EXTREMA: ENABLING AND EVALUATING AGENTIC RESEARCH ON EXTREME WEATHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding extreme weather is vital for climate risk management, but it demands scarce domain expertise. This barrier deepens global inequalities in climate resilience. AI agents offer a path toward democratizing such research. However, existing agents struggle with long-horizon inquiry, as error-free code cannot guarantee physical validity. We introduce Extrema, a system combining ExtremaWorld—which unifies 20 datasets and 100 validated tools—with ExtremaAgent, which pairs a scientific review loop and a persistent task graph to guard against methodological errors. To evaluate such investigations, we introduce ExtremaBench, which distills 58 tasks from 19 published studies covering five extreme weather categories. Withholding the source study, we require each agent to rediscover the study's findings independently, and grade the resulting conclusions and supporting figures against expert-authored rubrics. Powered by GPT-5.6-sol, Extrema scores 71.58/100, outperforming the strongest baseline by 10.07 points.

1 INTRODUCTION

Climate change is altering weather extremes worldwide, with substantial consequences for societies and ecosystems (Seneviratne et al., 2021; Pörtner et al., 2022). Understanding how such events form and evolve is essential for disaster preparedness and climate risk management. Yet these investigations draw on heterogeneous data sources and specialized meteorological methods, demanding a level of domain expertise that remains scarce in many developing countries (Philip et al., 2020; Camps-Valls et al., 2025; Lamptey et al., 2024). This disparity constrains the capacity of the most vulnerable regions to assess their own climate risks (Blicharska et al., 2017; Testani et al., 2025), motivating the development of AI agents that can conduct such studies end to end—from data analysis to scientific conclusions—and thereby broaden access to extreme weather research.

Existing agents target a range of meteorological tasks: Zephyrus (Varambally et al., 2026) supports weather question answering, forecasting, and simulation; Kim et al. (2025) automate routine data retrieval and visualization; and frameworks such as HVR-Met (Tang et al., 2026) and EWE (Jiang et al., 2025) rely on standard indices and straightforward plot interpretation. A representative HVR-Met task, for instance, asks the agent to identify extreme weather in Beijing from forecast data and to analyze anomalies relative to climatology (Tang et al., 2026). In contrast, we focus on long-horizon investigations that combine multiple data sources and diverse analytical methods to explain event evolution and the roles of different physical processes(Table 1). General-purpose coding agents can write arbitrary analysis scripts, but in scientific inquiry, successful code execution does not guarantee physical validity: a script may run without errors even when domain-specific methods are misapplied or when the empirical results do not support the conclusions. Consequently, realizing these in-depth analyses demands a unified scientific environment that equips agents with valid analytical tools and data to dynamically assemble comprehensive research workflows.

To bridge this gap, we present Extrema, an end-to-end autonomous research system designed for extreme weather investigation. Extrema combines a domain-grounded scientific computing environment (ExtremaWorld) with an evidence-driven research agent (ExtremaAgent). Guided by a systematic survey of 500 domain papers, ExtremaWorld unifies 20 primary meteorological datasets and encapsulates 100 validated scientific tools, abstracting complex workflows—from event tracking to numerical hypothesis testing—away from low-level data engineering. Operating within this environment, ExtremaAgent conducts investigations via a three-stage pipeline: planning, experimentation,

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.

Table 1: Comparison of task scope and supported analyses across weather agents.

System	Task source	Extreme-event Focus	Scientific Discovery	Long-horizon Tasks	Multi-source Analysis	Method Diversity	Numerical Experiments
Zephyrus (ICLR 2026)	Semi-synthetic	×	×	×	×	✓	✓
HVR-Met (ICML 2026)	Paper	✓	×	×	×	×	×
EWE (arXiv 2025)	Semi-synthetic	✓	×	×	×	×	×
Extrema (Ours)	Paper	✓	✓	✓	✓	✓	✓

and report generation. Crucially, to guard against methodological errors and unsupported conclusions, the agent pairs an independent scientific review loop with a persistent task graph. While the reviewer strictly verifies algorithmic implementations and empirical grounding, the task graph maintains execution dependencies, enabling selective, cost-effective recomputation whenever hypotheses or intermediate analyses are revised.

To evaluate such open-ended scientific investigations without subjective bias, we formulate event analysis as a scientific rediscovery challenge within ExtremaBench. Derived from 19 peer-reviewed papers spanning heatwaves, cold waves, torrential rains, cyclones, and droughts, ExtremaBench establishes 58 tasks evaluated against expert-curated rubrics measuring diagnostic accuracy, causal completeness, and evidence grounding under a blind rediscovery setting. Evaluated on this benchmark, Extrema demonstrates superior performance, achieving a score of 71.58/100 with GPT-5.6-sol—outperforming the strongest baseline by 10.07 points and improving upon Codex by 11.34 points (and 18.82 points with GPT-5.5) with consistent advantages across all five disaster categories.

In summary, our key contributions are:

- We support long-horizon extreme weather research through **Extrema**, which integrates a scientific computing environment with an autonomous agent. Validated tools and independent scientific review help guard against scientific hallucinations.
- We provide a common evaluation protocol for extreme weather research through **ExtremaBench**. It comprises 58 rediscovery tasks from 19 papers, with expert-curated rubrics for assessing scientific findings and supporting evidence.
- We demonstrate that **Extrema** outperforms Codex on **ExtremaBench**. With matched backbones, the improvements are 11.34 points using GPT-5.6-sol and 18.82 points using GPT-5.5

2 Related Work

Weather and Climate Science Agents. LLM agents support a range of tasks in weather and climate science, including data analysis, physical modeling, and numerical experimentation. Zephyrus (Varambally et al., 2026) combines tools and code execution for weather queries, forecasting, and counterfactual analysis, while ClimateAgent (Kim et al., 2025) automates climate data acquisition, processing, and visualization. For climate research, ClimAgent (Wang et al., 2026a) focuses on physical modeling and computational problem solving, and EarthLink (Guo et al., 2025) supports climate model diagnosis, projections, and studies of physical mechanisms. TianJi (Zhang et al., 2026a) autonomously drives numerical models to verify physical mechanisms.

EWE (Jiang et al., 2025) and HVR-Met (Tang et al., 2026) focus on extreme weather diagnosis, primarily computing meteorological indices and interpreting diagnostic figures. Their demonstrated tasks provide limited coverage of investigations that combine multiple data sources, diverse analytical methods, and numerical experiments. Extrema brings these capabilities together to support scientific discovery through long-horizon investigations. Its tasks are drawn from scientific papers and address questions about event evolution and physical mechanisms (Table 1).

Environments and Tools for Scientific Agents. Specialized computing environments are pivotal for grounding scientific agents in empirical physical laws. In biomedical science and chemistry, platforms such as Biomni (Huang et al., 2026), ToolUniverse (Gao et al., 2025), ChemCrow (M. Bran et al., 2024), and MolClaw (Zhang et al., 2026b) have been demonstrated to substantially migrate

code hallucination and increase agent problem solving capabilities. However, in atmospheric science, especially for extreme weather events, existing systems largely rely on unstandardized, on-the fly script with isolated Metpy function.Integrated environments for in-depth studies of the evolution and physical mechanisms remain limited.

ExtremaWorld is the first to integrates atmospheric datasets with validated tools for event characterization, physical diagnosis, and numerical experimentation, making these resources reusable across agent frameworks.

Benchmarks for Scientific Analysis and Rediscovery. Evaluating open-ended scientifc discovery remains a fundamental challenge due to lack of ground truth in novel findings. Scientific agent benchmarks cover data analysis, research reproduction, and the rediscovery of published findings. ScienceAgentBench (Chen et al., 2025) evaluates programs and their outputs on scientific analysis tasks, and DiscoveryBench (Majumder et al., 2025) assesses hypothesis discovery from datasets and research goals. For research reproduction, CORE-Bench (Siegel et al., 2024) provides code and data, whereas PaperBench (Starace et al., 2025) requires agents to implement methods and run experiments from published papers. FIRE-Bench (Wang et al., 2026b) evaluates rediscovery in machine learning by comparing agents' findings with published results at the claim level. Research-ClawBench (Xu et al., 2026) evaluates rediscovery across ten scientific domains with expert-curated rubrics, but includes only four Earth science tasks. Existing rediscovery benchmarks provide limited coverage of the evolution and physical mechanisms of extreme weather events. We introduce ExtremaBench to assess whether agents can recover published findings in this domain without access to the target papers, evaluating both scientific claims and supporting diagnostic figures.

3 METHOD

To automate complex meteorological research, we propose a framework comprising two tightly coupled components: a dedicated computing sandbox and an autonomous research agent (Figure 1). **ExtremaWorld** (Section 3.1) serves as the foundational environment, integrating diverse meteorological data and domain-specific tools. Operating within this environment, **ExtremaAgent** (Section 3.2) acts as the cognitive core, autonomously translating high-level scientific questions into executable workflows and synthesizing physical interpretations.

3.1 EXTREMAWORLD: SCIENTIFIC COMPUTING ENVIRONMENT

To reduce the effort required for data preparation and scientific software integration, ExtremaWorld combines meteorological data and reusable scientific tools in a sandboxed execution environment. Its data resources and software were selected through a survey of extreme weather studies.

Literature Survey. We surveyed 500 extreme weather papers from journals selected by four domain experts, including Nature-family journals, Monthly Weather Review, and Journal of Climate. MinerU parsed the full texts, and LLMs extracted datasets, methods, software, and code to guide resource curation. These papers do not overlap with ExtremaBench's target papers.

3.1.1 DATA RESOURCES

We selected 20 data resources based on usage frequency and research applications in the surveyed literature. The collection covers satellite observations, atmospheric reanalyses, ocean boundary conditions, climate model outputs, and extreme event products. These resources span diverse data representations, including time series, multivariate gridded fields, and event tracks. We downloaded and organized the data for direct access in the sandbox. The agent can select resources by browsing a data catalog and querying variable information and spatiotemporal coverage. A comprehensive inventory and detailed descriptions of all 20 datasets are provided in Appendix A.

3.1.2 SCIENTIFIC TOOLS

ExtremaWorld exposes reusable tools adapted from established scientific software as documented Python functions. It also provides agent-based tools that encapsulate complex workflows involving multiple tool calls.

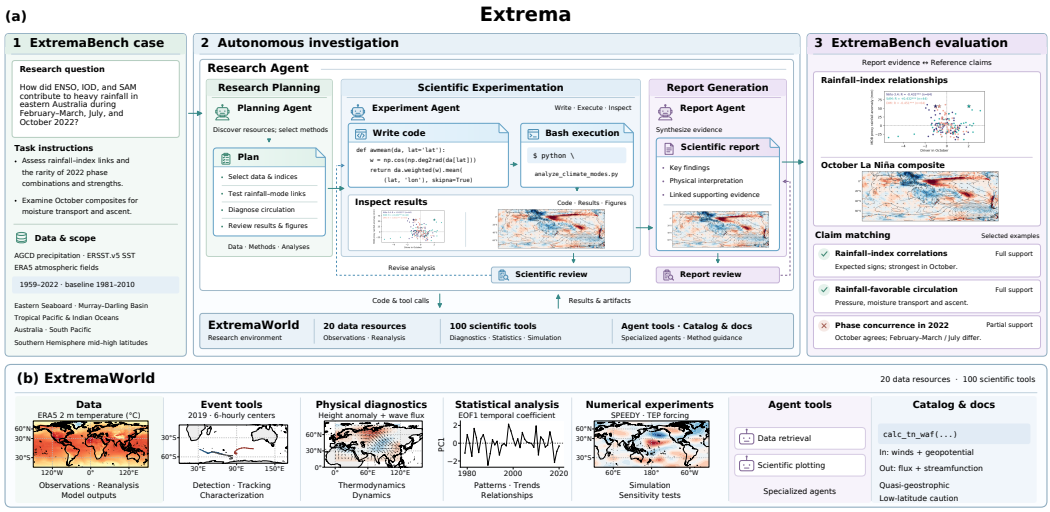


Figure 1: **Overview of Extrema.** (a) The research agent conducts planning, experimentation, and report generation in ExtremaWorld; ExtremaBench evaluates the findings against withheld references. (b) ExtremaWorld provides data resources, scientific and agent tools, and method documentation. Plots illustrate tool capabilities across different research cases.

Scientific Toolset. ExtremaWorld provides 100 reusable tools spanning data preparation, event characterization, physical diagnosis, and hypothesis testing through numerical experiments. Data preparation includes retrieval, reading, regridding, bias correction, and filtering. Extreme indices quantify event intensity, persistence, and anomalies using climatology and historical data. Detection and tracking tools identify atmospheric rivers and cyclones and trace their evolution.

Dynamical and thermodynamic diagnostics characterize circulation, moisture transport, and thermodynamic conditions underlying event formation and maintenance. Climate indices and ocean state diagnostics provide environmental context across scales, while statistical tools extract dominant modes of variability, identify typical weather patterns, and test for trends. Numerical simulation tools test physical explanations through control and sensitivity experiments with configurable external forcing. Agents can combine tools, write analysis code, and create new tools as needed (Appendix B).

Tool Construction. Each tool preserves the algorithms of the underlying scientific software while handling input preparation and execution. We reviewed source code and documentation to identify input and preprocessing requirements, then implemented wrappers that adapt inputs accordingly and retain spatial and temporal coordinates in the outputs applicable. For workflows involving multiple programs, the tools connect inputs and outputs across stages, manage intermediate files, and keep parameters configurable.

For sensitivity experiments, we wrapped the SPEEDY (Molteni et al., 1996) model so the agent can specify external forcing and its spatial extent. The tool prepares inputs on the model grid, generates configuration files, manages compilation and execution, and converts outputs for subsequent diagnostics and visualization.

Tool Validation. We validated the tool wrappers by comparing each tool's outputs with those of the original software under identical inputs and configurations. Only tools with consistent outputs were included in ExtremaWorld.

Tool Access. ExtremaAgent discovers tools through an automatically generated catalog of function signatures and summaries. Python docstrings detail scientific principles, applicability, computational steps, input/output structures, coordinate conventions, and preprocessing requirements. The agent writes code importing these functions and executes it through Bash, using returned results,

file paths, or errors for subsequent analysis and debugging. The tools also work with Claude Code, Codex, and other coding agents.

Specialized Agents. ExtremaWorld exposes subagents for data retrieval, downloading, and diagnostic plotting. Given a task description and file paths, each uses a dedicated prompt and toolset in a separate context, saves reusable artifacts in the working directory, and returns a summary with file paths. Keeping intermediate interactions in separate contexts reduces the execution details passed to ExtremaAgent.

3.2 EXTREMAAGENT: RESEARCH AGENT

ExtremaAgent conducts planning, scientific experimentation, and report generation, supported by a persistent task graph and independent scientific review.

3.2.1 RESEARCH WORKFLOW

Planning. The Planning Agent selects data and methods by aligning the research question with the data coverage and tool capabilities of ExtremaWorld. The plan justifies these choices, states method applicability conditions, and specifies task dependencies and expected outputs. To limit context size, the agent consults data and tool summaries before retrieving detailed documentation as needed.

Scientific Experimentation. The Experiment Agent develops and executes code using ExtremaWorld's data and tools to test and revise explanations regarding event processes. The plan and task states—classified as *pending*, *in progress*, or *completed*—are maintained in external files structured as a task graph. Tasks are executed strictly after their prerequisites are satisfied. If methods, assumptions, or intermediate results require revision, the affected tasks and their downstream dependencies are reset to *pending* and re-executed. Unaffected artifacts are preserved in the working directory for subsequent reuse. The system continuously monitors these task states, automatically triggering subsequent executions until the entire workflow is completed.

If the Experiment Agent fails to resolve a specific code error after three consecutive attempts, the Extrema system invokes a specialized Debug Agent to diagnose and repair the issue within an isolated context.

Report Generation. The Report Agent synthesizes experimental results into a scientific report, linking key findings and physical interpretations to supporting artifacts.

3.2.2 SCIENTIFIC REVIEW

The Review Agent works in a separate context, guided by the original research question and task requirements. It returns structured feedback with evidence for identified issues and recommended revisions.

Experiment Review. At the Experiment Agent's request, the reviewer examines code and experimental artifacts to check whether methods are implemented correctly and whether results support the physical interpretations. The Experiment Agent uses the feedback to correct code, add analyses, or revise the plan.

Report Review. The reviewer checks the draft for omitted findings, inconsistencies with computational results or diagnostic visualizations, and inferences that exceed the available evidence. The Report Agent revises the draft based on this feedback.

4 EXTREMABENCH

We introduce ExtremaBench to evaluate autonomous research on extreme weather events through scientific rediscovery. Tasks are based on published findings about event characteristics and physical mechanisms. Agents investigate the corresponding scientific questions without access to the target papers. Expert rubrics based on the papers' findings and evidence assess how well agents recover

the findings and support their conclusions. Figure 2 illustrates how published findings are converted into validated rediscovery tasks, with agent inputs separated from withheld evaluation references.

4.1 Task Definition

We define each task in ExtremaBench as a scientific rediscovery instance:

$$\tau_i = (Q_i, I_i, A_i), \tag{1}$$

Here, Q_i is the scientific question, I_i the task instructions, and A_i the evaluation reference. The instructions describe the data and spatiotemporal scope without revealing target conclusions or key results. The evaluation reference contains an expert rubric, relevant paper text, and reference figures. The target paper and A_i are withheld from the agent during task execution.

Given (Q_i, I_i), the agent independently plans and conducts research using the environment's data and tools, then submits analysis code and a scientific report:

$$y_i = (C_i, R_i). \tag{2}$$

Here, C_i is the analysis code and R_i the scientific report, which presents scientific conclusions, physical interpretations, and supporting evidence. Agents may choose any scientifically valid approach that meets the task requirements.

4.2 Task Construction

ExtremaBench contains 58 tasks from 19 target papers. We annotate two types of individually verifiable statements: text claims for scientific findings and image claims for key visual facts. The benchmark includes 244 text claims and 234 image claims.

Paper Selection. Domain experts selected papers from the same journals as the ExtremaWorld literature survey, with no overlap between the two collections. They prioritized recent extreme weather studies with clear questions and findings supported by data. Papers were excluded if key analyses were too computationally expensive or required unavailable data without a reasonable substitute.

Task Formulation and Annotation. Experts selected findings from each paper that were directly supported by at least one visualization. They examined the underlying analyses and evidence to formulate research questions. Following FIRE-Bench (Wang et al., 2026b) and RAGChecker (Ru et al., 2024), experts decomposed the paper text describing the target findings into atomic, verifiable claims. Text claims capture key scientific conclusions, while image claims capture visual facts supported by the reference figures. These annotations, together with the corresponding paper text and reference figures, form A_i. The benchmark includes 244 text claims and 234 image claims.

Task Validation. Domain experts tested each candidate task using the specified data and retained only those whose core findings could be reproduced. We allowed differences in figure presentation and reasonable numerical deviations due to data versions, sources, or justified substitutions. LLM tools may assist annotation and validation, but experts make the final decisions.

4.3 Evaluation Protocol

Evaluation Overview. We assess both the scientific conclusions in the report and the evidence in its cited figures. A multimodal LLM judges these two components separately, based on the scientific question, task instructions, and evaluation reference.

Text Evaluation. The judge scores the report against each text claim for correctness, completeness, and evidential support. For quantitative claims, numerical deviations are assessed against the annotated precision requirements. Claim scores are then combined using their importance weights to obtain the text score.

Figure Evaluation. Image claims identify the key visual elements to compare, including patterns, anomaly directions, and spatiotemporal relationships. The judge scores how well the cited figures match the reference figures on these elements. These scores are then combined with weights to obtain the figure score.

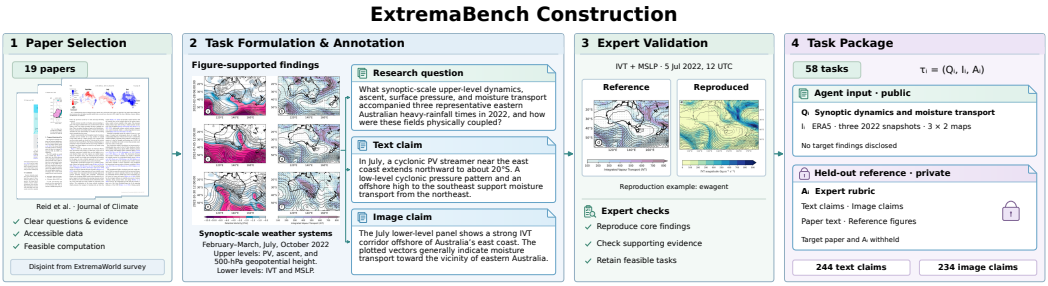


Figure 2: **Overview of ExtremaBench construction.** The pipeline comprises paper selection, task formulation and claim annotation, expert validation, and task packaging. An extreme-rainfall case illustrates text and image claims and reproduction of supporting evidence. Agents receive the research question and instructions (Q_i, I_i); the target paper and evaluation reference A_i are withheld.

Table 2: Main results on 58 ExtremaBench tasks. Scores are mean rubric scores on a 0–100 scale (higher is better). Overall averages all evaluated tasks with equal task weight; category scores average tasks within each category. Bold indicates the best score in each column. Differences reported in the text are computed from unrounded scores.

Agent	Backbone LLM	Overall	Cold waves	Cyclones	Drought	Extreme rainfall	Heatwaves
Codex	GPT-5.4	39.08	41.76	33.85	21.60	40.12	47.27
Codex	GPT-5.5	44.69	43.82	41.37	41.76	42.58	53.11
Codex	GPT-5.6-sol	60.23	70.80	53.64	61.06	50.86	63.94
Claude Code	Claude Opus 5	61.51	69.89	59.44	60.04	50.70	65.64
ExtremaAgent (ours)	GPT-5.5	63.50	67.59	60.77	55.24	65.42	64.59
ExtremaAgent (ours)	GPT-5.6-sol	**71.58**	**73.11**	**68.24**	**63.30**	**74.43**	**75.37**

Score Aggregation. We normalize the text score S_{text} and figure score S_{figure} separately and compute their weighted sum:

$$S_{\text{task}} = 0.6S_{\text{text}} + 0.4S_{\text{figure}}, \tag{3}$$

The overall benchmark score is the arithmetic mean of all task scores.

5 Experiments

5.1 Experimental Setup

We evaluate ExtremaAgent against the general-purpose coding agents Codex and Claude Code on ExtremaBench (Table 2). Comparisons between ExtremaAgent and Codex use matched GPT-5.5 or GPT-5.6-sol backbones. Each agent has a 48-hour limit per task; GPT-5.6-sol serves as the multimodal judge for reports and cited figures.

5.2 Main Results

ExtremaAgent with GPT-5.6-sol achieves the highest overall score of 71.58, exceeding Claude Code, the strongest baseline, by 10.07 points (Table 2). With GPT-5.5, it scores 63.50 and surpasses all evaluated baselines, including Codex with GPT-5.6-sol.

With the backbone held fixed, ExtremaAgent improves over Codex by 18.82 points with GPT-5.5 and 11.34 with GPT-5.6-sol, with gains across all five event categories. It scores higher on 50 of 58 and 40 of 58 tasks, respectively, with two ties for GPT-5.6-sol. For GPT-5.6-sol, the largest gain occurs on extreme rainfall (+23.57 points), with smaller gains on cold waves and drought. These comparisons assess the complete systems.

Stronger backbones can also correct implementation errors: Codex with GPT-5.6-sol corrected GPT-5.5's filtering of discontinuous January data by processing each winter separately. The larger gain with GPT-5.5 suggests that ExtremaAgent's scientific tools and agent design can partly compensate for limited backbone capabilities while remaining useful as models improve.

Token Efficiency. ExtremaAgent achieves higher scores with lower adjusted token usage (Table 3). It uses 16.5% fewer tokens than Codex with the same GPT-5.6-sol backbone and 67.7% fewer than Claude Code. In the inspected traces, ExtremaAgent reuses scientific tools and specialized data interfaces, whereas the baselines build and debug more of their own processing pipelines and repeatedly check intermediate results. This pattern is consistent with lower interaction overhead from tool reuse.

Table 3: Mean tokens per task (millions) and overall scores. Token counts include cached inputs and exclude confirmed redundant empty polling.

Agent	Backbone	Tokens (M) ↓	Score ↑
Codex	GPT-5.6-sol	3.96	60.23
Claude Code	Claude Opus 5	10.23	61.51
ExtremaAgent	GPT-5.6-sol	**3.30**	**71.58**

5.3 Error Analysis

We analyze 174 runs in total, covering each agent's strongest configuration on the same 58 tasks. LLM-assisted inspection of evaluation feedback and execution artifacts, supplemented by targeted verification, identifies five error types. *Methodological Error* covers errors in data processing, scientific method selection, or implementation. *Scientific Discovery Omission* denotes missing analyses or required findings, whereas *Interpretive Omission* denotes failure to identify or explain findings supported by the results. *Unsupported Inference* covers conclusions beyond the available evidence; *Task Completion Failure* covers incomplete execution or missing required artifacts. Multiple labels may apply to each run.

Failure Patterns and Performance. Methodological errors and missing discoveries are the most frequent failures (Figure 3(a)). ExtremaAgent has a lower methodological error rate (36.2%) than Codex (65.5%) and Claude Code (43.1%), with a modest reduction in scientific discovery omissions. The evaluator flags 20.1% of ExtremaAgent's evaluated claims, compared with 41.0% for Codex and 33.8% for Claude Code (Figure 3(b)). These claim-level flags provide a complementary perspective on overall performance (Figure 3(c)).

Scientific Tool Support. ExtremaAgent's WAF tool retains the omitted latitude factor, while its atmospheric-river detector provides vector-guided tracking with centroid adjustments confined to local cross-sections (Table 4), although some tracking details differ from the reference method. Codex's potential vorticity (PV) calculation also omits terms coupling vertical wind shear with horizontal potential-temperature gradients, leaving the diagnostic incomplete. Reusing these tools can reduce the need to reconstruct scientific methods during execution.

Interpretive Omission. Codex and ExtremaAgent correctly plotted May 2023 warming southeast of Madagascar and cooling northwest of Australia, but their reports focused on the Arabian Sea, equatorial Indian Ocean, and Bay of Bengal, omitting the dipole-like structure farther south. In the 2022 Eurasian circulation analysis, Claude Code identified the Iranian component of the subtropical high-pressure belt; Codex omitted it, and ExtremaAgent left it unspecified. These cases distinguish computational correctness from interpretive completeness: generating evidence does not ensure explicit recognition of its implications. This gap is analogous to the distinction between encoded knowledge and successful recall (Calderon et al., 2026).

5.4 Ablation Study

To assess the contributions of individual components, we conduct an ablation study by separately removing scientific review and access to the provided scientific tools from ExtremaAgent. To reduce evaluation cost, we evaluate these configurations on a task subset using GPT-5.6-sol, with Codex using the same backbone as an external baseline (Table 5).

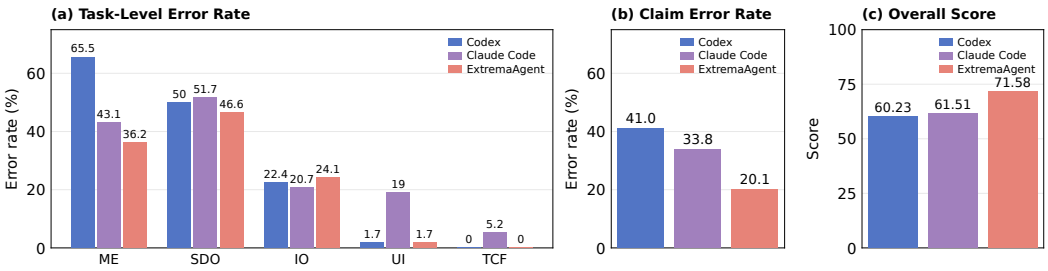


Figure 3: Error analysis of the strongest evaluated agent configurations. (a) Task-level error rates. (b) Percentage of evaluated claims flagged for method failure by the judge, excluding runs without claim-level evaluations. (c) Mean overall scores. Panels (a) and (c) include all 58 tasks per agent. ME: Methodological Error; SDO: Scientific Discovery Omission; IO: Interpretive Omission; UI: Unsupported Inference; TCF: Task Completion Failure.

Table 4: Representative scientific implementation issues verified in baseline execution artifacts. Baselines use Claude Opus 5 for Claude Code and GPT-5.6-sol for Codex.

Case	Implementation issue
Wave activity flux	Claude Code omits the latitude factor $\cos\phi$ from the flux formula, distorting the estimated strength of wave propagation and its variation with latitude.
Atmospheric-river detection	Codex downgrades the core trajectory-tracing algorithm to a geometric approximation. Claude Code introduces a serious implementation error in the trajectory update, which can underestimate river length and cause missed detections.
Potential vorticity	Codex omits the coupling terms between vertical wind shear and horizontal potential-temperature gradients, leaving out their joint contribution to potential vorticity.

Table 5: Ablation results on a task subset. Review and Tools denote the dedicated scientific review component and the provided scientific toolset, respectively

Configuration	Review	Tools	Score ↑
ExtremaAgent	✓	✓	**79.2**
w/o Review	×	✓	74.2
w/o Tools	✓	×	57.1
Codex	×	×	61.7

Removing scientific review lowers the score by 5.0 points, while removing the provided tools leads to a larger drop of 22.1 points. The review ablation suggests that access to scientific tools still benefits from independent checks of method implementation and evidential support. The larger decline without tools, even with review retained, highlights the importance of providing reusable implementations for scientific analyses. Together, these results support the inclusion of both components in ExtremaAgent. The system also exceeds Codex by 17.5 points.

6 Conclusion

Extrema supports long-horizon investigations of extreme weather by integrating scientific data and tools with an autonomous research agent. Its workflow combines planning, experimentation, and independent scientific review to study event evolution and physical mechanisms. ExtremaBench evaluates this capability through 58 rediscovery tasks from 19 papers. Experiments show improvements over the evaluated coding agents across all five event categories, along with fewer methodological errors. Incomplete findings and interpretations remain challenges, highlighting the need to better connect computational evidence with scientific conclusions. Together, the system and benchmark provide a basis for making extreme weather research more reliable and accessible.

AI USE STATEMENT

We used generative AI tools to generate plotting code, polish the manuscript, and assist with creating benchmark tasks and reproducing analyses from the source papers during the construction of ExtremaBench. These contributions were manually verified by the authors. We also used Codex to inspect the working directories of agent runs and assist in identifying and summarizing failure causes. The use of language models for literature information extraction, agent execution, and benchmark evaluation is described in the corresponding sections of the paper. The authors take responsibility for the final benchmark annotations, analyses, and all other content of this work.

ETHICS STATEMENT

Extrema aims to make extreme weather research more accessible by reducing the technical burden of scientific investigation. Automated analyses may produce incorrect physical interpretations or unsupported causal claims, potentially misleading assessments of climate risk. The system links scientific claims to computational evidence and uses an automated review agent to help identify errors. These mechanisms do not guarantee scientific validity, and findings should be independently assessed by domain experts before informing disaster preparedness or climate risk management. ExtremaBench is derived from a limited set of published studies, so benchmark performance should not be interpreted as establishing reliability across all regions, extreme events, or data conditions.

REPRODUCIBILITY STATEMENT

Our experiments use pretrained language models without additional training or fine-tuning. Sections 3.1 and 3.2 describe the scientific computing environment and agent workflow, respectively. The ExtremaBench section describes task construction, reference annotation, task validation, and the evaluation protocol. We will release the source code for Extrema, including the agent implementation and benchmark evaluation scripts, upon acceptance of the paper.

REFERENCES

Malgorzata Blicharska, Richard J Smithers, Magdalena Kuchler, Ganesh K Agrawal, José M Gutiérrez, Ahmed Hassanali, Saleemul Huq, Silvia H Koller, Sugata Marjit, Hassan M Mshinda, et al. Steps to overcome the north–south divide in research relevant to climate change policy and practice. *Nature Climate Change*, 7(1):21–27, 2017.

Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? recall is the bottleneck for parametric factuality. *arXiv preprint arXiv:2602.14080*, 2026.

Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature communications*, 16(1):1919, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, et al. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*, 2025.

Zijie Guo, Jiong Wang, Fenghua Ling, Wangxu Wei, Xiaoyu Yue, Zhe Jiang, Wanghan Xu, Jing-Jia Luo, Lijing Cheng, Yoo-Geun Ham, et al. A self-evolving ai agent system for climate science. *arXiv preprint arXiv:2507.17311*, 2025.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, et al. Autonomous biomedical research with an artificial intelligence agent. *Science*, pp. eadz4351, 2026.

Zhe Jiang, Jiong Wang, Xiaoyu Yue, Zijie Guo, Wenlong Zhang, Fenghua Ling, Wanli Ouyang, and Lei Bai. Ewe: An agentic framework for extreme weather analysis. *arXiv preprint arXiv:2511.21444*, 2025.

Hyeonjae Kim, Chenyue Li, Wen Deng, Mengxi Jin, Wen Huang, Mengqian Lu, and Binhang Yuan. Climateagent: Multi-agent orchestration for complex climate data science workflows. *arXiv preprint arXiv:2511.20109*, 2025.

Benjamin Lamptey, Salah Sahabi Abed, Masilin Gudoshava, Joseph Mutemi, Mary-Jane Bopape, Elijah Adesanya Adefisan, Moudi Pascal Igri, Ibrah Seidou Sanda, Ousmane Ndiaye, Douglas J Parker, et al. Challenges and ways forward for sustainable weather and climate services in africa. *nature communications*, 15(1):2664, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4556–4579, 2025.

Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529):73–119, 1996.

Sjoukje Philip, Sarah Kew, Geert Jan van Oldenborgh, Friederike Otto, Robert Vautard, Karin van Der Wiel, Andrew King, Fraser Lott, Julie Arrighi, Roop Singh, et al. A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2):177–203, 2020.

Hans-Otto Pörtner, Debra C Roberts, Helen Adams, Carolina Adler, Paulina Aldunce, Elham Ali, R Ara Begum, Richard Betts, R Bezner Kerr, Robbert Biesbroek, et al. Climate change 2022: Impacts, adaptation and vulnerability. 2022.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027, 2024.

Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, A Di Luca, Subimal Ghosh, Iskhaq Iskandar, James Kossin, Sophie Lewis, et al. Weather and climate extreme events in a changing climate. *Climate change 2021: The physical science basis: Working group I contribution to the sixth assessment report of the intergovernmental panel on climate change*, pp. 1513–1766, 2021.

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.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Shuo Tang, Jiadong Zhang, Gengxian Zhou, Qizhao Jin, Qinxuan Wang, Yi Hu, Ning Hu, Hongchang Ren, Lingli He, Shiming Xiang, et al. Hvr-met: A hypothesis-verification-replanning agentic system for extreme weather diagnosis. *arXiv preprint arXiv:2603.01121*, 2026.

Nadia Testani, Lucía M Cappelletti, Leandro B Díaz, Camila Prudente, Valentina Rabanal, Julia Mindlin, Reyk Börner, Divya David T, Ismaila Diallo, Inés M Leyba, et al. Balancing earth science careers in an unequal world. *Communications Earth & Environment*, 6(1):22, 2025.

Sumanth Varambally, Marshall Fisher, Jas Thakker, Yiwei Chen, Zhirui Xia, Yasaman Jafari, Ruijia Niu, Manas Jain, Veeramakali Vignesh Manivannan, Zachary Novack, et al. Zephyrus: An agentic framework for weather science. In *International Conference on Learning Representations*, volume 2026, pp. 156654–156704, 2026.

Hao Wang, Jindong Han, Wei Fan, and Hao Liu. Climagent: Llm as agents for autonomous open-ended climate science analysis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21219–21241, 2026a.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, et al. Fire-bench: Evaluating agents on the rediscovery of scientific insights. *arXiv preprint arXiv:2602.02905*, 2026b.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, et al. Researchclawbench: A benchmark for end-to-end autonomous scientific research. *arXiv preprint arXiv:2606.07591*, 2026.

Kaikai Zhang, Xiang Wang, Haoluo Zhao, Nan Chen, Mengyang Yu Jing-Jia Luo, Tao Song, and Fan Meng. Tianji: An autonomous ai meteorologist for discovering physical mechanisms in atmospheric science. *arXiv preprint arXiv:2603.27738*, 2026a.

Lisheng Zhang, Lilong Wang, Xiangyu Sun, Wei Tang, Haoyang Su, Yuehui Qian, Qikui Yang, Qingsong Li, Zhenyu Tang, Haoran Sun, et al. Molclaw: An autonomous agent with hierarchical skills for drug molecule evaluation, screening, and optimization. *bioRxiv*, pp. 2026–04, 2026b.