arXiv:2611.28477v1 [cs.AI] 21 Nov 2026

ESM-BENCH：评估人工智能代理是否理解地球系统模型物理学与和代码的基准测试

Koutian Wu[1†] **Yuxuan Cao**[1‡]
Gengchen Mai[1§] Zeping Liu[1¶]
[1]德克萨斯大学奥斯汀分校 *通讯作者：
ktwu@utexas.edu

摘要

修改地球系统模型（ESM）代码的人工智能编码代理，必须同时保留软件接口及其中编码的物理假设。我们推出 ESM-BENCH，这是一个基准测试，包含来自 6 个代码库的 60 项开发者提交任务：Noah-MP、CLM5、SUMMA、ParFlow、MOM6 和 VIC。v5 语料库包含 17 项基于物理的错误修复、16 项过程表示修改、13 项参数化方案选择，以及 14 项参数优化任务。每项任务都包含一个科学目标、父提交源文件、一份参考差异，以及三个存储的提示变体。制品审计确认，全部 60 份参考差异均可应用于所提供的初始文件。我们制定了包含八个条目的补丁审查评分标准，并依据已实现的本地检查器审计了五个已归档补丁夹具。五个夹具差异均无法应用于所提供的初始文件，但其中三个文本模式检查器却判定通过。ParFlow 表达式重放得出的结果与参考斜率不一致，而 MOM6 模式检查则发现遗漏了一项更新。检查还发现，VIC 中保留的一条参考赋值语句会覆盖预期的植被特定反照率。这些发现揭示了参考补丁和本地代理指标作为正确性证据的局限性。本次发布提供了可复现语料库和诊断工具；在 v5 上，完整的专家验证、代理性能比较以及完整模型执行情况仍未进行测量。

1 引言

地球系统模型（ESM）通过大型科学代码库来表示物理过程。对径流计算、积雪参数或海洋混合例程的修改，必须符合模型的接口、单位和物理假设。语法有效的编辑可能针对错误的处理路径，或遗漏配套更新。反过来，与开发者参考值不同的补丁，也可能实现了可接受的科学解决方案。这些特点使 ESM 维护成为研究 AI 如何修改代码、且不局限于与参考文本进行相似性比较的理想场景。

诸如 SWE-bench 之类的软件工程基准测试 [1] 以代码仓库中的任务为评估基础，而 SciCode [2] 和 ScienceAgentBench [3] 则将评估扩展到科学编程以及
工作流。ESM-BENCH 专注于一项更窄的任务：在提供相关父提交文件的情况下，为生产环境地球系统模型 Fortran 和 C 代码生成可供审查的补丁。它不测试

*通讯作者
[†]ORCID：0000-0002-8409-3579
[‡]ORCID：0000-0003-3565-906X[§]
ORCID：0000-0002-7818-7309
[¶]ORCID：0000-0003-2898-0023

预印本。

不受限制的仓库导航或自主科学发现。其科学动机在于，需要同时检查代码变更及其旨在修改的物理行为 [4]。
我们将 v5语料库呈现为一个明确且可审计的快照：包含来自 6 个代码仓库的 60 个任务，每个代码仓库 10 个任务，涵盖四类科学变更。本文报告的证据分为两部分。首先，我们审计存储的任务记录、源文件和参考差异。其次，我们通过代码仓库的确定性本地检查，重放五个现有补丁固定样例。此范围支持我们对语料库结构以及这些检查所观察到的行为作出判断，但不足以对当前模型进行排名，也不足以估算总体物理失效率。
本文的贡献如下：

1. 一个以提交记录为依据的语料库，涵盖陆面、水文学和海洋代码，并包含科学目标、源代码快照、参考补丁和存储的提示变体。2. 一套评估协议，区分补丁应用、物理学感知审查，以及针对具体任务的数值检查或源代码契约检查。3. 一项可复现的语料库审计和五次诊断性样例重放，并明确区分当前证据、历史汇总分数与尚未执行的实验性扩展。

实证审查发现了一种具体的评估失效模式：局部源代码模式检查器可能会接受无法应用的差异，而参考文本中也可能包含一条赋值语句，致使任务无法实现其既定目标。这些发现表明，在使用语料库进行智能体比较评分之前，必须设置明确的有效性门槛并开展独立的科学审查。

2 相关工作

基准测试发现与更广泛的智能体评估。 Benchmark Radar [5] 提供了一个动态数据库和搜索引擎，用于发现人工智能基准测试并审查其评估证据；我们借助它找到了本节讨论的若干基准测试。这类基础设施有助于将特定领域的基准测试置于更广阔的评估图景中。MatrAIx [6]通过大规模角色智能体，将评估扩展到与各种模拟用户的交互。ASI-Bench [7] 涵盖了 11 个科学领域的 60 项项目级研究任务，并逐步撤去方法论指导，以检验智能体在缺少指导的情况下能推进到何种程度。研究智能体基准测试采用分层评分标准，对机器学习论文复现（PaperBench [8]）、由人类专家评判的开放式研究工程（RE-Bench [9]）、研究扩展的实现（RExBench [10]）、国家实验室科学家提出的开放问题的后续研究（Into the Unknown [11]），以及图形化科学软件的使用（OSWorld-Science [12]）进行评分。这些工作致力于基准测试发现、面向人类的行为和开放式研究；ESM-BENCH 则聚焦于智能体生成的科学模拟代码修改在物理上的有效性。

代码仓库与科学代码评估。SWE-bench [1] 启发我们采用真实的代码仓库变更和可复现的初始状态。HumanEval [13] 为基于样例的代码生成指标提供了广泛使用的依据。SciCode [2] 和 ScienceAgentBench [3] 启发我们关注科学目标和针对具体任务的评估。ESM-BENCH 使用开发者提交获取具体的源代码变更，然后将受影响的源文件作为补丁合成工作区。它重点关注科学目标、实现变更与评估该变更所需证据之间的关系。ScienceIDE [14] 将科学代码仓库转化为可执行的智能体训练和评估环境，并以专家定义的科学案例和验收标准为指导。它支持任务生成、执行和科学验证。这一基础设施与 ESM-BENCH 针对地球系统模型中代码编辑和物理有效性的领域专属评估相辅相成。POSEIDON [15] 大规模地从科学代码仓库生成可执行的训练环境，而 DeployBench [16] 则评估智能体在全新机器上配置研究制品的能力。SWE-bench Science [17] 是面向科学软件工程的代码仓库级基准测试，包含来自 20 个领域、98 个科学代码仓库的 119 项任务；其失败机制包括缺乏科学知识、表层修复和修复覆盖范围不完整。两项基准测试都指出，通过功能测试不足以作为充分证据。SWE-Gate [18]加入了根据拉取请求审查得出的约束，而 PETSCAgent-Bench [19] 则从求解器选择、库规范和性能等方面对基于生产级高性能计算库构建的代码进行评分，并发现即使

代码能够编译和运行。ESM-BENCH 基于另一项标准认同这一观点：编辑是否保留了地球系统模型的物理假设。Noah-Agent [20]将多专家智能体框架应用于大型 Fortran 气候模型的自动参数化与验证，其应用场景与本文所用的地球系统代码十分接近。

地球科学、物理模拟与高性能计算代码评估。在地球科学领域，AtmosCoder- Bench [21] 会对环境科学计算中执行的代码进行评分，并报告模型常常无法一致地应用已知公式和约束，或在任务条件使规范方法不再适用时仍回退到这些方法。TerraBench [22] 评估智能体借助科学工具对网格数据、图像和模拟器输出进行推理的能力；ClimateViz [23] 则评估智能体对科学图表进行事实核查的能力。RealBench [24]使用业务分析场和站点观测，而非再分析数据，评估数据驱动的天气预报模型。在物理模拟方面，WorldSolver [25]要求智能体在固定的模拟框架中实现求解器，并检查其执行情况、视觉保真度和物理合理性；GPUPhysBench [26] 衡量GPU物理代码的正确性和速度；RSI-PDEbench [27] 则根据解数据评估方程识别和求解器配置。HPC-Bench [28] 评估高性能计算工作负载中的内核优化，一项综述回顾了语言模型在高性能计算编程中的应用 [29], , 而ADFD-Migrate [30]将遗留Fortran代码库移植到Python，并检查其行为是否与Fortran源代码一致。这些基准测试评估计算、数据推理、预测、新编写的求解器或核函数，或翻译后的代码；而ESM-BENCH考察的是对现有生产级地球系统模型进行修改，其正确性取决于周边代码中已经编码的物理假设。

物理模型验证。 科学软件测试必须关注数值和科学行为，也要关注实现错误 [4]。气候模型评估通常按组件和物理变量 [31, 32]汇总结果。我们采用按组件和任务类别报告的方式，以避免将异质的 ESM 编辑视为同质总体。本文的本地检查涵盖选定表达式和代码模式；这些检查可补充耦合模型验证，但不能取而代之。

溯源与污染。 LiveCodeBench [33]等时间基准测试设计所应对的需求，与可复现源代码快照所应对的需求不同。父提交和参考差异可以确立任务身份，但无法说明模型是否在训练期间接触过目标变更。研究人员也曾从科学研究轨迹的提交日志中挖掘可复用的代理技能 [34]。ESM-BENCH记录提交哈希和源代码内容。所选任务记录缺少明确的源提交时间戳，因此，针对特定模型进行时间过滤还需要额外的上游历史元数据。

3 ESM-bench设计

3.1 代码库和任务类别

v5语料库从六个代码库中抽样：用于陆面过程的Noah-MP [35, 36]和CLM5 [37]，用于水文过程的SUMMA [38], 、ParFlow [39], 和VIC [40] ，以及用于海洋过程的MOM6 [41] 。WRF和E3SM不在此快照范围内，因此既不包含任务，也不提供评估结果。

任务有四个类别标签。**基于物理的错误修复（PBF）**修复被描述为不符合物理规律的行为。**过程表示修改（PRM）**改变过程的已实现表示。**参数化方案选择（PSS）**改变方案或其选择逻辑。**参数优化（PO）**改变参数或相关默认值。PO 是数据集中的类别名称；这些记录属于源代码编辑任务，不一定涉及优化算法或校准实验。类别边界和科学描述属于待验证的标注，而不是根据存在开发者提交就能保证正确性的内容。

3.2 任务记录与构建

每项任务记录了源提交、其父提交哈希、开发者提交消息、科学目标及背景、物理约束、相关文件、参考统一差异，以及所提供父文件的完整内容。起始_状态是一组相关文件，而不是

一个已解析完整依赖关系的仓库检出。源代码快照足以检查参考补丁的应用情况；编译可能需要其他文件和上游构建环境。

记录显示，deepseek-v4-pro 是元数据生成器和筛选器，而 openai-codex-cli-0.136.0 是元数据验证器。这些字段记录了自动化准备过程，但不能证明经过专家批准。保留的抽查清单选取了六个任务进行领域审查，但经过审计的快照中没有完成的 v5 专家标注数据集，因此无法估计一致性或标注错误率。

3.3 提示变体

每个任务在 level_1、level_2 和 level_3 下分别存储三个提示词；评估接口将其称为 P1、P2 和 P3。直接提示词运行器会附加所提供的完整源文件，并要求生成完整的统一差异。预期的递进方式是先加入常规提示或行位置提示，然后提供更明确的、源自参考值的代码上下文和编码指导。这些属于预言机辅助条件，因为提示信息是利用目标变更构造的。P1 描述已经可以指明文件和例程，因此分层比较并不能干净地检验不受限制的代码导航能力。
定义各条件的是所存储的提示词字符串，而不是重构出的提示词描述。受控比较要求各层级具有相同的任务标识和源代码哈希。本文未报告这样完整的 v5 求解器比较。外部文档检索、提交记忆检索、流程谱系实验和轨迹评分仍属于协议的扩展，在此快照中没有测量结果。

3.4 特定任务的物理契约

五条任务记录包含一项物理_契约。其主题分别是地形坡度（ParFlow）、垂直单位重标度（MOM6）、外部城市耦合与叶片质量初始化（Noah-MP）、径流核算（SUMMA）以及新雪反照率（VIC）。相关工具会检查源代码或补丁模式，并针对四项任务，在配置的输入上评估提取的表达式或手工编写的场景计算。只有 ParFlow 工具会评估从候选代码中提取的表达式。Noah-MP、SUMMA 和 VIC 工具会在匹配代码模式后，选取类似参考值的输出；MOM6 工具则采用源代码模式契约。这些工具都不会编译并执行完整的已打补丁 ESM。

在全部 60 条记录中，更广泛的计算_测试和物理_测试字段均标记为仅_脚手架。尽管这些完整运行时字段只是占位字段，五项局部契约仍提供了范围更窄的检查。因此，门控检查通过仅表示其配置的检查已通过，并不意味着任务中列出的每项物理约束都已得到测试。

4 数据集统计与制品审计

表1 展示了完整的 v5语料库，而非经过评估的求解器样本。其 60 条记录对应 60 个不同的源提交，每个代码库包含十个任务。各类别在整个语料库中的分布大致均衡。保留的层级标签将 49 个任务归类为修改，将 11 个任务归类为实现；这些标签尚未根据专家评定的难度进行校准。

我们根据存储的文件内容重建每个起始工作区，并对参考差异运行 git apply –check。60 个差异全部通过检查。此检查确认了存储的补丁与起始文件之间的内部一致性，但不能验证文件是否与上游提交相符、科学描述是否准确，也不能验证生成的代码是否能够编译。

任务提供 1 至 10 个源文件（中位数为 2 个）。参考差异包含 5–497 行新增行和删除行（中位数为 90 行），不包括差异标头。因此，选定语料库并不局限于早期草案中所述的 5至150行筛选条件。所有任务都有三个存储的提示变体。其中五个任务具有本地物理契约。没有任何任务包含答案_类型标签或明确的源提交日期字段，并且所有60full-runtimetest记录仍然只是脚手架。表2 区分了这些证据级别。

表1：按代码库和类别划分的V5任务数量。每行包含60任务语料库中的十项任务。

代码库	组件	PO	PSS	PRM	PBF	总计
Noah-MP	Land	2	3	2	3	10
CLM5	Land	2	2	3	3	10
SUMMA	陆地水文学	2	2	3	3	10
ParFlow	水文学	3	2	2	3	10
MOM6	海洋	2	2	3	3	10
VIC	水文学	3	2	3	2	10
总计		14	13	16	17	60

表2：所选 v5快照中的观测到的工件覆盖率。计数表示可用记录或已完成检查的数量，而非模型性能。

工件或检查	任务
具有源哈希和父哈希的不同任务	60/60
参考差异适用于提供的父文件	60/60
三个存储的提示变体	60/60
任务特定的本地物理契约	5/60
非脚手架完整运行时测试元数据	0/60
明确的答案多重性标注	0/60
明确的源提交日期字段	0/60

5 评估框架

5.1 针对不同问题分别提供证据

该框架区分三个问题：补丁是否能够应用，其实现经审查后在科学上是否看起来合理，以及它是否满足可执行的或静态的任务特定检查？在某一层级检查成功，并不意味着其他层级也成功。尤其是，解析统一差异弱于应用统一差异，而应用统一差异又弱于构建修改后的程序。

ˆ ˆ**补丁应用**。 对于任务 τ和提交的差异 δ，当差异格式正确且能应用于提供的起始文件时，定义 $V(\tau, \delta) = 1$。语料库审计将此测试应用于全部 60 个参考补丁和保留的五个测试样例，并分别报告这两项的分母。该审计不报告求解器 ValidDiff 对比率。

物理学感知审查。 配置的 v2 评分标准包含八个评分项，分为 0（未通过）、1（部分通过）或 2（明确通过）。其中四项检查代码：目标是否正确、局部一致性、可编译性和完整性。另四项检查物理学：物理意图、量纲一致性、守恒意识和参数化有效性。在此评分标准中，“可编译性”是对静态可行性的判断，而非记录下来的编译器结果。守恒意识同样是对可见改动的审查；全局守恒可能需要执行超出局部补丁范围的代码才能验证。
对于已完成的审查，若评分为 r_1、.....、 r_8，则归一化分数和审查通过谓词定义如下：

$$R = \frac{1}{16}\sum_{i=1}^{8} r_i, \qquad A = \mathbf{1}\Big[\min_i r_i \geq 1\Big]. \tag{1}$$

代码和物理子分数可以分别按各自的 0–8 分量表报告。这些属于协议定义。现有证据并未提供完整的专家裁定 v5 审查表，也未校准这些分数与物理正确性之间的关系。历史上的六项评分标准和当前的八项评分标准输出不得混合汇总，除非保留各自的评分版本。

参考相似度。 完全匹配和结构差异相似度用于衡量与某个开发者补丁的一致性。科学上可接受的替代方案在结构上可能有所不同，因此相似度是一项诊断指标，而非物理正确性的判据。由于 v5 缺少答案多重性标签，本文不报告单一答案完全匹配结果。

表 3：五个归档测试样例的回放。通过表示配置的本地检查器返回 true；五个夹具差异均未通过应用检查。这些结果无法衡量 v5 上的全模型行为或求解器成功。

Task	Gate	观测证据
PF-PO-001	Fail	提取的斜率表达式与参考值在 512×512 网格上存在差异；均方根误差对于 x 为 0.1131，对于 y为 0.0790。
MOM6-PO-007	Fail	五项必需的源代码模式检查中有四项通过；整体混合层摩擦速度默认值更新缺失。
NOAH-PBF-003	Pass	五项源代码模式检查通过。报告的叶片质量一致性取决于所识别的模式，而非候选代码执行。
SUMMA-PBF-001	Pass	四项源代码模式检查通过。检查器会选择参考值径流表达式，当这些模式匹配时。
VIC-PO-001	Pass	五项源代码模式检查通过。检查器会将参考值反照率分配给候选分支，当模式匹配时。

局部契约门控。 测试夹具重放会针对每个可用的存档补丁运行已配置的关卡。关卡定义结合了源代码模式要求；对于 ParFlow，还包括提取的表达式与参考内核之间的数值一致性。其他数值场景输出是由识别器选定的条件性示例，而不是对候选代码的执行结果。数值阈值是保留配置中的固定容差，并非根据重复的编译器或离散化扰动估算得出。第 6 节分别报告每个关卡，以避免将源代码模式失败与测得的模拟偏差混淆。

5.2 报告与范围

比较性求解器研究应明确任务快照、准确的求解器和评判器配置、提示内容、采样程序以及评分任务分母。评分标准评估还需要针对领域专家和独立于供应商的评判器进行校准。重复采样可用于估算 pass@k；本文未报告多样本求解器结果。 [13]

本地夹具样本经过刻意挑选，规模较小。其结果无法估算60任务语料库、未纳入的代码库（CLM5）或某个模型系列的成功率。我们报告通过的已配置门控数量，作为夹具重放诊断；不提供总体置信区间，也不作排行榜解读。

6 结果：可复现的本地夹具检查

6.1 证据快照与重放设置

当前结果包括60任务制品审计，以及2026年9月6日对五个已归档补丁夹具的重放。这些夹具带有标识符 codex_solver；其中未保留固定的求解器模型标识符、生成提示、令牌用量和单次调用成本。我们不将其归因于任何特定的 Codex 或基础模型版本。本次重放未进行新的求解器或评判器调用。

命令`scripts/run_sim_compare.py` 使用 `configs/simulation_compare.yaml` 来选择五条任务记录、对应的已归档差异以及已注册的本地检查。一个检查器会评估提取出的 ParFlow 表达式；另外四个则会识别源代码模式。其中三个识别器还会根据识别出的分支计算示例性场景输出。回放不包括原生模型编译和完整的 ESM执行。这五项任务包含两项 PBF任务和三项 PO任务，每个代码仓库各提供一个夹具。它们并非随机样本。表 3列出了每种结果所依据的证据。

6.2 ParFlow夹具中的参考不一致

ParFlow 检查评估从父源代码、开发者差异和已归档测试夹具中提取的有效斜率赋值语句。参考值使用 0.10sin($x/5$) 和 0.10cos($y/5$)；测试夹具使用 sin(x)/8 和 cos(y)/20。测试夹具中的两个场在配置网格上均为有限值，但其波长

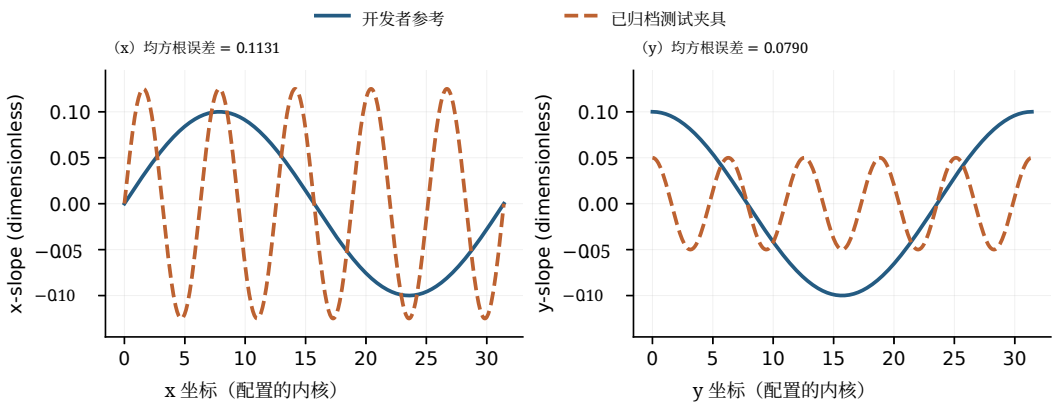


图 1：ParFlow夹具重放中评估的斜率表达式的一维切片。参考值曲线和测试夹具曲线根据已归档结果工件在配置的坐标上的数据重建。尽管两个场仍保持有限，波长差异依然存在。斜率无量纲；坐标单位遵循配置的内核。

和振幅均与参考值不同（图 1）。最大绝对差异分别为 0.2064 和 0.1322，均超过配置的 10^{-5} 最大误差容差；其均方根误差值超过 10^{-6} 容差。

这是一个可复现的、与参考内核不一致的结果。这并不能证明每个替代的周期性斜率场在物理上都无效。要确定某个替代方案是否可接受，需要足够明确的科学目标，并对契约进行独立审查。当前关卡会在选定表达式上强制执行参考一致性。

6.3 补丁应用与识别器结果

所有五个已归档夹具差异在针对各自对应的 v5 父文件执行 git apply –check 时均失败。Git 报告 MOM6、Noah-MP、SUMMA 和 VIC 的补丁已损坏；ParFlow 的差异块则无法匹配提供的源代码。这些失败与 60 项成功的参考补丁检查截然不同。本地检查运行器会读取差异文本，但不会强制应用补丁，因此其输出描述的是从文本中得出的诊断结果，而非经过验证的已提交程序。

MOM6 识别器找到了五种必需更新模式中的四种，但遗漏了混合层摩擦速度默认值。它会搜索差异块文本，包括已删除的行，因此即使匹配到某种模式，其解读也受到限制。Noah-MP、SUMMA 和 VIC 均通过本地检查。它们在各情景下的零误差结果以正则表达式识别为前提：例如，如果 SUMMA 工具匹配到其源代码模式，就会将候选径流设为手工编码的参考计算结果；VIC 则将参考反照率字典赋给候选分支。这种一致性是检查器预先设定的，不能为候选代码提供独立的数值验证。

因此，五个本地检查器中有三个接受其测试夹具文本，而五个测试夹具均不满足补丁应用要求。这是对评估仪器的审计，并非对模型成功率的估算。配置还额外设置了 judge_passed=true 和 plausible_patch=true；这些是预先提供的标志，而非实际观察到的审查结果。派生出的 silent_physics_failure 标签并未用作科学结果。

6.4 参考值与补丁的一致性问题

保留的 VIC-PO-001 开发者差异添加了一个分支，用于赋值植被特定的新雪反照率，但随后无条件执行的赋值语句又将其恢复为全局新雪值：

```
如果 (options.MAX_SNOW_ALBEDO){如果 (iveg!=
```

```
            雪->反照率   =参数。SNOW_NEW_SNOW_ALB;
        }
}
雪->最后_雪   =  0;
雪->反照率   =参数。SNOW_NEW_SNOW_ALB;
```

后续的赋值语句会在此分支中覆盖较早的值。因此，存储的参考值并不能证明局部契约所描述的植被特定行为。这一源代码层面的发现说明，仅凭开发者溯源和补丁应用无法证明科学正确性。我们保留原始任务记录以确保可审计性，并将该任务标记为不适合进行正确性评分，直到其目标和参考值经过裁定。语料库统计数据包含这条被标记的记录；这些统计数据是已发布任务记录的数量，而不是经过专家认证的任务数量。

7 分析与局限性

当前证据所能证明的内容。 在参考补丁应用层面，语料库内部一致：所有 60 个参考补丁均可应用于所提供的父文件。测试夹具重放揭示了补丁应用与局部识别器接受之间的分歧，并量化了 ParFlow 表达式不匹配。这些结果有助于复现，但不足以证明基准难度、经过校准的科学正确性或模型能力的相对表现。

专家验证与标注。 科学描述、类别标签和物理契约仍需完成专家裁定。此快照包含自动验证器元数据和审核筛选列表，而非经过测量的人类一致性研究。没有可靠依据可报告 Cohen κ系数、类别混淆矩阵或标注错误率。缺少答案多重性标签，也使得无法合理划分出单一答案的精确匹配子集。VIC 覆写这一发现明确表明，在将语料库作为评分基准之前，应先裁定目标和参考值。

契约覆盖率。 五个关卡使用任务特定识别器和简化计算。只有 ParFlow 会独立评估从候选文本中提取的表达式；另外三个工具则在模式识别后选择参考输出。识别器可能遗漏有效的替代实现，而通过识别模式也可能与补丁其他位置无关的错误并存。有限的输出以及少数场景中的参考一致性，并不能证明长期的水、能量或示踪剂守恒。要评估这些性质，还需要完整构建、更广泛的测试和科学审查。

采样与比较。 六个代码仓库各贡献十个选定任务；这种设计无法代表科学代码变更的自然频率。五个测试装置未涵盖 CLM5，也未涵盖 PRM 和 PSS 类别。它们的求解器身份和生成配置并不完整。因此，无法根据这些结果推断模型之间的比较结果，也无法推断提示变体、检索或智能体脚手架的影响。

公共历史暴露。 提交和源代码均为公开内容。哈希值可以使任务可识别，但不能证明已实施训练数据排除。在作出新鲜度或记忆化声明之前，需要时间戳元数据和经过审计的时间划分。已存储的记忆排除策略表达的是预期规则，而不是某次特定检索运行遵循这些规则的证据。

成本与可复现性。 本地回放不会调用付费模型。根据保留的测试样例，无法还原历史求解器和评判器成本、令牌数、提供商区域以及确切的调用配置，因此无法进行成本效率比较。未来的实测评估需要将这些信息与逐任务输出一并保留。此处的可复现性是指根据存档输入重新生成语料库统计数据和局部门控结果，而不是复现未明确说明的模型调用。

8 讨论与结论

ESM-BENCH提供了一个包含60项任务的v5语料库，涵盖六个ESM代码仓库中的物理感知代码修改任务。这些任务结合了开发者提交溯源、提供的源文件、参考

补丁、科学描述和提示变体。制品审计验证了每项任务的参考补丁应用情况。五夹具审计表明，本地识别器可能会接受不生效的差异。AParFlow表达式重放量化了参考不一致；而对VICsource的检查发现，有一条参考赋值语句与所述反照率目标相悖。

当前证据支持这些语料库和工具结果，但不支持以下主张：某个特定模型理解地球系统模型物理、一个模型优于另一个模型，或成功的审查系统性地掩盖了物理失败。完整的 v5 评估需要可追溯的求解器输出、经过校准的专家审查、匹配的提示条件，以及更广泛的可执行覆盖率。发布的协议明确了这些要求，同时保留了目前已经可以复现的更有限结果。

致谢

作者感谢 Zesen Huang 和 Hejia Geng 在社区发展与合作方面所做的贡献，以及他们围绕科学基准测试进行的讨论并提供的反馈。

我们感谢 Tian Zhou、Lingcheng Li、Weihao Liu、Mashrekur Rahman、Yuhan Wang、Cenlin He、Harsh Kamath、HydroML 2026 和第四届 Annual Good Systems Smart Cities and AI Innovations Symposium 的参与者，以及其他研究人员和合作者，感谢他们围绕地球系统建模、基准设计和物理评估展开讨论并提供反馈。

我们感谢 Noah-MP、CLM5、SUMMA、ParFlow、VIC 和 MOM6 的开发者以开放方式开发这些模型；基准测试任务基于它们的公开提交历史构建。

作者使用人工智能工具协助软件开发和编辑，并对本文内容负责。本项目得到了 NSF NCAR/CISL 分配项目 UTAA0012 的支持。部分模拟在 NCAR CISL 和德州高级计算中心（TACC）超级计算机系统上完成。

参考文献

[1] Carlos E Jimenez、John Yang、Alexander Wettig、Shunyu Yao、Kexin Pei、Ofir Press 和 Karthik Narasimhan。SWE-bench：语言模型能否解决现实世界中的 GitHub 问题? 载于 国际学习表征会议，2024 年。网址：https://arxiv.org/abs/2310.06770。

[2] Minyang Tian、Luyu Gao、Shizhuo Dylan Zhang、Xinan Chen、Cunwei Fan、Xuefei Guo、Roland Haas、Pan Ji、Kittithat Krongchon、Yao Li、Shengyan Liu、Di Luo、Yutao Ma、Hao Tong、Kha Trinh、Chenyu Tian、Zihan Wang、Bohao Wu、Yanyu Xiong、Shengzhu Yin、Minhui Zhu、Kilian Lieret、Yanxin Lu、Genglin Liu、Yufeng Du、Tianhua Tao、Ofir Press、Jamie Callan、Eliu Huerta 和 Hao Peng。SciCode：由科学家整理的研究型编程基准。载于神经信息处理系统进展，第 37 卷，2024 年。网址：https://arxiv.org/abs/2407.13168。

[3] Ziru Chen、Shijie Chen、Yuting Ning、Qianheng Zhang、Boshi Wang、Botao Yu、Yifei Li、Zeyi Liao、Chen Wei、Zitong Lu、Vishal Dey、Mingyi Xue、Frazier N. Baker、Benjamin Burns、Daniel Adu-Ampratwum、Xuhui Huang、Xia Ning、Song Gao、Yu Su 和 Huan Sun。《Sci- enceAgentBench：面向数据驱动科学发现的语言智能体严格评估》。发表于国际学习表征会议（*ICLR*），2025 年。网址：https://arxiv.org/abs/2410.05080。

[4] Thomas Clune 和 Richard Rood。气候模型开发中的软件测试与验证。*IEEE Software*，28(6):49–55，2011 年。网址：https://ieeexplore.ieee.org/document/5999647。

[5] Koutian Wu、Wenjun Zhang、Junjie Zhou、Pengqian Han、Qiming Bao、Ergan Shang、Jiayu Wang、Junkai Wang、Wanghan Xu、Songyuanyi Lu、Yixuan Li、Rudy Mizrahi Celekli、Lazizbek Ravshanov、Yaoliang Bian、Xudong Han 和 Lin Shi。《Benchmark Radar：人工智能基准测试与评估的动态数据库和搜索引擎》，2026年。URL https://arxiv.org/abs/2609.11115。

[6] Xiaomin Li、Yuexing Hao、Jianheng Hou、Jintao Huang、Qianfeng Wen、Shirley Huang、Yifan Liu、Xiaoyi Liu、Yilan Fan、Yijun Wang、Koutian Wu、Ruoqi Gao、Muhammad Ahmed Mohsin、Jing Tang、Brihi Joshi、Heming Liu、Zheyuan Deng、Zonglin Di、Sankalp Jajee、Jiuyao Lu、Zhiwei Zhang、Saksham Kapoor、Ishan Gupta、Yunhan Zhao、Chanwoo Park、Yucheng Lu、Bing Hu、Weihang Xiao、Aravind Mohan、Hanwen Xing、Runyu Zhang、Mihir Kulshreshtha、Yuanda Xu、Qianyu Zhu、Dianzhuo Wang、Yuxin Xiao、Bowen Jiang、Yongye Su、Wenhao Chai、Zuxin Liu、Lawrence Yunliang Chen、Xuandong Zhao、Ethan Ye、ShivamPatel、Jason Xie、Alex Martin Richmond、Weixiang Ding、Emre Okcular、Diya Mathew、Ziheng Wang、Rana M. Shahroz Khan、Zhejian Peng、Fang Wu、Fan Nie、Xinyang Han、Yubin Kim、Jiawei Zhang、Zhenting Qi、Huangyuan Su、Xu Pan、Abinitha Gourabathina、Hyewon Jeong、Hemanth Neelgund Ramesh、Kumail Alhamoud、Kimia Hamidieh、Zidi Xiong、Samuel Schmidgall、Pengrui Han、Yepeng Huang、Yongheng Wang、Bowen Yang、Alex Gu、Yuchu Wang、Akshay Paruchuri、Brenna Li、Hejie Cui、Jiayuan Ding、Chaosheng Dong、Jiahao Wang、Yixuan He、Chi Wang、Pamela Bhattacharya、Tianyi Peng、Paul Pu Liang、Mitchell Gordon、Yilun Du、Marinka Zitnik、James Zou、Prasanna Tambe、Philip Torr、Emily Fox、Asu Ozdaglar 和 Dawn Song。《MatrAIx：使用83亿个角色智能体模拟世界》，2026年。URL https://arxiv.org/abs/2608.04205。

[7] Junwei Zhou、Zhen Sun、Binyu Li、Jiangyu Zhou、Yuexi Pan、Hengyu Wang、Honghe Ren、Xiaohan Jia、Xueyang Zhou、Xiaoyu Cao、Yongchao Chen、Yuanning Feng、Junhao Wu、Cheng Zhang、Sijia Chen、Haoyu Xue、Chengsong You、Huan Wang、Koutian Wu、Peigan Gao、Jiakun Wu、Wenzhe Li、Ergan Shang、Qingyuan Zheng、Jingjing Zhou、Ruixuan Jia、Yan Xu、Hongrui Zhang、Xiao-Han Ma、Zhengxiang Cheng、Yuexing Hao、Liting Mai、Xianglin Ji、Wenjun Zhang、Zhuofan Chen、Yixiao Huang、Chi Wang、Wenyue Hua、Yilun Hao、Yuantao Zhai、Ziyan Zhao 和 Jingyan Xie。《ASI-Bench：人工超智能的黎明》，2026年。URLhttps://arxiv.org/abs/2608.17271。

[8] Giulio Starace、Oliver Jaffe、Dane Sherburn、James Aung、Jun Shern Chan、Leon Maksin、Rachel Dias、Evan Mays、Benjamin Kinsella、Wyatt Thompson、Johannes Heidecke、Amelia Glaese 和 Tejal Patwardhan。《PaperBench：评估人工智能复现人工智能研究的能力》，2025 年。网址 https://arxiv.org/abs/2504.01848。

[9] Hjalmar Wijk、Tao Lin、Joel Becker、Sami Jawhar、Neev Parikh、Thomas Broadley、Lawrence Chan、Michael Chen、Josh Clymer、Jai Dhyani、Elena Ericheva、Katharyn Garcia、Brian Goodrich、Nikola Jurkovic、Holden Karnofsky、Megan Kinniment、Aron Lajko、Seraphina Nix、Lucas Sato、William Saunders、Maksym Taran、Ben West 和 Elizabeth Barnes。《RE-Bench：基于人类专家评估前沿人工智能研发语言模型智能体能力》，2024 年。网址https://arxiv.org/abs/2411.15114。

[10] Nicholas Edwards、Yukyung Lee、Yujun Audrey Mao、Yulu Qin、Sebastian Schuster 和 Najoung Kim。RExBench：编码智能体能否自主实现人工智能研究扩展？，2025年。网址 https://arxiv.org/abs/2506.22598。

[11] Seungone Kim、Jiayi Geng、Young-Jun Lee、Bo Liu、Juyoung Suk、Nicholas Lubbers、Mitchell Wood、Marcus Noack、Yong-Xin Yao、Yuan Shi、Roxana Bujack、Cristina Garcia Cardona、Austin Rodriguez、Ilon Joseph、Woojin Cho、Zora Zhiruo Wang、Chuanyang Jin、Ian Wu、Apurva Gandhi、Tianjian Li、Weiwei Sun、Carolin Lawrence、Kiril Gashteovski、Dongyeop Kang、Jay Shin、Akari Asai、Graham Neubig 和 Sean Welleck。《Into the Unknown：在开放性科学研究问题上对人工智能代理进行基准测试》，2026年。网址 https://arxiv.org/abs/2610.19159。

[12] Dingyuan Dai、Heli Qi、Lei Liu、Yinxi Li、Baiding Chen、Zijun Dou、Qingcheng Zeng、Qi Kang、Oliver Sun、Eric Wang、Bo Zhou、Haixin Wang、Yufan Du、Shi Bo、Ruihan Lin、Mengqi Yuan、Dunjie Lu、Steven Dillmann、Yiming Shi、Tina Su、Amy Xin、Minghao Liu、Xi Wang、Xu Huang、Ge Zhang、Pengyu Nie、Zhen Yang、Jie Tang、Juanzi Li、Weihao Xuan 和 Tianyu Liu。《OSWorld-Science：用于学习和使用科学软件的计算机使用智能体基准测试》，2026年。URLhttps://arxiv.org/abs/2609.39903。

[13] Mark Chen、Jerry Tworek、Heewoo Jun、Qiming Yuan、Henrique Ponde de Oliveira Pinto、Jared Kaplan、Harri Edwards、Yuri Burda、Nicholas Joseph、Greg Brockman、Alex Ray、Raul Puri、Gretchen Krueger、Michael Petrov、Heidy Khlaaf、Girish Sastry、Pamela Mishkin、Brooke

Chan、Scott Gray、Nick Ryder、Mikhail Pavlov、Alethea Power、Lukasz Kaiser、Mohammad Bavarian、Clemens Winter、Philippe Tillet、Felipe Petroski Such、Dave Cummings、Matthias Plappert、Fotios Chantzis、Elizabeth Barnes、Ariel Herbert-Voss、William Hebgen Guss、Alex Nichol、Alex Paino、Nikolas Tezak、Jie Tang、Igor Babuschkin、Suchir Balaji、Shantanu Jain、William Saunders、Christopher Hesse、Andrew N. Carr、Jan Leike、Josh Achiam、Vedant Misra、Evan Morikawa、Alec Radford、Matthew Knight、Miles Brundage、Mira Murati、Katie Mayer、Peter Welinder、Bob McGrew、Dario Amodei、Sam McCandlish、Ilya Sutskever 和 Wojciech Zaremba。评估在代码上训练的大语言模型。*arXiv* 预印本 *arXiv:2107.03374*，2021年。URLhttps://arxiv.org/abs/2107.03374。

[14] Hejia Geng, Zesen Huang, Haoyang Li, Wenbin Li, Koutian Wu, Zihan Zhou, Yuanbo Pang, Weihao Liu, Zigong Xu, Zhiping Li, Zongzheng Zhang, Chuanfei Dong, Jiankai Sun, Tianzhe Zheng, Fengyu Xie, Yue Ma, Yueheng Shi, Tong Xie, Zonglin Di, Xianrong Liu, Qucheng Gao, Yimin Liu, Jiaming Pan, Sheng Huang, Xiao-Han Ma, Lanqing Yuan, Zhenlin Zhu, Ziang Liu, Ziyang Xu, Junkai Wang, Kangkai Liang, Jiayi Xian, Zehong Zhao, Liuwei Xu, Jingxu Xie, Peijin Zhang, Qiang Gao, Chengyi Xing, Zhe Zhao, Xi Wang, Yaopeng Xing, Xing Meng, Zhenfei Yin, Yingcheng Wu 和 Ling Yang。ScienceIDE：将全球科学代码库转化为智能体可学习环境，2026年。网址：https://arxiv.org/abs/2609.19134。

[15] Qifan Zhang、Yuhan Li、Nuo Chen、Bowen Liu、Xiao Hu、Miao Peng、Chenyi Zi、Lixin Liu、Lydell Huang、Jing Tang 和 Jia Li。POSEIDON：扩展科学智能体的训练环境，覆盖代码仓库，2026年。网址：https://arxiv.org/abs/2610.32448。[16] Yuanli Wang、Yaoyao Qian、Yue Zhang、Hanhan Zhou、Jindan Huang、Tianfu Fu、Qiuyang Mang、Huanzhi Mao、Wenhao Chai、Wendong Fan 和 Liqiang Jing。DeployBench：大语言模型智能体研究产物部署基准测试，2026年。网址：https://arxiv.org/abs/2606.05238。[17] Zhipeng Xu、Jiahao Lu、Yining Zheng、Yuxin Wang 和 Xipeng Qiu。SWE-bench Science：编码智能体能否解决科学领域的工程任务？，2026年。网址：https://arxiv.org/abs/2608.19799。[18] Xin He、Yanlin Wang、Mingwei Liu、Jiachi Chen、Hongyu Zhang 和 Guanbin Li。SWE-Gate：通过功能测试对软件工程智能体来说还不够，2026年。网址：https://arxiv.org/abs/2609.04167。[19] Hong Zhang、Barry Smith、Satish Balay、Le Chen、Murat Keceli、Lois Curfman McInnes 和 Junchao Zhang。PETSc中人工智能生成的科学代码的智能体评估框架，2026年。网址：https://arxiv.org/abs/2603.15976。[20] Koutian Wu。Noah-Agent：用于大型Fortran气候模型自动参数化与验证的多专家人工智能智能体框架。预印本，0.1版，Zenodo，2025年。网址：https://doi.org/10.5281/zenodo.17862049。[21] Maohao Ran、Chendong Ma、Yanting Zhang、Dailing Jiang、Yusen Huang、Meng Gao 和 Jun Song。基于执行的评估揭示了环境科学语言模型计算中的隐性失败，2026年。网址：https://arxiv.org/abs/2608.18726。[22] Dat Tien Nguyen、Thao Nguyen、Fadillah Adamsyah Maani、Huy M. Le、Muhammad Umer Sheikh、Numan Saeed、Muhammad Haris Khan 和 Salman Khan。TerraBench：智能体能否对异构地球系统数据进行推理？，2026年。网址：https://arxiv.org/abs/2606.13148。[23] Ruiran Su、Jiasheng Si、Zhijiang Guo 和 Janet B. Pierrehumbert。ClimateViz：用于科学图表统计推理和事实核查的基准测试，2025年。网址：https://arxiv.org/abs/2506.08700。[24] Ruize Li、Zhibin Wen、Tao Han、Hao Chen、Fenghua Ling、Wei Zhang、Song Guo 和 Lei Bai。RealBench：在业务运行条件和极端事件挑战下对数据驱动的数值天气预报进行基准测试，2026年。网址：https://arxiv.org/abs/2605.24945。[25] Siru Jiang、Yongzhe Lyu、Shuo Lu、Yubin Wang、Yuxiang Zhang、Yue Liao、Bin Wang、Jian Liang 和 Tieniu Tan。WorldSolver：大语言模型智能体能否通过生成求解器模拟物理动力学？，2026年。网址：https://arxiv.org/abs/2610.08720。

[26] Yuchen Sun、Jinjin He、Sinan Wang 和 Bo Zhu。GPUPhysBench：对编码智能体进行基准测试，评估其能否正确且高效地运行GPU物理模拟，2026年。网址：https://arxiv.org/abs/2609.35639。[27] Binghang Lu、Lang Cao、Changhong Mou、Yige Sun、Cindy Xiangrui Kong、Zheyuan Deng、Tian Yuan、Bing Hu、Runyu Zhang、Na Lu 和 Guang Lin。RSI-PDEbench：偏微分方程求解自我改进代理的基准测试，2026年。网址：https://arxiv.org/abs/2611.14949。[28] Bowen Cui、Junyu Yin、Tejas Ramesh、Leo Lim、Oscar Hernandez 和 Keren Zhou。HPC-Bench：高性能计算代码综合基准测试，2026年。网址：https://arxiv.org/abs/2611.02752。[29] Strahinja Ljaljevic、Josep Jorba 和 Sergio Iserte。探索大语言模型在高性能计算编程中的作用：综述，2026年。网址：https://arxiv.org/abs/2608.26110。[30] Shraddha Surana、Ashwin Srinivasan 和 Michael Bain。识别代码的潜在声明式表示，以辅助代码仓库迁移，2026年。网址：https://arxiv.org/abs/2608.23619。[31] Veronika Eyring、Sandrine Bony、Gerald A Meehl、Catherine A Senior、Bjorn Stevens、Ronald J Stouffer 和 Karl E Taylor。耦合模式比较计划第六阶段（CMIP6）实验设计与组织概述。《地球科学模型开发》，9(5)：1937–1958，2016年。网址：https://gmd.copernicus.org/articles/9/1937/2016/。[32] Mattia Righi、Bouwe Andela、Veronika Eyring、Axel Lauer、Valeriu Predoi、Manuel Schlund、Javier Vegas-Regidor、Lisa Bock、Björn Brötz、Lee de Mora 等。ESMValTool v2.0——技术概述。《地球科学模型开发》，13(3)：1179–1199，2020年。网址：https://gmd.copernicus.org/preprints/gmd-2019-226/gmd-2019-226.pdf。[33] Naman Jain、King Han、Alex Gu、Wen-Ding Li、Fanjia Yan、Tianjun Zhang、Sida Wang、Armando Solar-Lezama、Koushik Sen 和 Ion Stoica。LiveCodeBench：对大语言模型代码能力进行全面且无污染的评估。发表于《国际学习表征会议》，2025年。网址：https://arxiv.org/abs/2403.07974。[34] Koutian Wu。将提交日志作为技能矿藏：从科学研究轨迹中提取16项代理技能，2026年4月。网址：https://doi.org/10.5281/zenodo.19582289。[35] Guo-Yue Niu、Zong-Liang Yang、Kenneth E Mitchell、Fei Chen、Michael B Ek、Michael Barlage、Anil Kumar、Kevin Manning、Dev Niyogi、Enrique Rosero、Mukul Tewari 和 Youlong Xia。具有多参数化选项的社区Noah陆面模式（Noah-MP）：1. 模式描述及基于局地尺度测量的评估。《地球物理研究杂志：大气》，116(D12)：D12109，2011年。doi：10.1029/2010JD015139。网址：https://doi.org/10.1029/2010JD015139。[36] Koutian Wu、Lingcheng Li、Daniella Rempe、Ashley M. Matheny、Mahmoud Mbarak 和 Zong-Liang Yang。整合植物水力学如何改进Noah-MP陆面模式。第106届AMS年会，美国气象学会，2026年1月。网址：https://ams.confex.com/ams/106ANNUAL/meetingapp.cgi/Paper/473862。会议摘要及海报，2026年1月26日。[37] David M. Lawrence、Rosie A. Fisher、Charles D. Koven、Keith W. Oleson、Sean C. Swenson、Gordon Bonan、Nathan Collier、Bardan Ghimire、Leo van Kampenhout、Daniel Kennedy、Erik Kluzek、Peter J. Lawrence、Fang Li、Hongyi Li、Danica Lombardozzi、William J. Riley、William J. Sacks、Mingjie Shi、Mariana Vertenstein、William R. Wieder、Chonggang Xu、Ashehad A. Ali、Andrew M. Badger、Gautam Bisht、Michiel van den Broeke、Michael A. Brunke、Sean P. Burns、Jonathan Buzan、Martyn Clark、Anthony Craig、Kyla Dahlin、Beth Drewniak、Joshua B. Fisher、Mark Flanner、Andrew M. Fox、Pierre Gentine、Forrest Hoffman、Gretchen Keppel-Aleks、Ryan Knox、Sanjiv Kumar、Jan Lenaerts、L. Ruby Leung、William H. Lipscomb、Yaqiong Lu、Ashutosh Pandey、Jon D. Pelletier、Justin Perket、James T. Randerson、Daniel M. Ricciuto、Benjamin M. Sanderson、Andrew Slater、Zachary M. Subin、Jinyun Tang、R. Quinn Thomas、Maria Val Martin 和 Xubin Zeng。社区陆面模式第5版：新特性描述、基准测试及强迫不确定性的影响。《地球系统建模进展杂志》，11(12)：4245–4287，2019年。doi：10.1029/2018MS001583。网址：https://doi.org/10.1029/2018MS001583。

[38] Martyn P. Clark、Bart Nijssen、Jessica D. Lundquist、Dmitri Kavetski、David E. Rupp、Ross A. Woods、Jim E. Freer、Ethan D. Gutmann、Andrew W. Wood、Levi D. Brekke、Jeffrey R. Arnold、David J. Gochis 和 Roy M. Rasmussen。基于过程的水文建模的统一方法：1. 建模概念。《水资源研究》，51(4):2498–2514，2015年。doi: 10.1002/2015WR017198。网址：https://doi.org/10.1002/2015WR017198。

[39] Stefan J. Kollet 和 Reed M. Maxwell。地表水—地下水流动综合建模：并行地下水流动模型中的自由水面地表径流边界条件。《水资源进展》，29(7):945–958，2006年7月。doi: 10.1016/j.advwatres.2005.08.006。网址：https://doi.org/10.1016/j.advwatres.2005.08.006。

[40] Xu Liang、Dennis P. Lettenmaier、Eric F. Wood 和 Stephen J. Burges。面向大气环流模式的陆地表面水和能量通量简易基于水文过程的模型。《地球物理研究杂志：大气》，99(D7):14415–14428，1994年。doi: 10.1029/94JD00483。网址：https://doi.org/10.1029/94JD00483。

[41] Alistair Adcroft、Whit Anderson、V. Balaji、ChrisBlanton、MitchellBushuk、Carolina O. Dufour、John P. Dunne、Stephen M. Griffies、Robert Hallberg、Matthew J. Harrison、Isaac M. Held、Malte F. Jansen、Jasmin G. John、John P. Krasting、Amy R. Langenhorst、Sonya Legg、Zhi Liang、Colleen McHugh、Aparna Radhakrishnan、Brandon G. Reichl、Tony Rosati、Bonita L. Samuels、Andrew Shao、Ronald Stouffer、Michael Winton、Andrew T. Wittenberg、Baoqiang Xiang、Niki Zadeh 和 Rong Zhang。GFDL 全球海洋和海冰模型 OM4.0：模型描述和模拟特征。《地球系统建模进展杂志》，11(10):3167–3211，2019年。doi: 10.1029/2019MS001726。网址https://doi.org/10.1029/2019MS001726。

可复现性与证据清单

A.1 选定快照

选定语料库为 data/tasks_v5。论文的预印本_snapshot.json 列出了全部 60 个任务路径、任务和源提交标识、SHA-256 文件哈希、类别和层级元数据、源文件数量、参考差异应用结果以及契约覆盖率。其清单哈希标识的是任务文件内容，而非依赖可能变化的目录名称。论文使用根据此明确语料库生成的 preprint_metrics.tex。早先生成的 generated_metrics.tex 仍是历史工件，不纳入本预印本的数值结果。
可以在仓库根目录运行以下命令，重新执行语料库审计：

```
python3 脚本/ generate_预印本_快照 .py
```

该脚本会在临时目录中重建每个提供的起始工作区，并使用 git apply –check 检查参考补丁。它会将难度辅助文件排除在任务数量统计之外。它不会下载上游代码库，也不会查询模型 API。此检查验证的是存储源一致性，而非通过全新上游检出版本核实外部来源。

A.2 本地门控重放

该重放使用 Python 3.10 或更高版本、NumPy 和 PyYAML。可以使用以下命令重新生成：

```
python3 scripts / run_sim_compare .py \ -- config configs / simulation_compare . ya
```

输出目录必须位于仓库内，因为运行器会记录相对于仓库的路径。归档重放包含已解析的 YAML 配置、摘要以及每个任务对应的五个结果文件。输入差异位于 data/sim_co mpare_fixtures 中。该重放于 2026年9月6日运行，使用的是来自仓库提交 30c4d76 的评估代码；预印本修订版更改了文档并添加了审计/图表工具，但未更改该门控实现。
ParFlow 核函数在 $x \in [0.001, 10\pi]$ 和 $y \in [0, 10\pi]$上每个坐标采样 512 个点。均方根误差和最大误差容差分别为 10^{-6} 和 10^{-5}。MOM6 关卡检查五种源代码

代码模式，且不进行数值轨迹比较。Noah-MP 在两种情景下将叶质量与 $10^{-12}\,\mathrm{g\,m^{-2}}$绝对容差进行比较。SUMMA 在两种情景下使用 $10^{-12}\,\mathrm{m\,s^{-1}}$ 容差比较径流。VIC 在两种情景下使用 10^{-12} 反照率容差和 $10^{-9}\,\mathrm{Wm^{-2}}$ 吸收短波辐射容差。这些是已配置的工程阈值；本文未声称进行过容差校准实验。

A.3 InterpretingChecker 输出

存档字段 sim_pass 表示已注册的本地检查器接受了样例文本。补丁应用由语料库审计脚本单独检查。只有 ParFlow 会评估从样例中提取的数值表达式。MOM6 能识别差异模式，包括已删除的块行。Noah-MP、SUMMA 和 VIC 在识别预期文本后会计算示例性场景输出，因此它们与手工编码参考值之间的零差异并非独立的执行结果。存档保留了这些原始输出，但解读时必须考虑这些实现方面的限制。
任务清单包含参考差异应用结果以及单独记录的样例差异失败。第 6 节中指出的 VIC-PO-001 覆盖问题属于定性源代码审计，并非对整个语料库中参考值缺陷的完整估算。该任务仍保留在已发布的记录集中，并附有这一明确说明。

A.4 比较评估的要求

要对该语料库进行模型比较，需要提供任务清单和内容哈希；准确存储的提示词及附加的源代码；固定的模型标识符和提供方配置；采样参数和时间戳；完整响应、提取出的差异以及应用检查；包含评分标准和评审器版本的逐项审查评分；以及令牌数、延迟和成本记录。人工校准应报告其采样程序和实际的专家一致性。提示词比较需要使用匹配的任务。任何有关完整模型执行的声明，都需要提供可复现构建和运行时输出，而不能仅凭本地模式检查。

当前代码包含一套八项审查评分标准和审查资料包工具。它们可用并不意味着已完成人工验证。本预印本未评估上下文检索、基于提交记录的记忆、轨迹评分以及长时段物理审计。

B 数据与代码可用性

重现所报告检查所需的 v5 任务记录、夹具差异、审计和本地检查脚本、配置以及存档输出均由作者保留，未随本版本发布。本版本未声称提供公共代码或数据集存储库。由于底层求解器和评审器产物未包含在发布内容中，历史模型运行汇总数据未纳入所报告的结果。

C 更广泛影响

该基准测试有助于使对人工智能生成的科学代码的审查更加明确。其主要风险在于，将局部得分过度解读为补丁可安全用于生产级建模的证据。无论是参考补丁一致性，还是在选定内核上的成功，都不能替代完整科学工作流程中的领域审查和测试。所有任务记录均源自公开的代码库历史记录；基准测试用户应保留上游来源归属及适用的许可证。