独立验证的部署覆盖情况。图表所附的汇总值和编码定义可向通讯作者合理索取。

3. 从观测到决策

3.1. 区分产品与行动

图1对比了两种洪水处理流程。在事件后制图中，将光学或合成孔径雷达（SAR）获取的数据转换为地表水或淹没产品，并针对独立事件或区域进行核查。结果可用于确定现场核查的优先级。在洪峰前预警中，输入必须是在预报发布时间可获得的信息：降雨、上游流量、气象以及其他可接受的先前信息。预警只有满足所需提前期且具有可接受的可靠性时才有用。回溯地图即使具有很高的准确性，也无法单独证明这种预测能力。

任务定义决定了什么才算有效标签。卫星过境时观测到的水体不一定代表最大洪水淹没范围，永久性水体也不等于新近被淹没的陆地。数据获取时机、云、冠层、地形和混合像元都会改变标签与目标事件之间的关系。洪水制图和云感知光学制图系统的综述说明了为何传感器特征和参考数据构建必须纳入评估，而不能仅被视为预处理细节 [6, 9]。

3.2. 使观测与过程相匹配

编码研究最常使用多光谱光学影像（70）、数字高程模型或静态协变量（54）、合成孔径雷达（46）、再分析或气象信息（34）、现场观测（29）、粗分辨率光学数据（29）、超高分辨率光学影像（28）以及热红外数据（21）。这些是相互重叠的模态标签。它们的多样性反映了不同的测量问题：活跃火灾信号、冠层温度估计和雪水当量估计，并不是同一现象的可替代标签。

对于运行层面的比较，所有模型都应接收明确定义的信息集。使用更多波段、较晚观测数据或额外局部标签的模型具有信息优势，这种优势不应完全归因于其架构。在 H2O-Net 的 PlanetScope 评估中，RGB 模型的平均交并比（mIoU）达到60.92%，而 RGB U-Net 为42.63%。加入近红外后，H2O-Net 的结果提升至74.52%。匹配的RGB比较测试的是方法；额外波段的比较则改变了该方法可获得的信息[10]。这一点看似简单，但在解读关于多模态或基础模型优势的论断时至关重要。

4. 基础模型：能力与证据边界

4.1. 地球观测表征学习

图 2 概述了两种预训练机制，并非对其灾害表现进行排名。SatMAE 从掩码化的时序和多光谱卫星观测数据中学习。其表征经过训练，能够根据可见上下文恢复缺失信息。CROMA 通过多模态融合结合成对的光学和雷达编码器、跨模态对比学习与掩码重建。这些目标有助于学习实用的表征，但下游洪水或野火表现仍取决于适配数据和评估设置 [11, 12]。

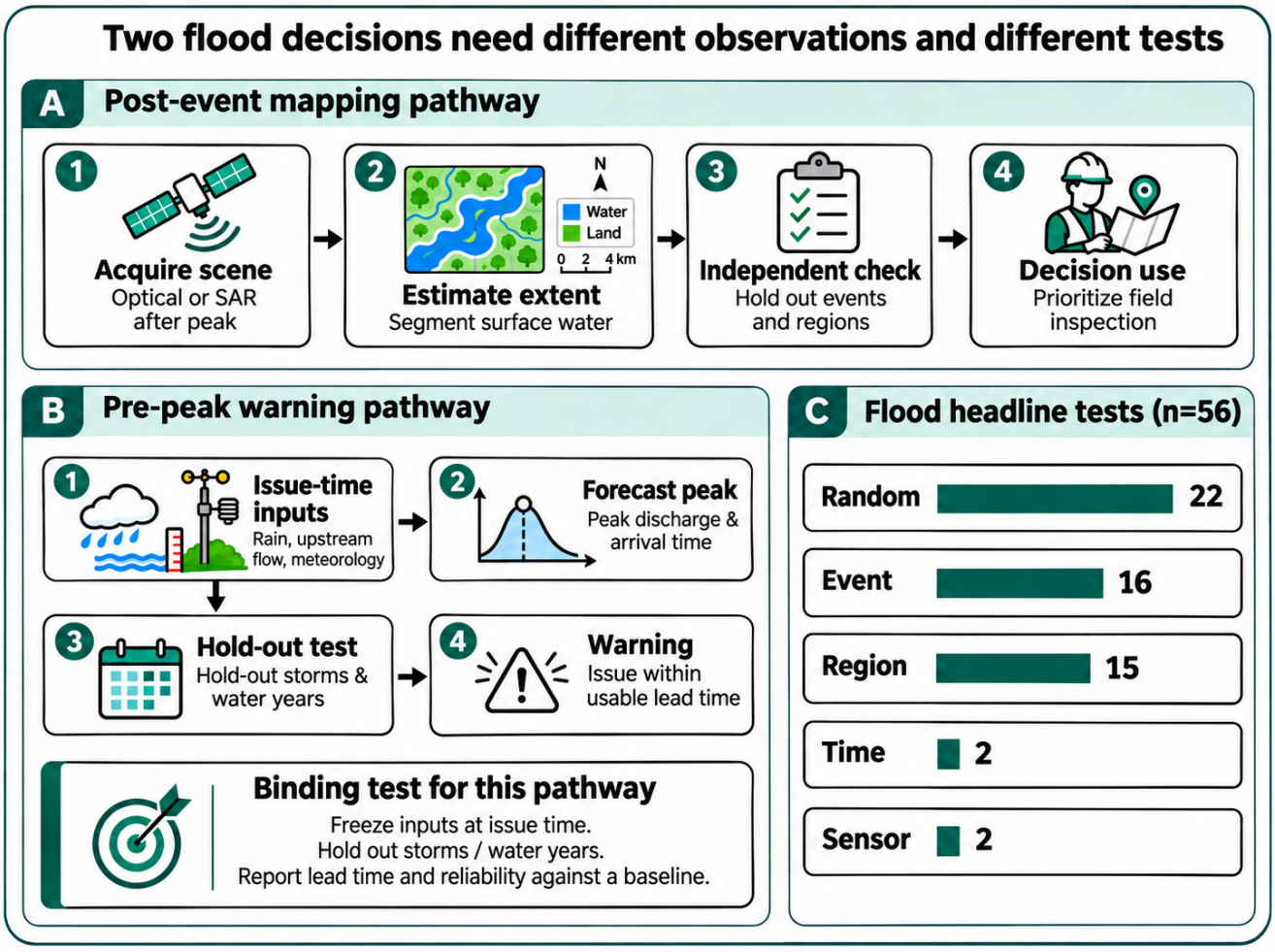

图1： 两种洪水决策需要不同的观测数据和不同的测试。(A) 灾后制图路径。(B) 洪峰前预警路径。(C) 可用审查中的洪水研究所采用的主要验证设计。

当业务传感器配置发生变化时，这一区别便显得尤为重要。模态内预训练或许适用于观测数据与训练输入相似的下游任务。当一种模态不可用时，跨模态对齐为研究迁移提供了理由，但它本身并不是对迁移效果的测量。为便于阅读，示意图中对 CROMA 的独立编码器和融合机制进行了简化。因此，缺失一种输入时的测试、使用新卫星进行的测试，以及获取目标传感器标签后进行适应的测试，都应分别报告。

近期模型拓展了输入和适配策略的范围。DOFA依据光谱信息调整处理方式，以适应异构传感器；AnySat应对不同的模态、分辨率和尺度；TerraMind将多模态学习与生成能力相结合；Prithvi-EO-2.0则拓展了多时相地球观测预训练 [13–16]。这些设计选择各不相同，其优势也各有不同，且可以通过测试加以验证。对于灾害应用用户，关键问题在于：在标签、输入可用性和计算资源相同的条件下，某种特定选择是否能提升指定任务的表现。生成或推断出的模态不得被视为对环境的独立测量。

表3按照代表性模型和基准测试的贡献以及仍需验证的内容进行了归类。这种做法取代了跨论文的准确率排行榜，转而比较各类机制及其所需的证据。它也避免将模型规模、已发布的嵌入和可下载的模型权重视为能力或可复现性的可互换指标。

4.2. 大气预报是一条并行路径

盘古气象和 GraphCast 展示了中期大气预测方面的进展；GenCast 生成概率预报；NeuralGCM 将学习得到的组件与大气动力学相结合；Aurora 则将基础模型方法应用于多种地球系统预测任务 [17–21]。它们的输出和验证设置不同于用于分割洪水水体或烧毁地表的编码器。两条技术路线互不构成对方的最终阶段。

贯通这些发展脉络有助于实现影响导向预警：大气预报为水文、火灾或暴露度模型提供信息，地球观测数据则用于更新陆地表面的状态。然而，每一项连接都会带来新的不确定性和时效要求。一次有用的耦合测试

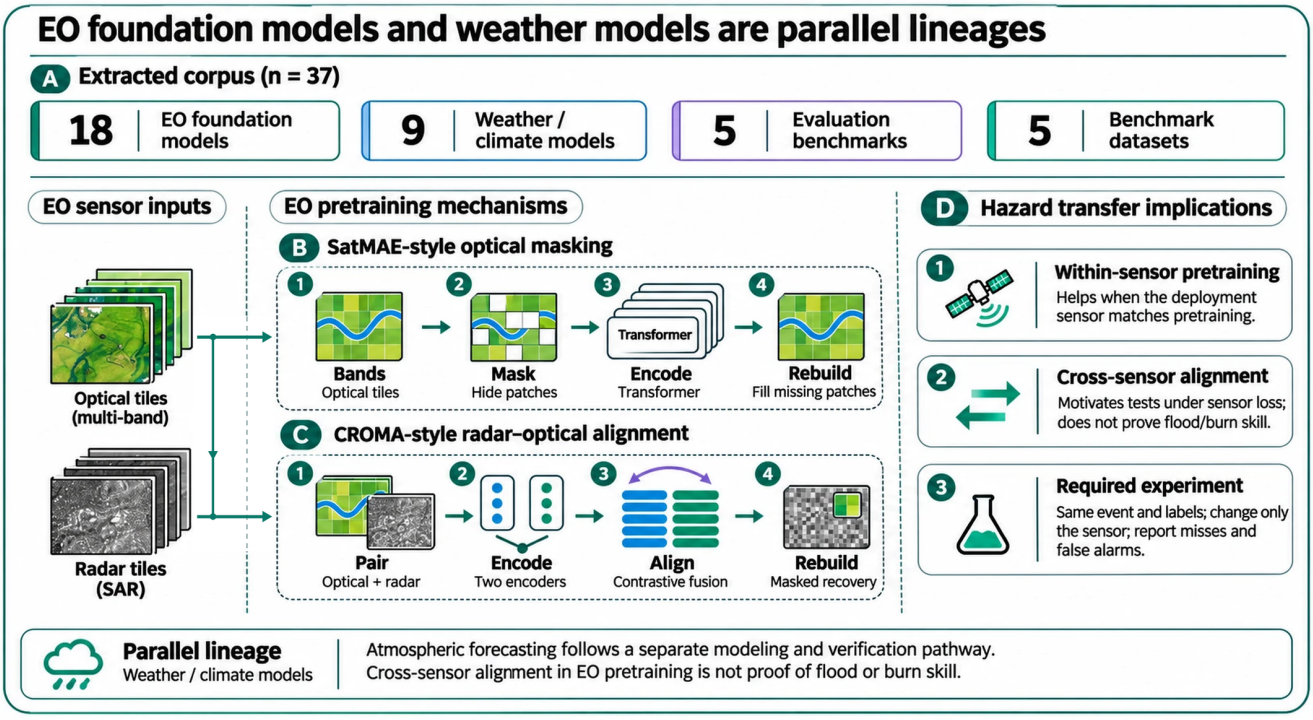

图 2： 地球观测基础模型与天气模型是并行发展的两条谱系。(A) 提取语料构成。(B) SatMAE 风格的光学掩码路径。(C) CROMA 风格的雷达–光学对齐路径。(D) 对灾害迁移测试的启示。

应追踪预报发布时间、驱动不确定性、空间分辨率、观测可用性以及下游决策阈值。Reichstein 及其同事提出的整合方案提供了一个研究方向，[3]但这并不意味着可以据此推断，某个耦合系统已针对每一种灾害完成验证。

4.3. 基准测试得分是有条件的证据

GEO-Bench 和 PANGAEA 对多样化的地球观测任务进行了组织，并揭示了一致的适配设置和强有力的监督基线的重要性 [22, 23]。基准测试结果取决于其标签、数据划分、解码器、调优预算和指标。这些选择之间的差异可能改变模型的排序。Corley 及其同事对基础模型报告的审查进一步强调，有必要检查底层比较，而不是从另一篇论文中借用“最先进”这一标签 [8]。

对于灾害应用，反复使用一个熟悉的事件也可能缩小证据的有效多样性。许多论文在一个留出洪水事件上进行评估，并不等同于在多个相互独立的洪水事件上进行测试。数据集级分离、事件级重复验证和预训练来源应一并报告。新模型的科学价值，只有在其所宣称的优势经受住受控比较，并且在数个相关变化中仍然可见，而不是依赖于一次有利的划分时，才最为清晰。

5. 不确定性、解释与运行可靠性

5.1. 不确定性报告并不等同于校准

图3展示了应用语料中的不确定性相关编码。16项研究（11.7%）记录了相关方法：其中5项使用了未经校准的概率输出，4项使用了集成，2项明确使用了校准概率，另外5项采用了其他与不确定性相关的方法。其余121项研究在该编码方案下未记录任何方法。这些类别描述的是报告实践，并不是对每个不确定性估计质量的评分。特别是，介于零和一之间的输出并不会自动成为可信的概率。

校准关注的是：被赋予某一给定概率的事件，是否以相应的频率发生。现代神经网络可能需要显式校准，而在分布偏移下，不确定性表现可能恶化 [24, 25]。因此，对于洪水地图，应在用于行动的阈值上，针对留出事件或区域检查可靠性。对于预警，概率验证应以提前期和事件严重程度为条件。总体校准统计量可能掩盖模型在罕见但后果严重的结果上的不良表现。

所需的不确定性摘要取决于产品。分割任务可能需要空间误差摘要和事件级置信区间；洪峰流量预报可能需要预测区间、阈值超越概率以及按提前期划分的覆盖率。集成离散度可能具有参考价值，但模型之间达成一致并不能证明

表3代表性模型机制及灾害应用所需证据

模型或基准	可复用能力	仍需的特定灾害证据
SatMAE（Cong 等人，2022年）	掩码时序与多光谱表征迁移学习。	事件/区域迁移 and 明确输入-可用性测试。
CROMA（Fuller 等人，2023年）	独立的雷达和光学编码器、对比对齐和融合重建。	区分缺失模态与未见传感器；测试每个运行输入集。
DOFA；AnySat（Xiong 等人，2024年；Astruc 等人，2025年）	对异构光谱输入的适配，模态或尺度。	匹配的标签预算，以及针对目标传感器和分辨率的明确测试。
Prithvi-EO-2.0；TerraMind (Szwarcman et al., 2026年; Jakubik 等，2025年)	多时序预训练以及多模态或生成式表征。	与强大的特定任务模型进行消融比较；任何推断模态的来源信息。
GEO-Bench；PANGAEA（Lacoste 等，2023年；Marsocci 等，2026年）	共享任务和标准化模型评估设置。	核实基准标签、数据划分和评估指标是否符合所提出的灾害应用。
盘古气象；GraphCast（Bi 等，2023；Lam 等，2023）	大气场的学习演变。	相关预报时效上的极端事件验证，遵循-随后进行下游影响评估。
GenCast（Price 等，2025）	概率大气预报。	针对特定情形的校准和决策价值影响阈值和区域。
NeuralGCM（Kochkov 等，2024年)	混合学习组件与大气动力学。	在时间尺度和过程上的验证应用相关的时间尺度和过程上进行验证。
Aurora（Bodnar 等人，2025年）	适应性地球系统预测表征。	在业务信息约束下进行特定任务的极端情况与迁移测试业务信息约束。

这是一项精选的方法学比较，并非对37条记录的语境语料库进行完整盘点。本文不对不同任务或规程的数值精度进行排名。

其正确性，因为它们可能共享输入、标签或结构偏差。校准数据必须与最终评估数据分开。重采样也应以独立单元为依据，例如事件或流域，而不是将相邻像素视为独立重复样本。

5.2. 解释应回答一个明确的问题

可解释性方法在27项研究（19.7%）中有所记录。视觉上合理的显著性图或特征排序有助于诊断模型，但并不能证明该解释忠实地代表了学习到的关系。显著性合理性检查表明，仅凭视觉吸引力是不够的 [26]。解释应与具体问题相联系：模型是否依赖不合理特征、其行为是否在不同环境中发生变化，或是否能在使用前识别输入故障。

对于影响研究，还必须将预测性解释与因果归因区分开来。例如，植被与结构损失之间的关系可能有助于进行风险筛查，但不能证明仅改变植被就会产生预测中的损失变化。在将可解释地理人工智能转化为规划建议时，空间验证和明确的观测性解释十分重要。Palisades火灾研究近期提供了一个例子，说明了基于模型的关联与干预性主张之间的区别 [27]。

5.3. 可复现性和延迟也是结果的一部分

52项研究（38.0%）明确提供了代码可用性。表4将这一结果与其他信任相关指标汇总在一起。可复现性所需的不仅仅是模型名称：预处理版本、波段定义、归一化、标签规则、数据划分标识符、适配设置和决策阈值共同决定了实验。FAIR 原则为组织这些资源提供了实用依据，便于其他用户查找和重复使用 [28]。

端到端延迟同样重要。推理时间只衡量了从数据获取或预报到交付产品这一流程中的一个环节。数据访问、质量控制、大气校正、配准、切片、后处理和审查都可能成为运行延迟的主要来源。用于部署的比较应报告完整延迟，以及测量该延迟时所使用的硬件和输入条件。当效率能够在资源受限用户所需的、与决策相关的性能不受影响的情况下带来优势时，它才是真正的优势；效率不能替代这种性能。

6. 应用审查揭示了什么

6.1. 模型家族与灾害覆盖率

审查涵盖了56个洪水标签、51个野火标签、16个干旱标签、14个损害标签、11个冰冻圈标签、9个高温标签、5个复合型灾害标签和5个其他灾害标签。由于研究可以对应多个标签，这些数字反映的是覆盖率，而非互斥类别。图4展示了相应的模型家族交叉统计。记录显示，编码器—解码器分割出现在71项不同的研究中，卷积网络出现在38项研究中，Transformer或视觉Transformer出现在31项研究中，经典机器学习比较模型则出现在29项研究中。明确命名的地理空间基础模型微调出现在10项研究中。

气泡矩阵中较大的行合计表示模型家族标签的分配次数，而非独立研究的数量。一篇比较 U-Net、Transformer 和经典模型的论文会计入多个单元格。同样的区别

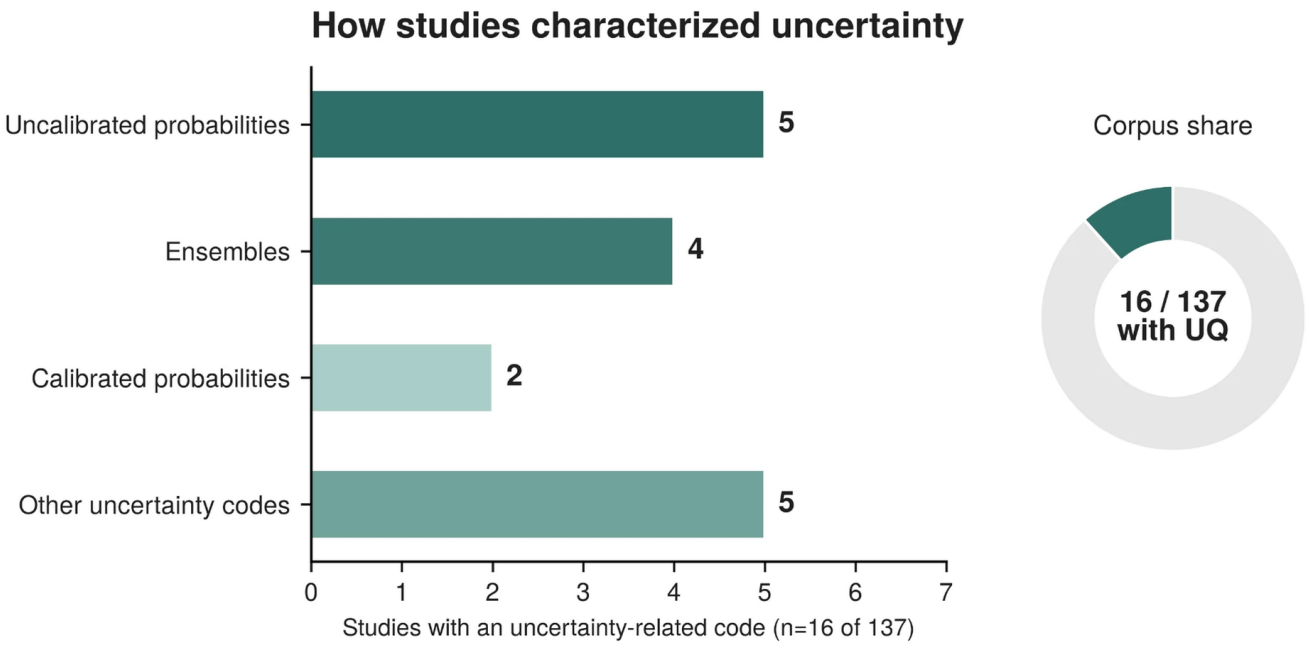

图3：137项研究语料库中的不确定性相关报告。

也适用于灾害：一篇同时涉及洪水和损害评估的论文会计入两行。该图描述的是文献库的方法构成；它无法证明常见的模型家族更准确，也无法证明较少见的模型家族不太适用。

6.2. 证据位于何处

图5绘出了110项可在地理位置上归属的研究所记录的位置。美国在各国提及次数中占比最高，其他地区的分布则不均衡。解读这一分布时，应同时考虑洪水和野火研究的突出地位，以及多国基准数据集的重复使用。一国可能出现在训练数据、测试数据或两者中；某个国家以填充多边形显示，并不能证明模型已在该国所有环境中接受独立测试。

实际应用中，应在进行外推之前评估地理适用性。即使在同一国家内部及跨越国界的地区，环境相似性、传感器特征和标签可用性也可能存在显著差异。适用范围分析有助于识别与训练数据不同的条件，但它只能补充、不能取代独立测试 [29]。因此，报告实际评估地点及其环境特征，比展示数据覆盖范围的全球地图并据此证明全球就绪度更有参考价值。

6.3. 哪些变化得到了重点测试?

98 项研究（71.5%）报告了至少一项分布外实验。具体的主要编码包括：48 项研究采用随机划分（35.0%），31 项采用留出事件（22.6%），27 项采用留出时间（19.7%），25 项采用留出区域（18.2%），7 项采用空间区块（5.1%），2 项采用留出传感器（1.5%）。图6展示了这些编码在不同灾害中的分布。这些编码彼此重叠，因此不能将其视为语料的划分，也不能将其理解为互不重叠的群体依次达到的不同等级。

随机划分是展示的类别中占比最大的单一类别，但并未构成语料库的大多数。对于在明确界定的数据分布内进行插值的问题，随机划分可能是合适的。当插值结果被用来支持有关新事件、新地点或新传感器的论断时，其局限性便显现出来。反过来，仅仅报告一项额外的迁移实验，并不能证明模型在特定部署场景下具有足够的性能。审查记录了是否进行了测试；测试结果的可靠性仍取决于留出条件、参考质量和比较设计 [4]。

重点传感器测试在展示的验证设计中最为缺乏，是最集中的缺口。这促使我们针对传感器变化开展实验，尤其是在运行连续性取决于云层遮挡时使用雷达，或以一个星座替代另一个星座的情形下。此处并不能据此断定所有模型都会在此类变化下失效，也不能说更广泛的文献中只有两篇论文曾考虑传感器差异。这里的推论仅涉及这一编码样本中的主要测试设计。

6.4. 基础模型研究与比较组

图7采用组内百分比而非原始计数，比较10项基础模型微调研究与其他127项应用。4项基础模型研究和27项其他研究采用了事件留出；时间留出分别为2项和25项；区域留出分别为2项和23项；传感器留出分别为0项和2项。相应的事件占比分别为40.0%和21.3%，但两组在规模和构成上差异很大。这些差异仅为描述性结果，

表4137项研究应用语料库中的报告与评估指标

指标	研究	占比（%）	指标含义
任何分布外测试	98	71.5	至少报告了一项事件、区域或时间或传感器测试的某些——研究中的某处。
Any 不确定性相关方法	16	11.7	包括未经校准的概率——性，集成及其他变异性/情景方法。
经过明确校准的概率值	2	1.5	16 种不确定性方法中的一部分相关记录。
可解释性方法	27	19.7	报告了一种方法；其忠实度或因果解释并不意味着这些方面得到了保证。
明确说明代码可用性	52	38.0	代码可用性声明或链接出现在编码论文中。
地理空间基础模型微调	10	7.3	A 命名 地球观测/地理空间基础模型经过适配以适应该任务。
标题随机划分	48	35.0	离散随机数据划分标题编码。
标题传感器留出	2	1.5	离散标题编码；并非对所有传感器相关实验的全面统计。

每个百分比分母均为 137。各项指标之间存在重叠，不应将计数相加。发表时可用性编码不同于当前的 y 代码仓库审计。

并不能证明微调会导致采用某种特定的验证实践。

微调组中传感器留出测试的数量为零，这表明存在证据缺口，并非对基础模型成功迁移概率的估计。要确定比较性的迁移性能，需要进行匹配实验，最好涵盖多个事件和传感器条件。因此，这项审查指向的是一个研究重点，而非对模型类别的定论。

7. 不同灾害需要不同的测试

表 5 将灾害覆盖范围与观测数据、目标和限制因素联系起来。这些因素是关于特定灾害—任务对的工作假设，并非整个灾害类别的永久属性。同一场野火可能涉及图像分割问题、蔓延预测和火灾后水文评估，而每项任务的信息需求各不相同。

7.1. 洪水：水体制图与流量预测

洪水应用最清楚地体现了表征学习与预测之间的差异。Sen1Floods11 支持了共享的洪水制图实验，而 Kuro Siwo 则扩展了多时相 SAR 观测和人工标注事件的范围 [30, 31]。这些数据集支持可复现的比较，但结果的适用范围仍取决于标签和事件的选择。云感知光学系统进一步表明，业务化制图涉及数据获取可用性和端到端处理，而不仅仅是分割 [9]。

基于 Prithvi 的研究表明，为何基线也必须受到同等重视。早期研究评估了地理空间基础模型在洪水淹没制图任务中的表现，而 U-Prithvi 则将预训练编码器与 U-Net架构相结合 [32, 33]。在后者强化基线的对比中，玻利维亚的平均交并比为：U-Net 87.77%，U-Prithvi 87.70%。两者几乎持平，并不意味着预训练无效；这表明，当任务专用对照模型得到加强时，结论也会随之改变。因此，模型的优势应归因于经过测试的具体配置，而不能仅凭基础模型标签来判断。

近期一篇关于紧凑模型的预印本，也在同一水体制图任务下比较了预训练、随机初始化、冻结特征、U-Net 和计算成本。其结果表明，预训练编码器体现的是准确率—效率权衡，而非普遍优越性 [34]。对于预警，相关证据则有所不同：无测站流域全球预测等预测研究，评估的是未来河川流量，而非卫星观测水体掩膜 [35]。在有实际预警价值的提前期内，地图基准无法取代流域与事件留出集测试。

7.2. 野火：区分检测、事后影响与蔓延

活动火灾探测用于识别正在发生的热信号或光谱信号；烧毁区域制图用于识别发生变化的地表；严重程度和结构损失分析则用于描述后果。火灾蔓延和危险性预测关注的是

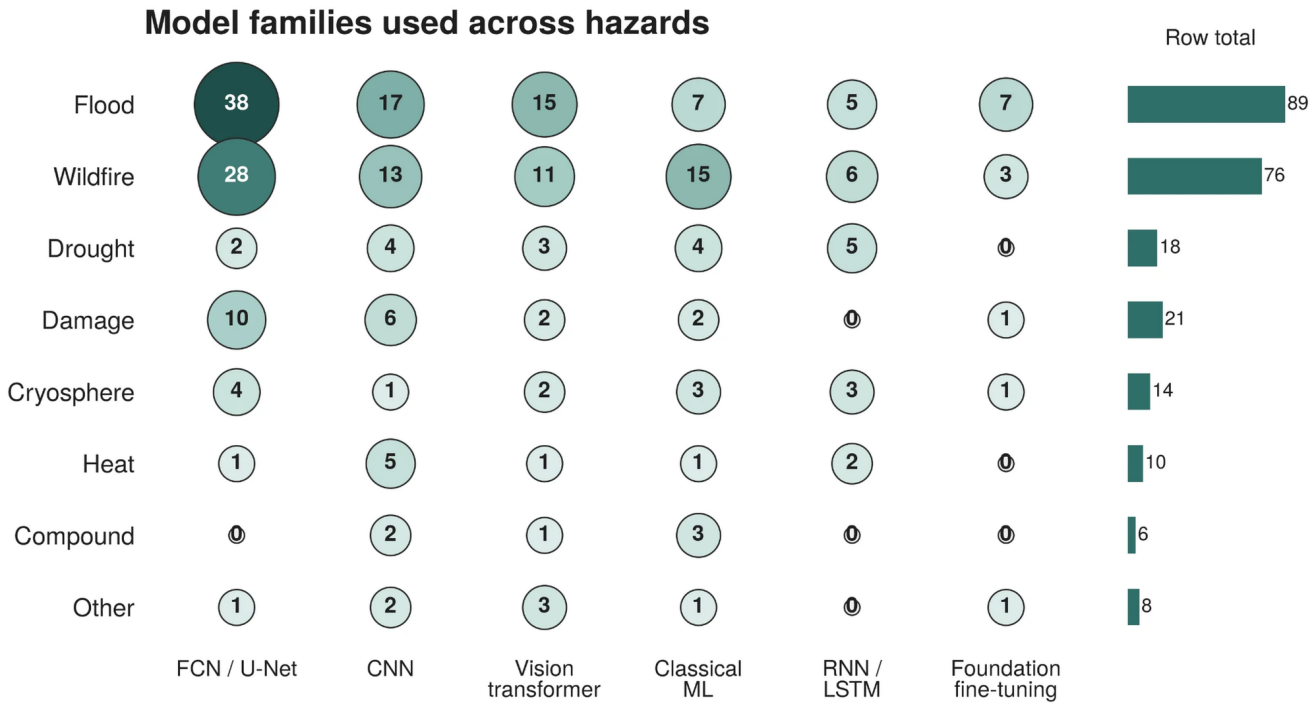

图 4：跨不同灾害使用的模型家族。

未来状况。混淆这些目标，可能会让一种方法看起来解决了比实际更广泛的业务问题。结合动态环境信息与地理背景的危险性预测模型表明，天气与位置必须结合起来考量，而不能用更大型的光学编码器取而代之（Shams Eddin 等，2023年）。

按事件划分的数据集很有价值，但标签生成过程仍是任务的一部分。涵盖 25 场加利福尼亚州野火的 Sentinel-2 活动火灾基准数据集，将每场野火分配至单一划分，并提供明确的元数据和基于规则生成的掩膜 [36]。此类数据划分可防止同一场野火同时出现在训练集和测试集中；但这并不意味着光谱规则就成了对每个火灾边界的独立观测。当论断超出复现标签规则的范围时，还需根据另行审核或独立观测的目标进行评估。

对于建筑损失筛查，事件前输入、空间分隔，以及关联与因果关系之间的区别尤为重要 [27]。即使预测判别能力具有价值，关于采取何种措施或改变土地利用的决策，也需要比模型解释更多的证据。业务化野火比较应报告复杂情境下的漏检情况，例如小型火灾或复杂的交界区域，同时也应报告总体分割或分类得分。

金凯德山火的葡萄园分析提供了一个相关的景观尺度案例：多源观测数据可以表征田块尺度的野火影响和暴露，但土地覆盖与火烧严重度之间的空间关联本身并不能作为管理干预效果的可迁移估计 [34]。当遥感证据用于支持农业和城市景观中的适应规划时，这一区别至关重要。

7.3. 干旱：参照定义了预测

干旱可通过若干彼此相关但各不相同的目标来表征，包括土壤水分、植被状况和分类干旱评估。DroughtCast 预测美国干旱监测，而植被健康和土壤湿度预测研究则预测不同的环境量 [37–39]。相对于一种参照，模型可能表现出较高的预测能力，却无法再现另一种参照。因此，报告的准确率在一定程度上取决于目标定义、空间支撑、聚合方式和预测时域。

若要声称具备业务化预测能力，时间留出集是必要的；但如果预期用于相关场景，还应补充针对罕见严重程度类别和地理迁移的测试。应筛查输入，确保其在发布时点可用，包括再分析数据和监测产品的发布时间。一个有用的实验是，在保持数据划分和评估尺度一致的同时，比较更改参照目标与更改模型所产生的影响。这样可以检验标签定义假设，而无需假定目标选择总是比架构更重要。

7.4. 高温：从大气预测能力到暴露

高温应用将大气预测、近地面气温、地表温度与人群暴露联系起来。这些变量彼此相关，但不能互换。卫星热红外观测提供地表信息；要理解人们所处的环境，还需要地面观测和大气模型。次日

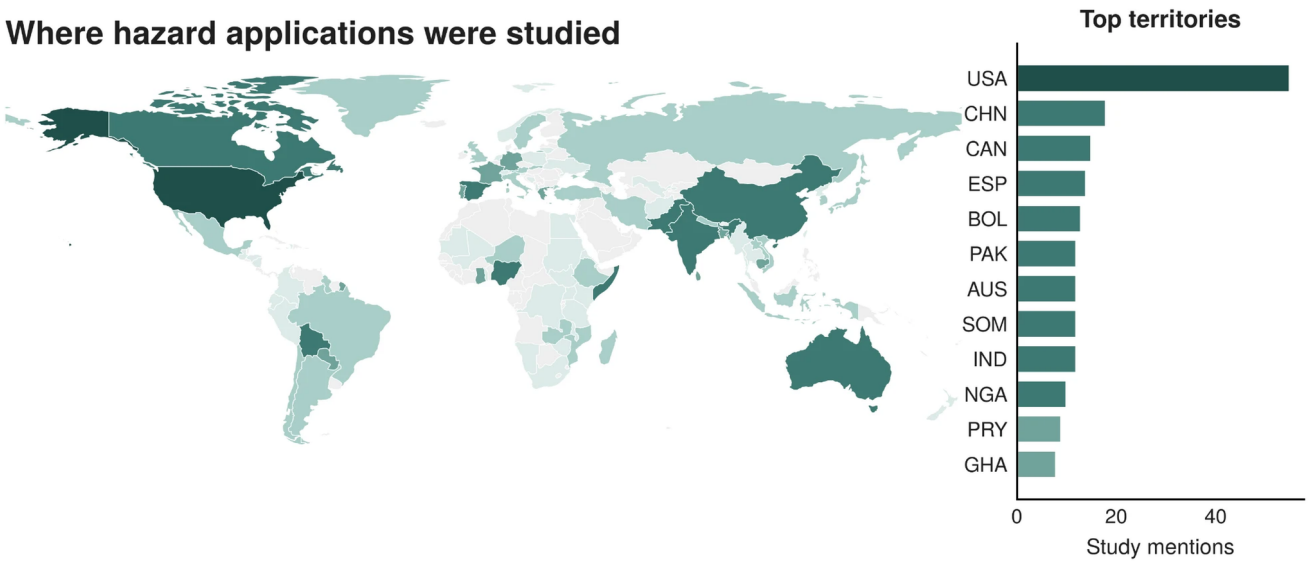

图5： 研究了哪些灾害应用。

结合卫星与地面测量进行的冠层城市热岛预测，体现了明确定义暴露相关目标的价值 [40]。

对近期高影响热事件的测试表明，较强的平均天气预报性能并不意味着在极端情况下也能取得同样出色的表现 [41]。这类结果揭示了评估系统和预报时效的局限；但单凭这些结果，并不能证明存在根本性的物理可预测性极限。业务化热风险评估应按强度、提前期以及相关人口或城市环境对误差进行分层，同时保留适当的预报基线。审查中提及的九个热事件标签只是一个小型证据子集，并非衡量热风险社会重要性的指标。

7.5. 冰冻圈：参照数据稀少且观测条件不断变化

雪水当量估算、冰川前缘 delineation 和冰川湖制图，对参考数据和观测的要求各不相同。将用于雪水当量估算的长短期记忆模型迁移到其他区域，与调整图像编码器以描绘湖泊或冰川前缘，是不同的问题 [42, 43]。因此，应明确指出独立区域、采集条件和参考数据来源，而不应笼统地用冰冻圈精度主张加以概括。

SSL4SAR 报告称，在 CaFFe 数据集上，冰川崩解前缘的平均误差为 293 m。在另一个多标注者子集中，其集成结果为 75 m，而人工标注之间的分歧为 38 m [42]。子集比较表明，相对于专家一致性，模型仍存在误差；不同的评估人群不得合并为单一的模型与人工对比。标注变异性是相关的参考数据属性，但不能证明误差下限不可避免。

7.6. 损害与级联影响

损害制图常常面临成像条件、建筑形态和灾害背景的变化。将全球建筑制图模型迁移到异质的灾后影像上的研究表明，在尝试进行任何后续损害分类之前，提取问题本身就可能发生变化 [44]。因此，留出灾害应保留业务目标，并明确说明是否使用了局部标签、事件前影像或传感器特定适配。

级联灾害为各个过程之间增添了时间联系。火烧严重度会影响后续泥石流灾害，而火灾相关冠层变化则会改变积雪持续性。将火灾前条件、火烧严重度与泥石流暴露联系起来的研究，以及近期关于野火严重度与积雪持续性的多源研究，展示了这些关联路径 [45, 46]。可信的级联测试必须追踪事件序列，并在各个阶段传递不确定性。第一张地图的准确率很高，并不能证明下游影响预测具备良好能力。审查中明确涉及复合灾害的五项研究，表明有必要建立共享的事件链基准测试，而不是据此得出级联风险本身发生频率的结论。

8. 从决策出发反向选择测试

8.1. 选择数据划分前，先明确要验证的主张

所提出的协议从行动出发，而不是从架构出发。说明谁将使用输出、输出会为哪项决策提供依据、该决策何时作出，以及届时有哪些信息可用。明确目标、空间支撑、相关事件严重程度、可接受延迟，以及漏报和虚警带来的成本或运营后果。随后，评估应重现可能导致该决策失效的变化：新事件、区域、季节、传感器或输入可用性条件。

表5灾害任务约束与优先级迁移测试

灾害（研究标签）	观测与目标	实际限制因素	优先级测试
洪水（56）	光学/SAR水体制图；降雨和流量，用于预警。	云、冠层和地图的标签定义；预测的驱动因子和提前期。	新事件/区域；传感器替换；签发时刻预测验证性。
野火（51）	热学/光学活跃火灾、烧毁区域、严重程度；用于危险性/蔓延的气象。	制图的可观测性；过程演况和天气预测。	区分目标；留出火灾和年份；评估小事件漏检。
干旱（16）	土壤水分、植被状况和类别型干旱参考数据。	参考定义、聚合方式和预测期限。	留出年份/区域；比较具有共同空间和时间支撑范围的目标。
损害（14）	甚高分辨率影像、建筑或基础设施参考数据。	标注成本、场景异质性和跨灾害差异。	留出灾害/站点；说明事前事件输入和本地适应。
冰冻圈（11）	积雪覆盖/雪水当量、冰川前缘和冰川湖。	稀疏独立参考数据和传感器/季节歧义。	区域和季节留出集；参考数据不确定性和采集—条件。
高温（9）	大气场、地表热观测数据和气温参照—数据。	预报提前期以及从场到暴露的转换。	高影响事件测试；针对提前期的特定可靠性和与暴露相关的目标。
复合型（5）	配对事件链，例如火灾随后发生降雪或水文变化。	耦合过程、时间滞后和对齐标签。	保留完整的事件链；传播 不确定性 跨越阶段。

这些约束是可检验的假 设，适用于特定任务，而非灾害的固定属性。危害标签存在重叠。另有五条记录包含其他灾害 d 标签。代表性证据将在第7节中讨论。

表6将这些主张与具体测试和报告要求联系起来。图8中的“迁移阶梯”是一种组织工具，并非对难度的通用排序。一个设计良好的

g定的术语p口头预报测试可能比

表6

决策 -与标题审计计数相关联的回溯测试 s

部署问题	标题研究（%）	评估要求	面向决策的报告
抽样分布内的插值分布	48（35.0）	采用明确抽样单位的随机划分，且不作更广泛的迁移主张。	按相关类别和环境条件统计误-差。
用于空间分离区域	7 (5.1)	留出空间区块；说明区块尺度和间隔距离。	空间变异、失效位置和适用性。
在未来季节或时段中使用	27 (19.7)	保留时间；将输入限制为可用的发布时点。	提前期误差、可靠性和时间漂移。
在新灾害事件期间使用	31 (22.6)	将完整事件排除在拟合、调以及校准。	事件级漏检、虚警和不确定性。
在新区域中的应用	25 (18.2)	留出一个区域，并记录环境心理和标签差异。	区域性能和适用性以外的条-件。
在传感器或模态变化	2（1.5）	指定未见传感器、消融或适配——性；匹配标签和预算。	传感器条件下的误差、可用—性以及从数据到产品的完整延迟。

计数采用图8中的离散标题编码；不会暗中加入随机交叉验证和空间交叉验证。编码之间存在重叠。对于所有迁移主张，保留匹配的特定任务/业务基线，记录预训练重叠情况，并在部署环境中评估不确定性。

8.2. 实验1：在传感器变化下保持制图

设想某个机构在云层阻碍光学数据采集时，仍必须继续开展洪水制图。受控实验应定义相同的目标事件和参考标签，匹配适应标签与调优预算，并比较可用的传感器配置。除预训练模型外，还应保留特定任务基线。在可能的情况下，应使观测时间、空间分辨率以及各传感器所承载的信息保持一致；否则，应将这些差异作为运行比较的一部分予以报告。

应将两个问题区分开来。受控传感器比较关注的是：在条件高度匹配的情况下，观测配置发生变化时，性能会如何变化。业务替代测试关注的是：当无法实现完美的输入一致性时，现有产品中哪一种能提供最佳决策支持。如果使用了目标传感器标签或模型更新，就不应将其描述为零样本传感器迁移。

地理位置遥远但环境相似的图像测试更具挑战性。同样，传感器测试不能替代事件测试。所需的组合取决于部署声明。必须披露预训练重叠以及任何目标域适应情况。

应在最终测试前选定阈值。报告漏判的淹没区域、虚警、事件级性能和不确定性，并按与响应相关的条件对失效情况进行分层，例如云层、森林、密集开发区和狭窄水体。报告从数据可用到地图交付所需的延迟。只有在平均重叠度的提升不会掩盖妨碍预期巡查或响应决策的失效模式时，这种提升才有意义。

8.3. 实验2：在峰值到来前发出预警

对于预警服务，应在每个发布时点冻结信息集，并评估未用于模型拟合、调优和校准的风暴或水文年。历史数据应按照其在业务运行中实际可获得的情况进行重建，而不是补充之后的观测数据或事后修订的产品。基线应

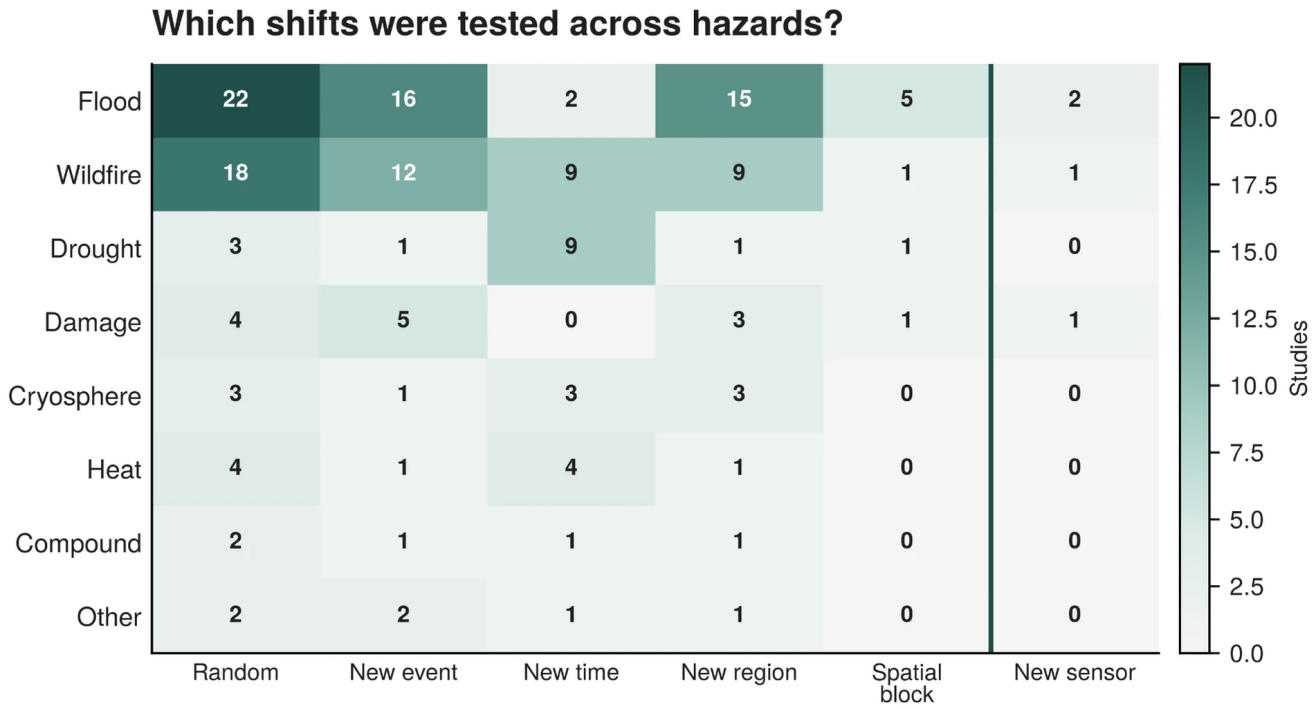

图6：在不同灾害中测试了哪些变化?

反映该服务实际可采用的替代方案，例如业务预报、水文模型或适当的简化统计方法。

评估峰值幅度、到达时间、是否超过阈值、可用提前期和概率可靠性。按事件严重程度报告漏报和虚警，而不要只报告许多平静时段的平均误差。比较时必须确保各竞争方法使用相同的发布时点和允许输入。对于天气到影响的耦合系统，应将大气不确定性传递到下游组件，而不是把单条最佳大气轨迹当作确定结果来评估。

8.4. 评估预训练是否会改变决策

最小化的归因实验应比较预训练架构、随机初始化的对应模型、强大的特定任务模型，以及与行动相关的业务或经典基线。应匹配输入波段、时空覆盖范围、标签和调优预算，并记录无法匹配的差异。改变可用下游标签的数量，以检验标签效率，而不是仅凭一个训练集推断。使用不同随机种子进行重复实验，可量化训练变异性；重复留出事件则可量化另一种变异来源。应分别识别这两种来源，而非不加区分地合并。

实际输出是在与决策相关的误差、提前期、成本和覆盖率之间进行权衡。当紧凑型模型支持在有限硬件上部署时，它可能更可取；而当可复现的性能提升能够改变行动方案时，采用更大型号也可能是合理的。结果应明确模型在哪些条件下能发挥作用，而不是简单地宣布某个模型在从未同时评估过的各种场景中胜出。

9. 讨论与研究重点

9.1. 从模型能力到有效迁移的证据

审查发现，潜在部署条件的广度与作为主要评估所报告的测试的具体性之间存在失衡。这并非反对基础模型，而是主张开展能够明确其贡献的实验。当可复用编码器能够提升明确界定的任务表现、减少所需标签或资源，或在有意义的分布偏移下扩大可靠覆盖范围时，它便具有价值。应直接证明这些成果。

由此提出四项研究重点。第一，构建涵盖多种事件且明确标注传感器的基准，以区分观测水体、淹没、活动火灾、烧毁区域和影响目标。第二，报告预训练来源，并同时说明下游留出集，以便将表征复用与评估独立性区分开来。第三，在与准确率相同的分布偏移条件下评估校准情况和失效模式。第四，在实际预警决策所需的发布时点和预报时效上，测试耦合天气与地球观测系统。这些重点与关于综合预警和透明基准评测的更广泛倡议相辅相成 [3, 8]。

9.2. 测试限制因素假设

该框架提出，当观测数据包含所需信号，而主要难点在于利用有限的下游标签学习可迁移表征时，表征学习最为有用。这预示着

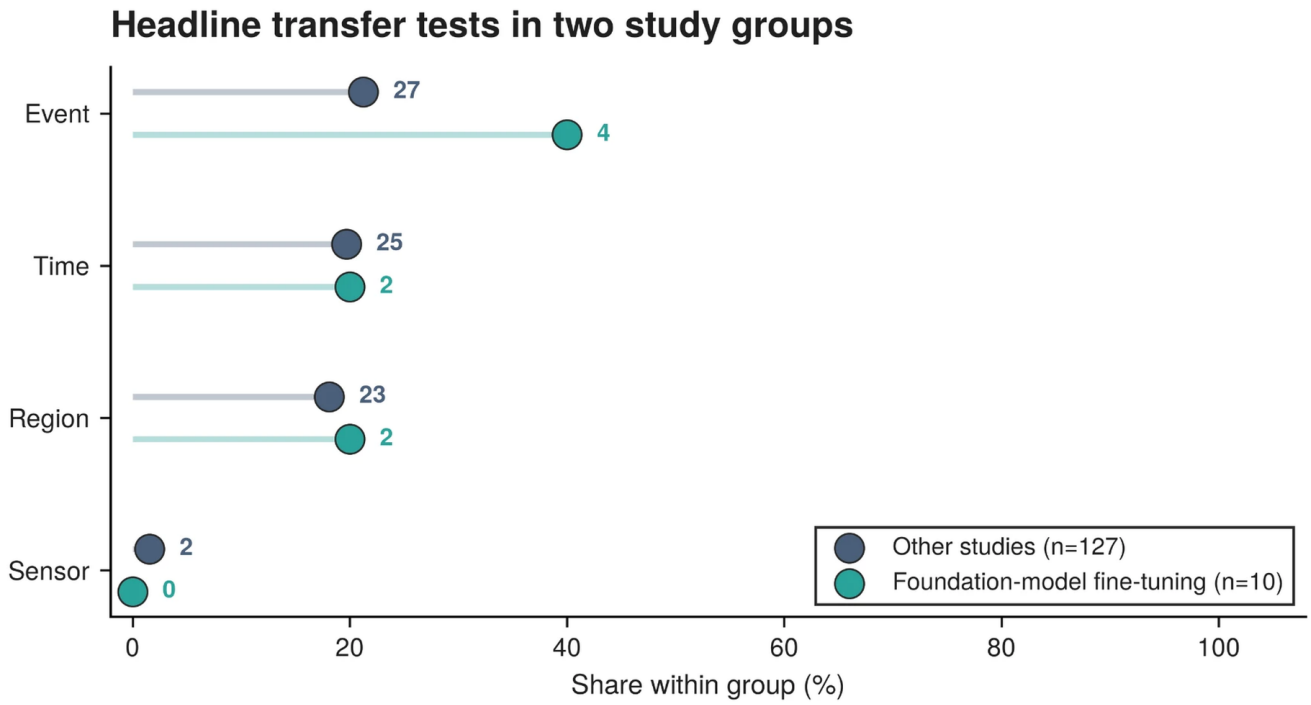

图7： 两个研究组中的标题迁移测试。

这种优势应在不同标签预算和独立留出的环境中进行衡量。当可见性是主要限制因素时，增加一项有信息量的观测数据，可能比扩大编码器更重要。当挑战在于未来的过程演变时，驱动因子信息和预报提前期可能比映射架构更为关键。当决策依赖校准概率时，仅有更高的重叠分数并不足够。

这些命题需要采用干预式比较，而不是断言某项约束始终控制着某种灾害。尽可能保持事件和参照不变，同时更换传感器；在保留预测任务的同时改变提前期；或者在保留模型和划分的同时更换参照指标。由此产生的性能差异可以揭示在该情境下哪项限制最为重要。多项约束可能会同时起作用，而其中一项的改善可能会暴露出另一项约束。

9.3. 农业与城市监测中的迁移经验

相邻的环境应用有助于阐明这些评估选择，但不应被计作额外的灾害研究。哨兵-2号作物产量综述强调，解读影像时需结合作物生长和预期的预测尺度 [47]。基于无人机和卫星的分枝列当时间序列研究展现了另一种目标识别问题：必须将观测到的作物胁迫信号与参照条件及观测时间联系起来 [48, 49]。这些研究关注的是生物胁迫，而非对干旱或气候极端事件预测的直接验证。它们的意义在于方法层面：视觉上相似的胁迫响应若无支持性证据，不应将其归因于同一种原因。

空间制图研究提供了另一种有用的区分。基于NAIP影像的农田范围和可见边界制图涉及可观测的土地覆盖几何，而非法律意义上的地籍边界 [50]。光学城市树冠制图和韧性筛查将遥感特征与规划背景联系起来，但不会将树冠检测转变为经过验证的热暴露预测 [51]。这些例子进一步说明，在选择模型之前必须先定义输出；当部署问题要求如此时，应保持场景或地理独立性；并将制图代理指标与其旨在提供信息支持的决策变量区分开来。

9.4. 证据范围及其对应用的影响

描述性百分比刻画的是可获取且经过编码的语料。全文可获取性、检索来源以及灾害的分布都会影响其构成。这些数字并不估计某种做法在全球范围内的普遍程度，也不提供模型家族之间的因果比较。规模较小的微调子集和传感器留出子集也使得得出广泛的组级结论并不恰当。一种透明的解读方式是利用审查来识别具体缺失的测试，并利用应用文献来设计这些测试。

同一原则也适用于气候相关性。洪水或火灾地图可以为适应提供支持，但这并不能证明气候变化归因。短期天气预报并非气候预测。与当前损害的关联并不意味着减缓效益。明确保持这些区分，能让这篇综述同时为研究人员和从业者提供帮助：指出某项输出能够支持什么、某项应用还需要哪些额外证据，以及哪些方面

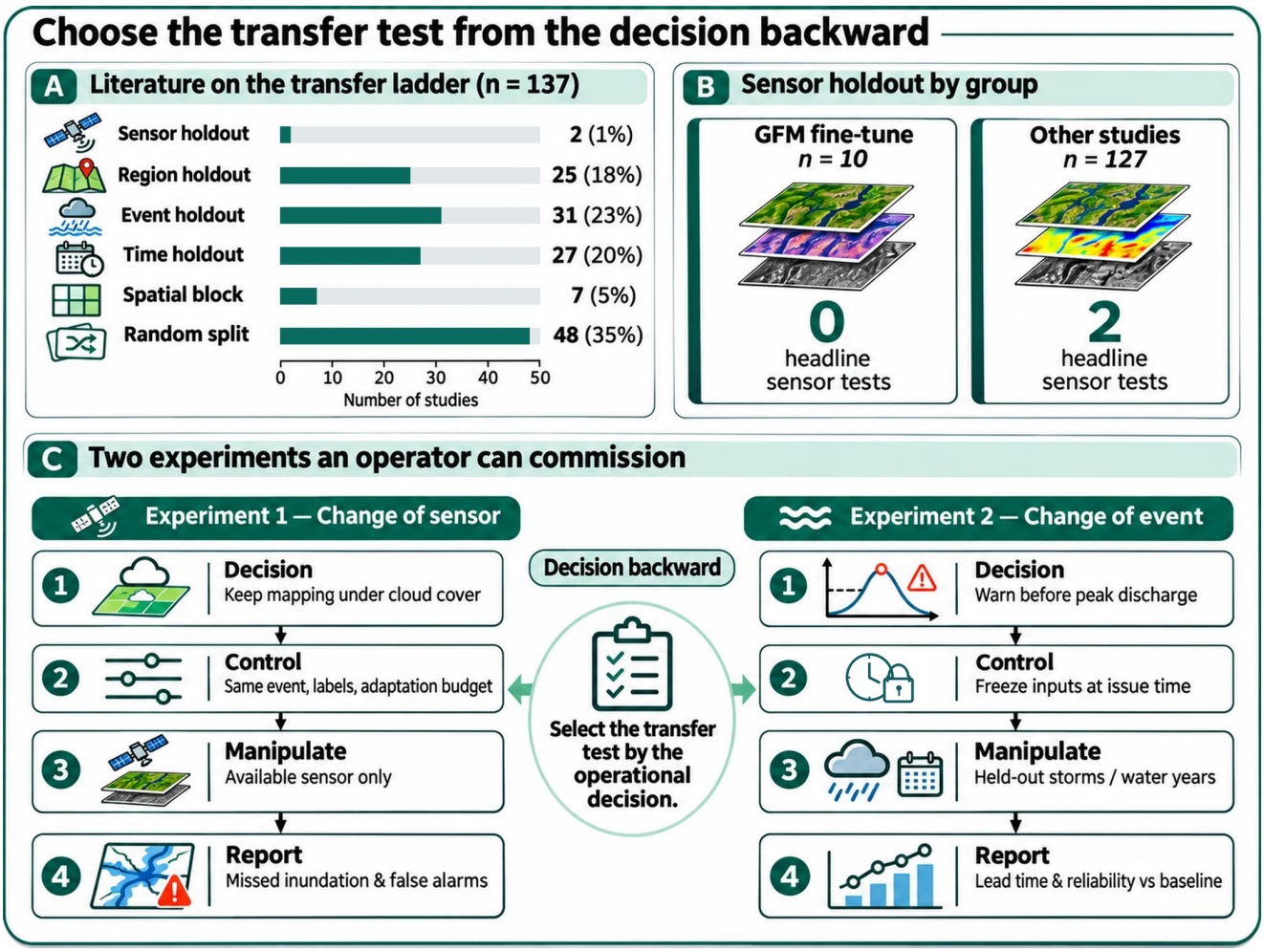

图8：从决策出发，反向选择迁移测试。

10. 结论

地理空间基础模型拓展了地球观测信息再利用的方式，但气候韧性需要有证据表明，在决策所处的条件下能够实现迁移。137项研究审查显示，分布外实验得到了有意义的应用，但对某些备受关注的变化类型，研究覆盖仍不均衡。分别有31项和25项研究记录了留出事件和区域；仅有两项研究采用传感器留出，而在10项基础模型微调研究中，没有一项采用这种方法。不确定性相关报告也较为有限，明确标注校准概率的研究仅有两项。

这些发现确立了一项评估议程，而非断言预训练模型无效。地球观测表征学习与大气预报各具互补优势，应围绕信息可用性、不确定性的传播以及决策时机，对二者的整合进行测试。对于不同灾害，标签定义、可观测性、预测时域和运行要求决定了可能实现何种改进，以及应如何衡量改进。

核心建议是从决策出发，反向选择测试。明确行动，在匹配的信息条件下重现相关变化，并比较模型和适应预算，同时报告决定实际应用的误差、可靠性和延迟。当模型在这些条件下改善了决策，而非仅以熟悉的基准测试得分作为其已准备好应对陌生世界的证据时，才算取得进展。

数据可用性声明

支持本综述的筛选与编码材料可根据合理请求向通讯作者索取。经许可的全文出版物不予再分发。

作者贡献

PF：概念构思、方法学、调查研究、数据整理、正式分析、软件、可视化、撰写——初稿、撰写——审阅与编辑。

MN：概念构思、方法学、调查研究、数据整理、正式分析、软件、可视化、项目管理、撰写——初稿、撰写——审阅与编辑。

资助

本研究未获得公共、商业或非营利部门任何资助机构的专项资助。

利益冲突

作者声明不存在需要披露的竞争性利益。

伦理声明

本研究使用遥感、环境、基础设施、地块及汇总公共统计数据。本研究不涉及人类参与者，也不使用任何可识别的个体层面记录。

生成式人工智能声明

在本项工作准备期间，作者使用 ChatGPT 改进语法准确性、优化句子结构并增强可视化效果。作者对所有由人工智能生成的修订内容进行了全面审阅和编辑，以确保其相关性和准确性。作者对科学内容、解读、引文和最终呈现承担全部责任。

参考文献

[1] M. Reichstein、G. Camps-Valls、B. Stevens、M. Jung、J. Denzler、N. Carvalhais 等，“用于数据驱动的地球系统科学的深度学习与过程理解”，《自然》566，195–204，2019年。 [ ] 10 1038 41586 019 0912 1 在线。可访问：https://doi.org/ 。 /s - - - G. Camps-Valls、M. Á. Fernández-Torres、K. H. Cohrs、A. Höhl、A. Castelletti、A. Pacal 等，“用于模拟和理解极端天气与气候事件的人工智能”，《Nature Communications》16，1919，2025年。 [在线]。可访问：https://doi.org/ 10 1038 41467 025 56573 8 。 /s - - - M. Reichstein、V. Benson、J. Blunk、G. Camps-Valls、F. Creutzig、C. J. Fearnley 等，“利用集成式人工智能对复杂气候风险进行早期预警”，《Nature Communications》16，2564，2025年。[在线]。可访问：https://doi.org/10.1038/s41467-025-57640-w [4] D. R. Roberts、V. Bahn、S. Ciuti、M. S. Boyce、J. Elith、G. Guillera-Arroita 等，“针对具有时间、空间、层级或系统发育结构数据的交叉验证策略”，《生态地理学》40，913–*929*, 2017 [ ] 10 1111 02881 。在线。可访问：https://doi.org/ 。 /ecog。H. Meyer、C. Reudenbach、T. Hengl、M. Katurji 和 T. Nauss，“使用前向特征选择和目标导向验证改进时空机器学习模型的性能”，《环境建模与软件》101，1–9，2018年。 [在线]。可访问：https://doi.org/ 10 1016 2017 12 001 。 /j.envsoft。。。。R. Bentivoglio、E. Isufi、S. N. Jonkman 和 R. Taormina，“洪水制图的深度学习方法：现有应用与未来研究方向综述”，《水文学与地球系统科学》26，4345–4378，2022年。 [在线]。可访问：https://doi.org/ 10 5194 26 4345 2022 。 /hess- - - S. Illarionova、U. Tasuev、P. Tregubova、I. Rubin、D. Shadrin、A. Zaytsev 等，“从光谱指数到基础模型：基于卫星的环境灾害监测中的人工智能革命综述”，《地球系统与环境》，2026年。 [在线]。可访问：https://doi.org/ 10 1007 41748 026 01328 8 。 /s - - - I. Corley、N. Lehmann、C. Robinson、G. Tseng、A. Fuller、H. Alemohammad 等，“没有人知道地理空间基础模型的技术现状”，arXiv 预印本 arXiv: 2605 12678, 。。在线。可访问：2026 [ ] 10 48550 2605 12678 。 /arXiv。 。

[9] E. Portalés-Julià、G. Mateo-García、C. Purcell 和 L. Gómez- Chova，“光学卫星图像中的全球洪水范围分割”，《科学报告》13，20316，2023年。 [在线]。可访问：https://doi.org/ 10 1038 41598 023 47595 7 。 /s - - - P. Akiva、M. Purri、K. Dana、B. Tellman 和 T. Anderson，“H2o-net：通过对抗式领域自适应和标签细化进行自监督洪水分割”，载于 IEEE/CVF冬季计算机视觉应用会议论文集，111-120，2021年。 [ ] 10 1109 48630 2021 00016 在线。可访问：https://doi.org/ 。 /WACV 。 . Y. Cong、S. Khanna、C. Meng、P. Liu、E. Rozi、Y. He 等，“Satmae：针对时序多光谱卫星影像的 Transformer 预训练”，载于神经信息处理系统进展 *35*, 2022 [ ] 10 52202 068431 。在线。可访问：https://doi.org/ 。 / - 0015 A. Fuller、K. Millard 和 J. Green，“CROMA：采用对比式雷达-光学掩码自编码器的遥感表征”，载于神经信息处理系统进展 36，2023年。 [ ] 10 52202 075280 0241 在线。可访问：https://doi.org/ 。 / - Z. Xiong、Y. Wang、F. Zhang、A. J. Stewart、J. Hanna、D. Borth 等，“受神经可塑性启发的地球观测多模态基础模型”，arXiv 预印本 arXiv:2403.15356，2024年。 [在线]。可访问：https://doi.org/ 10 48550 2403 15356 。 /arXiv。 . G. Astruc、N. Gonthier、C. Mallet 和 L. Landrieu，“AnySat：适用于多种分辨率、尺度和模态的地球观测模型”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，19530-19540，2025年。 [在线]。可访问：https://openaccess.thecvf.com/content/CVPR2025/html/Astruc_Any Sat_One_Earth_Observation_Model_for_Many_Resolutions_Scale s_and_CVPR_2025_paper.html [15] J. Jakubik、F. Yang、B. Blumenstiel、E. Scheurer、R. Sedona、S. Maurogiovanni 等，“TerraMind：面向地球观测的大规模生成式多模态”，载于 IEEE/CVF国际计算机视觉会议论文集，7383-7394，2025年。 [ ] 20 在线。可访问：https://openaccess.thecvf.com/content/ICCV /html/Jakubik_TerraMind_Large-Scale_Generative_Multimodali ty_for_Earth_Observation_ICCV_2025_paper.html [16] D. Szwarcman、S. Roy、P. Fraccaro、Þ. E. Gíslason、B. Blumenstiel、R. Ghosal 等，“Prithvi-EO-2.0：面向地球观测应用的通用多时相基础模型”，《IEEE地球科学与遥感汇刊》64，4400120，2026年。 [在线]。可访问：https://doi.org/ 10 1109 2025 3642610 。 /TGRS。 . K. Bi 、 L. Xie 、 H. Zhang 、 X. Chen 、 X. Gu 和 Q. Tian，“ 利用三维神经网络实现准确的中期全球天气预报”，《自然》619，533-538，2023年。 [在线]。可访问：https://doi.org/ 10 1038 41586 023 06185 3 。 /s - - - R. Lam 、 A. Sanchez-Gonzalez 、 M. Willson 、 P. Wirnsberger 、 M. Fortunato、F. Alet 等，“学习高技巧中期全球天气预报”，《科学》382，1416-1421，2023年。 [在线]。可访问：https://doi.org/ 10 1126 2336 。 /science.adi I. Price、A. Sanchez-Gonzalez、F. Alet、T. R. Andersson、A. El- Kadi、D. Masters 等，“利用机器学习进行概率天气预报”，《自然》637，84-90，2025年。 [在线]。可访问：https://doi.org/ 10 1038 41586 024 08252 9 。 /s - - - D. Kochkov、J. Yuval、I. Langmore、P. Norgaard、J. Smith、G. Mooers 等，“用于天气与气候的神经环流模式”，《自然》632，1060-1066，2024年。 [在线]。可访问：https://doi.org/10.1038/s41586-024-07744-y [21] C. Bodnar、W. P. Bruinsma、A. Lucic、M. Stanley、A. Allen、J. Brandstetter 等，“地球系统基础模型”，《自然》641，1180-1187，2025年。 [在线]。可访问：https://doi.org/10.1038/s41586-025-09005-y [22] A. Lacoste、N. Lehmann、P. Rodriguez、E. D. Sherwin、H. Kerner、B. Lütjens 等，“GEO-Bench：迈向地球监测基础模型”，载于神经信息处理系统进展 *36*, 2023 [ ] 10 52202 075280 2223 。在线。可访问：https://doi.org/ 。 / - V. Marsocci、Y. Jia、G. Le Bellier、D. Kerekes、L. Zeng、S. Hafner 等，“PANGAEA：通过全球包容性基准评估地理空间基础模型能力”，《IEEE地球科学与遥感杂志》14(1)，2026年。 [在线]。可访问：https://doi.org/10.1109/MGRS.2025.3628194

[24]C. Guo、G. Pleiss、Y. Sun 和 K. Q. Weinberger，“现代神经网络的校准”，载于机器学习研究论文集 70，1321-1330，2017年。 [在线]。可访问：https://proceedings.mlr.press/v70/guo17a.html [25] Y. Ovadia、E. Fertig、J. Ren、Z. Nado、D. Sculley、S. Nowozin 等，“你能相信你的模型不确定性吗？数据集偏移下预测不确定性的评估”，载于神经信息处理系统进展 32，2019年。 [在线]。可访问：https://research.google/pubs/can-you-trust-your-models-uncertainty-evaluating-predictive-uncertainty-under-dataset-shift/ [26] J. Adebayo、J. Gilmer、M. Muelly、I. Goodfellow、M. Hardt 和 B. Kim，“显著性图的合理性检查”，载于神经信息处理系统进展 31，2018年。 [在线]。可访问：https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html[27] P. Farajpoor 和 M. Narimani，“用于城市野火结构损失的可解释地理人工智能：2025年帕利塞兹山火中的植被与建成环境”，《可持续发展前沿》7，1975724，2026年。 [在线]。可访问：https://doi.org/ 10 3389 2026 1975724 。 /frsus。 . M. [28] D. Wilkinson、M. Dumontier、I. J. Aalbersberg、G. Appleton、M. Axton、A. Baak 等，“科学数据管理与监管的 FAIR指导原则”，《科学数据》3，160018，2016年。 [ ] 10 1038 2016 18 在线。可访问：https://doi.org/ 。 /sdata。 . H. Meyer 和 E. Pebesma，“向未知空间进行预测？空间预测模型适用区域的估计”，《生态学与进化方法》12，1620-1633，2021年。 [在线]。可访问：https://doi.org/ [29] 10 1111 2041 210 13650 。 / - X. D. Bonafilia、B. [30] Tellman、T. Anderson 和 E. Issenberg，“Sen1Floods11：用于训练和测试哨兵-1号洪水算法的地理配准数据集”，载于 IEEE/CVF计算机视觉与模式识别会议研讨会，2020年。 [在线]。可访问：https://doi.org/ 10 1109 50498 2020 00113 。 /CVPR。 . N. I. [31] Bountos、M. Sdraka、A. Zavras、I. Karasante、A. Karavias、T. Herekakis 等，“Kuro Siwo：水下330亿平方米——用于快速洪水制图的全球多时相卫星数据集”，载于神经信息处理系统进展 37，2024年。 [ ] 10 52202 079017 1204 在线。可访问：https://doi.org/ 。 / - W. Li、H. Lee、S. Wang、C. Y. Hsu 和 S. T. Arundel，“用于洪水淹没制图的新型地理空间人工智能基础模型评估”，载于第六届 ACM SIGSPATIAL地理知识发现人工智能国际研讨会，2023年。 [在线]。可访问：https://doi.org/ 10 1145 3615886 3627747 。 / . V. Kostejn、Y. Essus、J. Abrahamson 和 R. R. Vatsavai，“U-Prithvi：融合基础模型与U-Net 以增强洪水淹没制图”，载于莱布尼茨国际信息学论文集 346，18:1-18:17，2025年。 [在线]。可访问：https://doi.org/ [32] [33] 10 4230 2025 18 。 /LIPIcs.GIScience。 . P. [34] Farajpoor 和 M. Narimani，“洪水期间紧凑型基础模型用于地表水制图的准确率与效率”，arXiv预印本 arXiv:2610.34904，2026年。 [在线]。可访问：https://doi.org/ 10 48550 2610 34904 。 /arXiv。 . G. [35] Nearing、D. Cohen、V. Dube、M. Gauch、O. Gilon、S. Harrigan 等，“无测站流域极端洪水的全球预测”，《自然》627，559-563，2024年。 [在线]。可访问：https://doi.org/ 10 1038 41586 024 07145 1 。 /s - - - S. [36] Mitra、M. Narimani 和 P. Farajpoor，“用于2020年7月至2026年8月25场加利福尼亚州野火深度学习活跃火点分割的哨兵-2号基准数据集”，《数据简报》69，113357，2026年。 [ ] 10 1016 2026 113357 在线。可访问：https://doi.org/ 。 /j.dib。 . C. Brust、J. Kimball、M. Maneta、K. Jencso 和 R. Reichle，“DroughtCast：美国干旱监测的机器学习预测”，《大数据前沿》4，773478，2021年。 [在线]。可访问：https://doi.org/10.3389/fdata.2021.773478 [37]

[38]T. Lees、G. Tseng、C. Atzberger、S. Reece 和 S. Dadson，《利用深度学习预测植被健康：肯尼亚案例研究》，《Remote Sensing》14，698，2022年。 [在线]。链接：https://doi.org/ 10 3390 14030698 。/rs K. [39] Lesinger 和 D. Tian，《利用深度学习动态模型实现精准的次季节土壤水分干旱预报》，《自然》通讯 16，7461，2025年。 [在线]。链接：https: //doi.org/ 10 1038 41467 025 62761 3 。/s - - - S. Ge、W. Zhan、J. Li、L. [40] Li、P. Dong、X. Li 等，《结合多块卷积神经网络与卫星及地面观测数据预测次日1公里分辨率的冠层城市热岛》，《Sustainable Cities and Society》131，106701，2025年。 [ ] 10 1016 2025 106701 在线。链接：https://doi.org/。/j.scs.。O. C. Pasche、J. Wider、Z. Zhang、J. Zscheischler 和 S. Engelke，《在近期高影响极端事件中验证深度学习天气预报模型》，《Artificial Intelligence for the Earth Systems》4，e240033，2025年。 [在线]。链接：https://doi.org/ 10 1175 24 0033 1 。/AIES-D- -。 [41] [42] N. Gourmelon、M. Dreier、M. Mayr、T. Seehaus、D. Pyles、M. Braun 等，《SSL4SAR：利用自监督学习从合成孔径雷达影像中提取冰川崩解前缘》，《IEEE地球科学与遥感汇刊》，2025年。 [在线]。链接：https://doi.org/ 10 1109 2025 3580945 。/TGRS.。D. Jiang、S. Li、I. [43] Hajnsek、M. A. Siddique、W. Hong 和 Y. Wu，《利用遥感地理基础模型进行冰川湖泊制图》，《International Journal of Applied Earth Observation and Geoinformation》136，104371，2025年。 [在线]。链接：https://doi.org/ 10 1016 2025 104371 。/j.jag.。Y. Hu 和 H. Tang，《论全球模型在多种自然灾害情景下异质卫星影像快速建筑物制图中的泛化能力》，《Remote Sensing》13，984，2021年。 [ ] 10 3390 13050984 在线。链接：https://doi.org/。/rs A. G. Wells、C. B. Yackulic、J. Kostelnik、A. Bock、R. E. Zuellig、D. M. Carlisle 等，《火灾发生前：预测火烧严重度及火后泥石流灾害对科罗拉多河割喉鳟保护种群的潜在威胁（Oncorhynchus clarkii pleuriticus）》，《International Journal of Wildland Fire》33，WF23199，2024年。 [在线]。链接：https://doi.org/ 10 1071 23199 。/WF P. Farajpoor 和 M. [44] [45] [46] Narimani，《从野火严重度到积雪持续性：对2020年克里克山火的多源地理人工智能研究》，《Frontiers in Forests and Global Change》9，2016172，2026年。 [在线]。链接：https://doi.org/ 10 3389 2026 2016172 。/ffgc.。M. Narimani、A. Pourreza、A. Moghimi 和 P. Farajpoor，《哨兵-2号用于作物产量估算：系统综述》，《Smart Agricultural Technology》14，102405，2026年。 [在线]。链接：https://doi.org/ 10 1016 2026 102405 。/j.atech.。 [47] [48] M. Narimani、A. Pourreza、A. Moghimi、M. Mesgaran、P. Farajpoor 和 H. Jafarbiglu，《无人机多光谱成像与深度学习：及时检测番茄农场中的分枝列当》，载于《SPIE会议论文集》13053，1305304，2024年。 [在线]。链接：https://doi.org/ 10 1117 12 3021219 。/。M. Narimani、A. Pourreza、A. Moghimi、P. Farajpoor、H. Jafarbiglu 和 M. Mesgaran，《利用卫星影像和时间序列分析检测番茄农场中的分枝列当》，载于《SPIE会议论文集》13475，134750U，2025年。 [在线]。链接：https://doi.org/ 10 1117 12 3059998 。/。M. Narimani、V. Anand 和 P. [49] [50] Farajpoor，《利用残差U-Net和文本提示SAM 3 2607 21881, 细化，从1米分辨率NAIP影像中提取农田范围并绘制可见边界》，arXiv预印本 arXiv:。。在线。链接：https://doi.org/。/arXiv.。 2026 [ ] 10 48550 2607 21881 [51] M. Narimani、S. Mitra 和 P. Farajpoor，《光学地理人工智能与空间建模助力戴维斯（加利福尼亚）的城市树冠制图与韧性筛查》，arXiv预印本 arXiv:2608.13856，版本3，2026年。[在线]。链接：https://doi.org/10.48550/arXiv.2608.13856