EARTHARENA: BENCHMARKING AGENTS ON REAL-WORLD, LONG-HORIZON SCIENTIFIC TASKS IN EARTH OBSERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific investigations in Earth observation combine heterogeneous observations, geospatial processing, and interdisciplinary analyses into long-horizon workflows with interdependent stages. Agents must apply scientific methods appropriately and maintain the scientific validity of intermediate artifacts and final conclusions. However, whether they can meet these requirements throughout the workflow remains to be systematically evaluated. To this end, we introduce EARTHARENA, an executable benchmark for real-world, long-horizon scientific tasks in Earth observation. We construct 101 tasks from 30 peer-reviewed publications spanning diverse areas of Earth observation. Each case organizes tasks linked by intermediate artifacts around a scientific objective, with reproduced reference results and human-reviewed scoring rubrics. EarthArena links scientific conclusions to supporting artifacts and execution evidence to assess agents' ability to maintain scientific validity throughout the workflow. Across six language models and three agent harnesses, the best-performing configuration achieves a case-averaged Rubric score of 44.3. Model rankings vary across harnesses, highlighting the importance of evaluating scientific task performance at the level of complete model–harness configurations. EarthArena provides a testbed for systematically evaluating and improving agents' ability to conduct real-world, long-horizon scientific investigations in Earth observation.

1 INTRODUCTION

AI4Science seeks to connect scientific questions with evidence-supported results (Wang et al., 2023). Earth observation makes this goal concrete: investigations of glacier retreat, urban cooling, and fire emissions connect observations, geospatial processing, statistical analysis, and physical interpretation across disciplines. Observations vary across locations, times, and sensors, with differences in spatial resolution, temporal coverage, and spectral measurements (Rolf et al., 2024). Integrating these observations requires reconciling spatial and temporal scales (Reichstein et al., 2019) and ensuring that the chosen methods are applied under conditions consistent with their scientific assumptions. Scientific validity must therefore be maintained throughout the workflow, from data acquisition and processing to analysis and interpretation. Masks, baselines, and incomplete downloads can alter the measured population or comparison.

Language-model agents plan geospatial analyses and operate EO tools (Shabbir et al., 2025; Feng et al., 2026; Shabbir et al., 2026). Can they sustain an investigation by selecting suitable products, acquiring observations, applying defensible methods, and preserving these decisions across dependent analyses? A workflow is long-horizon because earlier choices constrain later evidence, not simply because it contains many tool calls. Evaluation must therefore connect intermediate products to the conclusions they support.

Scientific benchmarks assess paper-derived programming and research replication (Chen et al., 2025; Starace et al., 2025); EO benchmarks evaluate tool-driven analysis and disaster-response workflows (Shabbir et al., 2025; Feng et al., 2026; Shabbir et al., 2026; Wang et al., 2026). EarthArena focuses this evaluation on paper-grounded geoscience investigations with linked tasks, reproduced scientific references, and inspectable evidence. EARTHARENA contains 30 cases and 101 tasks. Realism comes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

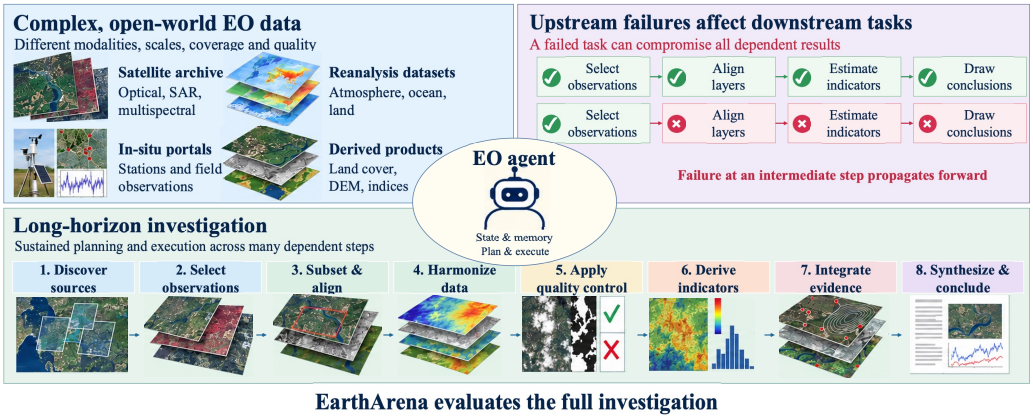

Figure 1: Motivation for EarthArena.

from published research objectives, data requirements, and methods; long-horizon structure comes from necessary artifact dependencies. Reference execution and human review establish expected results. Public packages specify questions, permitted inputs, outputs, and execution constraints; private packages hold reference artifacts and frozen rubrics. Scoring answers, artifacts, and execution evidence exposes both partial progress and unsupported conclusions.

We evaluate the model–harness system, including tool access, state, process management, and recovery. Interface design affects agent performance; geoscience additionally requires large-data access and persistent scientific settings. Across six models and three harnesses, GPT-6-Luna with DeepSeek attains the highest Rubric score, 44.3, and the leading model changes with the harness. Scientific agents should therefore be designed and evaluated as integrated model–harness systems (Yang et al., 2024).

EarthArena connects benchmark measurement to an AI4Science research agenda. We call for harnesses that acquire data at the required scale, preserve scientific parameters, and recover along artifact dependencies. The next evaluation frontier is to start from a scientific objective, discover compatible evidence across data providers, and verify the resulting analysis. Progress should mean more scientifically supported work under a fixed budget.

2 Related Work

2.1 Scientific Agents and Scientific Task Benchmarks

PaperBench evaluates research replication with hierarchical rubrics (Starace et al., 2025); ScienceAgentBench tests individual paper-derived, data-driven tasks (Chen et al., 2025). BixBench evaluates multi-step biological data analysis and interpretation (Mitchener et al., 2025). ScienceWorld studies reasoning in a simulated scientific environment (Wang et al., 2022), while MLAgentBench, MLE-bench, and RE-Bench assess machine-learning experimentation and engineering (Huang et al., 2023; Chan et al., 2024; Wijk et al., 2024). CORE-Bench evaluates computational reproducibility from published artifacts (Siegel et al., 2024). EarthArena adopts paper grounding and outcome-based assessment while preserving dependencies among heterogeneous geoscience data and products.

2.2 Earth Observation Agents and Benchmarks

The EO benchmarks in Table 1 emphasize tool-grounded reasoning. ThinkGeo evaluates multi-step planning over satellite and aerial imagery, with step-wise and final-answer assessment (Shabbir et al., 2025). Earth-Bench covers quantitative analysis across RGB, spectral observations, and EO products (Feng et al., 2026). OpenEarthAgent couples queries with imagery, GIS layers, and rasters for structured geospatial tool use (Shabbir et al., 2026). DORA grounds disaster perception, spatial analysis, evacuation planning, and report synthesis in real-world disaster events (Wang et al., 2026).

Table 1: Comparison with representative agent benchmarks.

Benchmark	Domain	Task sources	Data acquisition	Multi-turn	User simulation	Avg.Steps	Avg.Tokens	EO rubric	Evaluation
PaperBench	ML	20 publications	✓	×	×	–	–	×	Answer + Process
BixBench	Biology	Research workflows	×	×	×	–	–	×	Answer
GAIA2	General	Simulated environments	×	✓	×	22.5	10K	×	Action + Process
ScienceAgentBench	Science	44 publications	×	×	×	–	–	×	Answer + Process
ThinkGeo	EO	EO datasets	×	×	×	3.65	–	×	Answer + Process
Earth-Bench	EO	EO datasets	×	×	×	5.4	414.5K	×	Answer + Process
OpenEarthAgent	EO	EO datasets	×	×	×	6.04	–	×	Answer + Process
DORA	EO	Real-world disaster events	×	×	×	6.8	–	×	Answer + Process
EarthArena (ours)	EO	30 EO publications	✓	✓	✓	50.4	1M	✓	Answer + Process

TerraBench combines EO imagery, gridded environmental data, GIS, and simulation in 403 tasks, including document-grounded verification, with separate tool-use metrics and tolerance-aware numerical scoring (Nguyen et al., 2026). EarthArena emphasizes paper-derived investigations organized into dependent tasks, with rubrics linking scientific conclusions to reproduced references, intermediate artifacts, and execution evidence.

2.3 Long-Horizon Evaluation and Data Discovery

Mind2Web studies cross-website generalization; WebArena evaluates long-horizon web tasks, and GAIA tests browsing and tool use (Deng et al., 2023; Zhou et al., 2024; Mialon et al., 2024). GAIA2 evaluates adaptation to asynchronous events under time constraints, with action-level verification (Froger et al., 2026). EarthArena targets scientific dependencies: downstream analyses must preserve an artifact's spatial support, temporal coverage, and physical meaning. NASA-EO-Bench evaluates EO data discovery (Yu et al., 2026); EarthArena connects retrieval to computation and interpretation, requiring a relevant resource to become an accessible, scientifically compatible input. Table 1 compares benchmark settings.

3 The EarthArena Benchmark

3.1 Benchmark Structure and Task Definition

A *case* defines a bounded scientific objective from a source paper; a *task* is an independently issued and scored unit within it. An *artifact* is an inspectable product, such as a quality-controlled time series, gridded measurement, spatial boundary, or structured summary.

3.2 Paper-Grounded Construction and Quality Control

Tasks form a directed acyclic dependency graph. Each edge identifies an upstream artifact or result needed downstream. For example, one task delineates glacier boundaries from satellite images acquired in two different years; a downstream task uses these boundaries to calculate glacier area loss. Revising either boundary changes the estimated loss.

Each public task specifies its query, available inputs, required outputs, and execution constraints. Submissions contain an answer, requested artifacts, and a structured execution record. Private references remain outside candidate workspaces. The task contract defines the evaluated objective; a case need not reproduce every analysis in its source paper.

Construction has five stages: paper selection, task design, reference execution, rubric construction, and human review. We start from 10,636 publications represented in NASA-EO-Bench and linked through the NASA Earth Observation Knowledge Graph (Yu et al., 2026). LLM-assisted screening selects 500 candidates with concrete objectives and recoverable data and methods, excluding inaccessible proprietary inputs and undocumented manual intervention.

For each candidate, we define the objective and necessary tasks. Agent-assisted reference execution tests data access, implementation, and result reconstruction, retaining 100 papers for human review.

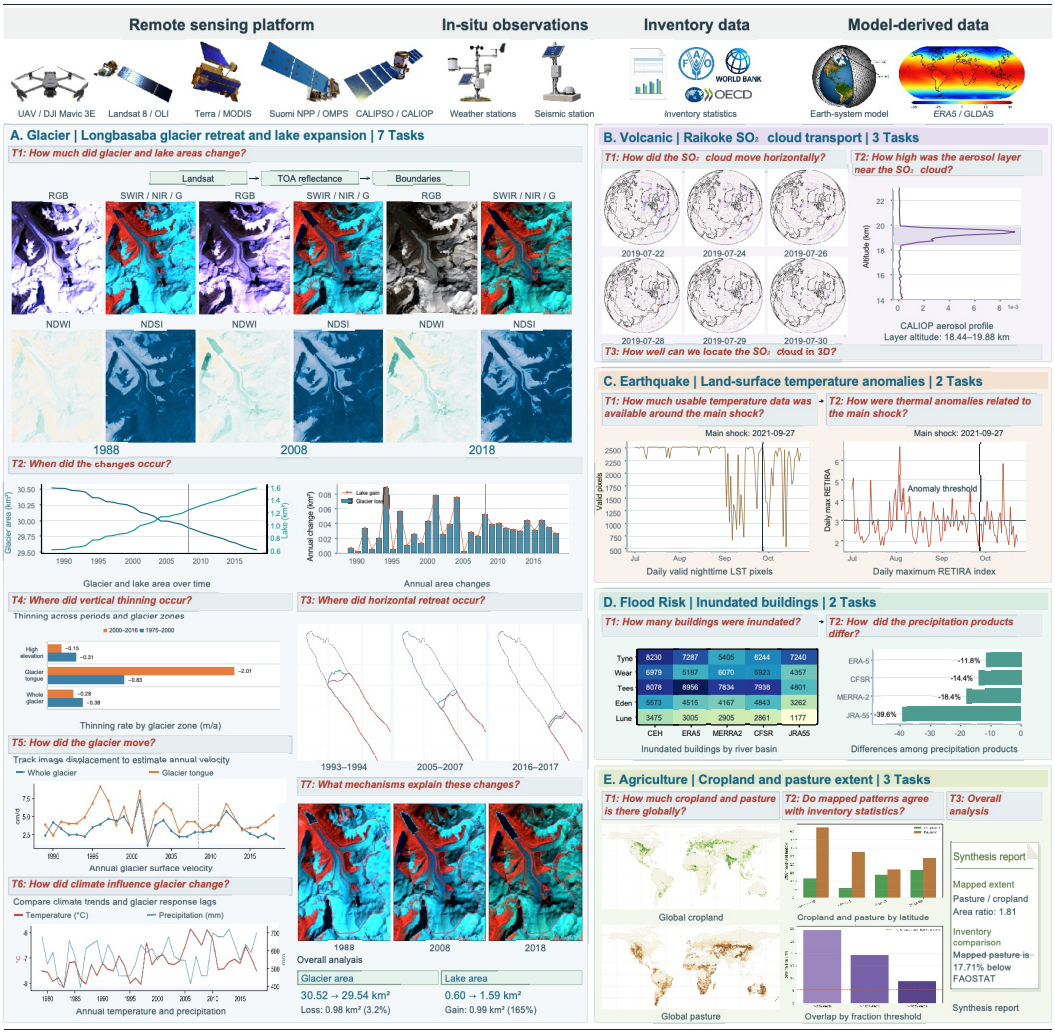

Figure 2: Representative cases of EarthArena.

Reproduced quantities are stored separately from paper-reported values. Discrepancies due to data revisions, rounding, or implementation choices are documented; tolerances reflect units, precision, and reference resolution.

Human review checks scientific correspondence, executable scope, references, and rubrics. Geoscience results depend on quality controls, spatial masks, aggregation, forcing data, calibration, and boundary conditions. Simulation or socioeconomic studies are excluded when missing resources prevent validation, rather than on disciplinary grounds.

Review also separates legitimate inputs from shortcuts. An archived model output can support a task about that archive, but cannot establish model execution when execution is required. Copying a published intermediate measurement likewise does not demonstrate its computation. Accepted packages freeze queries, references, output contracts, and scoring criteria before evaluation. The final benchmark contains 30 cases and 101 tasks. Figure 3 summarizes the pipeline; Appendix A provides the review details.

3.3 Coverage and Task Characteristics

Cases span NASA GCMD topics including atmosphere, oceans, cryosphere, land surface, biosphere, agriculture, solid Earth, and terrestrial hydrosphere. They range from glacier–lake systems and

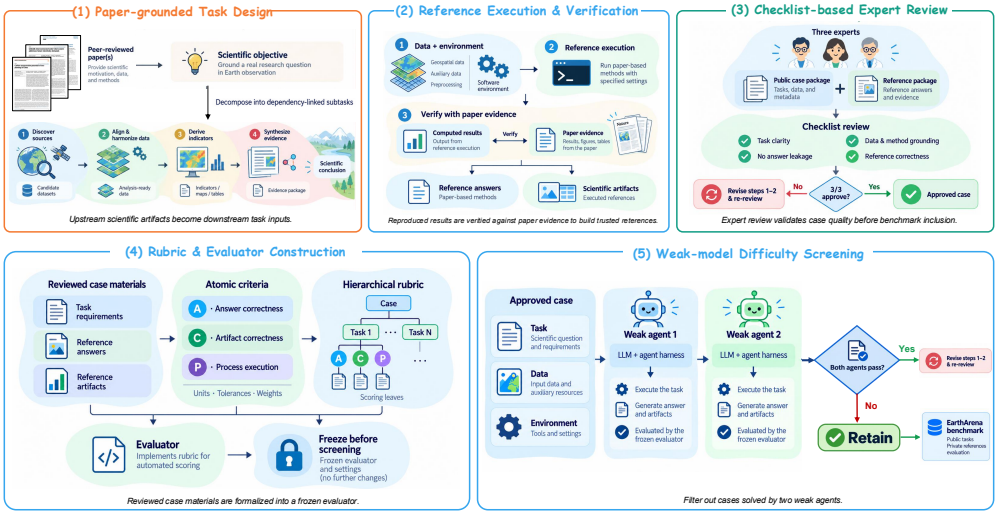

Figure 3: Paper-grounded construction and quality control.

cities to continental inventories and global land-use maps, with short-lived fire and volcanic events, multidecadal climate records, and biome reconstructions spanning 1700–2000.

Inputs include satellite observations, reanalysis, in-situ measurements, environmental inventories, and published model outputs. Acquisition, geospatial processing, statistical comparison, sensitivity analysis, and synthesis form workflows whose upstream choices determine the evidence available downstream.

Starting conditions distinguish primary-data acquisition, derived-product analysis, and public-record verification. Stefansson Sound audits in-situ radiation records; the sea-ice albedo case analyzes archived simulations. These evaluate record verification and simulation-output analysis, respectively, rather than open-world discovery or model execution. Appendix B lists source papers.

3.4 Representative Scientific Workflows

Shanghai links harmonized temperature, vegetation, population, elevation, land-cover, and built-environment data to cooling estimates and city-level checks. Galapagos links annual SST grids to trends and regional synthesis. The seven-task Longbasaba case combines glacier and lake change, retreat, surface lowering, velocity, and climate evidence. Raikoke combines SO_2 tracking with lidar constraints; Crete compares thermal anomalies with event dates without treating coincidence as seismic causation. Figure 2 illustrates the case setting; Appendix F preserves the detailed workflows.

4 Evaluation Protocol and Metrics

4.1 Execution Environment and Interaction Protocol

A run pairs a language model with a harness that invokes tools, manages processes, exposes outputs, and maintains state. Agents receive the public package and permitted upstream artifacts in the case environment. They submit `answer.md`, required files, and JSONL records of inputs, operations, parameters, outputs, status, and provenance. Records link claimed computations to inspectable evidence; a narrative alone does not establish execution. Scores apply to the submitted snapshot.

Retries consume the run budget. Downstream tasks use designated candidate artifacts. Runs that exceeded the prescribed execution-time limit and failed attempts remain in the denominator under the applicable scoring policy. Private-gold access violates this separation and is handled as a protocol exclusion. Reproduction requires run-specific model, harness, environment, and budget settings.

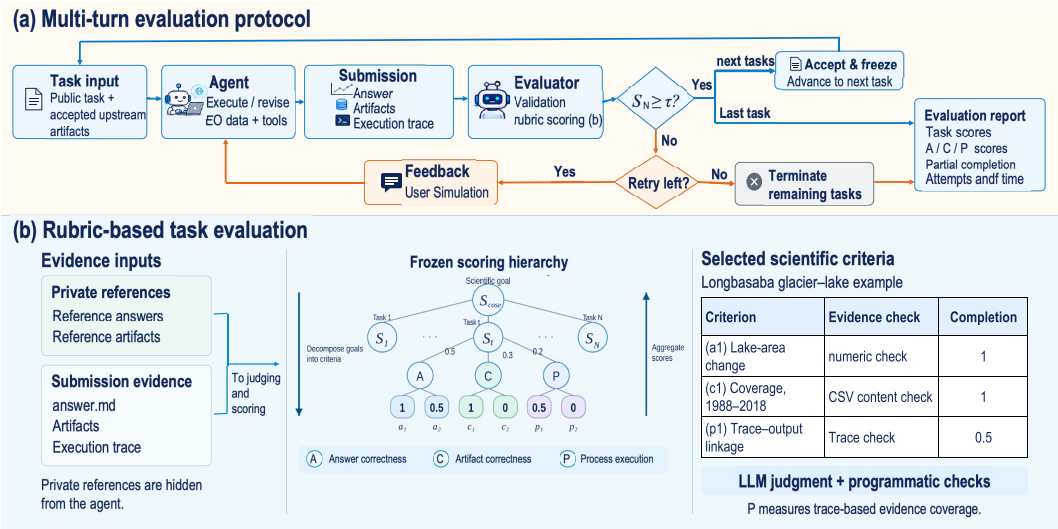

Figure 4: EarthArena evaluation pipeline. (a) Multi-turn execution and feedback. (b) Rubric-based evaluation.

4.2 Rubric-Based Evaluation

LLM-based evaluation supports criteria-driven scoring but remains susceptible to bias (Liu et al., 2023; Zheng et al., 2023). Our judges apply frozen task-specific rubrics. Each atomic criterion specifies a claim or artifact property, reference evidence, scoring rule, and numerical tolerance where relevant. Judges inspect answers, artifact content, and execution evidence, returning criterion-level decisions and rationales.

Answer criteria assess quantities and interpretations; artifact criteria assess fields, values, units, coverage, and consistency; process criteria assess evidence of data access and transformation. JSONL records document execution but do not establish scientific correctness by themselves. Alternative implementations receive credit when they satisfy the same scientific requirements. References define evaluation targets, not a compulsory tool sequence.

4.3 Metrics and Aggregation

For task t, normalized criterion weights w_{ti} and awarded ratios $r_{ti} \in [0, 1]$ define

$$S_t = 100 \sum_{i \in R_t} w_{ti} r_{ti}, \qquad \sum_{i \in R_t} w_{ti} = 1. \tag{1}$$

Case aggregation follows frozen task allocations (Appendix C). Continuous scores are distinct from success rates, which additionally require thresholds and submission gates.

The reported performance and cost columns are unweighted case-level macro-averages. The Rubric metric combines normalized answer, artifact, and process scores with weights 0.5, 0.3, and 0.2; tasks without a process branch use the answer score directly under the supplied evaluation rule. This task-level fallback precedes aggregation, so the reported Rubric column should not be reconstructed by weighting the displayed macro-averages.

Tokens, elapsed time, and interaction rounds measure execution cost separately from scientific credit. Precise judge inputs, normalization, partial-completion rules, and cost-accounting definitions remain required for reproduction (Appendix H).

Table 2: Performance and execution cost of six models across three agent harnesses on EarthArena.

Model	Tokens ($\times 10^4$)	Time (min)	Rounds	Artifact correctness	Process execution	Partial Completion	Score answer	Rubric
Codex harness								
DeepSeek-V4.1-Flash	202.1	11.0	24.6	**29.8**	**20.4**	31.8	36.1	34.6
GLM-5.3-Flash	60.9	19.2	10.9	17.4	12.2	18.0	22.4	21.4
DeepSeek-V4-Pro	139.7	15.3	35.4	28.2	18.0	**35.6**	**43.3**	**38.1**
GPT-6-Luna	22.3	16.9	10.0	19.9	12.8	27.0	32.9	27.2
Qwen3.6-35B-A3B	16.7	12.2	7.2	0.0	0.0	0.0	0.0	0.0
Qwen3.8-27B	51.5	18.0	5.1	5.1	0.7	8.4	10.7	10.7
Claude Code harness								
DeepSeek-V4.1-Flash	172.6	12.8	64.2	**23.0**	**14.3**	**35.2**	**42.9**	**36.5**
GLM-5.3-Flash	51.1	20.2	24.7	9.8	6.4	10.3	15.9	15.2
DeepSeek-V4-Pro	143.2	16.7	55.9	20.9	11.8	23.6	32.8	30.4
GPT-6-Luna	22.6	20.7	27.9	11.8	3.1	8.0	10.9	9.9
Qwen3.6-35B-A3B	50.6	17.2	26.7	6.2	1.0	9.0	11.4	10.8
Qwen3.8-27B	75.1	17.7	35.1	3.3	2.6	7.9	11.0	10.7
DeepSeek harness								
DeepSeek-V4.1-Flash	362.1	13.3	81.7	**30.8**	**22.3**	41.6	47.7	42.0
GLM-5.3-Flash	168.2	18.4	39.8	13.9	14.3	17.1	21.3	19.5
DeepSeek-V4-Pro	268.6	13.8	62.5	30.7	18.8	37.6	44.4	40.2
GPT-6-Luna	40.8	10.0	27.9	30.3	19.7	**42.1**	**51.1**	**44.3**
Qwen3.6-35B-A3B	98.3	16.7	52.2	9.7	6.1	9.5	12.0	12.0
Qwen3.8-27B	84.5	18.5	38.9	9.6	5.8	8.6	11.5	13.0

5 Experiments and Analysis

5.1 Experimental Setup

The evaluation covers all 30 cases and 101 distinct tasks in the benchmark. The corrected summary compares six models with Codex, Claude Code, and DeepSeek harnesses, giving 18 configurations. DeepSeek denotes a harness as well as part of two model names; these are separate identifiers. All 18 configurations use the same 30-case, 101-task benchmark.

5.2 Overall Performance

Table 2 reports corrected performance and cost metrics. GPT-6-Luna with DeepSeek has the highest Rubric score (44.3) and normalized answer score (51.1). These measure different aspects of awarded credit; neither is an end-to-end success percentage.

Model rankings depend on the harness. DeepSeek-V4-Pro leads Rubric under Codex (38.1), DeepSeek-Flash under Claude Code (36.5), and GPT-6-Luna under DeepSeek (44.3). Luna's Rubric score ranges from 9.9 to 44.3, a 34.4-point spread. These descriptive differences support reporting the complete model–harness configuration; controlled ablations are needed to isolate individual harness mechanisms.

5.3 Execution Cost and Reliability

Cost and scientific credit need to be read together. Under DeepSeek, Luna attains Rubric 44.3 with reported mean time 10.0 minutes and 40.8×10^4 tokens, compared with Flash's 42.0, 13.3 minutes, and 362.1×10^4 tokens. These are case-level macro-averages, not matched causal efficiency estimates. Provider accounting, cache treatment, and the unit of elapsed-time aggregation must be specified before converting token differences into monetary costs.

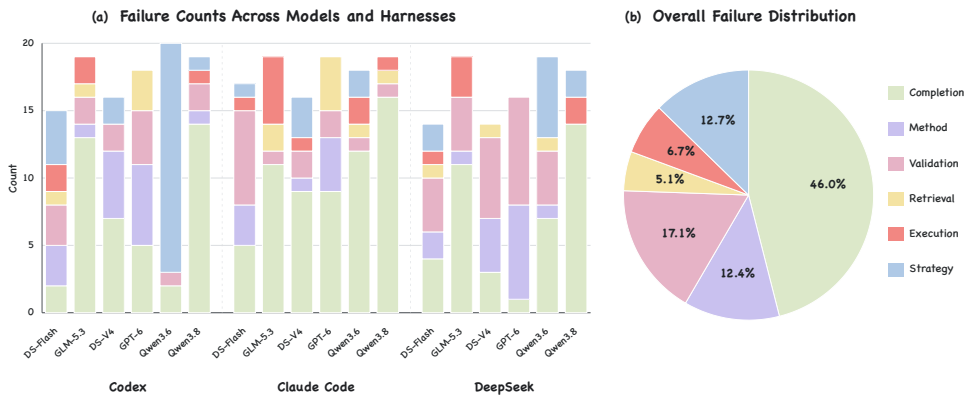

Figure 5: **Failure modes.** Primary failure counts by model–harness pair (a) and aggregate distribution after evaluator repair and integrity exclusion (b). Each run receives its earliest decisive cause. The six categories are mutually exclusive; completion dominates, followed by validation, strategy, scientific method, execution, and data retrieval.

Artifact and process evidence expose dimensions that answer metrics alone miss. Under DeepSeek, Flash leads artifact correctness (30.8) and process execution (22.3), while Luna leads the normalized answer branch (51.1) and Rubric (44.3). This difference motivates inspecting criterion-level decisions and the supporting files rather than identifying one answer metric with complete scientific execution.

5.4 Failure Modes and Data Retrieval

Completion is the dominant primary failure category (Figure 5), making sustained delivery of scientific outputs a central challenge. Retrieval errors include unavailable assets, connection failures, missing dependencies, and incompatible payloads (Appendix D, Figure 7). They require different responses: partitioning a large request may address transfer limits, but cannot recover a nonexistent resource. Harnesses should diagnose the failure before choosing a retry or an alternative source.

These aggregates motivate the case-grounded interventions below; they do not identify individual causes or dependency-depth effects. Domain breakdowns and judge–human agreement still require the records and checks in Appendix D.

6 Towards Reliable Earth Observation Agents

6.1 Case-Grounded Challenges and Harness Design

Extending ReAct's reasoning–action loop and Reflexion's feedback (Yao et al., 2023; Shinn et al., 2023), case requirements motivate three testable harness improvements, detailed in Appendix G.

Plan acquisition at the required scale. Shanghai and Longbasaba require heterogeneous, multi-temporal evidence. Estimate data volume, request spatial and temporal subsets, and partition large transfers. Resumable downloads and chunk manifests should preserve coverage and aggregation weights without double-counting; unavailable products require source discovery.

Keep scientific parameters explicit. Crete requires RETIRA strictly greater than 3; the inundation comparison uses an inclusive 0.3 m depth threshold. Persist units, grids, masks, baselines, quality filters, missing-data rules, and threshold operators. Resolve unspecified choices from permitted evidence, never private references.

Recover along artifact dependencies. Galapagos trends depend on annual SST grids; Longbasaba combines preceding measurements. Track artifact inputs and producing operations, validate coverage and schemas, invalidate affected descendants, and resume from the earliest inconsistent step.

Test these proposals with fixed models, tasks, and budgets, reporting scientific credit, acquisition completeness, parameter violations, downstream consistency, and cost; their gains remain to be measured.

6.2 Open-World Data Discovery for Geoscience

Dataset selection must match variables, region, period, resolution, processing level, and access conditions, supported by machine-actionable metadata and provenance (Wilkinson et al., 2016). EarthArena covers supplied files and named products requiring acquisition; the next step is discovery across catalogues from scientific objectives. Agents should justify sources, test samples before large downloads, and verify substitutions against the intended comparison rather than a preferred URL. Report evidence gaps when no compatible source exists. Pair replayable and live-access evaluation, recording source versions, retrieval dates, and service failures to distinguish analytical progress from changing data availability.

7 Limitations

EarthArena favors accessible inputs and reconstructable methods, underrepresenting field campaigns, proprietary imagery, and simulations with unavailable calibration or forcing data. Derived-product tasks evaluate their stated analysis, not reproduction of the underlying model. The evaluation covers the full benchmark across all configurations. Repeated-run uncertainty, domain breakdowns, and judge–human agreement remain incomplete. Aggregate scores cannot isolate reasoning, harness behavior, and infrastructure; LLM grading also requires independent auditing.

8 Conclusion

EarthArena evaluates real, long-horizon geoscience investigations through 30 paper-grounded cases and 101 linked tasks. Reproduced references and multidimensional rubrics connect scientific claims to data, artifacts, and execution evidence. Six-model results yield a best Rubric score of 44.3 and harness-dependent rankings. We call for scale-aware acquisition, persistent scientific parameters, dependency-aware recovery, and more open-world data discovery. The goal is measurable AI4Science progress: agents that sustain valid evidence from a research question to a scientifically supported conclusion.

Reproducibility Statement

The benchmark packages specify public queries, permitted inputs, output requirements, execution constraints, and reference-based evaluation criteria. Reproducing the result tables additionally requires the evaluated case manifest, exact model and harness identifiers, budgets, judge configuration, and the underlying task-level scores. Reference answers and rubrics must remain inaccessible to candidate agents during held-out evaluation.

Ethics Statement

The benchmark uses published research and public or appropriately licensed data. Redistribution must respect source licenses and access restrictions. Geospatial analyses can inform consequential environmental and infrastructure decisions, but a benchmark score does not establish readiness for autonomous hazard response or policy use. Reusing candidate outputs requires inspection of their data provenance, scientific scope, and supporting evidence.

AI Use Statement

The authors used large language models and other AI tools to aid and polish writing; support retrieval and discovery, including finding related work and source materials; assist research ideation and execution, including candidate-paper screening and reference execution as described in Section 3.2; and generate synthetic datasets used during benchmark development. These tools were not used to draft complete sections of the paper or to prove mathematical claims. The authors reviewed and verified all AI-assisted literature records, prose, code, synthetic data, and research analyses against the cited sources and executed artifacts, and take full responsibility for the paper's content, results, and conclusions.

References

M. Aun, K. Lakkala, R. Sanchez, E. Asmi, F. Nollas, O. Meinander, L. Sogacheva, V. De Bock, A. Arola, G. de Leeuw, V. Aaltonen, D. Bolsée, K. Cizkova, A. Mangold, L. Metelka, E. Jakobson, T. Svendby, D. Gillotay, and B. Van Opstal. Solar UV radiation measurements in marambio, antarctica, during years 2017–2019. *Atmospheric Chemistry and Physics*, 20:6037–6054, 2020. doi: 10.5194/acp-20-6037-2020. URL https://acp.copernicus.org/articles/20/6037/2020/.

A. Ayik et al. Preliminary wind resource assessment in south sudan using reanalysis data and statistical methods. *Renewable and Sustainable Energy Reviews*, 2021. doi: 10.1016/j.rser.2020.110621.

Rebecca R. Buchholz et al. New seasonal pattern of pollution emerges from changing north american wildfires. *Nature Communications*, 2022. doi: 10.1038/s41467-022-29623-8.

Ken O. Buesseler et al. High-resolution spatial and temporal measurements of particulate organic carbon flux using thorium-234 in the northeast pacific ocean during the EXport Processes in the Ocean from RemoTe Sensing field campaign. *Elementa: Science of the Anthropocene*, 2020. doi: 10.1525/elementa.030.

Jun Shern Chan et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024. doi: 10.48550/arXiv.2410.07095.

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, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Yosef Mengistu Darge et al. Detection of geothermal anomalies using Landsat 8 TIRS data in tulu moye geothermal prospect, main ethiopian rift. *International Journal of Applied Earth Observation and Geoinformation*, 2019. doi: 10.1016/j.jag.2018.08.027.

Xiang Deng et al. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1220.

D. J. du Preez, J. V. Ajtić, H. Bencherif, N. Bègue, J.-M. Cadet, and C. Y. Wright. Spring and summer time ozone and solar ultraviolet radiation variations over cape point, south africa. *Annales Geophysicae*, 37:129–141, 2019. doi: 10.5194/angeo-37-129-2019. URL https://angeo.copernicus.org/articles/37/129/2019/.

E. C. Ellis, K. Klein Goldewijk, S. Siebert, D. Lightman, and N. Ramankutty. Anthropogenic transformation of the biomes, 1700 to 2000. *Global Ecology and Biogeography*, 19:589–606, 2010. doi: 10.1111/j.1466-8238.2010.00540.x. URL https://doi.org/10.1111/j.1466-8238.2010.00540.x.

Peilei Fan et al. Recently constructed hydropower dams were associated with reduced economic production, population, and greenness in nearby areas. *Proceedings of the National Academy of Sciences*, 2022. doi: 10.1073/pnas.2108038119.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, and Weijia Li. Earth-Agent: Unlocking the full landscape of Earth observation with agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dkIXAbWuxO.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.11964.

Gamil Gamal et al. Remote sensing surveillance of NO2, SO2, CO, and AOD along the suez canal pre- and post-COVID-19 lockdown periods and during the blockage. *Sustainability*, 2023. doi: 10.3390/su15129362.

Nick Gorkavyi et al. Tracking aerosols and SO2 clouds from the raikoke eruption: 3d view from satellite observations. *Atmospheric Measurement Techniques*, 2021. doi: 10.5194/amt-14-7545-2021.

Aslak Grinsted, Christine S. Hvidberg, Néstor Campos, and Dorthe Dahl-Jensen. Periodic outburst floods from an ice-dammed lake in east greenland. *Scientific Reports*, 7:9966, 2017. doi: 10.1038/s41598-017-07960-9. URL https://www.nature.com/articles/s41598-017-07960-9.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023. doi: 10.48550/arXiv.2310.03302.

Roshan Jha, Arpita Mondal, Anjana Devanand, M. K. Roxy, and Subimal Ghosh. Limited influence of irrigation on pre-monsoon heat stress in the indo-gangetic plain. *Nature Communications*, 13:4275, 2022. doi: 10.1038/s41467-022-31962-5. URL https://www.nature.com/articles/s41467-022-31962-5.

Liam Kiely et al. Assessing costs of indonesian fires and the benefits of restoring peatland. *Nature Communications*, 2021. doi: 10.1038/s41467-021-27353-x.

Maria-Elissavet Koukouli et al. Sudden changes in nitrogen dioxide emissions over greece due to lockdown after the outbreak of COVID-19. *Atmospheric Chemistry and Physics*, 2021. doi: 10.5194/acp-21-1759-2021.

Katja Kowalski, Cornelius Senf, Patrick Hostert, and Dirk Pflugmacher. Characterizing spring phenology of temperate broadleaf forests using Landsat and Sentinel-2 time series. *International Journal of Applied Earth Observation and Geoinformation*, 92:102172, 2020. doi: 10.1016/j.jag.2020.102172. URL https://doi.org/10.1016/j.jag.2020.102172.

Baofu Li et al. Recent fall eurasian cooling linked to north pacific sea surface temperatures and a strengthening siberian high. *Nature Communications*, 2020. doi: 10.1038/s41467-020-19014-2.

Yuxiang Li et al. Green spaces provide substantial but unequal urban cooling globally. *Nature Communications*, 2024. doi: 10.1038/s41467-024-51355-0.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

F. McClean, R. Dawson, and C. Kilsby. Intercomparison of global reanalysis precipitation for flood risk modelling. *Hydrology and Earth System Sciences*, 27:331–347, 2023. doi: 10.5194/hess-27-331-2023. URL https://hess.copernicus.org/articles/27/331/2023/.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025. doi: 10.48550/arXiv.2503.00096. URL https://arxiv.org/abs/2503.00096.

Deependra Neupane et al. Solar and wind energy potential assessment at provincial level in nepal: Geospatial and economic analysis. *Renewable Energy*, 2022. doi: 10.1016/j.renene.2021.09.027.

Dat Tien Nguyen, Thao Nguyen, Fadillah Adamsyah Maani, Huy M. Le, Muhammad Umer Sheikh, Numan Saeed, Muhammad Haris Khan, and Salman Khan. TerraBench: Can agents reason over heterogeneous earth-system data? *arXiv preprint arXiv:2606.13148*, 2026. doi: 10.48550/arXiv.2606.13148. URL https://arxiv.org/abs/2606.13148.

Homero A. Paltan et al. Climate and sea surface trends in the galapagos islands. *Scientific Reports*, 2021. doi: 10.1038/s41598-021-93870-w.

Bolivar Paredes-Beltran, Alvaro Sordo-Ward, and Luis Garrote. Dataset of georeferenced dams in south america (DDSA). *Earth System Science Data*, 13:213–229, 2021. doi: 10.5194/essd-13-213-2021. URL https://essd.copernicus.org/articles/13/213/2021/.

Sofia Peleli et al. Satellite-observed thermal anomalies and deformation patterns associated to the 2021, central crete seismic sequence. *Remote Sensing*, 2022. doi: 10.3390/rs14143413.

Navin Ramankutty et al. Farming the planet: 1. geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 2008. doi: 10.1029/2007GB002952.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566:195–204, 2019. doi: 10.1038/s41586-019-0912-1. URL https://www.nature.com/articles/s41586-019-0912-1.

Benjamin Heikki Redmond Roche and Martin D. King. Quantifying the effects of background concentrations of crude oil pollution on sea ice albedo. *The Cryosphere*, 2022. doi: 10.5194/tc-16-3949-2022.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical – satellite data is a distinct modality in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42691–42706. PMLR, 2024. URL https://proceedings.mlr.press/v235/rolf24a.html.

David E. Rupp et al. Mapping an observation-based global solar irradiance climatology across the conterminous united states. *Journal of Applied Meteorology and Climatology*, 2022. doi: 10.1175/JAMC-D-21-0236.1.

Ju-Mee Ryoo, Leonhard Pfister, Rei Ueyama, Paquita Zuidema, Robert Wood, Ian Chang, and Jens Redemann. A meteorological overview of the ORACLES (ObseRvations of Aerosols above CLouds and their intEractionS) campaign over the southeastern atlantic during 2016–2018: Part 1 – climatology. *Atmospheric Chemistry and Physics*, 21:16689–16707, 2021. doi: 10.5194/acp-21-16689-2021. URL https://acp.copernicus.org/articles/21/16689/2021/.

Akashah Shabbir, Muhammad Akhtar Munir, Akshay Dudhane, Muhammad Umer Sheikh, Muhammad Haris Khan, Paolo Fraccaro, Juan Bernabe Moreno, Fahad Shahbaz Khan, and Salman Khan. ThinkGeo: Evaluating tool-augmented agents for remote sensing tasks. *arXiv preprint arXiv:2505.23752*, 2025. doi: 10.48550/arXiv.2505.23752. URL https://arxiv.org/abs/2505.23752.

Akashah Shabbir, Muhammad Umer Sheikh, Muhammad Akhtar Munir, Hiyam Debary, Mustansar Fiaz, Muhammad Zaigham Zaheer, Paolo Fraccaro, Fahad Shahbaz Khan, Muhammad Haris Khan, Xiao Xiang Zhu, and Salman Khan. OpenEarthAgent: A unified framework for tool-augmented geospatial agents. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2602.17665.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024. doi: 10.48550/arXiv.2409.11363.

Rakesh Kumar Singh et al. Satellite-derived photosynthetically available radiation at the coastal arctic seafloor. *Remote Sensing*, 2022. doi: 10.3390/rs14205180.

Giulio Starace et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025. doi: 10.48550/arXiv.2504.01848.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. doi: 10.1038/s41586-023-06221-2. URL https://www.nature.com/articles/s41586-023-06221-2.

Junjue Wang, Weihao Xuan, Heli Qi, Pengyu Dai, Kunyi Liu, Hongruixuan Chen, Zhuo Zheng, Junshi Xia, Stefano Ermon, and Naoto Yokoya. Can LLM agents respond to disasters? benchmarking heterogeneous geospatial reasoning in emergency operations. *arXiv preprint arXiv:2605.11633*, 2026. doi: 10.48550/arXiv.2605.11633. URL https://arxiv.org/abs/2605.11633.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Junfeng Wei et al. Longbasaba glacier recession and contribution to its proglacial lake volume between 1988 and 2018. *Journal of Glaciology*, 2021. doi: 10.1017/jog.2020.119.

Christine Wiedinmyer et al. The fire inventory from NCAR version 2.5: an updated global fire emissions model for climate and chemistry applications. *Geoscientific Model Development*, 2023. doi: 10.5194/gmd-16-3873-2023.

Hjalmar Wijk et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024. doi: 10.48550/arXiv.2411.15114.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618.

Y. Xu, L. Yu, W. Li, P. Ciais, Y. Cheng, and P. Gong. Annual oil palm plantation maps in malaysia and indonesia from 2001 to 2016. *Earth System Science Data*, 12:847–867, 2020. doi: 10.5194/essd-12-847-2020. URL https://essd.copernicus.org/articles/12/847/2020/.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.15793.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Minghan Yu, Youran Sun, Chugang Yi, Yixin Wen, and Haizhao Yang. Bringing agentic search to Earth observation data discovery. *arXiv preprint arXiv:2607.02387*, 2026. doi: 10.48550/arXiv.2607.02387. URL https://arxiv.org/abs/2607.02387.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Benchmark Construction Details

The construction record links each source paper to a bounded case objective, required data, target measurements, and a task dependency graph. Reference execution produces answers, artifacts, code, and execution records. Human review checks that the task scope matches the evidence and that the reference follows the computation being evaluated.

A task can legitimately start from a released derived product if its question concerns that product. For example, the sea-ice case compares archived albedo scenarios, and the flood case compares modeled building depths. Neither establishes that an agent reproduced the underlying simulation. This distinction is retained in the query and evaluation criteria. Conversely, a task that requires producing an intermediate measurement must not award execution credit for copying that measurement from the source paper.

Review covers the following properties:

- the scientific question and required outputs match the case objective;
- source data, release versions, access conditions, and spatial and temporal scope are recoverable;
- units, masks, thresholds, denominators, and missing-data conventions are specified when they affect correctness;
- reference computations have inspectable inputs and outputs;
- downstream artifacts preserve their upstream meaning;
- rubric criteria refer to available evidence and use justified numerical tolerances;
- public packages do not expose private reference answers or rubrics.

B Selected Source Papers

Table 3 lists the source papers for all 30 cases in the benchmark. Their frozen rubrics contain 1726 atomic scoring leaves in total.

Table 3: Source papers for the complete 30-case benchmark. *Source* gives the journal name and quartile. *Topic* assigns each case a primary top-level NASA GCMD Earth Science topic according to its scientific focus. *Nodes* counts atomic scoring leaves across all tasks in the available rubric version, excluding parent nodes and repeated tree references.

Paper	Source	Topic	Nodes
Climate and sea surface trends in the Galapagos Islands (Paltan et al., 2021)	Scientific Reports(Q1)	Oceans	62
Green spaces provide substantial but unequal urban cooling globally (Li et al., 2024)	Nature Communications(Q1)	Land surface	46
Recent fall Eurasian cooling linked to North Pacific sea surface temperatures and a strengthening Siberian high (Li et al., 2020)	Nature Communications(Q1)	Atmosphere	37
Preliminary wind resource assessment in South Sudan using reanalysis data and statistical methods (Ayik et al., 2021)	Renewable and Sustainable Energy Reviews(Q1)	Atmosphere	44
New seasonal pattern of pollution emerges from changing North American wildfires (Buchholz et al., 2022)	Nature Communications(Q1)	Atmosphere	21
Satellite-Derived Photosynthetically Available Radiation at the Coastal Arctic Seafloor (Singh et al., 2022)	Remote Sensing(Q1)	Oceans	8

Continued on next page

Table 3 – continued from previous page

Paper	Source	Topic	Nodes
The Fire Inventory from NCAR version 2.5: an updated global fire emissions model for climate and chemistry applications (Wiedinmyer et al., 2023)	Geoscientific Model Development(Q1)	Atmosphere	55
Mapping an Observation-Based Global Solar Irradiance Climatology across the Conterminous United States (Rupp et al., 2022)	Journal of Applied Meteorology and Climatology(Q2)	Atmosphere	32
Satellite-Observed Thermal Anomalies and Deformation Patterns Associated to the 2021, Central Crete Seismic Sequence (Peleli et al., 2022)	Remote Sensing(Q1)	Solid Earth	31
Tracking aerosols and SO_2 clouds from the Raikoke eruption: 3D view from satellite observations (Gorkavyi et al., 2021)	Atmospheric Measurement Techniques(Q1)	Atmosphere	44
Quantifying the effects of background concentrations of crude oil pollution on sea ice albedo (Roche & King, 2022)	The Cryosphere(Q1)	Cryosphere	39
Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000 (Ramankutty et al., 2008)	Global Biogeochemical Cycles(Q1)	Agriculture	92
Longbasaba Glacier recession and contribution to its proglacial lake volume between 1988 and 2018 (Wei et al., 2021)	Journal of Glaciology(Q1)	Cryosphere	115
High-resolution spatial and temporal measurements of particulate organic carbon flux using thorium-234 in the northeast Pacific Ocean during the EXport Processes in the Ocean from RemoTe Sensing field campaign (Buesseler et al., 2020)	Elementa: Science of the Anthropocene(Q1)	Oceans	45
Sudden changes in nitrogen dioxide emissions over Greece due to lockdown after the outbreak of COVID-19 (Koukouli et al., 2021)	Atmospheric Chemistry and Physics(Q1)	Atmosphere	40
Remote Sensing Surveillance of NO_2, SO_2, CO, and AOD along the Suez Canal Pre- and Post-COVID-19 Lockdown Periods and during the Blockage (Gamal et al., 2023)	Sustainability(Q1)	Atmosphere	31
Assessing costs of Indonesian fires and the benefits of restoring peatland (Kiely et al., 2021)	Nature Communications(Q1)	Atmosphere	22
Solar and wind energy potential assessment at provincial level in Nepal: Geospatial and economic analysis (Neupane et al., 2022)	Renewable Energy(Q1)	Atmosphere	24
Recently constructed hydropower dams were associated with reduced economic production, population, and greenness in nearby areas (Fan et al., 2022)	Proceedings of the National Academy of Sciences(Q1)	Biosphere	24
Detection of geothermal anomalies using Landsat 8 TIRS data in Tulu Moye geothermal prospect, Main Ethiopian Rift (Darge et al., 2019)	International Journal of Applied Earth Observation and Geoinformation(Q1)	Solid Earth	30
Dataset of Georeferenced Dams in South America (DDSA) (Paredes-Beltran et al., 2021)	Earth System Science Data(Q1)	Terrestrial hydrosphere	27
Characterizing spring phenology of temperate broadleaf forests using Landsat and Sentinel-2 time series (Kowalski et al., 2020)	International Journal of Applied Earth Observation and Geoinformation(Q1)	Biosphere	162
Annual oil palm plantation maps in Malaysia and Indonesia from 2001 to 2016 (Xu et al., 2020)	Earth System Science Data(Q1)	Agriculture	48
Periodic outburst floods from an ice-dammed lake in East Greenland (Grinsted et al., 2017)	Scientific Reports(Q1)	Cryosphere	101

Continued on next page

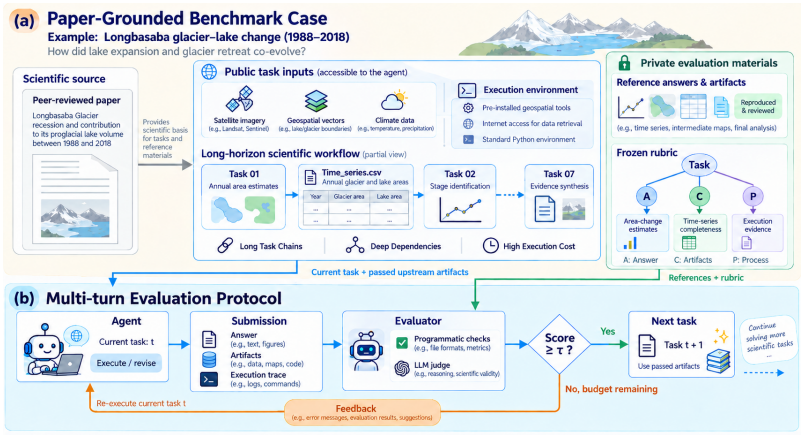

Figure 6: EarthArena case structure and evaluation. Public queries and permitted inputs define the objective; candidate artifacts connect tasks, and private references support rubric-based assessment.

Table 3 – continued from previous page

Paper	Source	Topic	Nodes
Intercomparison of global reanalysis precipitation for flood risk modelling (McClean et al., 2023)	Hydrology and Earth System Sciences(Q1)	Terrestrial hydrosphere	84
Anthropogenic transformation of the biomes, 1700 to 2000 (Ellis et al., 2010)	Global Ecology and Biogeography(Q1)	Biosphere	65
Solar UV radiation measurements in Marambio, Antarctica, during years 2017–2019 (Aun et al., 2020)	Atmospheric Chemistry and Physics(Q1)	Atmosphere	149
Limited influence of irrigation on pre-monsoon heat stress in the Indo-Gangetic Plain (Jha et al., 2022)	Nature Communications(Q1)	Atmosphere	55
A meteorological overview of the ORACLES campaign over the southeastern Atlantic during 2016–2018: Part 1 – Climatology (Ryoo et al., 2021)	Atmospheric Chemistry and Physics(Q1)	Atmosphere	118
Spring and summer time ozone and solar ultraviolet radiation variations over Cape Point, South Africa (du Preez et al., 2019)	Annales Geophysicae(Q2)	Atmosphere	75

C Evaluation Rubrics

Rubrics define criterion identifiers, score allocations, reference evidence, and the required properties of answers, artifacts, and execution records. Answer criteria are frozen before candidate outputs are inspected. For each task, the answer judge receives the task query, the frozen Answer branch, the private gold `answer.md`, and the candidate `answer.md`. It returns a criterion identifier, a cited candidate span, any extracted value and unit, a satisfaction decision, and a short rationale. The judge does not supply claims omitted by the candidate. Numerical comparisons and score aggregation are performed by the evaluator using the frozen tolerances and weights.

When a leaf contains several independent checks, its completion ratio is the frozen weighted mean of those checks, allowing partial credit without changing the leaf budget. Answer, artifact, and process dimensions are then aggregated with task-level weights 0.5, 0.3, and 0.2 when all three are applicable. Execution records are assessed alongside the files and results they reference.

For a case c with tasks T_c, the general case score is

$$S_c = \sum_{t \in T_c} b_{ct} S_t, \qquad \sum_{t \in T_c} b_{ct} = 1, \tag{2}$$

where allocations are fixed in the case rubric. Table 2 reports answer, artifact, and process metrics aggregated over cases.

Table 4: Answer-criterion decision rules. Wording may differ from the gold answer when the scientific meaning is equivalent.

Criterion type	Frozen decision rule
Numerical	First align the object, condition, metric, and unit. After any predeclared unit conversion, compare the candidate value with the gold value using the frozen absolute tolerance, or $\lvert x - y \rvert \leq \max(\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \lvert y \rvert)$ when both tolerances apply. A numerically close value assigned to the wrong scientific quantity receives no credit.
Directional or categorical	Check the named objects, metric, direction, ordering, conditions, and negation. Scientifically equivalent statements receive the same credit; merely listing the objects does not establish the required relationship.
Interpretive and scope	Check the required scientific proposition and its permitted scope. Paraphrases are accepted, but unsupported causal claims, overgeneralization, or a conclusion that contradicts the frozen criterion loses credit.
Missing or conflicting	An omitted requirement receives no credit. If the answer gives incompatible conclusions without an explicit correction, the judge cannot select only the favorable statement for full credit.

D Additional Experimental Results

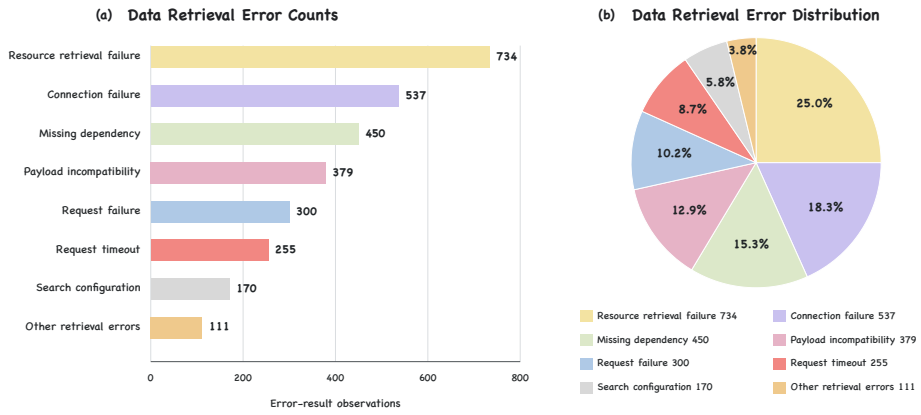

Figure 7: **Data retrieval failures.** Counts (a) and distribution across eight categories (b) of retrieval-error observations, retaining repeats within runs. Resource retrieval failure combines unavailable or obsolete assets, version-path failures, and invalid asset URLs. Authentication and permission events are excluded.

D.1 Performance across Domains and Task Complexity

Domain and complexity analysis requires linking scores to online data access, source count, artifact size, parameter choices, and dependency position. Missing upstream inputs, incorrect inherited measurements, and newly introduced analysis errors should be distinguished. The available aggregates do not establish domain rankings or dependency-depth effects; Appendix D specifies the required records.

D.2 Evaluation Reliability

Reference validation and candidate-score validation are separate stages. Candidate-level judgments were reviewed by human evaluators against the frozen rubric, private gold, and the evidence cited from each submission. The review protocol covers numerical and interpretive criteria, scientifically valid alternative formulations, and incomplete execution evidence. The manuscript reports this candidate-score review separately from the expert review used to approve each case.

Task-level records are needed to supplement the aggregate tables with per-topic results, dependency-stage comparisons, and execution-cost measurements. Every subgroup should report its case and task

counts. Uncertainty estimates should respect within-case dependencies rather than treat correlated downstream tasks as independent samples.

Failure analysis should distinguish data-selection errors, access failures, incomplete acquisition, invalid scientific parameter choices, upstream artifact failures, downstream processing errors, unsupported interpretation, and execution-time-limit exceedances. A category should be assigned from inspectable evidence, not inferred from a low final score. For runs that exceeded the prescribed execution-time limit, stage-level timing would help separate network, computation, retries, and inference.

A matched harness study can test the proposals in Section 6 by changing one component under the same model, task subset, and budgets. Candidate interventions include partitioned acquisition with resumable transfers, persistent scientific parameter specifications, artifact validation, and dependency-aware recovery. These are proposed experiments; no performance gains from these interventions are reported here.

The candidate-score audit is recorded separately from reference validation. Reviewers inspect the criterion definition, cited candidate evidence, numerical extraction and tolerance where applicable, and the assigned completion ratio. For each reviewed item, the audit record preserves the reviewer decision and any correction when one is made. Sample size and agreement statistics should accompany the released evaluation record.

D.3 Human Candidate-Answer Review

We conducted a human review of candidate-answer scoring on a fixed sample of one submitted task from each of 10 cases. The sample contained 10 answers, 54 answer-branch (A) scoring leaves, and 143 adapter checks. Case, task, attempt, candidate answer, and rubric versions were recorded before review. The candidate submissions and previously reported benchmark results were not edited during this process.

Three human reviewers received the same task statement, public instructions, reference answer, frozen rubric, and scoring manual. The automatic score, candidate model identity, and other reviewers' judgments were hidden during independent review. Each reviewer scored every check, cited candidate evidence, and separately recorded problems in the rubric or supporting evidence. Numerical items were aligned by scientific object, condition, and unit before applying the stated tolerance. Categorical, directional, and ranking items were assessed by the stated relationship, accepting semantically equivalent expressions. Interpretive items were assessed by the conclusion, supporting evidence, and scope rather than exact wording; contradictory, out-of-scope, or unsupported claims were handled by the corresponding criterion. Items that could not be resolved from the frozen materials were recorded as unresolved rather than assigned zero.

The initial review identified disagreements about rubric interpretation and cases in which the public task, reference implementation, tolerance, or scoring adapter did not provide a sufficiently clear decision rule. We revised the affected rubric definitions, retained the pre-revision records, and froze the revised version before re-reviewing the affected items. The first evidence packets also omitted some public instructions referenced by the tasks. We retained that round, added the missing public materials, and asked the same reviewers to repeat their independent checks without seeing the automatic scores or one another's judgments. The table reports the review results after the rubric and evidence package were frozen. All 143 check scores from each reviewer were unchanged after the public materials were added, while the number of leaves carrying at least one rule or evidence flag increased from 26 to 38.

Table 5: Agreement in the human candidate-answer review after the revised rubric and evidence package were frozen.

Measure	Result
A leaves scoreable by all three reviewers	54/54
Exact three-way agreement on A-leaf scores	54/54 (100%)
Exact three-way agreement on check scores	143/143 (100%)
Three-category Fleiss' κ on A leaves	1.000
Automatic–human exact A-leaf agreement	46/54 (85.19%)
Leaves with automatic–human score differences	8
Leaves with at least one rule/evidence flag	38/54

The three-category κ groups scores into zero, partial, and full credit and is reported together with exact agreement. Agreement does not by itself establish that a scoring rule is scientifically valid. The rule and evidence flags were therefore retained even when all three reviewers assigned the same score.

The review exposed four recurrent issues. First, frozen numerical tolerances can be narrower than the reporting precision. In one example, both the candidate answer and the textual reference reported cropland area as 14.950 million km^2, while the numeric checker compared it with 14.949630906937461 using a default tolerance of 10^{-6} and assigned zero. Second, reference-implementation choices can become hidden requirements: a 120 m export step and one treatment of Oceania records were not specified as the only permissible methods in the public tasks. Third, missing spatial boundaries can make candidate and reference statistics incomparable when they use different grid extents. Fourth, semantic equivalence in interpretive answers still requires domain judgment. The eight automatic–human score differences were retained for expert adjudication rather than treated automatically as scoring errors.

This review covered only the A branch for one task in each sampled case; it did not review artifact (C) or process (P) branches, all tasks in those cases, or complete case scores. Rubric changes are versioned, and all submissions affected by a changed rule are evaluated under the same frozen revision.

D.4 Checklist-based Expert Review

After reference execution, three experts use a shared checklist to review `case_input`, provided to candidate agents, and the private `case_gold` reference materials. Cases that fail review are rewritten, revised, and reviewed again by the same experts. After these materials pass review, the rubric and evaluator are constructed. The same three experts then review the final `case_rubric.json` and `evaluate.py`, including the atomic criteria, gold references, weights, tolerances, submission gates, partial-credit anchors, and score aggregation implemented by the evaluator. Final case acceptance requires unanimous approval of the input, gold, rubric, and evaluator from all three experts.

A separate difficulty screen uses two weaker models: a case is excluded from the final benchmark if both models pass it. Expert review assesses scientific quality, whereas this screen assesses task difficulty; neither replaces the other.

Tables 6 and 7 specify the review criteria; Table 8 specifies the supporting evidence records. The Gold–paper comparison rules below govern reference acceptance during benchmark construction. They do not automatically define scoring tolerances for candidate submissions.

Table 6: Case Input review checklist. Three experts inspect the public query, supplementary inputs, access restrictions, and deliverable requirements.

ID	Dimension	Review criterion
I01	Query clarity	The query specifies the study subject, period, and geographic extent.
I02	Query clarity	The query and subtasks specify the quantities to compute or analyze and the questions to answer.

Continued on next page

Table 6: Case Input review checklist (continued).

ID	Dimension	Review criterion
I03	Query clarity	Definitions and statistical conventions that affect the answer are explicit and understandable without consulting the source paper.
I04	Query clarity	The task can be completed using the query, attachments, and permitted public resources.
I05	Answer leakage	The query and attachments do not directly supply the paper results or Gold answers that agents must compute.
I06	Attachments	If additional input files are provided, the query identifies each file and its purpose.
I07	Attachments	All required attachments are present and readable.
I08	Blacklist	The source paper that agents must not access is explicitly identified by title or DOI.
I09	Blacklist	The restriction covers the source paper itself; it does not automatically extend to supplementary materials, public data, or data documentation.
I10	Blacklist	Completing the task does not require accessing the prohibited source paper.
I11	Tasks and outputs	The case contains clearly defined subtasks.
I12	Tasks and outputs	Each subtask specifies its required deliverables.

For directly comparable numerical quantities with nonzero paper results, we use relative error:

$$\text{Relative Error} = \frac{|\text{Gold result} - \text{Paper result}|}{|\text{Paper result}|} \times 100\%. \tag{3}$$

Zero or near-zero results, and quantities for which relative error is inappropriate, use predetermined alternative comparison criteria as specified in Table 7.

Table 7: Case Gold review checklist. Review covers independent computation, source-paper evidence, numerical discrepancies, scientific comparability, execution evidence, and deliverables.

ID	Dimension	Review criterion
G01	Independent computation	Data come from publicly accessible sources, and the specific sources are recorded.
G02	Independent computation	Intermediate data, results, and figures from the paper are not used directly as answers to questions requiring computation.
G03	Independent computation	When using an author-released data package, distinguish raw inputs, intermediate artifacts, and final results; verify that precomputed answers are not used directly.
G04	Independent computation	The key Gold results are recomputed by executing code.
G05	Data alignment	Compare the paper and Gold datasets, versions, variables, temporal and spatial coverage, spatiotemporal resolution, and processing levels item by item, documenting each difference. Cases using different data are labeled public-data projections rather than direct reproductions of the paper.
G06	Method evidence	For each major step affecting the result, provide the source-paper location and the corresponding code and data locations.
G07	Method evidence	Key metrics, thresholds, spatial partitions, baseline periods, and statistical rules have explicit supporting evidence, and their required inputs are available.
G08	Method evidence	Any alternative method is accompanied by its rationale, justification, and implications for the result.
G09	Result evidence	For every comparable subtask and final result, record the Gold result, paper result, comparison basis, and relative error.

Continued on next page

Table 7: Case Gold review checklist (continued).

ID	Dimension	Review criterion
G10	Result evidence	For zero or near-zero paper values, or quantities such as categories, directions, and spatial overlap, use an appropriate alternative rather than forcing a relative-error calculation. Alternatives include reported uncertainty, absolute error, spatial overlap, categorical agreement, or a discipline-specific criterion. Verify that the criterion was documented in the execution plan before the first run.
G11	Result evidence	Core numerical results with relative error at most 10% pass numerical reproduction, with the direction of change and physical meaning also checked.
G12	Result evidence	For relative error above 10% and at most 20%, verify that the methodological details have been reviewed once. If no omission or error is found, the result may be retained as an acceptable discrepancy, with the review, evidence, and residual difference documented.
G13	Result evidence	Relative error above 20% cannot be accepted directly. Revisit the paper's Methods, figure and table captions, supplementary materials, author-released code, and data-product documentation. Where supported by these sources, revise the execution plan and rerun affected tasks and their downstream dependents.
G14	Result evidence	A tolerance specified in advance for a particular metric, including one based on the paper's uncertainty, takes precedence. Do not relax it after observing the Gold–paper discrepancy.
G15	Result evidence	Methodological revisions must be grounded in the paper, attachments, author-released code, or public data-product documentation. Do not tune parameters to match the paper's result or replace computed Gold values with reported values.
G16	Result evidence	If relative error remains above 20% after the required review, retain the independently computed result and document the failure to reproduce and its possible causes. Successful code execution alone does not establish acceptance when core evidence fails the predetermined criteria.
G17	Result evidence	Every paper result has an exact locator, such as a section, PDF page, table, or figure number, and its corresponding source evidence is retained.
G18	Comparability	Gold and paper results refer to the same metric, with matching or correctly converted units.
G19	Comparability	Temporal coverage, spatial coverage, and statistical conventions match: for example, totals are compared with totals and annual means with annual means.
G20	Comparability	Directions of change, signs, and percentage baselines match. Numerical proximity alone does not establish acceptance.
G21	Execution and outputs	Gold includes executed code, commands, logs, inputs, and output records sufficient to verify that the results were produced by execution.
G22	Execution and outputs	Every required subtask deliverable is complete and can be opened.
G23	Execution and outputs	Key Gold results are traceable to their deliverables and the code that produced them.

Table 8: Supporting evidence records for expert review. Record individual data comparisons, method steps, and subtask metrics rather than substituting an overall pass flag for item-level evidence.

Record type	Fields to record for each item
Data alignment	Comparison item; data used in the paper; data used in Gold; agreement and differences. Items cover dataset, version, variable, temporal and spatial coverage, spatiotemporal resolution, and processing level.
Method alignment	Method step; paper method and exact source location; Gold implementation and code location; differences and rationale.

Continued on next page

Table 8: Supporting evidence records for expert review (continued).

Record type	Fields to record for each item
Result acceptance	Subtask or metric; Gold result; paper result; units; comparison basis; relative or absolute error; predetermined tolerance; methodological review; source-paper evidence location; acceptance decision.

E Case Domain Distribution

Each case is assigned one primary top-level NASA GCMD Earth Science topic using the frozen topic labels in Table 3. Table 9 summarizes the coverage of all 30 cases. Percentages are rounded to one decimal and therefore may not sum exactly to 100%.

Table 9: Distribution of the 30 EarthArena cases across their primary Earth Science domains.

Domain	**Cases**	**Share (%)**
Atmosphere	14	46.7
Oceans	3	10.0
Cryosphere	3	10.0
Biosphere	3	10.0
Agriculture	2	6.7
Solid Earth	2	6.7
Terrestrial hydrosphere	2	6.7
Land surface	1	3.3
Total	30	100.0

F Representative Scientific Workflows

The Shanghai case combines satellite temperature and vegetation observations with population, elevation, land-cover, and built-environment data. Its harmonized pixel dataset supports local cooling estimates, city-level aggregation, and consistency checks. The Galapagos case uses annual sea-surface-temperature grids for spatial trends and regional synthesis. Both require downstream analyses to preserve the upstream data support.

The seven-task Longbasaba case connects glacier and lake time series, stage identification, terminus retreat, surface lowering, ice velocity, and climate variability before an integrated interpretation. The Raikoke case combines horizontal SO_2 tracking with a lidar profile to constrain plume location. The Crete case evaluates night-time land-surface-temperature anomalies and event-date overlap, without treating temporal coincidence as seismic causation. These workflows expose different scientific requirements within a shared executable evaluation interface.

G Case-Grounded Challenges and Harness Design

The reported scores and case requirements motivate closer study of execution reliability. The cases make several intervention targets concrete. The following examples are drawn from task requirements; they are proposed harness improvements, not measured explanations of individual failures in the reported runs.

Large-data access should be planned as part of the analysis. The Shanghai workflow combines products with different resolutions over a common urban boundary; the Longbasaba workflow combines multitemporal glacier measurements with additional climate and surface-change evidence. Downloading complete archives when only a bounded region or period is needed can consume the execution budget before analysis begins. A scale-aware harness could inspect metadata and estimate the required volume, request spatial or temporal subsets where supported, and partition remaining transfers. Resumable downloads and a persistent chunk manifest would allow progress to survive

interruptions. Partitioning must preserve coverage, avoid duplicated observations at boundaries, and use appropriate weights when merging summaries.

Scientific parameters should be persistent, inspectable state. For Crete, the task requires a fixed grid, a defined background period, quality-based scene retention, and a RETIRA threshold strictly greater than 3. For the inundation comparison, modeled building depth is counted with an inclusive 0.3 m threshold. A program may execute successfully while changing either rule. A harness could store units, masks, reference periods, missing-data rules, and threshold operators in a structured analysis specification, then check each transformation against it. Unspecified methodological choices should be resolved from permitted documentation and scientific reasoning, with sensitivity analysis when appropriate. Evaluation references should never be used to tune those choices.

Recovery should follow scientific dependencies. Galapagos trend estimates depend on the annual SST grids produced earlier, while Longbasaba's later interpretation combines evidence from several preceding tasks. A generic retry that silently changes an input product or time interval can invalidate downstream files. An artifact registry could track input identity, units, spatial support, and producing operations, invalidate affected descendants, and resume from the earliest inconsistent step. Small executable checks of coverage, dimensional consistency, and required schemas could detect such problems before an expensive analysis continues.

These proposals can be tested with matched harness interventions: hold the model, task subset, and resource allowance fixed, then change data-access planning, parameter checks, or recovery behavior. The relevant outcome is increased scientific credit or task success under the same budget, together with fewer unsupported or inconsistent artifacts. The current model–harness comparison motivates these experiments but does not substitute for them.

G.1 Open-World Data Discovery for Geoscience

Real geoscience investigations often start with an information need, not a prepared directory of suitable observations. Finding a dataset requires matching a physical variable, region, period, resolution, processing level, and access conditions. The next step is to establish that the selected release can support the intended comparison. A relevant dataset title is insufficient when the required variable is absent, the temporal record has gaps, or access is restricted.

EarthArena already contains tasks for which agents must obtain public inputs, but many contracts name products or constrain acceptable alternatives. We encourage extending evaluation along an explicit gradient of data-access difficulty: supplied files, named products requiring acquisition, and scientific objectives requiring discovery across catalogues. Reporting these settings separately would distinguish analytical competence from autonomous source selection.

An open-world extension should evaluate the resulting scientific evidence rather than reward matching one preferred URL. Alternative sources require checks of variable definitions, coverage, resolution, units, and suitability for the method. Successful adaptation to an unavailable source should document what changed and whether the original comparison remains valid. When the evidence cannot support the requested conclusion, a documented data gap is more useful scientifically than an unjustified substitute.

Increasing realism also requires controlling environmental variation. We propose pairing a stable, replayable evaluation with a live-access condition that records source versions, retrieval dates, and service failures. This would separate improvements in scientific analysis from changes in data availability. Together with dependency-aware harnesses, such evaluations can move AI4Science toward agents that acquire, maintain, and scrutinize the evidence needed for real geoscience investigations.

H Metric Definitions and Reporting Details

This appendix specifies the calculations underlying the reported metrics. For a fixed model–harness configuration, let c, t, and i index cases, tasks, and atomic rubric leaves. Branch scores lie in $[0, 1]$; performance columns are reported on a 0–100 scale. Task allocations b_{ct} follow the frozen rubric, as in Eq. 2.

Leaf and branch scores. Let q_{cti} be the awarded leaf ratio and $v_{cti} > 0$ its weight. The evaluator reads the first available ratio from `score`, `completion`, or `earned_points`/`allocated_points`, and clips it to $[0, 1]$. The weight is the positive allocated-point value, otherwise a positive explicit weight, otherwise one. For branch $B \in \{A, C, P\}$,

$$B_{ct} = \frac{\sum_{i \in \mathcal{L}_{ct}^B} v_{cti} q_{cti}}{\sum_{i \in \mathcal{L}_{ct}^B} v_{cti}}, \qquad q_{cti} = \min\{1, \max\{0, r_{cti}\}\}. \tag{4}$$

Here r_{cti} is the raw awarded ratio; A, C, and P denote answer, artifact correctness, and process execution, respectively. Artifact and process scores use content and execution-evidence leaves, respectively. Missing or empty evaluation outputs and submission-gated failures receive zero credit; a branch explicitly marked not applicable is excluded, not scored as zero.

Task-level Rubric and case aggregation. The answer-only fallback is applied before aggregation:

$$R_{ct} = \begin{cases} A_{ct}, & \text{if the frozen rubric has no process branch,} \\ 0.5A_{ct} + 0.3C_{ct} + 0.2P_{ct}, & \text{otherwise.} \end{cases} \tag{5}$$

For $X \in \{A, C, P, R\}$, let T_c^X contain the applicable tasks and $\mathcal{C}^X$ the cases with an applicable score. Retaining the frozen allocations gives

$$X_c = 100 \frac{\sum_{t \in T_c^X} b_{ct} X_{ct}}{\sum_{t \in T_c^X} b_{ct}}, \qquad \overline{X} = \frac{1}{|\mathcal{C}^X|} \sum_{c \in \mathcal{C}^X} X_c. \tag{6}$$

Equal task allocations reduce the first expression to a task mean. The displayed Score answer, Artifact correctness, Process execution, and Rubric columns correspond to $\overline{A}$, $\overline{C}$, $\overline{P}$, and $\overline{R}$. Different applicability sets and the task-level fallback prevent reconstructing Rubric from the displayed branch averages. All configurations cover 30 cases and 101 tasks; valid metric denominators are retained in the reporting manifest.

Partial completion. Unlike weighted branch scores, partial completion averages awarded ratios uniformly over the scoring leaves selected for each case:

$$\mathrm{PC}_c = \frac{100}{|\mathcal{L}_c|} \sum_{i \in \mathcal{L}_c} q_{ci}, \qquad \overline{\mathrm{PC}} = \frac{1}{|\mathcal{C}^{\mathrm{PC}}|} \sum_{c \in \mathcal{C}^{\mathrm{PC}}} \mathrm{PC}_c. \tag{7}$$

At task level, if at least 60% of expected frozen leaf identifiers match the evaluator output, the expected leaf set is used and unmatched leaves receive zero; otherwise the returned evaluator leaves are used. Missing, empty, or gated submissions receive zero on all expected leaves. Answer-only tasks contribute only answer leaves. Parent nodes and repeated tree references are not additional scoring leaves.

Execution cost. Costs include all actual attempts, including unsuccessful ones. For resource $K \in \{\text{tokens}, \text{seconds}, \text{rounds}\}$ and attempt set $\mathcal{A}_{ct}$,

$$K_{ct} = \sum_{a \in \mathcal{A}_{ct}} K_{cta}, \qquad K_c = \frac{1}{|T_c|} \sum_{t \in T_c} K_{ct}, \qquad \overline{K} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} K_c. \tag{8}$$

Displayed token counts are $\overline{K}_{\text{tokens}}/10^4$ and minutes are $\overline{K}_{\text{seconds}}/60$; rounds are unscaled. When token coverage is below 90%, missing counts are estimated using tokens per round from the same model–harness configuration, or the same model under another harness when that stream is unavailable. Unresolved token counts remain not available; time and rounds are not imputed this way. Token accounting (including cached or reasoning tokens), timing endpoints, and round boundaries must be retained with the execution records. Monetary costs are not reported.