EARTHDATABENCH: BENCHMARKING LLM AGENTS FOR LONG-HORIZON SCIENTIFIC DATA PRODUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transforming heterogeneous Earth data into reliable scientific products requires complex, labor-intensive processing, limiting the pace and scale of research. Although large language model (LLM) agents have demonstrated success in automating a range of scientific tasks, their ability to produce research-quality Earth-science data remains insufficiently assessed. We introduce EarthDataBench, a benchmark of 40 executable, paper-derived tasks spanning seven Earth-science topics and diverse data sources, processing mechanisms, and product types. Tasks require agents to complete long-horizon scientific workflows, integrating heterogeneous data and implementing domain-specific methods to produce research-quality datasets. Unlike existing QA benchmarks that rely on matching reference answers, our evaluation combines comparisons with expert products and validation against independent observations, enabling assessment of whether agent-generated products match or surpass expert-reference quality. Across nine LLMs and ten model–harness configurations, 52.0% of outcomes pass scientific-consistency checks, but only 19.3% reach expert-reference quality or higher and 0.8% outperform expert references on independent validation. Our empirical results suggest that reliable execution remains a key bottleneck in scientific data production, even when methods are prescribed. EarthDataBench provides a testbed for advancing research-ready data production in Earth science and assessing improvements beyond expert references.

1 INTRODUCTION

Earth science studies the atmosphere, oceans, land, ice, and living systems, and how they interact. Understanding these processes supports weather and climate prediction, disaster risk reduction, and the management of water, ecosystems, and mineral resources (NASA, 2026; U.S. Geological Survey, 2021). These activities depend on reliable datasets that describe the state and changes of the Earth. However, satellite measurements, field observations, reanalysis products, and model outputs often require substantial processing before they can be used together. Producing high-quality datasets requires researchers to clean and combine data, align measurements in space and time, and apply suitable scientific methods. Automating this work could expand the scale of research, provided that the resulting data remain scientifically reliable.

LLM agents offer a promising way to automate parts of this process by combining scientific reasoning, programming, and tool use. Codex and Claude Code have supported scientific software development (OpenAI, 2026d), while Claude Science brings analysis tools and computing resources into multi-step research workflows (Anthropic, 2026). Within Earth science, OpenEarthAgent and OpenEarth-Agent support geospatial analysis through tool use and tool creation (Shabbir et al., 2026; Zhao et al., 2026), while EarthLink combines planning and code execution with climate analysis, including an investigation of Atlantic Niño precursors (Guo et al., 2025). More directly, SciDataCopilot automates scientific data preparation and reports efficiency gains over manual workflows in selected cases, including polar meteorological data processing (Rao et al., 2026). These advances make systematic evaluation of scientific data-production quality increasingly important.

Many Earth-science benchmarks focus on question answering or short analytical tasks, leaving complete data-production workflows from published research underexplored (Xu et al., 2025; Wang et al., 2025; Feng et al., 2025). Moreover, reference-matching scores measure agreement with expert answers or products, but cannot establish improvements beyond them. This leaves a central

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.

question: can LLM agents produce research-quality Earth-science datasets that match or surpass expert-produced references?

We introduce EarthDataBench, a benchmark of 40 data-production tasks derived from published Earth-science workflows. Built through expert-guided human–AI collaboration, it covers seven scientific topics, five input source families with 18 subtypes, and ten processing mechanisms. Each task requires agents to transform supplied inputs into a complete scientific product, testing their ability to carry out the processing steps needed in real research.

To assess product quality, EarthDataBench combines expert references with independent validation data in every task. A staged, quantitative scoring scheme checks product usability, scientific consistency, and reference-level quality. By evaluating agent and expert products against the same independent observations, it also measures improvements beyond the expert reference under predefined validation criteria.

We evaluate nine LLMs in ten configurations using Codex and Claude Code. In current evaluation records, 52.0% of outcomes reach the scientific-check threshold, but only 19.3% reach the reported reference-level threshold. Method guidance does not consistently close this gap, highlighting the challenge of translating scientific methods into reliable data products.

Our contributions are threefold:

1. We introduce EarthDataBench, a benchmark of 40 paper-derived, long-horizon tasks across seven Earth-science topics for evaluating LLM agents' scientific data-production capabilities.
2. We develop a systematic evaluation protocol combining expert products and independent observations to assess whether agent-generated products match or surpass expert-reference quality.
3. We evaluate nine frontier LLMs using two agent harnesses, revealing limitations of general-purpose agents in scientific data production and identifying directions for improving their reliability.

2 Related Work

LLM and MLLM evaluation in Earth science. Text-based benchmarks probe domain knowledge and scientific reasoning through questions and research discussions, as in EarthSE, and geographic and spatiotemporal understanding, as in GeoBenchLLM (Xu et al., 2025; Rodrigues et al., 2026). Multimodal evaluations extend this focus to reasoning over weather observations and Earth imagery (Ma et al., 2024; Wang et al., 2025; Xu et al., 2026), as well as interpreting scientific figures and signals in MSEarth and the Earth-science tasks of Scientists' First Exam (Zhao et al., 2025; Zhou et al., 2025). These evaluations largely use question answering or related prediction tasks to assess knowledge, perception, and reasoning. They establish prerequisites for scientific work, but do not test whether models can turn source data into scientifically reliable datasets, the capability targeted by EarthDataBench.

Agent evaluation for Earth-science workflows. Agent benchmarks extend model evaluation to planning, tool execution, and result delivery. ThinkGeo, Earth-Bench, OpenEarthAgent, and Earth-Bench-Pro evaluate tool execution and the correctness of answers to geospatial questions (Shabbir et al., 2025; Feng et al., 2025; Shabbir et al., 2026; Lv et al., 2026). OpenEarth-Bench checks outputs against expert-verified references using metadata criteria and numerical tolerances (Zhao et al., 2026), while TerraBench scores numerical results by their deviation from reference values (Nguyen et al., 2026). Because expert-generated scientific products may contain errors (Merchant et al., 2017; Loew et al., 2017), comparison with them alone makes it difficult to determine whether an agent's product surpasses expert quality. EarthDataBench addresses this gap by combining scientific validity checks with evaluation of agent and expert products against the same independent observations. It also targets more demanding, paper-derived data-production tasks involving multi-stage scientific processing. Completing an individual task is estimated to require days to weeks of expert work, compared with the minutes-to-days estimates for existing benchmarks (Table 1), indicating a greater workload per task.

Scientific validation of Earth-science data products. Quality control and evaluation are essential to the scientific reliability of Earth-science data. The CF conventions, NASA's Data Product

Table 1: Earth-science benchmarks grouped by evaluation setting.

Benchmark	Domain	Task formulation	Effort (est.)	Exec.	Paper	Sci.	Ind.
Model evaluation							
EarthSE (2025)	ES	Text QA	mins	×	×	×	×
OmniEarth-Bench (2025)	ES	Multimodal QA	mins	×	×	×	×
Agent evaluation							
ThinkGeo (2025)	EO	Tool-based QA	mins–hrs	✓	×	×	×
Earth-Bench (2025)	EO	Tool-based QA	mins–hrs	✓	×	×	×
Earth-Bench-Pro (2026)	EO	Tool-based QA	mins–hrs	✓	×	×	×
OpenEarthAgent (2026)	EO/GIS	Tool-based QA	mins–hrs	✓	×	△	×
OpenEarth-Bench (2026)	EO	Workflow execution	hrs–days	✓	×	△	×
DORA (2026)	Disaster	Tool-based QA	mins–hrs	✓	×	△	×
TerraBench (2026)	ES	Tool-based QA	hrs–days	✓	△	△	×
EarthDataBench (ours)	ES	**Scientific data production**	**days–weeks**	✓	✓	✓	✓

Labels. ES: Earth sciences; EO: Earth observation. Scientific data production constructs reusable datasets from heterogeneous inputs under structural, physical, and statistical constraints. Exec.: tool/code execution; Paper: paper-derived workflows; Sci.: artifact-level scientific validity beyond reference matching; Ind.: agent and expert products tested on the same withheld observations. ✓: reported; △: partial or related support (numerical/spatial checks or scientific rubrics for Sci.); ×: not reported under these definitions. **Effort (est.):** coarse estimates of expert active time per task (mins: minutes; hrs: hours; days/weeks: working time), based on workflow scope and available runtime evidence.

Development Guide, QARTOD, and QA4EO provide complementary guidance on data semantics, plausibility checks, and traceable uncertainty information (Hassell et al., 2017; Ramapriyan et al., 2024; U.S. IOOS, 2020; Committee on Earth Observation Satellites, 2010; Merchant et al., 2017). Independent observational validation assesses accuracy against external measurements, accounting for measurement uncertainty and differences in spatial and temporal scales (Loew et al., 2017; Gruber et al., 2020; Muñoz-Sabater et al., 2021). These practices support EarthDataBench's product-structure and semantic checks, task-specific scientific checks, expert-product comparison, and evaluation of both agent and expert products against the same withheld observations. Improvements are defined on the specified validation data and metrics, subject to their uncertainty and coverage limits.

3 EarthDataBench: Tasks, Evaluation, and Construction

3.1 Benchmark Overview

EarthDataBench comprises 40 Earth-science data-production tasks derived from published research. The tasks span seven scientific topics based on the Topic level of NASA's GCMD Science Keywords (Global Change Master Directory (GCMD), 2026): Atmosphere, Biosphere, Land Surface, Cryosphere, Terrestrial Hydrosphere (Hydrology), Oceans, and Solid Earth (Figure 1a). Each task involves a long, multi-stage workflow that transforms heterogeneous Earth-science inputs (Figure 1b), through a combination of data-processing methods (Figure 1c), into a target scientific data product (Figure 1d).

Input labels describe scientific source provenance rather than file formats. The inputs span 18 subtypes across five source families: satellite remote sensing, in-situ observations, reanalysis products, model outputs, and laboratory data. Optical/multispectral data (23 tasks), atmospheric reanalysis (14), and land-station observations (11) are the most common subtypes. Half of the tasks use inputs from more than one source family, requiring agents to work with different types of data within the same workflow.

The tasks involve ten processing mechanisms, with 194 task–mechanism assignments supported by 162 concrete workflow steps. Each label is grounded in the full task prompt, public output contract, or referenced starter code. Quality control covers all 40 tasks, including input decoding, support masks, physical-range checks, and reproducibility audits. Multi-source fusion (29 tasks), aggregation

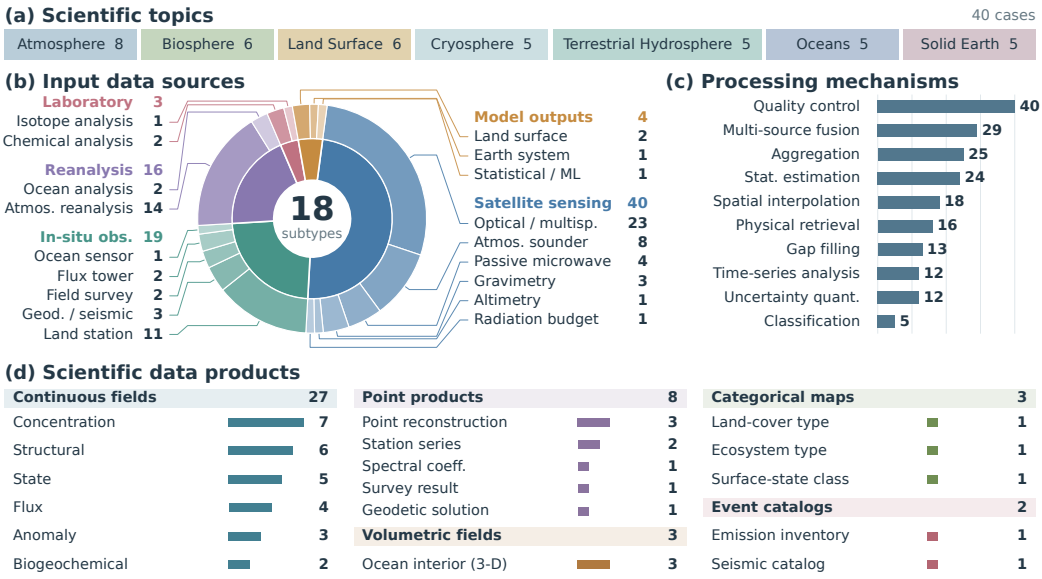

Figure 1: **Coverage of the 40 EarthDataBench tasks.** The panels show how tasks are distributed across (a) scientific topics, (b) input data sources, (c) processing mechanisms, and (d) target scientific products. Each task has one primary topic but may contribute to several categories in the other panels. Counts indicate tasks carrying each label, and source and product family totals sum their subtype counts. Supporting code, reports, and visualizations are excluded from panel (d).

(25), and statistical estimation (24) are the next most frequent. Labels are non-exclusive: each task can involve several mechanisms, and one workflow step can support several labels.

The target products cover five families and 17 subtypes. Continuous fields are the most common (27 tasks), including estimates of concentrations, physical states, and fluxes. Other targets include point products (8 tasks), categorical maps (3), volumetric fields (3), and event catalogs (2).

Sections 3.2, 3.3, and 3.4 describe the task formulation, scientific quality evaluation, and human–AI benchmark construction, respectively. Figure 2 provides an overview of these components. Appendix A illustrates the task interface and case-specific scoring with concrete task briefs for Atmosphere, Cryosphere, Biosphere, Hydrology, Land Surface, Oceans, and Solid Earth.

3.2 Task Formulation

For each task c, we provide input data, a task specification, and a controlled execution environment. The agent uses these resources to produce the requested scientific data product and supporting materials (Figure 2a). This process is written as

$$(\mathcal{O}_{c,r}, \tau_{c,r}) = \text{Agent}(\mathcal{A}_c, \mathcal{D}_c; \mathcal{R}_c). \tag{1}$$

Here, $\mathcal{A}_c$ denotes the input data, $\mathcal{D}_c$ the task specification, and $\mathcal{R}_c$ the execution environment. The outputs are the scientific product $\mathcal{O}_{c,r}$ and supporting materials $\tau_{c,r}$, where r indexes the agent run. Supporting materials include reproducible code and execution logs, with method reports and figures when required. Expert reference products and independent validation observations are withheld from the agent and used for the evaluation described in Section 3.3.

The task specification has three required components: a scientific production goal, a description of the supplied data, and output requirements. The goal defines the target variables and spatial and temporal coverage. The input description explains what datasets are available and how they are organized. Output requirements specify the files to submit, their formats and scientific metadata, and the supporting materials needed to reproduce the product. The specification may also provide methodological guidance, such as a recommended scientific procedure. The corresponding Base and Open variants are introduced in Section 4.1.

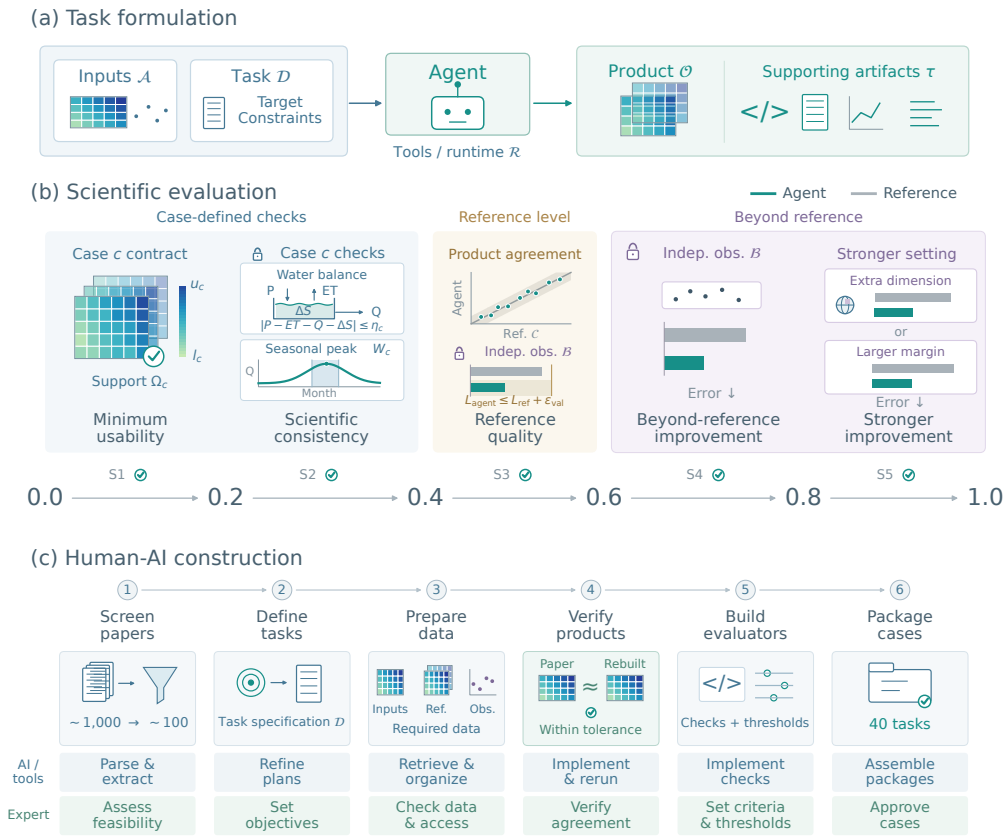

Figure 2: **Task definition, evaluation, and construction.** (a) The agent uses the supplied data, task instructions, and runtime environment to produce a scientific data product and supporting materials. (b) Evaluation progresses from minimum usability and scientific consistency to expert-reference quality, assessed using both product $\mathcal{C}$ and independent observations $\mathcal{B}$. Higher levels require improvement over the reference under predefined validation settings. Level 5 adds improvement on another validation dimension or a larger margin under the same setting. Water-budget and seasonal checks, error bars, and validation settings are illustrative. (c) Experts construct benchmark cases with AI assistance, following six stages from paper screening to packaging. Reconstructed products are scientifically verified before evaluator implementation. Criteria and validation settings are fixed before model comparison.

3.3 Scientific Quality Evaluation

We assess scientific data products and supporting materials using a cumulative five-level scoring scheme. The evaluation first checks whether a submission is usable and scientifically consistent, then determines whether its quality matches or exceeds an expert reference using independent observations (Table 2). For task c and agent run r, the score is

$$s_{c,r} = \mathcal{E}_c(\mathcal{O}_{c,r}, \tau_{c,r}; \mathcal{C}_c, \mathcal{B}_c) = 0.2\, k^\star_{c,r}. \tag{2}$$

Here, $\mathcal{E}_c$ is the task-specific evaluator, $\mathcal{O}_{c,r}$ is the submitted scientific product, and $\tau_{c,r}$ contains its supporting materials. The evaluator uses the expert reference product $\mathcal{C}_c$ and independent observations $\mathcal{B}_c$, which are withheld from the agent. The integer $k^\star_{c,r} \in \{0, \ldots, 5\}$ is the highest level for which all requirements up to that level are satisfied. A submission that fails the first level receives 0.0.

The five levels distinguish minimum usability (0.2), scientific consistency (0.4), expert-reference quality (0.6), improvement beyond the reference (0.8), and stronger improvement (1.0). The protocol requires task-specific criteria, thresholds, and the validation evidence needed for higher levels to be fixed before model comparison through the construction and calibration process in Section 3.4.

Table 2: Cumulative scientific-quality levels. Each level requires all preceding levels to pass.

Score	Level	Requirement
0.0	—	**No usable submission:** no submission, or failure to meet minimum usability requirements.
0.2	S1	**Minimum usability:** required supporting materials are complete, and the scientific product meets basic readability, structure, coverage, and numerical requirements.
0.4	S2	**Scientific consistency:** the product passes task-specific checks of scientific meaning, physical relationships, and coarse spatial or temporal patterns.
0.6	S3	**Expert-reference quality:** both agreement with the expert product and performance on independent observations meet predefined tolerances.
0.8	S4	**Beyond-reference improvement:** independent validation error is lower than the expert reference by at least a predefined positive margin.
1.0	S5	**Stronger improvement:** in addition to meeting Level 4, the product improves on an additional validation dimension or exceeds a larger improvement margin under the same setting.

Minimum usability (0.2) has two components. First, all supporting materials required by the task, such as code, logs, reports, and figures, must be present and readable. Second, the scientific product must contain the required variables and coordinates, provide interpretable units and missing-value information, and meet minimum spatial and temporal coverage and valid-data requirements. Empty or numerically degenerate outputs do not qualify. **Scientific consistency (0.4)** then checks whether the product satisfies the scientific relationships and broad patterns required by the task. For example, the hydrological illustrations in Figure 2b check whether the water-budget residual lies within a specified tolerance and whether the seasonal peak falls within an expected time window. These metrics and acceptance ranges depend on the task.

Expert-reference quality (0.6) requires agreement with the expert product and comparable-or-better performance against independent observations. The criteria are

$$\begin{aligned} d_c(\mathcal{O}, \mathcal{C}) &\le \epsilon_{\mathrm{ref},c}, \\ L_c(\mathcal{O}, \mathcal{B}) &\le L_c(\mathcal{C}, \mathcal{B}) + \epsilon_{\mathrm{val},c}. \end{aligned} \tag{3}$$

Here, d_c measures disagreement on the prescribed scientific quantities, while L_c measures validation error against the independent observations, with lower values indicating better performance. Both products use the same observations, spatial and temporal matching, and coverage rules. The nonnegative tolerances $\epsilon_{\mathrm{ref},c}$ and $\epsilon_{\mathrm{val},c}$ specify the allowed product discrepancy and validation-error gap, respectively. When a task specifies multiple required metrics, each must meet its own threshold. Improvement in one metric cannot compensate for failure in another.

Beyond expert-reference quality (0.8 and 1.0) is assessed under two predefined validation settings. Both use the same criterion: the agent's validation error must be lower than the reference by a setting-specific positive margin,

$$\begin{aligned} \Delta_{c,k} &= L_{c,k}(\mathcal{C}, \mathcal{B}_{c,k}) - L_{c,k}(\mathcal{O}, \mathcal{B}_{c,k}), \\ \Delta_{c,k} &\ge \delta_{c,k} > 0, \qquad k \in \{4, 5\}. \end{aligned} \tag{4}$$

Here, $k = 4$ and $k = 5$ correspond to the 0.8 and 1.0 levels, respectively. For each setting, $\mathcal{B}_{c,k}$ denotes the designated independent observations, $L_{c,k}$ the validation-error measure, $\Delta_{c,k}$ the improvement over the expert reference, and $\delta_{c,k}$ the required improvement margin. Level 4 evaluates improvement in the primary validation setting. Level 5 requires improvement on an additional validation dimension or a larger improvement under the same setting. In the latter case, the observations and error measure remain the same, with $\delta_{c,5} > \delta_{c,4}$. Validation settings and thresholds are fixed before model comparison.

3.4 Human–AI Collaborative Construction

We construct EarthDataBench through a six-stage human–AI workflow (Figure 2c). We first download approximately 1,000 papers published in *Earth System Science Data* (ESSD) over the past five years and parse all papers into a machine-readable format for AI-assisted analysis. We shortlist around 100 papers based on the availability of production and validation data and the feasibility of processing them at the required scale. Experts then define research objectives, which AI refines into

task plans. We prepare the required data, reconstruct and scientifically verify reference products, implement evaluators, and package the accepted cases for release. This process yields the 40 tasks in EarthDataBench.

Candidate paper screening. AI extracts each paper's production goal, data sources, methods, and validation approach. Experts assess scientific relevance, check the reported data-access conditions, and estimate the data volume and computational requirements. This screening identifies studies that are suitable for task construction.

Task definition. Experts select a scientific production goal from each candidate paper and define its spatial and temporal scope. AI develops a preliminary plan and drafts the task specification $\mathcal{D}$, including input descriptions, output requirements, and optional method guidance. Experts review the draft to ensure that the goal is clear and the proposed workflow is feasible.

Data preparation. AI and tools help retrieve and organize production inputs $\mathcal{A}$, available published reference products, and independent validation data $\mathcal{B}$. Experts check data quality, coverage, provenance, and permitted use. If essential data cannot be obtained or license restrictions prevent their use in the benchmark, we replace the candidate.

Scientific verification. AI assists with implementing and running the paper's method to reconstruct the target product. Experts examine the scientific assumptions and physical or statistical validity of the reconstruction, then compare it with the published product. A case passes only if the reconstructed product agrees with the published product within task-specific tolerances and repeated runs produce stable results. Accepted cases proceed to evaluator implementation.

Evaluator implementation. AI translates the task-specific scientific checks into the executable evaluator $\mathcal{E}$. Experts define the metrics and thresholds for all five scoring levels, covering minimum usability, scientific consistency, expert-reference quality, and the two levels of improvement beyond the reference. Reference runs, baseline products, and deliberately flawed outputs are used to check whether the evaluator accepts valid results and rejects failures at the intended level. The criteria and thresholds are fixed before model comparison.

Packaging and release. Each accepted case is packaged with public inputs, task instructions, and a runtime environment. Reference products, independent validation data, and evaluators are retained for private evaluation. Supporting code, provenance, and construction records are archived with a fixed case version.

4 Experiments and Results

4.1 Experimental Setup

Models and agent harnesses. We evaluate nine frontier LLMs: GPT-6-Astra (OpenAI, 2026b), GPT-6-Luna (OpenAI, 2026c), GPT-5.6-Sol (OpenAI, 2026a), GLM-5.3 (Z.ai, 2026a), GLM-5.3-Flash (Z.ai, 2026b), GLM-5-Turbo (Z.ai, 2026c), DeepSeek-v4-Flash (DeepSeek-AI, 2026), Qwen3.8-27B (Qwen Team, 2026), and Kimi-K2.6 (Moonshot AI, 2026). Each model is tested with Codex, and GLM-5.3-Flash is additionally tested with Claude Code, yielding ten model–harness configurations (Table 3). All configurations are evaluated on our 40 tasks. The main comparison uses the method-specified **Base** condition, which provides the scientific objective, input data, constraints, and a prescribed method. Section 4.2 examines the **Open** condition, where no specific method is prescribed.

Evaluation metrics. Table 3 summarizes scientific quality and resource use. **Score** is the mean task score multiplied by 100. We report the **scientific-consistency pass rate** ($s \geq 0.4$), the proportion **matching or exceeding expert-reference quality** ($s \geq 0.6$), and the proportion **exceeding expert-reference quality** ($s > 0.6$), following the scoring levels in Section 3.3. Resource metrics comprise total tool calls, total input and output tokens, and summed execution time across tasks.

4.2 Overall Data-Production Performance

Frontier agents remain far from expert-level data production. Table 3 shows that only 12.5–25.0% of tasks per configuration reach the expert-reference threshold. Across all 400 task–configuration outcomes, 52.0% pass scientific consistency, but only 19.3% reach expert-reference quality and 0.8% exceed it. Figure 3a shows that score 0.4 is the most frequent outcome. The largest drop between adjacent attainment thresholds occurs from 0.4 to 0.6. Thus, passing basic scientific checks frequently fails to translate into expert-quality products.

Table 3: Overall performance and resource usage in the Base setting. Bold denotes the highest result in each quality column. N/A indicates unavailable tool-call counts.

Harness	Model	Score (0–100)	Sci. cons. ≥ 0.4 (%)	Expert level ≥ 0.6 (%)	Beyond expert > 0.6 (%)	Tool calls (total)	Tokens (M)	Time (h)
Codex	GPT-6-Astra	33.5	**62.5**	20.0	**2.5**	828	38.94	9.43
Codex	GPT-6-Luna	29.0	47.5	17.5	**2.5**	758	39.13	8.50
Codex	GPT-5.6-Sol	30.0	45.0	20.0	0.0	699	38.86	8.07
Codex	GLM-5.3	**35.0**	60.0	22.5	0.0	2,213	271.85	29.02
Codex	GLM-5.3-Flash	30.0	52.5	17.5	0.0	2,103	242.93	35.03
Claude Code	GLM-5.3-Flash	32.0	55.0	**25.0**	**2.5**	N/A	302.54	65.73
Codex	GLM-5-Turbo	29.0	50.0	20.0	0.0	1,967	216.11	29.51
Codex	DeepSeek-v4-Flash	32.5	57.5	20.0	0.0	2,532	247.27	25.47
Codex	Qwen3.8-27B	25.5	45.0	12.5	0.0	2,746	199.75	48.03
Codex	Kimi-K2.6	26.0	45.0	17.5	0.0	5,875	583.25	29.82
Pooled (10 configurations)		30.3	52.0	19.3	0.8	—	—	—

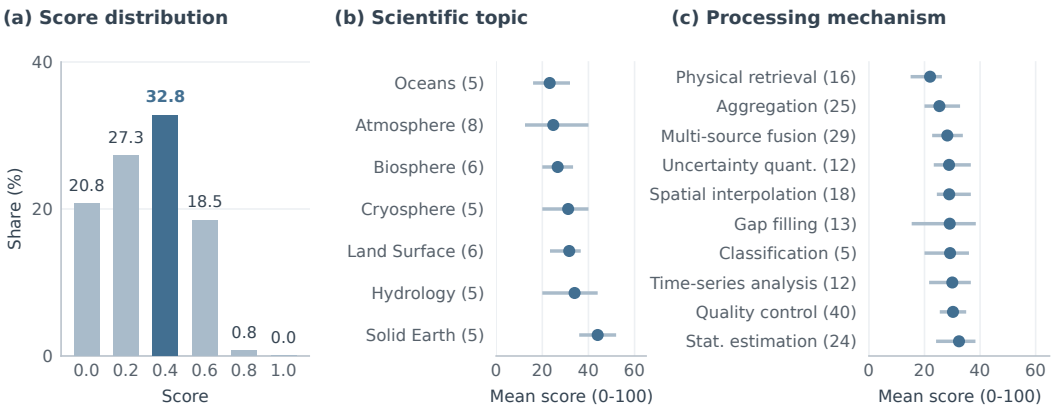

Figure 3: **Overall performance across task categories.** (a) Percentage at each score level across 400 Base outcomes from the same 40 tasks and ten configurations. (b) Scores by scientific topic. (c) Scores by processing mechanism. In (b,c), points average the ten configuration-specific category means and horizontal lines show their minimum–maximum range. Processing labels overlap because a task can involve multiple operations; quality control covers all 40 tasks and reproduces the overall mean.

Stronger general-purpose models also perform well on scientific tasks. Under Codex, GLM-5.3 and GPT-6-Astra lead in mean score (35.0 and 33.5) and scientific-consistency pass rate (60.0% and 62.5%). GLM-5.3 outperforms GLM-5.3-Flash (30.0) and GLM-5-Turbo (29.0), while GPT-6-Astra outperforms GPT-5.6-Sol (30.0) and GPT-6-Luna (29.0). These within-family comparisons suggest that stronger general reasoning and coding capabilities can benefit scientific data production.

The harness also affects expert-level attainment. With GLM-5.3-Flash, Claude Code achieves a mean score of 32.0 and expert-level attainment of 25.0%, compared with 30.0 and 17.5% under Codex. This 7.5-percentage-point difference in expert-level attainment highlights the relevance of the execution framework when evaluating scientific agents.

Table 4: Effect of method guidance.

Model	Scientific $s \geq 0.4$ (%)	Expert $s \geq 0.6$ (%)	Δ Score
GPT-6-Astra	47.5 → 62.5 **(+15.0)**	22.5 → 20.0 **(-2.5)**	+1.5
GPT-6-Luna	32.5 → 47.5 **(+15.0)**	17.5 → 17.5 **(0.0)**	+5.5
GLM-5.3	57.5 → 60.0 **(+2.5)**	25.0 → 22.5 **(-2.5)**	0.0
GLM-5.3-Flash	52.5 → 52.5 **(0.0)**	15.0 → 17.5 **(+2.5)**	-0.5
GLM-5-Turbo	45.0 → 50.0 **(+5.0)**	15.0 → 20.0 **(+5.0)**	+0.5
All five	47.0 → 54.5 **(+7.5)**	19.0 → 19.5 **(+0.5)**	+1.4

Open → Base (Base − Open).
Rate changes are in percentage points.

Method guidance mainly benefits scientific consistency. Across five Codex configurations on our 40 tasks, method guidance increases the scientific-consistency pass rate by about 8 percentage points, while expert-level attainment remains around 20% (Table 4). This suggests that guidance helps agents meet scientific requirements, but that expert-level production faces bottlenecks beyond access to a prescribed method. Tasks involving classification and aggregation show the largest observed mean-score gains, with aggregation improving all five models. One possible explanation is that explicit decision rules and aggregation procedures reduce implementation ambiguity, helping agents translate method descriptions into executable steps. Appendix B.2 provides the full breakdown.

4.3 Experimental Analysis

Most zero-score failures concern format compliance. Under Base, 20.8% of outcomes receive zero. Of these failures, 61.4% concern *format compliance*: required deliverables are missing, or submitted files and reports do not follow the specified structure. The remaining 38.6% are classified as failures of *scientific semantics*, including invalid values, insufficient coverage, or unmet minimum scientific requirements. Appendix C provides the detailed breakdown.

Scientific task execution remains a weakness. Physical-retrieval tasks receive lower scores than other task groups under Base (Figure 3c). Even among nonzero outcomes, only about 1% reach expert-level quality, compared with about 24% overall (Appendix B.1). This gap may reflect the difficulty of implementing complex, multi-step scientific workflows, which require agents to translate physical relationships into numerical calculations and apply data corrections consistently across stages. These results suggest that reliably executing scientific details remains a key challenge in turning prescribed methods into expert-quality data products.

GPT-series agents achieve the highest recorded token efficiency. The GPT configurations achieve scores around 30 using roughly 40M tokens, whereas Kimi-K2.6 scores lower despite consuming nearly 600M tokens (Table 3). GPT agents also make far fewer tool calls and have the shortest total execution times. This efficiency may partly reflect combining operations into fewer calls, reducing repeated context processing.

5 Conclusion

In this paper, we introduce EarthDataBench, a benchmark of 40 paper-derived, long-horizon data-production tasks spanning seven Earth-science topics. Its evaluation combines expert products and independent observations to assess whether agent-generated datasets match or surpass expert-reference quality. Evaluations of nine frontier LLMs using two harnesses reveal a persistent gap between passing scientific checks and producing expert-quality datasets, which methodological guidance alone does not substantially narrow. We hope EarthDataBench will serve as a standardized testbed for tracking progress, identifying capability gaps, and guiding the development of more reliable LLM agents for Earth-science data production.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the paper. We reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

EarthDataBench is intended to advance research on LLM agents for scientific data production in Earth science. We aim to evaluate whether these agents can produce scientific data products that match or surpass expert references, with quality assessed against independent observations. We plan to release the benchmark upon paper acceptance, subject to applicable data licenses and access restrictions. The authors declare no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We will release the full dataset and accompanying evaluation script upon paper acceptance to support reproducible evaluation. We hope that EarthDataBench will serve as a standardized testbed for tracking progress and guiding the development of LLM agents for scientific data production in Earth science.

REFERENCES

Anthropic. Claude Science, an AI workbench for scientists, is now available. https://www.anthropic.com/news/claude-science-ai-workbench, 2026. June 30, 2026; accessed September 25, 2026.

Committee on Earth Observation Satellites. A quality assurance framework for earth observation: Principles. Technical report, Committee on Earth Observation Satellites, 2010. URL https://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf. Version 4.0, 14 January 2010.

DeepSeek-AI. DeepSeek-V4-Flash model card. https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash, 2026. Accessed 2026-09-26.

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. arXiv:2509.23141, 2025. URL https://arxiv.org/abs/2509.23141.

Global Change Master Directory (GCMD). GCMD keywords, version 24.8. https://www.earthdata.nasa.gov/data/tools/gcmd-keyword-viewer, 2026. Accessed 2026-09-25.

A. Gruber, G. De Lannoy, C. Albergel, A. Al-Yaari, L. Brocca, J.-C. Calvet, A. Colliander, M. Cosh, W. Crow, W. Dorigo, et al. Validation practices for satellite soil moisture retrievals: What are (the) errors? *Remote Sensing of Environment*, 244:111806, 2020. doi: 10.1016/j.rse.2020.111806. URL https://doi.org/10.1016/j.rse.2020.111806.

Zijie Guo, Jiong Wang, Fenghua Ling, Wangxu Wei, Xiaoyu Yue, Zhe Jiang, Wanghan Xu, Jing-Jia Luo, Lijing Cheng, Yoo-Geun Ham, Fengfei Song, Pierre Gentine, Toshio Yamagata, Ben Fei, Wenlong Zhang, Xinyu Gu, Chao Li, Yaqiang Wang, Tao Chen, Wanli Ouyang, Bowen Zhou, and Lei Bai. A self-evolving AI agent system for climate science. arXiv:2507.17311, 2025. URL https://arxiv.org/abs/2507.17311.

David Hassell, Jonathan Gregory, Jon Blower, Bryan N. Lawrence, and Karl E. Taylor. A data model of the climate and forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). *Geoscientific Model Development*, 10:4619–4646, 2017. doi: 10.5194/gmd-10-4619-2017. URL https://gmd.copernicus.org/articles/10/4619/2017/.

Alexander Loew, William Bell, Luca Brocca, Claire E. Bulgin, Jörg Burdanowitz, Xavier Calbet, Reik V. Donner, Darren Ghent, et al. Validation practices for satellite-based earth observation data across communities. *Reviews of Geophysics*, 55, 2017. doi: 10.1002/2017RG000562. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017RG000562.

Zhutao Lv, Chenhao Dang, Yi Feng, Yanpei Gong, Xiaolei Wang, Junyan Ye, Conghui He, and Weijia Li. Earth-Agent-Pro: Towards real-world full-chain earth observation with agents. arXiv:2609.12533, 2026. URL https://arxiv.org/abs/2609.12533.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. WeatherQA: Can multimodal language models reason about severe weather? arXiv:2406.11217, 2024. URL https://arxiv.org/abs/2406.11217.

Christopher J. Merchant, Frank Paul, Thomas Popp, Michael Ablain, Sophie Bontemps, Pierre Defourny, Rainer Hollmann, Thomas Lavergne, Alexandra Laeng, Gerrit de Leeuw, Jonathan Mittaz, Caroline Poulsen, Adam C. Povey, Max Reuter, Shubha Sathyendranath, Stein Sandven, Viktoria F. Sofieva, and Wolfgang Wagner. Uncertainty information in climate data records from earth observation. *Earth System Science Data*, 9:511–527, 2017. doi: 10.5194/essd-9-511-2017. URL https://essd.copernicus.org/articles/9/511/2017/.

Moonshot AI. Kimi-K2.6 model card. https://huggingface.co/moonshotai/Kimi-K2.6, 2026. Accessed 2026-09-26.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, María Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13:4349–4383, 2021. doi: 10.5194/essd-13-4349-2021. URL https://essd.copernicus.org/articles/13/4349/2021/.

NASA. Earth system science research. https://science.nasa.gov/earth-science/research/, 2026. Updated September 24, 2026; accessed September 25, 2026.

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:2606.13148, 2026. URL https://arxiv.org/abs/2606.13148.

OpenAI. GPT-5.6 Sol model. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026a. Accessed 2026-09-26.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, 2026b. Accessed 2026-09-26.

OpenAI. GPT-6 Luna model. https://developers.openai.com/api/docs/models/gpt-6-luna, 2026c. Accessed 2026-09-26.

OpenAI. Scientific computing in the age of agentic AI. https://openai.com/index/scientific-computing-agentic-ai/, 2026d. July 28, 2026; accessed September 25, 2026.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026. Accessed 2026-09-26.

H. K. Ramapriyan, P. J. T. Leonard, E. M. Armstrong, S. J. S. Khalsa, D. K. Smith, L. F. Iredell, D. M. Wright, G. J. Huffman, and T. R. Walker. Data product development guide for data producers. Technical Report ESDS-RFC-041, NASA Earth Science Data and Information System Standards Office, 2024. URL https://doi.org/10.5067/DOC/ESCO/RFC-041VERSION2. Version 2.0.

Jiyong Rao, Yicheng Qiu, Jiahui Zhang, Juntao Deng, Shangquan Sun, Fenghua Ling, Hao Chen, Nanqing Dong, Zhangyang Gao, Siqi Sun, Yuqiang Li, Dongzhan Zhou, Guangyu Wang, Lijun Wu, Conghui He, Xuhong Wang, Jing Shao, Xiang Liu, Yu Zhu, Mianxin Liu, Qihao Zheng, Yinghui Zhang, Jiamin Wu, Xiaosong Wang, Shixiang Tang, Wenlong Zhang, Bo Zhang, Wanli Ouyang, Runkai Zhao, Chunfeng Song, Lei Bai, and Chi Zhang. SciDataCopilot: An agentic data preparation framework for AGI-driven scientific discovery. arXiv:2602.09132, 2026. URL https://arxiv.org/abs/2602.09132.

Rodrigo Ferreira Rodrigues, Karim Radouane, Jose G. Moreno, and Lynda Tamine. GeoBenchLLM: A comprehensive benchmark for evaluating LLMs on geo-related tasks. arXiv:2608.07411, 2026. URL https://arxiv.org/abs/2608.07411.

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:2505.23752, 2025. 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. arXiv:2602.17665, 2026. URL https://arxiv.org/abs/2602.17665.

U.S. Geological Survey. 120.1—creation, authority, mission, and functions. https://www.usgs.gov/survey-manual/1201-creation-authority-mission-and-functions, 2021. Survey Manual; issued January 15, 2021; accessed September 25, 2026.

U.S. IOOS. Manual for real-time quality control of in-situ temperature and salinity data. Technical report, U.S. Integrated Ocean Observing System, National Oceanic and Atmospheric Administration, 2020. URL https://cdn.ioos.noaa.gov/media/2020/03/QARTOD_TS_Manual_Update2_200324_final.pdf. Version 2.1.

Fengxiang Wang, Mingshuo Chen, Xuming He, Yi-Fan Zhang, Yueying Li, Feng Liu, Zijie Guo, Zhenghao Hu, Jiong Wang, Jingyi Xu, Zhangrui Li, Junchao Gong, Di Wang, Fenghua Ling, Ben Fei, Weijia Li, Long Lan, and Wenjing Yang. Omniearth-bench: Towards holistic evaluation of earth's six spheres and cross-spheres interactions with multimodal observational earth data. arXiv:2505.23522, 2025. URL https://arxiv.org/abs/2505.23522v3. Version 3, revised 2026.

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:2605.11633, 2026. URL https://arxiv.org/abs/2605.11633.

Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, Xiaoyu Yue, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. EarthSE: A benchmark for evaluating earth scientific exploration capability of LLMs. arXiv:2505.17139, 2025. URL https://arxiv.org/abs/2505.17139.

Zelin Xu, Yupu Zhang, Saugat Adhikari, Saiful Islam, Tingsong Xiao, Zibo Liu, Shigang Chen, Da Yan, and Zhe Jiang. EarthSpatialBench: Benchmarking spatial reasoning capabilities of multimodal LLMs on earth imagery. arXiv:2602.15918, 2026. URL https://arxiv.org/abs/2602.15918.

Z.ai. GLM-5.3 model card. https://huggingface.co/zai-org/GLM-5.3, 2026a. Accessed 2026-09-23.

Z.ai. GLM-5.3-Flash model card. https://huggingface.co/zai-org/GLM-5.3-Flash, 2026b. Accessed 2026-09-23.

Z.ai. GLM-5-Turbo: Overview. https://docs.z.ai/guides/llm/glm-5-turbo, 2026c. Accessed 2026-09-26.

Sijie Zhao, Feng Liu, Xueliang Zhang, Hao Chen, Xinyu Gu, Zhe Jiang, Fenghua Ling, Ben Fei, Wenlong Zhang, Junjue Wang, Weihao Xuan, Pengfeng Xiao, Naoto Yokoya, and Lei Bai. OpenEarth-Agent: From tool calling to tool creation for open-environment earth observation. arXiv:2603.22148, 2026. URL https://arxiv.org/abs/2603.22148.

Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. MSEarth: A multimodal benchmark for earth science phenomenon discovery with MLLMs. arXiv:2505.20740, 2025. URL https://arxiv.org/abs/2505.20740v3. Version 3, revised 2026.

Yuhao Zhou, Yiheng Wang, Xuming He, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, et al. Scientists' first exam: Probing cognitive abilities of MLLM via perception, understanding, and reasoning. arXiv:2506.10521, 2025. URL https://arxiv.org/abs/2506.10521.

A TASK EXAMPLES BY SCIENTIFIC TOPIC

The seven task briefs illustrate the scientific goal, production inputs, output specification, optional method guidance, and case-specific scoring rules. Output summaries distinguish the scientific product, its coverage or grid and values, and supporting files. Evaluation details are provided for readers; private validation resources and thresholds are not part of the agent-visible task.

A.1 ATMOSPHERE: ANNUAL POINT-SOURCE NOX EMISSIONS

EARTHDATABENCH / TASK EXAMPLE

Annual NOx Emissions

Six German point-source regions · 2018–2020

Atmosphere

Input data types	Processing mechanisms	Scientific product type
Satellite remote sensing Reanalysis products Model outputs	→ Physical retrieval Multi-source fusion Aggregation	→ Emission inventory Annual regional NOx emissions 6 regions × 3 years

A Task

01 Scientific goal

Estimate total NOx (NO + NO_2) for six fixed non-urban German regions, integrating all emissions within 15 km. Produce 18 site/year values in kg NO_2-equivalent s^{-1} for May–December 2018 and the full available years 2019–2020.

02 Input data

TROPOMI: 2,410 native-swath L2 NO_2 files (RPRO v2.4.0), May 2018–Dec 2020; columns, pixel geometry, quality fields, averaging kernels and air-mass factors.
ERA5: 0.25° meteorology on 37 pressure levels (6-hourly) + hourly surface pressure.
GEOS-CF: near-overpass ozone mixing ratios for NO_2-to-NOx conversion.
Context: fixed site centers, surface reflectivity, power-plant and city datasets.

03 Output specification

Scientific product: 18 annual NOx emission estimates (6 regions × 3 years) in CSV.

Coverage and values: fixed sites and years; finite, nonnegative kg NO_2-equivalent s^{-1}.

Supporting files: annual_point_source_nox.png plots the estimates; methodology.md describes processing and quality checks.

04 Method hints (optional; Base prompt)

Filter native swaths → interpolate winds at 500 m → apply AMF and NOx corrections → compute native-swath advection → correct topography → average on a 0.025° grid → integrate within 15 km → correct chemical loss. Document missing-data handling.

B Cumulative scoring

Each level requires all lower levels. Scores ≤0.6 = min(product tier, validation cap).

0.0	**No valid candidate** Hard failure (artifacts, schema, records, values or figure consistency), or S1 unmet.
0.2	**Valid candidate (S1)** Contract checks; ≥4 method topics; signal correlation ≥0.2.
0.4	**Scientific consistency** Expert-product r ≥0.5, NRMSE ≤1.0, mean ratio 0.25–4; ≥6 topics; cap below.
0.6	**Expert-level agreement** Expert-product r ≥0.9, NRMSE ≤0.35; ≥8 method topics; tighter cap below.
0.8	**Beyond expert** Overall gains (r, N, L) ≥(.005, .04, .03) ≥4 regions must have N gain ≥.02 and r/L deficit ≤.02.
1.0	**Strong improvement** Overall gains (r, N, L) ≥(.02, .12, .10) ≥5 regions must have N gain ≥.08 and r/L deficit ≤0.

Independent PRTR inventory: deficit caps relative to the expert baseline

Score / support	r	N	L	RMSE	|bias|	MAE
0.4 overall / region	.15 / .35	.40 / .40	.60 / .60	.20 / .35	.20 / .35	.16 / .35
0.6 overall / region	.02 / .15	.03 / .06	.05 / .08	.01 / .05	.02 / .05	.01 / .05

At each cap, ≥5/6 regions must be within tolerance. All metrics use the submitted emission target.
r: correlation; N: NRMSE; L: |log(mean ratio)|. Gains/deficits use the evaluator's baseline-relative scale.

A.2 CRYOSPHERE: FRACTIONAL SNOW COVER

EARTHDATABENCH / TASK EXAMPLE

ChinaAI-FSC: Fractional Snow Cover

Cryosphere

Four study regions · 2021–2022 · Spatially held-out prediction

Input data types	Processing mechanisms	Scientific product type
Satellite remote sensing ↳ Optical / multispectral **In-situ observations** ↳ Land station†	**Statistical estimation** **Multi-source fusion** **Quality control**	**Continuous field** ↳ State Fractional snow cover (%)

A Task

01 Scientific goal

Learn continuous fractional snow cover (FSC) from all 1,878 Training and 913 Validation tiles, then predict all 895 spatially held-out Testing tiles. Preserve the exact 128 × 128 grids and FSC range [0,100]%. Scope: the four-region benchmark.

02 Input data

One public HDF5: chinaai_fsc_2021_2022.h5. Each tile has 20 predictor bands: coordinates, terrain, land cover, LST, FTC, day of year, NDVI, NDSI and spectral bands. Training/Validation include FSC labels; target contains predictors and metadata only. Features: (tiles,20,128,128); CRS: EPSG:4326; public tile IDs and affine transforms.

† Land-station provenance is listed in the inventory but unspecified in the task card.

03 Output specification

Scientific product: 895 fractional snow-cover maps in chinaai_fsc_predictions.h5.

Grid and values: 128 × 128 float32 tiles in EPSG:4326; preserve tile IDs, order and georeferencing. All FSC values must be finite and within 0–100%.

Supporting files: representative_fsc.png maps the first target tile (testing_0000); method_report.md explains the method and quality checks.

04 Method hints (optional; Base prompt)

Pair features/labels by raw index → deterministic preprocessing → fuse input features → fit/select a continuous model using public labels → freeze one fitted processor/model → predict all target tiles → clip final FSC to [0,100]% → check geometry and finite values.

B Cumulative scoring

All lower gates are required; failure retains the last passed score. Errors use FSC percentage points.

0.0	**No valid candidate**	Delivery, schema, support, physical range or non-degeneracy gate fails.
0.2	**Valid candidate**	Exact files/schema; readable PNG/report; global FSC SD ≥5 percentage points; ≥100 tiles have spatial SD ≥1 percentage point. File-size and metadata checks apply.
0.4	**Scientific consistency**	r(FSC, NDSI) ≥ρ_min; maximum absolute correlation with any single feature on the fixed probes <ρ_copy. Reject feature-inconsistent or near-copy products.
0.6	**Baseline-level validation skill**	On hidden Testing FSC: RMSE_A ≤RMSE_C +ε_R; MAE_A ≤MAE_C +ε_M; r_A ≥r_C −ε_r. All three non-inferiority conditions must hold.
0.8	**Overall and blockwise improvement**	Mean tile-MAE difference Δ ≤−ε; bootstrap upper quantile U_q(Δ) <0. Meet paper U-Net overall RMSE/MAE/r bounds; Δ ≤−ε in every required region × label-decile stratum meeting the frozen minimum tile count.
1.0	**Stronger stratified generalization**	Meet Table-9 U-Net RMSE/MAE/r bounds in all 10 label deciles, macro RMSE/MAE caps and the global r floor. Require Δ ≤−ε in every region and every eligible region × decile stratum defined for this level.

A: agent; C: MB3U comparison. Δ = tile MAE_A − tile MAE_C. Validation is spatially held out within the dataset. ρ, ε, q and paper/stratum bounds are contract-defined; numeric values await evaluator_contract.json.

A.3 BIOSPHERE: SIERRA NEVADA ECOSYSTEM FUNCTIONAL TYPES

EARTHDATABENCH / TASK EXAMPLE

Sierra Nevada Ecosystem Functional Types

Biosphere

2001–2018 · 414 MOD13Q1 EVI composites · Fixed 164 × 410 source grid

Input data types	Processing mechanisms	Scientific product type
Satellite remote sensing ↳ Optical / multispectral **MOD13Q1 EVI**	**Aggregation** **Time-series analysis** **Quality control** **Classification**	**Categorical map** ↳ Ecosystem type **64 possible EFT classes**

A Task

01 Scientific goal

Reconstruct annual ecosystem functional attributes (EFAs), 64-class ecosystem functional types (EFTs), richness and rarity, plus interannual diversity and stability over the Sierra Nevada window.

02 Input data

Public a_evi_aoi/: 18 annual NPZ/JSON pairs, each with 23 16-day EVI composites (164,410), plus timestamps, native MODIS sinusoidal indices, provenance and a completeness manifest. EVI uses a 10,000 scale; fill/negative values are zero, excluded from positive-EVI statistics.

03 Output specification

Scientific product: sierra_nevada_eft.npz contains annual EFAs, EFTs, richness and rarity, plus interannual diversity and stability maps for 2001–2018.

Grid and values: 18 annual 164 × 410 native MODIS grids and eight summary maps; preserve grid and year order. thresholds.json records the fixed classification breaks.

Supporting files: build_product.py reproduces the product; methodology.md explains the method and QC; representative.png shows six summary maps.

04 Method hints (optional; Base prompt)

Read all composites → positive-EVI mean/population SD + earliest-maximum month → average annual spatial quartiles → fixed breaks → 4 × 4 × 4 EFT classes → 4 × 4 richness + rarity (n_max − n_i)/n_max → interannual means, circular month, mode (lowest-code ties) and variability → mean (1 − Jaccard) over all 153 year pairs. n_i: class area; n_max: largest class area. Use decreasing-seasonality classes. Anchor full windows top-left; zero the trailing 3 rows/columns.

B Evaluation & scoring

Valid delivery → test 1.0, then 0.8; otherwise take min(C-similarity tier, B-fidelity tier).

0.0	**Invalid delivery or product** Any formal file, schema, range, non-degeneracy, figure or code-validity gate fails.
0.2	**Valid candidate** Exact files/14 arrays; finite, in-range, ≥2 unique values per array. In the low branch, S <0.72 OR F <0.75 gives 0.2.
0.4	**Moderate product and field agreement** Low branch: S ≥0.72 AND F ≥0.75; the 0.6 conditions are not both met.
0.6	**Expert-level agreement** Low branch: S ≥0.92 AND F ≥0.95. Expert replay anchors the score at 0.6.
0.8	**Improved independent phenology** S ≥0.90 AND Δ_combined ≥0.25 month, after the 1.0 test fails.
1.0	**Stronger gains across held-out partitions** S ≥0.98 AND gains ≥0.25 month in BOTH site-disjoint and site-year-disjoint partitions.

S	Mean of five exact-match fractions (DMAX/EFT/richness/mode/variability) and four MAE terms. Each MAE term = max[0, 1 − MAE/k]; k = 500 (EVI mean), 300 (SD), 1 (rarity/dissimilarity).
F, Δ	$F = \min[1, (e_C + \varepsilon)/(e_A + \varepsilon)]$ on combined B; $\Delta = e_C - e_A$; $\varepsilon = 10^{-12}$. e: mean circular month error; A: agent; C: expert; B: withheld field phenology.
Note	Base allows zero targets when no positive EVI exists; the verifier rejects zero annual DMAX/EFT.

A.4 HYDROLOGY: GLOBAL TERRESTRIAL WATER STORAGE

EARTHDATABENCH / TASK EXAMPLE

Terrestrial Hydrosphere

TWSTORE: Global Terrestrial Water Storage

1984-2020 · 444 monthly fields · 0.5° global land grid · TWSA + variance

Input data types	Processing mechanisms	Scientific product type
Satellite remote sensing ↳ Gravimetric	**Statistical estimation** **Time-series analysis** **Multi-source fusion** **Uncertainty quantification** **Quality control** **Gap-filling**	**Continuous field** ↳ Anomaly **Monthly TWSA** **With variance estimates**

A Task

01 Scientific goal

Reconstruct monthly terrestrial-water-storage anomaly (TWSA) and variance over global land, excluding Greenland and Antarctica. Effective resolution is several hundred kilometres.

02 Input data

GRACE terrestrial-water-storage principal components provide training targets (2003-2010). Regional EOF spatial loadings support grid reconstruction; climate predictors and SLR-DORIS geodetic modes cover 1984-2020. Land masks and cell areas define the spatial domain and area weights.

03 Output specification

Scientific product: twstore_candidate.nc contains monthly TWSA (mm) and variance (mm²) for January 1984–December 2020 (444 months).

Grid and values: fixed 0.5° global grid (360 × 720); nonnegative variance; both fields finite on ≥95% of valid land-month cells and NaN outside the land mask.

Supporting files: twstore_validation_overview.png shows the trend map and global deseasonalized series with ±1σ; report.md explains the method, QC and uncertainty.

04 Method hints (optional; Base prompt)

Contiguous 2003-2010 validation blocks → train-only scaling → compare regularized climate and calibrated SLR-DORIS predictors → uncertainty-aware combination → EOF reconstruction.
TWSA = Σ(PC × loading × √area_weight). Fixed seeds; variance from blocked residuals or model disagreement → inspect seasonal/deseasonalized series, trends, ranges and coverage.

B Cumulative scoring

Each level requires every earlier gate; stop at the highest level passed. Conditions within rows are AND.

0.0	**Invalid delivery or product** Any file-set, NetCDF schema/coordinate/unit/range, PNG or report gate fails.
0.2	**Valid delivery** All delivery and schema/range gates pass; NetCDF ≥100,000 bytes.
0.4	**Coverage, variability and quality control** Coverage $\geq q_c$; no finite values outside the mask; 444 complete global months. Global SD $\geq q_s$; variance span $\geq q_v$; P99|Δ(global monthly TWSA)| $\leq q_j$.
0.6	**Agreement with the author product** Global r $\geq \tau_g$ and E $\leq \tau_E$; trend-map r $\geq \tau_t$. Regions: median r $\geq \tau_m$; P20 r $\geq \tau_{20}$; median(E/σ_C) $\leq \tau_n$.
0.8	**Improved independent global-mass agreement** Against ESA land-water storage (1993-2016): $r_A \geq r_C + \delta_r$ AND $E_A \leq E_C(1 - \delta_E)$.
1.0	**Gains beyond training and calibrated uncertainty** Both gains hold in 1993-2002 AND 2011-2016, with uncertainty noninferiority: $\lvert p_A - 0.68\rvert \leq \lvert p_C - 0.68\rvert + \delta_p$.

Metrics A: candidate; C: author; r: correlation; E: centered RMSE (mm); σ_C: regional author SD.
Scope Area-weighted series need ≥95% global / ≥90% regional common support; ESA tests use deseasonalized TWSA.
p Fraction of centered ESA residuals within ±√(area-mean candidate variance + ESA variance).
Config q, τ and δ denote private thresholds from thresholds.json (not supplied).

A.5 LAND SURFACE: ALL-WEATHER LAND SURFACE TEMPERATURE

EARTHDATABENCH / TASK EXAMPLE

Land Surface

All-Weather Land Surface Temperature

North America · 9 UTC dates × 5 MODIS tiles · 45 daily 1 km LST maps

Input data types	Processing mechanisms	Scientific product type
Satellite remote sensing ↳ Optical / multispectral **Reanalysis products** ↳ Atmospheric reanalysis	**Aggregation** **Physical retrieval** **Multi-source fusion** **Spatial interpolation** **Quality control** **Gap-filling**	**Continuous field** ↳ State

A Task

01 Scientific goal

Produce daily mean LST (K), including cloudy pixels within fixed albedo-derived land support. Dates: 2010-01-01 and Jan/Apr/Jul/Oct 1 of 2019 and 2020; each day uses the full UTC day.

02 Input data

MODIS C6.1 Terra/Aqua TOA radiance + geolocation (312 pairs; 1 km); GLASS albedo/QA (45 files; 500 m) and daily downward shortwave/longwave radiation (9 files each; 0.05°); ERA5-Land hourly skin temperature (9 days × 24 hours; 0.1°).

03 Output specification

Scientific product: 45 daily mean LST maps (GeoTIFF), one for each of the 9 dates × 5 tiles.

Grid and values: 1200 × 1200 float32, fixed MODIS sinusoidal grid, units K. Valid land pixels (including cloudy pixels): finite 180-360 K; all others: NoData −9999.

Supporting files: process.py reproduces the maps; report.md explains the method and QC; validation_overview.png shows representative maps and diagnostics.

04 Method hints (optional; Base prompt)

Pair/calibrate MODIS → QA + bands 31/32 → geolocate and aggregate UTC overpasses. QA-filter albedo → 2 × 2 aggregation; radiation → nearest neighbour.
ERA5-Land → 24-hour mean + bilinear reprojection → fuse thermal/albedo/radiation constraints → spatial-block or overpass validation. Fix seeds and processing rules.

B Cumulative scoring

Award the highest passed level; each positive level requires all earlier positive gates.

0.0 Invalid delivery or minimum science failure
Any file-set, safety, size, readability, grid/schema or minimum science gate fails.

0.2 Complete, physically usable raster set
Delivery gates pass. Every raster: coverage $\geq c_{min}$; no finite values outside support; $T_{min} \leq LST \leq T_{max}$; unique values $\geq u_{min}$; spatial SD $\geq s_{min}$.

0.4 Coarse consistency with the expert product
Pool matched pixels across all 45 rasters: $D \leq D_2$ AND $r \geq r_2$.

0.6 Closer agreement with the expert product
The same pooled comparison meets the S3 limits: $D \leq D_3$ AND $r \geq r_3$.

0.8 Independent station-day improvement with tile guards
B4: $n_A = n_C$; overall $E_A < E_C(1 - \delta_4)$;
every baseline tile t satisfies $E_{A,t} \leq E_{C,t} + \varepsilon$.

1.0 Further validation with strict improvement on every tile
B5: $n_A = n_C$; overall $E_A < E_C(1 - \delta_5)$;
every baseline tile t satisfies $E_{A,t} < E_{C,t}$.

Metrics D: agent-expert RMSE; r: Pearson correlation; E: station-day RMSE (K); n: valid station-day matches.
B4/B5 Configured validation-year subsets. A: agent; C: reference baseline; t: tile. All row conditions are AND.
Config Threshold symbols and validation years come from evaluation_contract.json (not supplied).

A.6 Oceans: Interior Dissolved Oxygen and Total Uncertainty

EARTHDATABENCH / TASK EXAMPLE

Oceans

GOBAI-O_2: Ocean Interior Oxygen

Near-global · March & August 2021 · 1° grid × 58 pressure levels

Input data types	Processing mechanisms	Scientific product type
In-situ observations ↳ Ocean sensor **Model outputs** ↳ Earth system model **Satellite remote sensing** ↳ Optical / multispectral†	**Aggregation** **Physical retrieval** **Statistical estimation** **Multi-source fusion** **Uncertainty quantification** **Quality control**	**Volumetric field** ↳ Ocean interior (3D)

A Task

01 Scientific goal

Reconstruct near-global monthly dissolved oxygen and total uncertainty for March and August 2021, from 2.5 to 1975 dbar, using the paper's regional ensemble.

02 Input data

BGC-Argo and GLODAPv2.2022 oxygen/hydrography train the ensemble; RG09 supplies March/August temperature and salinity; ETOPO2 supplies bottom depth. GFDL-ESM4 oxygen, temperature and salinity (2004–2022) support the algorithm-uncertainty experiment.

03 Output specification

Scientific product: gobai_o2_2021.nc (NetCDF) with monthly 3D oxygen and total-uncertainty fields for both months, in µmol kg^{-1}.

Grid and values: 2 × 58 × 145 × 360 (time, pressure, latitude, longitude), 1° float32 grid. Oxygen ≥0; uncertainty >0; both finite on ocean/bottom support and −9999 elsewhere.

Supporting files: build_product.py reproduces the product; methodology.md explains methods and QC; representative.png shows both fields and months at 150 dbar.

04 Method hints (optional; Base prompt)

All finite observations → 11 predictors → seven regional RF + three-member FNN ensembles → distance-weighted overlaps / Arabian-Sea exception → monthly RG09 prediction.
Measurement (3% of oxygen) + gridding + GFDL algorithm uncertainty → quadrature sum → support and quality checks.

B Evaluation & scoring

Acceptance → test 0.8, then 1.0; if 0.8 fails, use combined gap G for 0.6 / 0.4 / 0.2.

0.0	**Invalid delivery or scientifically unusable product** Delivery/schema, range/texture or support coverage/Jaccard gates fail. Ranges: oxygen 0–600; uncertainty (0,250] µmol kg^{-1}. Coverage uses expert and independent B.
0.2	**Accepted product with large combined gap** Acceptance passes; 0.8 fails; G $>\tau_6$ AND G $>\tau_4$.
0.4	**Moderate combined product and validation gap** Low branch: G $>\tau_6$ AND G $\leq\tau_4$.
0.6	**Small combined product and validation gap** Low branch: G $\leq\tau_6$.
0.8	**Pooled improvement on independent ship observations** c $\geq c_8$; $L_{pool} > \delta_{pool}$; $RMSE_A \leq \alpha\ RMSE_C$; $MAE_A \leq MAE_C$; $\lvert bias_A\rvert \leq \lvert bias_C\rvert + \varepsilon_b$; $NLS_A \leq NLS_C$. All conditions must hold.
1.0	**Month, depth and cruise robustness** 0.8 plus EACH month: $L_{month} > 0$ and $NLS_A < NLS_C$; at 100–1000 dbar: $L_{depth} > \delta_{depth}$. EVERY cruise: coverage $\geq c_{cruise}$ and $RMSE_A / \max(RMSE_C, 10^{-12}) \leq \alpha_{cruise}$.

G	Weighted sum of independent RMSE deficit, oxygen/uncertainty product gap and coverage shortfall.
L	Profile-bootstrap lower bound of median(|C−B|−|A−B|), resampling whole station/cast profiles.
Metrics	A: agent; C: expert; B: ship observations; c: coverage; NLS: Gaussian negative log score.
Config	Thresholds, weights and bootstrap settings: evaluator_contract.json (not supplied).
†	Optical input is inventory-listed; no corresponding dataset is specified in the task card.

A.7 Solid Earth: Tonga Marine Gravity-Gradient Tensor

EARTHDATABENCH / TASK EXAMPLE

Solid Earth

CUGB2023GRAD: Tonga Gravity-Gradient Tensor

Tonga marine region · Six tensor components · Native 1-arc-minute grid

Input data types	Processing mechanisms	Scientific product type
In-situ observations† ↳ Geodetic / seismic network **Satellite remote sensing†** ↳ Gravimetric ↳ Optical / multispectral	**Aggregation** **Quality control**	**Continuous field** ↳ Structural

A Task

01 Scientific goal

Reconstruct the six-component marine gravity-gradient tensor for Tonga on its native 1-arc-minute grid, preserving source values and their common finite support.

02 Input data

Six CUGB2023GRAD v2-corrected source grids, one per tensor component, with shared latitude/longitude coordinates. Each provides values in Eotvos for Tonga (176–172°W, 26–20°S).

03 Output specification

Scientific product: tensor.nc (CF-1.8 NetCDF) with Txx, Tyy, Tzz, Txy, Txz and Tyz in Eotvos.

Grid and values: 360 × 240 float32, EPSG:4326; retain native coordinates and the common finite-support mask, with NaN elsewhere. Axes: x north, y west, z radially outward.

Supporting files: report.md describes methods and provenance; representative.png maps Tzz with units and missing support. Store reproducible processing code outside output/.

04 Method hints (optional; Base prompt)

Check six source grids → intersect finite masks → map source z to each tensor component → write CF NetCDF → plot Tzz → record methods and provenance.
Copy native values directly; no interpolation, resampling, smoothing or gap filling.

B Cumulative scoring

Each positive level requires every earlier gate; return the highest level passed.

0.0	**Invalid delivery or tensor structure** File set, NetCDF schema/coordinates/mask, units, report headings or PNG checks fail.
0.2	**Valid six-component product** Delivery checks pass; all six fields are finite on the common mask and NaN elsewhere.
0.4	**Source fidelity and physical closure** Across six components: median source r ≥0.999 and median source NRMSE ≤0.03; RMS(Txx + Tyy + Tzz) <0.2 Eotvos on the common valid support.
0.6	**Agreement with the author tensor** Across six components: median author r ≥0.9999 AND median author NRMSE ≤0.02.
0.8	**Independent-block coherence gains** Exactly three B blocks: min(Δ) ≥0.0002; median(Δ) ≥0.0005; median(R^2_A) ≥median(R^2_C) +0.0005. All three conditions must hold.
1.0	**Larger gains with a bootstrap lower bound** 0.8 plus min(Δ) ≥0.00024; median(Δ) ≥0.0008; the 2.5% bootstrap quantile of mean block gain, $Q_{0.025}$, is >0.0002.

Metrics r: Pearson correlation; NRMSE: RMSE / (reference-component SD + 10^{-12}).
Δ, R^2 $\Delta = R^2_A - R^2_C$; R^2 is squared Pearson correlation with independent B. All gains are dimensionless.
Q $Q_{0.025}$ uses 4,096 resamples of the three block gains. A: agent Tzz; C: frozen Tzz baseline.
† Inventory labels retained; upstream sensor/network provenance is not documented in the task card.

B Processing-Mechanism Analysis

B.1 Base Performance by Processing Mechanism

Table 5 pools the same ten Base configurations over our 40 tasks. Each task contributes ten outcomes to every applicable category. Rows are ordered by mean score. Quality control covers all tasks and therefore serves as the overall baseline. These overlapping groups describe whole-task performance, rather than isolated processing skills.

Table 5: Base performance by processing mechanism.

			Score distribution (%)					Attainment (%)	
Mechanism	Tasks	Score	0.0	0.2	0.4	0.6	0.8	≥ 0.4	≥ 0.6
Physical retrieval	16	22.0	30.0	30.6	38.8	0.6	0.0	39.4	0.6
Aggregation	25	25.4	27.2	30.4	31.6	10.0	0.8	42.4	10.8
Multi-source fusion	29	28.2	21.7	33.4	27.2	17.2	0.3	44.8	17.6
Uncertainty quantification	12	28.8	20.0	37.5	23.3	16.7	2.5	42.5	19.2
Spatial interpolation	18	28.9	20.0	25.6	44.4	10.0	0.0	54.4	10.0
Gap filling	13	29.1	25.4	23.8	33.1	15.4	2.3	50.8	17.7
Classification	5	29.2	12.0	36.0	46.0	6.0	0.0	52.0	6.0
Time-series analysis	12	30.0	17.5	31.7	36.7	11.7	2.5	50.8	14.2
Statistical estimation	24	32.4	16.3	30.0	29.6	23.8	0.4	53.8	24.2
Overall	40	30.3	20.8	27.3	32.8	18.5	0.8	52.0	19.3

Each task contributes ten Base outcomes. Score uses 0–100. No outcome reaches 1.0. Labels overlap. Quality control covers all tasks and is represented by the Overall row.

Physical retrieval is difficult across quality thresholds. Across 16 tasks, physical retrieval has the lowest mean score (22.0) and the highest zero-score rate (30.0%, compared with 20.8% overall). Its expert-level attainment remains low after zero-score outcomes are excluded. Only one of the 112 nonzero outcomes reaches 0.6 (0.9%), compared with 77 of 317 nonzero outcomes overall (24.3%). The other 111 physical-retrieval outcomes remain at 0.2 or 0.4. Thus, the quality gap persists among products that already pass the minimum-usability gate. These conditional rates use nonzero outcomes as the denominator, while Table 5 reports rates over all outcomes.

Aggregation remains difficult despite benefiting from guidance. Aggregation has the second-lowest mean score (25.4), with 27.2% zero-score outcomes and 10.8% expert-level attainment. All ten configurations score below their own overall means on both aggregation and physical retrieval. The consistent aggregation gains in the five-configuration paired comparison (Appendix B.2) therefore coexist with low absolute performance under Base.

Classification has few zero-score outcomes but limited expert-level attainment. Classification has the lowest zero-score rate (12.0%), yet only 6.0% of outcomes reach expert level across its five tasks. Statistical estimation instead has the highest mean score (32.4) and expert-level attainment (24.2%). These contrasts show why category difficulty should be assessed at multiple quality thresholds. Task composition and overlapping labels prevent attributing the differences to any single operation.

B.2 Effect of Method Guidance

Method guidance mainly improves scientific consistency. Across five paired Codex configurations on our 40 tasks, Base increases scientific-consistency attainment by 7.5 percentage points, while expert-level attainment changes by only 0.5 percentage points (Table 4). Table 6 breaks down the effects by processing mechanism.

Table 6: Aggregated guidance gains by processing mechanism. Score changes are Base minus Open on the same five Codex configurations. $+/=/-$ counts configurations with higher/equal/lower category mean scores under Base.

Processing mechanism	n	Open	Base	Δ	$+/=/-$	95% interval
Classification	5	22.4	29.6	+7.2	3/2/0	[-1.6, +15.2]
Aggregation	25	23.2	26.2	+3.0	5/0/0	[-0.6, +6.7]
Quality control	40	29.9	31.3	+1.4	3/1/1	[-1.9, +4.6]
Multi-source fusion	29	28.3	28.3	0.0	3/0/2	[-3.0, +3.0]
Spatial interpolation	18	29.6	29.6	0.0	2/1/2	[-4.0, +4.2]
Statistical estimation	24	33.3	33.2	-0.2	2/0/3	[-3.2, +3.0]
Uncertainty quantification	12	31.0	30.3	-0.7	2/0/3	[-4.7, +3.7]
Time-series analysis	12	32.3	31.3	-1.0	2/1/2	[-8.7, +6.0]
Physical retrieval	16	25.7	23.3	-2.5	0/1/4	[-7.3, +1.5]
Gap filling	13	31.4	28.0	-3.4	2/0/3	[-9.5, +2.5]

n counts distinct tasks, not task–configuration rows. Mechanisms overlap; quality control covers the full task set. Scores use the 0–100 scale. Intervals use 20,000 paired task resamples, preserving all configurations; they are unadjusted exploratory intervals, not run-to-run uncertainty.

Classification and aggregation show the clearest observed benefits. Classification has the largest mean-score gain (+7.2), with three configurations improving and two tying. Aggregation improves all five (+3.0). Guidance does not help every group, however. Physical retrieval declines in four configurations and ties in one, while gap filling declines by 3.4 points on average.

All category-level 95% task-bootstrap intervals include zero. These patterns are therefore descriptive, and overlapping task labels prevent attributing the changes to isolated processing mechanisms.

C Failure Diagnostics

Failure to meet format requirements is the more common source of zero scores. Of the 83 Base zero-score outcomes (20.8% of all 400 outcomes), 51 concern *format compliance* and 32 concern *scientific semantics*, accounting for 61.4% and 38.6%, respectively (Table 7).

Table 7: Zero-score failures in the Base setting.

Failure category	Count	Share (%)
Format compliance	**51**	**61.4**
Missing deliveries	8	9.6
Noncompliant files, schemas, or reports	43	51.8
Scientific semantics	**32**	**38.6**
Invalid content or insufficient coverage	4	4.8
Below minimum scientific requirements	28	33.7
All zero outcomes	83	100.0

Percentages are within the 83 Base zero-score outcomes. Bold rows sum their two subcategories. Assignments follow the diagnostic taxonomy.

Format compliance covers both complete delivery and adherence to the required file and report specifications. Of its 51 failures, 43 involve noncompliant submissions: 18 concern file format, 15 other output requirements, and ten reports. The remaining eight are *Missing deliveries*. Scientific semantics concerns the validity, coverage, and minimum scientific quality of the product's content. Of the 32 failures assigned to this group, four involve invalid content or insufficient coverage, and 28 are labeled as falling below minimum scientific requirements.