

EarthLink: A Self-Evolving AI Agent for Climate Science

Zijie Guo^{1,2†}, Jiong Wang^{1,2†}, Xiaoyu Yue^{1,3}, Wangxu Wei^{1,4},
 Zhe Jiang^{1,2}, Wanghan Xu^{1,5}, Ben Fei^{1,6}, Wenlong Zhang¹,
 Xinyu Gu¹, Lijing Cheng⁴, Jing-Jia Luo⁷, Chao Li⁸, Yaqiang Wang⁹,
 Tao Chen², Wanli Ouyang¹, Fenghua Ling^{1*}, Lei Bai^{1*}

¹Shanghai Artificial Intelligence Laboratory.

² Fudan University.

³The University of Sydney.

⁴Institute of Atmospheric Physics, Chinese Academy of Sciences.

⁵Shanghai Jiao Tong University.

⁶The Chinese University of Hong Kong.

⁷Nanjing University of Information Science and Technology.

⁸East China Normal University.

⁹Chinese Academy of Meteorological Sciences.

*Corresponding author(s). E-mail(s): lingfenghua@pjlab.org.cn;
bailei@pjlab.org.cn;

†These authors contributed equally to this work.

Abstract

Modern Earth science is at an inflection point. The vast, fragmented, and complex nature of Earth system data, coupled with increasingly sophisticated analytical demands, creates a significant bottleneck for rapid scientific discovery. Here we introduce EarthLink, the first AI agent designed as an interactive copilot for Earth scientists. It automates the end-to-end research workflow, from planning and code generation to multi-scenario analysis. Unlike static diagnostic tools, EarthLink can learn from user interaction, continuously refining its capabilities through a dynamic feedback loop. We validated its performance on a number of core scientific tasks of climate change, ranging from model-observation comparisons to the diagnosis of complex phenomena. In a multi-expert evaluation, EarthLink produced scientifically sound analyses and demonstrated an analytical competency that was rated as comparable to specific aspects of a human junior researcher's workflow. Additionally, its transparent, auditable workflows and natural language interface empower scientists to shift from laborious manual execution to strategic oversight and hypothesis generation. EarthLink marks a pivotal step towards an efficient, trustworthy, and collaborative paradigm for Earth system research in an era of accelerating global change. The system is accessible at our website <https://earthlink.intern-ai.org.cn>.

1 Introduction

Modern Earth science is fundamentally underpinned by physics with the indispensable supply of rapidly increasing data [1, 2]. The data comes from a wide range of sources, including global model simulations, satellite remote sensing, and in-situ measurements [3–5]. However, these datasets are often fragmented, residing in disparate formats that require specialized analytical tools [6]. This fragmentation, combined with the growing complexity of the scientific questions being addressed, creates a significant bottleneck, slowing the pace of innovative discovery [7].

This challenge is particularly acute in climate change research, where researchers need to extract precise scientific insights from vast amounts of data to guide mitigation and adaptation strategies [8–12]. Central to these efforts are Earth System Models (ESMs) [13, 14], which simulate the complex interactions among the atmosphere, oceans, cryosphere, and biosphere. These models form the backbone of our understanding of climate dynamics and future projections, and are systematically evaluated through the Coupled Model Intercomparison Project (CMIP) [15]. With each successive phase, CMIP has grown in scope and ambition, and CMIP6 alone comprises over a hundred models and thousands of ensemble members, generating petabytes of output data that capture processes ranging from internal variability to forced responses to external factors on regional and global scales [16, 17].

The rapid growth of climate data has greatly advanced scientific understanding, but it has also created a serious bottleneck in analysis. Traditional workflows remain largely manual and fragmented. As data sets become larger and more complex, these workflows struggle to keep pace. Tasks such as comparing models with observations or detecting externally-driven signals now require substantial effort and specialized coding skills. For example, selecting models to estimate equilibrium climate sensitivities can involve exploring hundreds of model simulations from different experiments [18, 19]. This process is often tedious and prone to errors. Such limitations are particularly evident in large-scale assessment efforts, such as those of the Intergovernmental Panel on Climate Change (IPCC), which require extensive manual analyses and coordination across thousands of scientists over several years or even a decade [20].

In response to these analysis challenges, the climate science community has developed powerful diagnostic toolkits, such as ESMValTool [21], CCMVal Diags [22], PCMDI metrics [23], and ILAMB [24]. These tools have played a crucial role in standardising workflows and promoting transparency and reproducibility. Their predefined diagnostic scripts ensure comprehensive and consistent evaluations across models and datasets [19]. However, this predefined nature also limits their flexibility and makes them difficult to adapt to new scientific questions or rapidly changing data streams. Additionally, modifying or extending these tools often requires significant programming expertise and deep familiarity with their architecture. As a result, while they provide standardized and transparent analyses, they remain less suited to the agile and exploratory approaches increasingly needed in contemporary climate research.

The recent rise of Large Language Models (LLMs) and intelligent agents offers a new path forward [25, 26]. By integrating external tools and knowledge sources, LLMs can extend their capabilities beyond language tasks to solving scientific problems, while also

reducing hallucinations and improving output accuracy [27, 28]. These new paradigms have already achieved great successes in automating complex scientific workflows across a range of fields, such as biomedicine [29], chemistry [30], and materials science [31]. However, in Earth sciences, although foundation LLMs have shown preliminary success in domain-specific question answering and tool integration [32–34], the development of agent systems capable of fully automating the workflow of Earth system research remains an open challenge.

Here, we introduce EarthLink, an AI-driven multi-agent system designed to function as an evolving research assistant for Earth system science. EarthLink accepts natural language input to autonomously plan analyses, generate executable code, and interpret scientific results for a number of core tasks such as model-observation comparison. Its dialogue-driven and modular design enables scientists to iteratively refine workflows and extend the system capabilities over time, allowing the agent to continuously evolve alongside user needs. Critically, EarthLink outputs all intermediate scripts, results, and reasoning steps in a transparent manner, transforming scientists from manual executors into supervisors who can focus on formulating original hypotheses and exploring new scientific questions. This shift not only accelerates analysis and validation but also fosters a fundamentally more interactive and efficient research paradigm, substantially advancing the pace and depth of discovery in Earth system science.

2 Results

2.1 Overview of EarthLink

EarthLink is a multi-agent platform that integrates knowledge, data, and computational tools to automate and enhance climate science workflows (Fig. 1). The process begins with intelligent planning stage of human–machine collaboration (Fig. 1A), where the system parses a user’s natural language query to understand their scientific intent. It then consults an expanding Knowledge Library containing scientific literature, domain expertise, and previous analysis records. Based on this information, it generates multiple candidate workflows. A planning summary module selects the optimal analytical pathway and links it to suitable datasets from the Data Library, which includes CMIP6 datasets and multi-domain observations. Throughout this process, scientists are encouraged to supervise and refine the proposed plan to ensure it aligns with established scientific standards.

Following planning, a self-adaptive scientific laboratory translates the plan into executable code (Fig. 1B). It references existing algorithms and tools from a Tool Library and creates new, task-specific scripts. The agent autonomously manages the entire pipeline, from data retrieval and processing to visualization, while also correcting runtime errors and incorporating user feedback to refine outputs. Crucially, every successful task, including the query-code-result triplet, is fed back into the Knowledge and Tool Libraries, creating a virtuous cycle of continuous improvement. Finally, in the synthesis and interpretation stage (Fig. 1C), EarthLink synthesizes the results into coherent, human-readable scientific narratives and visualizations (see Methods for more details).

2.2 Confronting Model Simulations and Observational Data

To systematically evaluate the scientific capabilities of EarthLink, we designed a multi-level benchmarking framework, testing the system against tasks of increasing complexity that are foundational to climate science research (see Extended Data Tables 1 to 3 and Method section for more details). Our results show that EarthLink not only correctly executes standard diagnostic tasks but also demonstrates emergent capabilities in complex physical reasoning and literature-grounded synthesis, revealing its potential as a powerful scientific copilot.

At the first level, EarthLink was evaluated based on its ability to analyze the statistical characteristics of model simulations compared to observations (Fig. 2). In this level, tasks are further divided into three subcategories. The first involves basic climatological analyses, such as computing and visualizing the spatial distribution of surface temperature climatology, interannual variability, mean bias in model simulations, and annual cycles (Extended Data Fig. 1). These tasks demonstrate the EarthLink’s understanding of fundamental climate definitions. The second category requires simple multivariable computations, for example, evaluating cloud radiative effects by combining short-wave and long-wave components to calculate net cloud forcing (Extended Data Fig. 2). This type of task relies on basic data filtering and statistical capabilities. The third category focuses on the integration of multidimensional information within a single domain, exemplified by the calculation of the ocean heat content in different vertical layers (Extended Data Fig. 3). Across these subcategories, EarthLink correctly understood the tasks, produced accurate results, and generated standard diagnostic plots and data products semantically consistent with established scientific literature. Although the aesthetics of the visualizations are still somewhat crude, they do not prevent users from quickly verifying their ideas.

At the second level, we examined EarthLink’s capacity for tasks requiring integration of physical concepts, experimental design, and statistical analysis. For example, we selected a core diagnostic in climate change research, estimating equilibrium climate sensitivity (ECS) and transient climate response (TCR), which assesses how strongly climate models respond to a doubling of atmospheric carbon dioxide (Fig. 2B). EarthLink correctly identified the necessary CMIP6 experiments (e.g., abrupt4xCO₂, 1pctCO₂, and piControl), executed standard regression analyses or metrics calculations, and produced ECS and TCR values that fall within the ranges reported by the IPCC AR6 report. Interestingly, when explicitly instructed to estimate ECS without regression, EarthLink adopted a simple computational approach. It estimated ECS directly from the global temperature change during the quasi-equilibrium period by assuming a linear scaling of radiative forcing with CO₂ concentration. Although the resulting estimate was less accurate than the regression approach, the attempt indicates an understanding of the underlying physical relationship and reveals a physical intuition like humans.

Moving beyond the second level, we assessed EarthLink’s capacity for nuanced scientific reasoning by tasking it with a multifaceted diagnosis of the El Niño–Southern Oscillation (ENSO) in CMIP6 models. The first challenge was focused on ENSO diversity, which is relatively hard to simulate correctly. When prompted with established classification methods (e.g., as in [35, 36]), EarthLink demonstrated a conceptual understanding of the distinction between Central-Pacific (CP) and Eastern-Pacific (EP) types. It then correctly implemented

the core logic of each method and successfully reproduced the characteristic spatial patterns associated with each ENSO flavor. The second challenge tests its ability to analyze ENSO periodicity. EarthLink generated custom code beyond its existing tool library based on its own reasoning. This generated code correctly identified the 2–7 year periodicity of ENSO. These complex and multifaceted tasks highlight EarthLink’s potential to tackle nuanced scientific research questions.

To make the evaluation of EarthLink more objective, we conducted a formal multi-expert review. Five independent climate scientists scored the agent’s outputs using a predefined rubric (see Methods). We assessed three core competencies: the accuracy of the experiment plan, the correctness of the generated code, and the quality of the final visualization. We established a threshold for practical utility based on this rubric. An output scoring 4/5 or higher was deemed practically useful, analogous to the work of a junior researcher, though it may lack efficiency or aesthetic polish. Across the 36 benchmark tasks, EarthLink achieved this level of performance in 16 cases. A breakdown of the scores revealed a clear hierarchy in its current abilities. The agent’s strategic planning is its strongest attribute, followed by code generation, and then visualization quality (Fig. 2D). This result further proves that the EarthLink is suitable as a co-pilot for scientists, especially cross-domain experts, to quickly test ideas.

2.3 More open-ended questions in future projections

To further evaluate the capabilities of EarthLink on open-ended scientific questions, we tested its performance on tasks where ground truth is unavailable, such as future climate projections and impact assessments. These tasks necessitate exploratory analysis of the vast CMIP6 database and a deeper understanding of the basics of climate change.

We first tasked EarthLink with climate change detection and attribution under multiple experiments in DAMIP [37] (Fig. 3A). EarthLink produced outputs consistent with established physical understanding. Its analysis correctly distinguished the roles of natural forcings and anthropogenic drivers like greenhouse gases and aerosols. When attributing global warming since 1901 based on observational records, it identified greenhouse gases as the dominant contributor, with aerosols exerting a partially offsetting cooling effect. Furthermore, in the context of future projection, EarthLink correctly processed the multi-dataset inputs under different Shared Socioeconomic Pathways (SSPs) [38], automatically identifying and visualizing key scenarios including SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5. It presented the results using a multi-model ensemble mean, with shaded regions effectively representing inter-model spread, yielding a clear and scientifically informative visualization. These results demonstrate that EarthLink comprehends a number of core tasks of climate science and can flexibly and efficiently select appropriate data and tools for complex inquiries.

We next tested EarthLink with a more complex and open task involving future projections with constrained regional data. We focus on temperature changes from 2041 to 2060 for multiple cities in different global regions under the SSP2-4.5 scenario. The system was instructed to apply two distinct methodologies to constrain projection uncertainty. The first was a hierarchical emergent constraints (HEC) approach, and the second involved a technique

of aggregating spatial points. Impressively, EarthLink clearly distinguished the two methods and selected appropriate tools for the task. It successfully constrained the model uncertainty and corrected both the ensemble mean and range of projections (Fig. 3B). Compared to the original multi-model distribution, this provided a more refined risk assessment. Interestingly, the HEC script was generated autonomously by EarthLink based on its existing knowledge base. It first derived the appropriate formula from the literature and then generated the corresponding code implementation (Extended Data Fig. 4). Verification showed that the code was highly consistent with the derived formula and closely matched expert-developed scripts as in [39, 40]. This demonstrates EarthLink’s ability to understand geo-spatial data and its potential to deploy complex algorithms to address scientific problems.

Finally, to explore its potential for cross-domain synthesis, we tasked EarthLink for connecting our constrained climate projections (2041-2060) to potential downstream societal impacts. Leveraging the general knowledge of its underlying foundation model, the agent generated qualitative narratives outlining sector-specific risks, such as threats to agricultural yields or stress on public health infrastructure (Fig. 3C). While this initial analysis is illustrative (see Methods for more details), it demonstrates a crucial capability: the potential to bridge the gap between quantitative climate data and policy-relevant discourse. Future work will focus on integrating specialized impact models to enable comprehensive, quantitative risk assessments for climate adaptation planning.

3 Discussion

In this study, we introduce EarthLink, a new AI agent designed to function as an interactive co-pilot for climate science. By integrating a domain-specific knowledge base with dynamic planning and code generation, EarthLink aims to fundamentally alter or speed up the research workflow. Our multi-level evaluation demonstrates its capacity to reproduce foundational scientific analyses and tackle complex, open-ended problems in a manner analogous to a human researcher. This work represents a crucial step towards augmenting human scientific capabilities. It elevates the scientist’s role from a hands-on data processor to a strategic director of research, enabling them to pose more ambitious questions and explore the vast climate data landscape at a high speed previously unattainable. The immediate consequence is a dramatic compression of the research lifecycle, from months to days for complex analyses, thereby accelerating the pace of science discovery itself.

In contrast to the established diagnostic toolkits like ESMValTool, EarthLink introduces a paradigm of flexibility and interactivity. Our vision is not to replace these invaluable tools, but rather to unlock their full potential through a new layer of intelligent orchestration. The fundamental limitation of static toolkits is that their components, while excellent, are siloed within a rigid framework. EarthLink’s architecture offers a solution by treating the trusted, community-vetted functions within these toolkits as individual, callable tools in its library. This transforms a monolithic program into a flexible suite of capabilities. Scientists can dynamically invoke these standardized routines, chain them together, combine them with newly generated code, and apply them to tackle novel scientific questions. As correct scripts are continuously contributed to the system, its analytical repertoire becomes richer and more robust. This vision reframes the future of scientific software, not as a competition between

old and new, but as a composable ecosystem where the reliability of established tools is fused with the agility of an AI agent, creating a sustainable and community-driven path forward.

Despite its capabilities, it is critical to understand EarthLink’s limitations. Its reasoning is fundamentally interpolative, excelling at synthesizing existing knowledge and applying known methods in novel combinations. However, it cannot perform true extrapolative reasoning to formulate entirely new physical theories from first principles. Furthermore, its proficiency is directly tied to the quality of its knowledge base and the clarity of user prompts. A significant risk is the generation of “plausibly wrong” outputs, code that runs without error but produces scientifically incorrect results due to a subtle misinterpretation of a complex request. This risk underscores why our emphasis on transparent and auditable workflows is not just a feature, but a prerequisite for trustworthy AI in science. It reaffirms that the goal is not an infallible virtual scientist, but a powerful yet imperfect tool that demands a deep partnership with a human expert, whose critical judgment in validating results remains the ultimate arbiter of scientific truth.

Finally, our long-term vision extends beyond creating an open-access platform. We propose that EarthLink’s core architecture offers a path to address one of the most persistent challenges in Earth sciences, which is the fragmentation of data. At present, model simulation, satellite, radar and in-situ observation exist in disparate formats and require specialized toolsets, resulting in data silos that hinder holistic analysis. EarthLink’s natural language interface acts as a universal translator. It enables scientists to pose queries such as “Compare the sea ice extent from the NSIDC satellite record with the multi-model mean from the CMIP6 historical simulations”. In this workflow, EarthLink rather than the human bears the burden of finding, accessing, and harmonizing these heterogeneous data sources. Critically, as EarthLink successfully fulfills more of these cross-domain requests, it builds an internal semantic map of the global climate data ecosystem and becomes progressively more efficient in future harmonization tasks. Ultimately, our goal is to develop EarthLink into a global, open, and continuously learning resource that empowers the scientific community to understand and respond to our changing planet.

4 Methods

4.1 Implementation of EarthLink

EarthLink is an LLM-driven agent system that integrates a knowledge base, database, and tool library to tackle complex tasks in Earth science. The system is organized into three core modules: the Planning Module, the Self-Evolving Scientific Lab, and the Multi-Scenario Analysis Module. Each module independently handles a specific aspect of the research workflow, ensuring that complex scientific tasks are executed efficiently and systematically. While the modules employ different foundation LLMs tailored to suit their respective functionalities, they are all based on the OpenAI GPT-4.1 [41] or o4-mini [42] models.

4.1.1 Core Modules

Planning Module. To accomplish complex scientific tasks accurately, detailed and professional planning is essential. The Planning Module in EarthLink is designed to address

this need through a structured, three-stage workflow: input processing, knowledge retrieval, and plan generation. In the input processing stage, the module accepts both natural language queries and scientific literature. When users provide scientific documents, OCR tools such as MinerU [43] are used to parse and convert them into structured formats, enabling effective downstream processing.

During the knowledge retrieval stage, LLMs generate concise summaries of user requirements, which are then embedded as vectors using advanced embedding models. These vectors serve as keys to query the knowledge base, enabling the efficient retrieval of existing, mature plans that match the user’s objectives.

In the plan generation stage, the plan agent not only references the retrieved previous mature plans but also leverages integrated web search tools to access up-to-date scientific definitions and computational workflows from the internet. Additionally, the agent can invoke customized data retrieval tools to query the information about available datasets, including CMIP6 model simulations and relevant observational data. To ensure the diversity of the scientific plan, the agent utilizes stochastic sampling (i.e., a non-zero temperature setting) to generate multiple, diverse candidate plans in parallel. Optionally, these plans can be reviewed and refined by domain experts. If multiple plans are retained, a dedicated plan summary agent aggregates and synthesizes them into a comprehensive final plan.

Through this systematic approach, the Planning Module generates a plan that details the datasets to be utilized, the required data preprocessing steps, the diagnostic methods and computational procedures to be applied, as well as strategies for data visualization. This plan provides clear and thorough guidance for every stage of subsequent scientific analysis.

Self-Evolving Scientific Lab. Once the final plan is produced by the Planning Module, the next step is to translate it into an executable scientific workflow. The Self-Evolving Scientific Lab module is designed to fulfill this purpose by acting as an expert engineer in the Earth science domain. This module consists of two main stages: data preprocessing and scientific diagnosis with visualization.

In the data preprocessing stage, a dedicated preprocessing agent takes charge of loading and processing raw datasets in accordance with the generated plan. The agent leverages the ESMValTool framework to access both CMIP6 and observational data from our database. It performs a range of standardized preprocessing tasks, including regridding all datasets to a common spatial resolution, unifying units, selecting specific spatial or temporal subsets, and applying basic statistical operations such as mean, variance, and anomaly calculations.

Following data preprocessing, a coding agent takes over to perform scientific diagnostics and generate visualizations. The agent begins by parsing the current plan to extract a concise description of the specific diagnostic tasks, i.e., what computations to perform and what figures to plot. This description is embedded into a vector representation, which is then used to query a library of algorithm templates. Based on the retrieved references, the agent composes executable scripts tailored to the task.

During execution, if any errors occur, the agent receives detailed feedback and attempts to automatically debug the code. Once the code runs successfully, it produces both the result data and the figures. These outputs are validated by two additional agents: one for verifying the scientific accuracy of the data, and another for assessing the aesthetic appeal

and informativeness of the visualizations. If the outputs do not meet the required standards, feedback is sent back to the coding agent, which then iteratively revises the code until satisfactory results are achieved. When multiple diagnostic tasks are specified and are not sequentially dependent, the system executes them concurrently to maximize efficiency. The coding agent also has access to web search tools and API documentation retrievers (e.g., for ESMValTool and Iris), enabling retrieval-augmented generation (RAG) [44] to further enhance their performance.

Through this tightly orchestrated and adaptive workflow, the Self-Evolving Scientific Lab module produces high-quality scientific outputs, including interpretable diagnostic results, professional-level plots, and reproducible scripts. Importantly, when outputs are validated by domain experts, the associated codes can be stored in the algorithm template library, thereby improving the system’s ability to handle similar future tasks, exemplifying the self-evolving capability of the framework.

Multi-Scenario Analysis Module. After producing diagnostic results and visualizations, a crucial step in the scientific workflow is transforming these visual outputs into structured textual insights. The Multi-Scenario Analysis Module is designed to complete this final stage by converting computational outcomes into comprehensive scientific interpretations, facilitating downstream applications in decision-making and research dissemination.

This module leverages the image interpretation capabilities of modern LLMs to analyze the visual results generated during the diagnostic phase. For each diagnostic task, a dedicated image analysis agent inspects the corresponding figure and generates a textual description that highlights key patterns, anomalies, or trends, such as temperature increases, regional differences, or inter-model spreads. These per-task reports are produced independently and reflect the specific scientific focus of each visualization.

Once individual image analyses are completed, a report summarization agent integrates all the partial reports into a unified, coherent scientific report. This summary synthesizes findings across tasks, aligns them with the original research objectives, and provides an overall interpretation of the scientific results. The final report can be adapted to various application scenarios, such as climate risk assessment, energy and agricultural planning, environmental policy support, or public health evaluation.

By automating this critical interpretive step, the Multi-Scenario Analysis Module bridges the gap between data analysis and domain-relevant insights, enhancing both the accessibility and impact of the system’s outputs. This ensures that results are not only computationally accurate but also scientifically interpretable and practically useful across diverse decision-making contexts.

4.1.2 Resource Library

Knowledge Library. The Knowledge Library is specifically designed to support the Planning Module by providing a structured and retrievable knowledge base for generating scientific workflows. It semantically indexes important established diagnostic frameworks and expert-validated workflow templates, enabling the planning agents to rapidly access definitions, methodological guidance, and prior analysis plans relevant to user queries. Through continuous incorporation of validated plans and user feedback, the Knowledge

Library evolves over time, improving its ability to match and adapt proven research strategies to novel problems. By integrating both static domain knowledge and dynamically accumulated experience, the Knowledge Library enables the Planning Module to generate thorough, scientifically grounded, and context-aware analysis plans for a wide range of climate research tasks.

Data Library. The current database covers the three main MIPs in CMIP6, including CMIP, DAMIP and ScenarioMIP, covering 33 experiments and more than 70 climate models. Each model contains at least one ensemble member, with monthly and annual data. For observational data, in addition to all the data of obs4MIPs, it also includes data such as HadISST [45], HadCRUT5 [46], GPCP-SG [47], and ERA5 [48]. The total data volume has exceeded 1.5 PB and is continually expanding to include the complete CMIP6 archive and the upcoming CMIP7 datasets. And it will eventually be open to the world to support community-wide collaboration.

Tool Library. The Tool Library integrates a broad suite of open-source tools widely used in climate science for data processing, evaluation, and visualization. It includes established diagnostic and analysis packages such as ESMValTool, PCMDI metrics, CDO, and common Python libraries like xarray, cartopy, Iris, eof, and scikit-learn. The library also supports the incorporation of advanced, expert-developed algorithms—such as Bayesian emergent constraints—and facilitates seamless translation of workflows across multiple programming languages (e.g., NCL, R, MATLAB) into standardized Python scripts using LLM-based code generation. This diverse collection allows EarthLink to flexibly select and orchestrate tools in response to specific analytical requirements. New expert-validated scripts and analytic methods can be continuously added to the library, ensuring that EarthLink remains capable of handling both standardized diagnostics and cutting-edge scientific analysis tasks.

4.2 Evaluation Framework

4.2.1 Task Design

To comprehensively evaluate the scientific capabilities of EarthLink, we established a hierarchical task framework that reflects the increasing complexity of real-world climate research. This framework systematically divides climate analysis workflows into five levels, each targeting specific aspects of scientific reasoning, data processing, and analytical challenge (see Extended Data Tables 1 to 3). The tasks at each level were carefully designed to test EarthLink’s ability to understand, reason, and operate across a spectrum of scientific scenarios.

Level 1: Simple statistical analysis. The first level focuses on essential climatological tasks, such as data retrieval, preprocessing, calculation of annual means, spatial distributions, and interannual variability. These tasks serve as foundational exercises, testing EarthLink’s proficiency in handling basic data structures, performing standard computations, and generating visualizations to support initial model evaluation. With 23 tasks at this level, the emphasis is on routine yet fundamental operations necessary for preliminary climate data analysis and model-observation comparison.

Level 2: Mechanistic diagnosis. The second level introduces moderately complex scientific problems that require mechanistic understanding and integration of multiple datasets. For example, tasks include estimating Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR), which involve understanding the physical diagnostic framework, selecting relevant experiments, and applying statistical tools such as regression analysis. Six tasks were designed at this level to assess EarthLink’s ability to synthesize information across experiments and perform physically grounded diagnostics that move beyond simple descriptive analysis.

Level 3: Complex scientific reasoning. At the third level, tasks demand advanced scientific reasoning and methodological rigor. Here, EarthLink must decompose complex climate phenomena—such as the diversity and periodicity of the El Niño–Southern Oscillation (ENSO)—into logical subtasks. This often requires integrating advanced analytical techniques (e.g., Empirical Orthogonal Function decomposition, composite analysis) and specialized domain knowledge. The six tasks in this category challenge EarthLink to construct and execute extended reasoning chains, demonstrating both technical proficiency and conceptual understanding of complex climate dynamics.

Level 4: Semi-open scientific problem. The fourth level addresses semi-open scientific problems, which more closely simulate the open-ended questions encountered in practical climate research and policy assessment. In these cases, EarthLink is expected to autonomously select appropriate datasets, combine physical insight with adaptive workflows, and apply constraint methods (such as emergent constraints approaches) to refine projections and deliver decision-oriented recommendations. This level consists of a single, highly integrative task designed to probe EarthLink’s capacity for innovation, uncertainty quantification, and workflow adaptation in addressing climate-change imposed challenges.

Level 5: Fully open scientific problems. Finally, the fifth level represents fully open scientific problems—an aspirational category that underscores the long-term potential of autonomous scientific agents. Here, the agent would be expected to independently integrate literature, generate novel ideas, design experimental plans, and solve frontier problems without predefined guidance or ground truth. While EarthLink did not attempt such tasks in the current study, this level sets a vision for future developments in AI-driven scientific discovery.

By structuring the evaluation in this way, we are able to systematically investigate EarthLink’s strengths and limitations across a wide range of research scenarios. This graded approach not only benchmarks current performance but also highlights the pathways for future system enhancements and the evolution of AI agents towards more autonomous, creative scientific inquiry.

4.2.2 Basic Evaluation and Scoring Criteria

To ensure rigorous and reproducible assessment of EarthLink’s performance, we established a comprehensive scoring rubric spanning three core dimensions of scientific practice: (1) Experimental Planning and Method Design, (2) Coding Implementation, and (3) Result Synthesis and Visualization. Each task was evaluated independently by expert reviewers, who assigned scores using a 5-point Likert scale, with clear, criterion-based descriptors for

each level (see Extended Data Table 4). This structured approach emphasizes both technical correctness and scientific reasoning, ensuring alignment with community standards and best practices.

Experimental Planning and Method Design. This dimension evaluates the clarity, scientific rigor, and feasibility of the proposed workflow. A score of 5 indicates that the plan is complete, logically sound, and scientifically executable with no evident flaws. Lower scores reflect increasing levels of methodological ambiguity, omission of key steps, or misunderstanding of core scientific concepts.

Coding Implementation. This dimension assesses whether the generated code is syntactically correct, functionally complete, and aligned with the proposed plan. Top scores are awarded when the system produces robust, ready-to-run scripts that require minimal or no debugging. Points are deducted for significant reliance on self-correction cycles or logical errors in tool usage.

Result Synthesis and Visualization. This dimension focuses on the interpretability, clarity, and presentation quality of outputs—ranging from diagnostic plots to explanatory text. High-scoring outputs exhibit clean visual aesthetics, accurate labeling (units, axis titles, captions), and scientifically coherent narratives suitable for reports or publications. Lower scores reflect graphical or textual inconsistencies, poor formatting, or incomplete result delivery.

Each task was evaluated by three independent domain experts, with scores averaged across reviewers. Reviewers followed the comprehensive rubric shown in Extended Data Table 4, which specifies explicit standards for each score level. This rubric ensures fairness and objectivity in the evaluation of EarthLink across diverse scientific scenarios. Tasks with composite average scores above 4 were categorized as “expert-level”; those between 2.5 and 4 were “research-ready with minor oversight”; and scores below 2 signified either methodological or technical deficiencies that require significant expert intervention.

4.2.3 Estimated Climate Impacts Under Different Scenarios

Benefiting from the powerful foundation model, EarthLink demonstrates the potential to transform quantitative climate data into policy-relevant discourse. Despite the complexity and unpredictability of climate change’s impact on policy, EarthLink employs a divide-and-conquer methodology to simplify the task by decomposing policy demands into domain-specific sub-problems, which are then addressed by specialized domain expert agents.

The resulting dynamic multi-agent system operates akin to a committee, where a “chair” agent first retrieves relevant information from the knowledge base, synthesizes it with domain expertise, determines which domain-specific agents are required, and designs appropriate prompts for each sub-expert agent. For instance, in addressing the topic “Agricultural Development in the Moscow Region from 2041 to 2060,” agents such as an “Environmental Scientist” and an “Agricultural Scientist” would be deemed necessary. In our implementation, the chair agent generates 10 sub-expert agents for each topic. These dynamically generated sub-expert agents take reports from the Multi-Scenario Analysis Module as input, analyze the impacts of climate change within their respective domains based on retrieved domain knowledge, and provide quantitative assessments categorized as either positive or negative.

Finally, the chair agent's primary role is to synthesize the reports from all sub-expert agents into a structured final assessment. This output summarizes key findings and highlights areas of consensus and disagreement (for example, an agronomist might view a longer growing season as positive, while an environmental scientist may warn of increased water stress), and also identifies key uncertainties. For a high-level, heuristic overview, the system can also compute a coarse sentiment score (Fig. 3C), ranging from -1 (predominantly negative assessments) to 1 (predominantly positive). However, the core scientific output remains the detailed report, which provides the rich, multi-faceted analytical landscape needed to inform nuanced policy and decision-making.

Figures

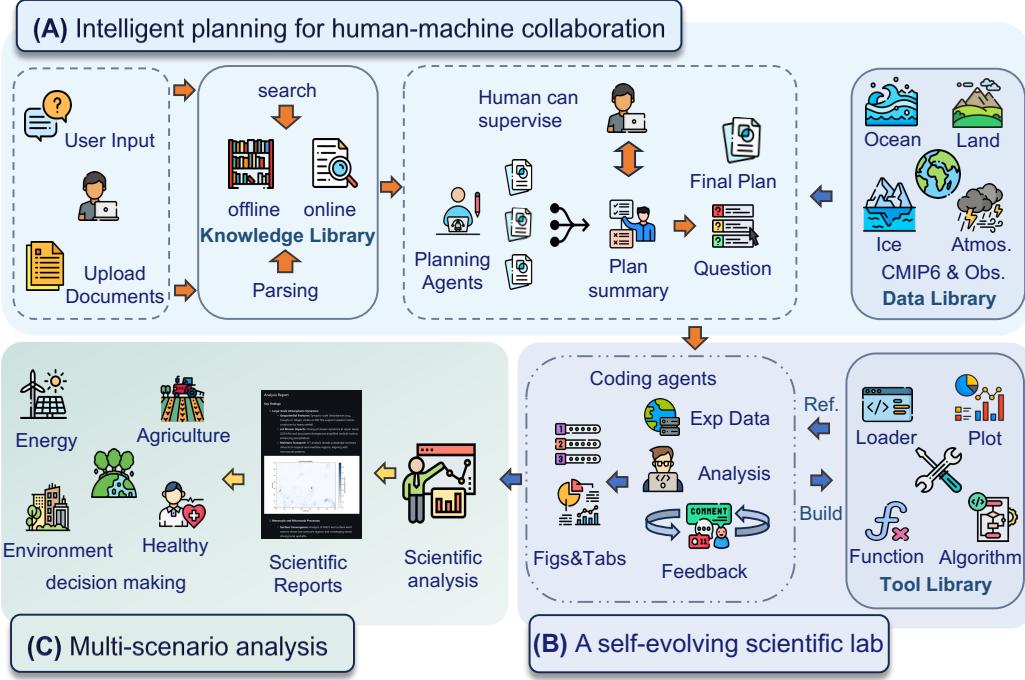
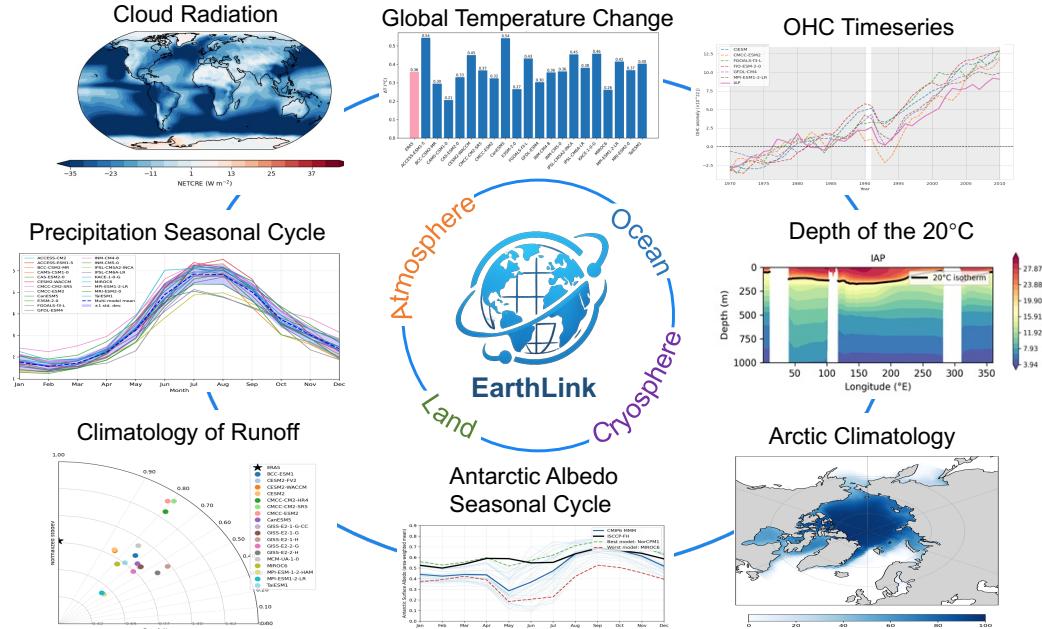
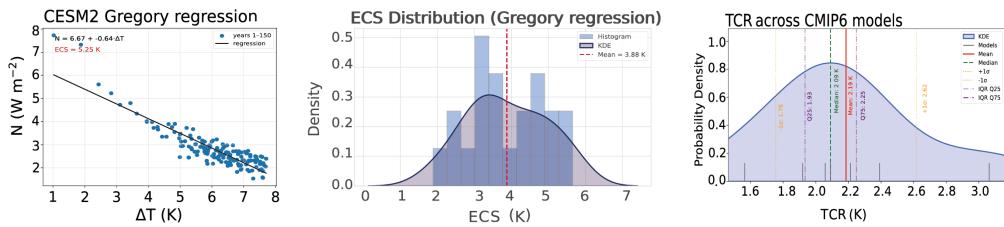


Fig. 1 The EarthLink platform workflow for automated climate data analysis. **(A)** Intelligent planning for human–machine collaboration. Users upload scientific queries or documents that are parsed to extract relevant concepts and goals. Planning agents generate candidate analysis workflows using resources from the Knowledge Library, which compiles published literature and expert knowledge. These plans are iteratively reviewed and refined, with human oversight encouraged to ensure scientific accuracy and task alignment. **(B)** A self-evolving scientific laboratory. The selected experimental plan is transformed into executable code, which autonomously handles data retrieval from the Data Library (including CMIP6 simulations and multi-domain observations), preprocessing, scientific analysis, and visualization. Algorithms from the Tool Library are dynamically composed, and the system applies an autonomous feedback loop for error correction and output refinement, with successful scripts contributing back to the Knowledge and Tool Libraries. **(C)** Multi-scenario analysis and synthesis. The final step involves transforming computational outputs and visualizations into structured, human-readable reports. This process provides scientific interpretations across various domains, including energy, agriculture, environment, and insurance, while delivering insights relevant to policy-making.

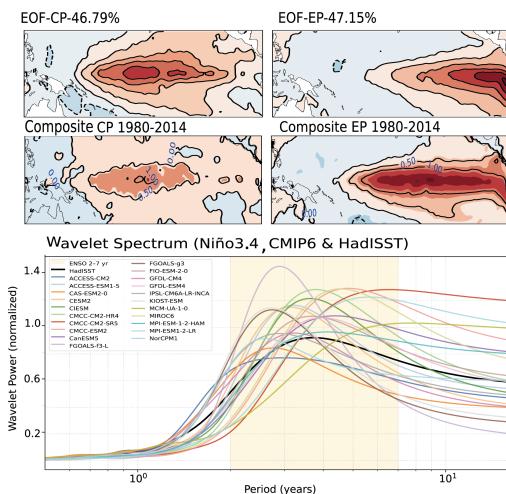
A Level 1: Multisphere Statistical Feature Comparison



B Level 2: Mechanism Diagnosis



C Level 3: Physical Process Diagnosis



D Differentiated Task Scorecard

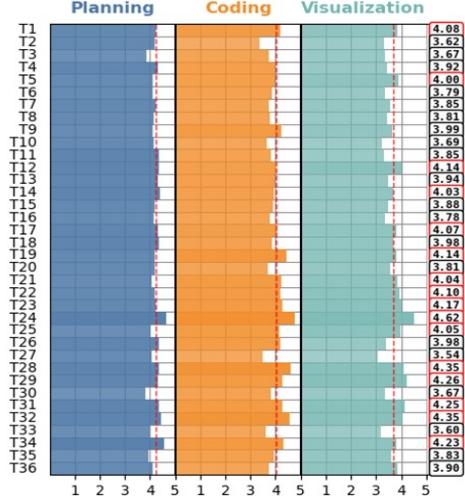
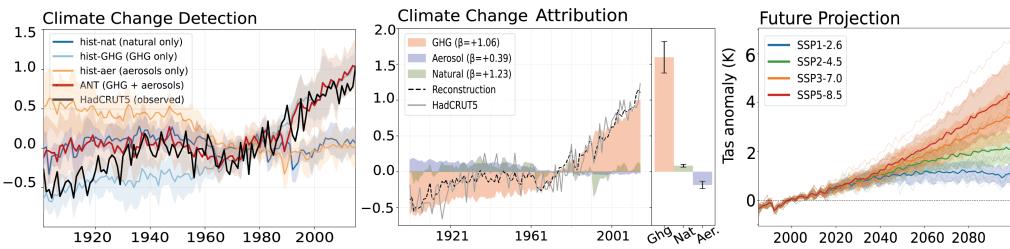


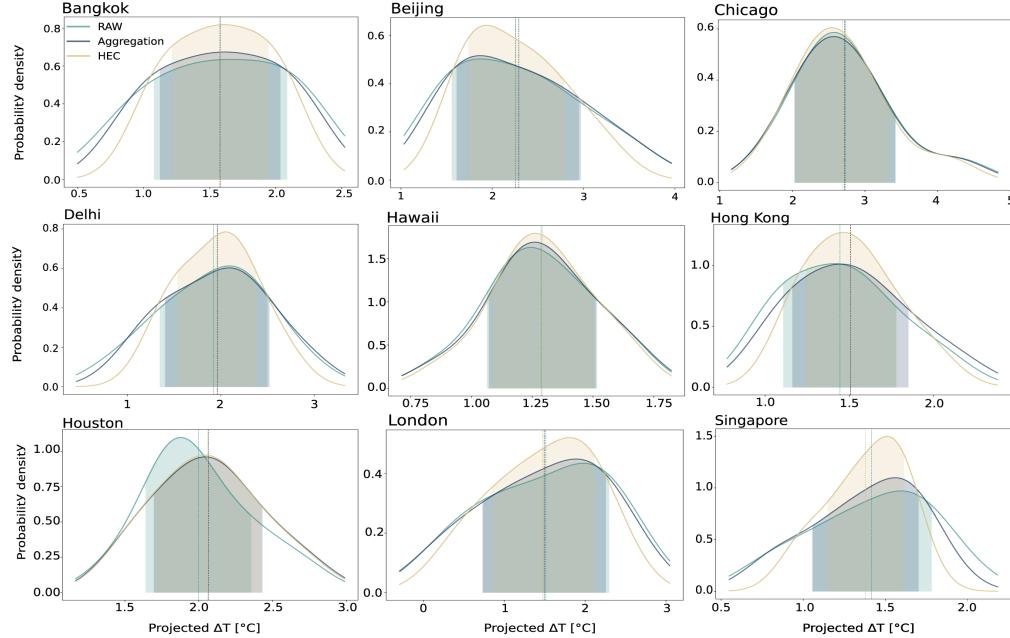
Fig. 2 Multi-level evaluation of EarthLink on a number of core climate analysis tasks. (A)

Level 1: Multisphere statistical feature comparison. EarthLink conducts diagnostic analyses across domains by comparing the CMIP6 simulation of climatological features, such as spatial patterns and variabilities with observations. Examples include seasonal cycles of precipitation, cloud radiative effects, global temperature change, ocean heat content (OHC) timeseries, Arctic mean climatology, 20°C isotherm depth, Antarctic surface albedo, and runoff patterns. (B) Level 2: Mechanistic diagnosis. EarthLink estimates scenario-driven metrics such as equilibrium climate sensitivity (ECS) and transient climate response (TCR), demonstrating its ability to extract relevant datasets and implement standard diagnostic methods. (C) Level 3: Physical process diagnosis. The platform performs advanced analyses such as ENSO diversity classification and period detection, displaying emergent capacity in physical reasoning and chain-of-thought synthesis. (D) Differentiated task scorecard. The system's performance across evaluation tasks is summarized, highlighting relative strengths in planning, coding, and visualization. Note that most of the image elements in (A-C) are directly produced by EarthLink, and the others are only slightly adjusted in layout.

A Future projection & climate change detection & attribution



B Constrained projections of future surface temperature for selected grid



C Impacts of climate change from 2041 to 2060

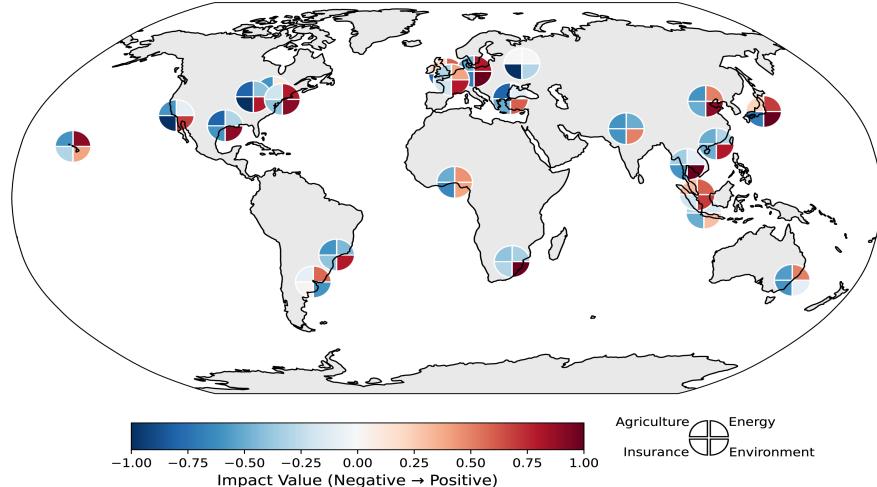
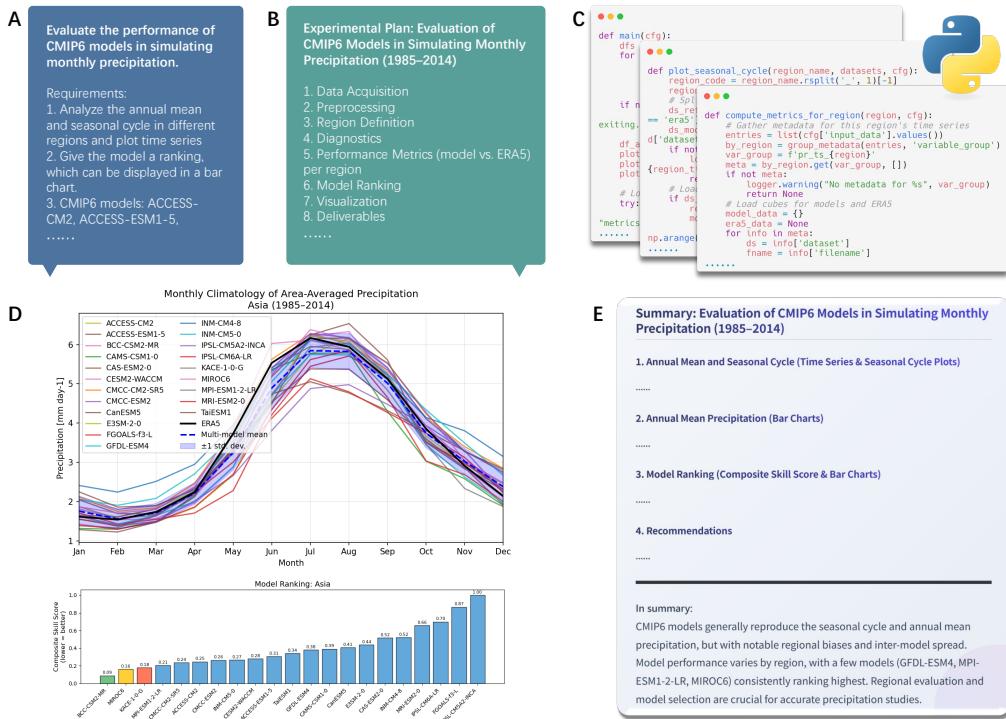
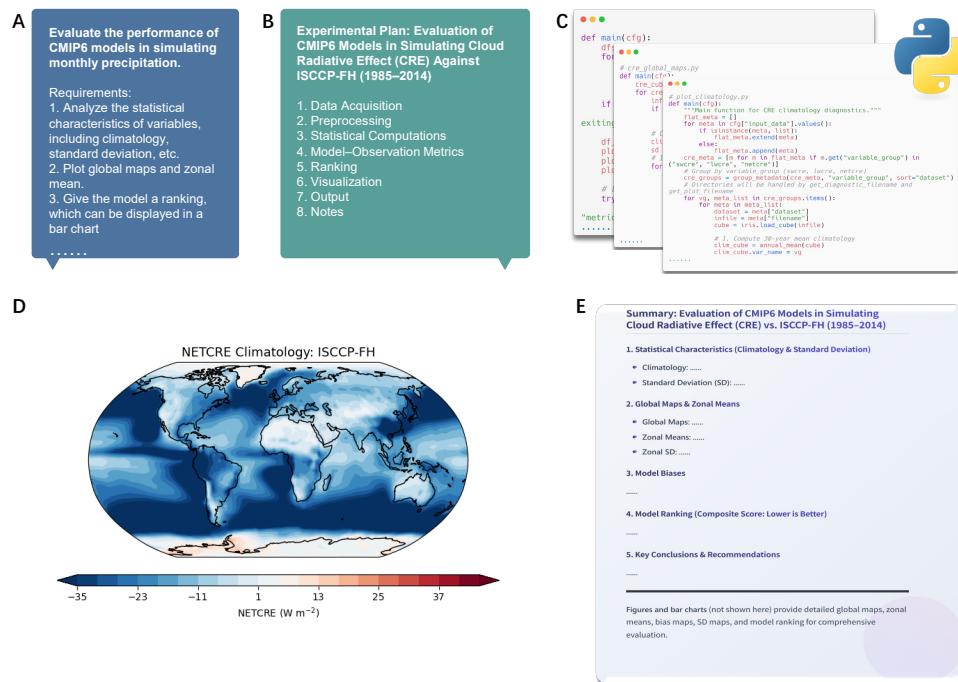


Fig. 3 Application of EarthLink to tackle open-ended and future-oriented climate research challenges. (A) Climate change detection, attribution, and future projection. EarthLink processes multi-model CMIP6 simulations under various experiments, accurately distinguishing between the effects of natural and anthropogenic forcings and generating global temperature anomaly timeseries. (B) Constrained projections of future surface temperature for selected regions. Using hierarchical emergent constraints (HEC) and spatial aggregation approaches, EarthLink reduces projection uncertainty for city-level temperatures under the SSP2-4.5 scenario (2041–2060), demonstrating autonomous code synthesis and the selection of literature-based methods. (C) Impacts of climate change over the next 20 years. The system connects quantified climate projections to sectoral impacts, providing preliminary risk assessments and qualitative narratives for agriculture, energy, insurance and environment, illustrating its ability to bridge scientific data with policy-relevant discussion. Note that most of the image elements in (A–B) are directly produced by EarthLink, and the others are only slightly adjusted in layout.

Extended Data



Extended Data Fig. 1 Evaluation of CMIP6 models in simulating seasonal cycles of monthly precipitation. **(A)** Task definition and diagnostic requirements, including annual mean and seasonal cycles and regional time series analyses. **(B)** Automated planning output from EarthLink, detailing end-to-end workflow from data acquisition to deliverables. **(C)** Example code snippet generated by the system for data processing and analysis. **(D)** Modelled and observed precipitation seasonal cycles over selected regions, with annual mean ranking and model performance comparison. **(E)** Automated textual interpretation of the results, providing a plain-language summary generated by the system.



Extended Data Fig. 2 Benchmarking CMIP6 model simulation of cloud radiative effects (CRE) against ISCCP-FH observations. (A) Task setup, requiring statistical mapping of climatology and variability and model ranking based on performance metrics. (B) End-to-end automated planning by EarthLink, including data acquisition, computation, ranking, and visualization steps. (C) Representative code snippet for automated CRE analysis. (D) Global maps and zonal means of CRE climatology and variance, with bar chart model ranking for comparative evaluation. (E) Automated textual summary of findings, demonstrating EarthLink’s interpretative and reporting capabilities.



Extended Data Fig. 3 Automated evaluation of ocean heat content (OHC) simulation in CMIP6 models. **(A)** Task framing for OHC evaluation using only thetao data to ensure comparability with observed datasets. **(B)** Full experimental plan drafted by EarthLink for preprocessing, calculation, anomaly/time trend analysis, and visualization. **(C)** Example code generated by the system for OHC calculation and plotting. **(D)** Time series of modeled and observed OHC anomalies and spatial patterns of OHC trends. **(E)** Automated text interpretation and summary table, including biases and quantitative comparisons between models and observations.

```

1 def calculate_HEC(obs, sigma_obs, mu_model, xf_model, alpha=0.1):
2     if len(xh_model) != len(xf_model):
3         raise ValueError('HEC: xh_model and xf_model dimension mismatch.')
4
5     obs = np.asarray(obs)
6     xh_model = np.asarray(xh_model)
7     xf_model = np.asarray(xf_model)
8
9     # Step 1: Verify data availability
10    mu_xh = np.nanmean(xh_model, axis=0)
11    mu_xf = np.nanmean(xf_model, axis=0)
12    sigma_xh = np.nanstd(xh_model, ddof=1)
13    sigma_xf = np.nanstd(xf_model, ddof=1)
14
15    valid_ids = ~np.isnan(mu_xh) & ~np.isnan(xf_model)
16    xh_model_valid=xh_model[valid_ids]
17    xf_model_valid=xf_model[valid_ids]
18    rho = np.corrcoef(xh_model_valid, xf_model_valid, rowvar=False)[0, 1]
19
20    # Step 2: Mean and standard deviation of HEC
21    mu_hec = mu_xh + ((rho * sigma_xh * sigma_xf) / (sigma_xh ** 2 + sigma_obs ** 2) * (obs - mu_xh))
22    mu_hec_low = mu_hec - (alpha/2) * (1 - rho ** 2 / (1 + sigma_obs ** 2 / sigma_xh ** 2))
23    mu_hec_high = mu_hec + (alpha/2) * (1 - rho ** 2 / (1 + sigma_obs ** 2 / sigma_xh ** 2))
24    sd_hec = np.sqrt((sigma_xh ** 2 + rho ** 2 * sigma_xf ** 2) / (1 - rho ** 2))
25
26    # Step 3: Constrained future climate estimates
27    xf_hec = mu_hec
28    cl_hec_low = norm.ppf(alpha/2, loc=mu_hec, scale=sd_hec)
29    cl_hec_high = norm.ppf(1 - alpha/2, loc=mu_hec, scale=sd_hec)
30    cl_hec = pd.DataFrame({'cl_hec_low': cl_hec_low, 'cl_hec_high': cl_hec_high})
31
32    return {
33        'xf_hec': xf_hec,
34        'cl_hec': cl_hec,
35        'mu_hec': mu_hec,
36        'sd_hec': sd_hec
37    }
38
39
40
41

```

Data Preprocess

Core Calculation of HEC

Core Calculation of HEC

Extended Data Fig. 4 Comparison of code generated by human expert and EarthLink.

Extended Data Table 1 The task complexity levels and representative capabilities of the EarthLink in this study.

Levels	Description	Number of Tasks
Level 1: Simple statistical analysis	Performs basic climatological tasks, including data retrieval, preprocessing, calculation of annual means, spatial distributions, and interannual variability, with visualizations supporting initial model evaluation.	23
Level 2: Mechanistic diagnosis	Solve moderately complex climate problems, such as estimating Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR), by understanding the physical diagnostic framework, invoking common analyses of multiple experiment datasets and applying simple mathematical tools.	6
Level 3: Complex scientific reasoning	Decomposes complex climate analyses into clear, logical subtasks. Integrates advanced analytical methods (e.g., Empirical Orthogonal Function (EOF), composite analysis) with specialized knowledge to study complex phenomena such as El Niño-Southern Oscillation (ENSO) diversity, requiring rigorous methodology and extended reasoning chain.	6
Level 4: Semi-open scientific problem	Automatically selects appropriate datasets based on detailed problem descriptions, combining physical understanding with adaptive workflows to address open-ended climate problems. Applies constraint methods (e.g., emergent constraints) to identify the constraint factor and produce constrained forecasts and preliminary decision-oriented recommendations.	1
Level 5: Fully open scientific problem	Independently integrates literature based on the given topic or question, generates new ideas, designs experimental plans, and solves problems without requiring predefined guidance.	0

Extended Data Table 2 The evaluation L1 Tasks for EarthLink in our study.

Task ID	Task Level	Description	Variable
Task 1	L1		Cloud radiative effect
Task 2	L1		Evaporation
Task 3	L1	Comparison of global multi-model	Precipitation
Task 4	L1	simulated climatology and variability	Sea ice concentration
Task 5	L1	for 1985–2014 with observations.	Surface air temperature
Task 6	L1		Surface full wind
Task 7	L1		Thermocline
Task 8	L1		Evaporation
Task 9	L1	Evaluation of global multi-model	Precipitation
Task 10	L1	simulations of annual mean and	Runoff
Task 11	L1	seasonal cycles.	Sea ice concentration
Task 12	L1		Surface air temperature
Task 13	L1		Ocean heat content
Task 14	L1	Comparison of global multi-model	Precipitation
Task 15	L1	simulated changes and trends from	Runoff
Task 16	L1	1991–2010 to 1970–1990 with	Sea ice concentration
Task 17	L1	observations.	Surface air temperature
Task 18	L1		Surface full wind
Task 19	L1	Comparison of CMIP simulations	Antarctica surface albedo
Task 20	L1	in specific regions with	Sea surface temperature
Task 21	L1	visualizations for selected areas.	Runoff
Task 22	L1	Comparison of differences between	Cloud radiative effect
Task 23	L1	various observations.	Precipitation

Extended Data Table 3 The evaluation tasks for EarthLink in our study from L2 to L4.

Task ID	Task Level	Description	Variable
Task 24	L2	Evaluate Equilibrium Climate Sensitivity (ECS) in different models using methods of self-choice.	ECS
Task 25	L2	Evaluation of ECS across different models using given methods, such as Gregory regression [49].	ECS
Task 26	L2	Evaluate Transient Climate Response (TCR) in different models	TCR
Task 27	L2	Comparison of climate changes	Precipitation
Task 28	L2	under different future scenarios	Surface air temperature
Task 29	L2	Detection of global climate change using DAMIP Experiments.	Surface air temperature
Task 30	L3	Evaluation of Atlantic Meridional Overturning Circulation (AMOC) simulation capability in CMIP6 models	AMOC
Task 31	L3	Evaluation of ENSO diversity simulation in CMIP6 using ENSO classification method in [35].	ENSO diversity
Task 32	L3	Evaluation of ENSO diversity simulation in CMIP6 using ENSO classification method in [36].	ENSO diversity
Task 33	L3	Evaluation of ENSO period simulated in CMIP6 using wavelet analysis.	ENSO periods
Task 34	L3	Quantification of contributions from different forcing factors to global mean temperature warming from 1901 to 2015.	Climate change attribution
Task 35	L3	Heat budget analysis	Heat budge
Task 36	L4	Using emergent constraints to constrain temperature trends in Africa over the next 20 years.	Future projection constraint

Extended Data Table 4 Comprehensive Evaluation Reference Table (Full Score 5 Points).

Score	Experimental Planning and Method Design	Code Implementation	Result Synthesis and Visualization
5 points	Complete planning, scientifically rigorous, clear logic, with practical feasibility; all steps closely tied to objectives, no redundancy or errors; accurate method description; comprehensive output of results.	Able to accurately call data and tools, results fully conform to planning and experimental design; high code quality, no debugging or only minimal debugging is needed for successful execution.	High consistency between text and figures, visually pleasing and standardized charts, rich textual information and clear logic; correct axis, units, scope, and annotation in charts; overall presentation meets publication or reporting standards, effectively supporting research conclusions.
4 points	Overall structure is reasonable, task decomposition is clear, but some non-core redundant calculations, inefficient processes, or steps weakly related to objectives exist, without affecting overall results.	Able to complete all tasks, but requires relatively more debugging (e.g., debug rounds >15), final results are correct and do not affect execution effect; non-fatal errors exist but can be corrected.	Text and figures are consistent, content is correct, but minor issues exist (e.g., slight color mismatches, inconsistent fonts, missing individual labels), which do not affect understanding; text expression is relatively complete but can be further optimized; overall quality is good, needing only slight polishing to reach an excellent level.
3 points	Able to design correct tools or implement correct experimental steps, complete main tasks, but lacks detailed explanation or sufficient interpretation of selection rationale, possibly affecting result rigor or reproducibility.	Able to generate intermediate data, and intermediate data basically meet expectations; core code tools also basically meet expectations, but only partial task goals are completed, or unable to fully implement the entire process.	Text and figures are basically consistent, but image aesthetics and textual information amount are low; obvious errors appear in images including axis labels, drawing scope, units, etc.
2 points	Able to find data needed for corresponding tasks, understand basic data format and rules, but cannot form effective analysis process or experimental pathway.	Able to load data and generate some tools or intermediate variables, but output results obviously do not meet requirements or contain logical errors, unable to support further analysis.	Able to provide images and reports, but reports contain obvious problems of inconsistency between text and figures.
1 point and below	Planning is incomplete or contains major errors, difficult to guide actual execution; task goals are misunderstood; data or methods seriously do not meet requirements.	Unable to load data, unable to generate required tools or functions; code has major errors, unable to run or continue the task.	Unable to provide images or comprehensive summary reports, unable to provide images meeting requirements.

Acknowledgements

We acknowledge the World Climate Research Programme (WCRP), which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access and the open source scripts repository from Earth System Modelling and Observation (ESMO) group. In addition, we would like to express our sincere gratitude to the domain experts and research scientists in the climate community, whose insights helped us to conduct the evaluation. We also thank the website development team (Shaowei Hou and Zheng Nie, et al.) and the data reserve team (Dong Zheng and Qihao Zheng, et.al.) of Shanghai Artificial Intelligence Laboratory for helping us quickly build the platform and database. J.-J L is supported by National Natural Science Foundation of China (Grant No. 42088101 and 42030605).

References

- [1] Overpeck, J.T., Meehl, G.A., Bony, S., Easterling, D.R.: Climate data challenges in the 21st century. *Science* **331**(6018), 700–702 (2011)
- [2] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, F.: Deep learning and process understanding for data-driven earth system science. *Nature* **566**(7743), 195–204 (2019)
- [3] Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., Blaschke, T.: Big earth data: disruptive changes in earth observation data management and analysis? *International Journal of Digital Earth* **13**(7), 832–850 (2020)
- [4] Dobler, A., Benestad, R.E., Lussana, C., Landgren, O.: Cmip6 models project a shrinking precipitation area. *Npj Climate and Atmospheric Science* **7**(1), 239 (2024)
- [5] Xie, S.-P.: Satellite observations of cool ocean–atmosphere interaction. *Bulletin of the American Meteorological Society* **85**(2), 195–208 (2004)
- [6] Guo, H., Liu, Z., Jiang, H., Wang, C., Liu, J., Liang, D.: Big earth data: A new challenge and opportunity for digital earth’s development. *International Journal of Digital Earth* **10**(1), 1–12 (2017)
- [7] Park, M., Leahey, E., Funk, R.J.: Papers and patents are becoming less disruptive over time. *Nature* **613**(7942), 138–144 (2023)
- [8] Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., *et al.*: Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change **2**(1), 2391 (2021)
- [9] Team, C.W., Lee, H., Romero, J., *et al.*: IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC

Geneva, Switzerland (2023)

- [10] Cheng, L., Abraham, J., Trenberth, K.E., Reagan, J., Zhang, H.-M., Storto, A., Von Schuckmann, K., Pan, Y., Zhu, Y., Mann, M.E., et al.: Record high temperatures in the ocean in 2024. *Advances in Atmospheric Sciences*, 1–18 (2025)
- [11] Kennedy, J., Trewin, B., Betts, R., Thorne, P., Foster, P., Siegmund, P., Ziese, M., Mishra, S., Uhlenbrook, S., Alvar-Beltran, J., et al.: State of the climate 2024. update for cop29 (2024)
- [12] World Meteorological Organization (WMO): State of the Global Climate 2024, Wmo-no. 1368 edn. World Meteorological Organization (WMO), Geneva (2025). <https://library.wmo.int/records/item/69455-state-of-the-global-climate-2024>
- [13] Stute, M., Clement, A., Lohmann, G.: Global climate models: Past, present, and future. *Proceedings of the National Academy of Sciences* **98**(19), 10529–10530 (2001)
- [14] Heinze, C., Eyring, V., Friedlingstein, P., Jones, C., Balkanski, Y., Collins, W., Fichefet, T., Gao, S., Hall, A., Ivanova, D., et al.: Esd reviews: Climate feedbacks in the earth system and prospects for their evaluation. *Earth System Dynamics* **10**(3), 379–452 (2019)
- [15] Meehl, G.A., Boer, G.J., Covey, C., Latif, M., Stouffer, R.J.: Intercomparison makes for a better climate model. *Eos, Transactions American Geophysical Union* **78**(41), 445–451 (1997)
- [16] Taylor, K.E., Stouffer, R.J., Meehl, G.A.: An overview of cmip5 and the experiment design. *Bulletin of the American meteorological Society* **93**(4), 485–498 (2012)
- [17] Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E.: Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development* **9**(5), 1937–1958 (2016)
- [18] Andrews, T., Gregory, J.M., Webb, M.J., Taylor, K.E.: Forcing, feedbacks and climate sensitivity in cmip5 coupled atmosphere-ocean climate models. *Geophysical research letters* **39**(9) (2012)
- [19] Meehl, G.A., Senior, C.A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R.J., Taylor, K.E., Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models. *Science Advances* **6**(26), 1981 (2020)
- [20] Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P.W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., et al.: Ipcc, 2023: Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change [core writing team, h. lee and j. romero (eds.)]. ipcc, geneva, switzerland. (No Title) (2023)
- [21] Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat,

- O., Brötz, B., Caron, L.-P., *et al.*: Earth system model evaluation tool (esmvaltool) v2.0—an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of earth system models in cmip. *Geoscientific Model Development* **13**(7), 3383–3438 (2020)
- [22] Gettelman, A., Eyring, V., Fischer, C., Shiona, H., Cionni, I., Neish, M., Morgenstern, O., Wood, S., Li, Z.: A community diagnostic tool for chemistry climate model validation. *Geoscientific Model Development* **5**(5), 1061–1073 (2012)
- [23] Lee, J., Gleckler, P.J., Ahn, M.-S., Ordonez, A., Ullrich, P.A., Sperber, K.R., Taylor, K.E., Planton, Y.Y., Guilyardi, E., Durack, P., *et al.*: Systematic and objective evaluation of earth system models: Pcmdi metrics package (pmp) version 3. *Geoscientific Model Development* **17**(9), 3919–3948 (2024)
- [24] Collier, N., Hoffman, F.M., Lawrence, D.M., Keppel-Aleks, G., Koven, C.D., Riley, W.J., Mu, M., Randerson, J.T.: The international land model benchmarking (ilamb) system: design, theory, and implementation. *Journal of Advances in Modeling Earth Systems* **10**(11), 2731–2754 (2018)
- [25] Wang, Z., Chu, Z., Doan, T.V., Ni, S., Yang, M., Zhang, W.: History, development, and principles of large language models: an introductory survey. *AI and Ethics* **5**(3), 1955–1971 (2025)
- [26] Zhang, Q., Ding, K., Lv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., *et al.*: Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys* **57**(6), 1–38 (2025)
- [27] Wang, Z., Cheng, Z., Zhu, H., Fried, D., Neubig, G.: What are tools anyway? a survey from the language model perspective. arXiv preprint arXiv:2403.15452 (2024)
- [28] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., Li, Q.: A survey on rag meeting llms: Towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6491–6501 (2024)
- [29] Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Zhang, J., Di, Y., *et al.*: Biomni: A general-purpose biomedical ai agent. bioRxiv, 2025–05 (2025)
- [30] Boiko, D.A., MacKnight, R., Kline, B., Gomes, G.: Autonomous chemical research with large language models. *Nature* **624**(7992), 570–578 (2023)
- [31] Kang, Y., Kim, J.: Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature communications* **15**(1), 4705 (2024)
- [32] Bi, Z., Zhang, N., Xue, Y., Ou, Y., Ji, D., Zheng, G., Chen, H.: Oceangpt: A large language model for ocean science tasks. arXiv preprint arXiv:2310.02031 (2023)
- [33] Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Xu, Y., Fu, L., Zhang, W., Wang,

- X., Zhou, C., *et al.*: K2: A foundation language model for geoscience knowledge understanding and utilization. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, pp. 161–170 (2024)
- [34] Zhang, Y., Wei, C., He, Z., Yu, W.: Geogpt: An assistant for understanding and processing geospatial tasks. International Journal of Applied Earth Observation and Geoinformation **131**, 103976 (2024)
- [35] Kao, H.-Y., Yu, J.-Y.: Contrasting eastern-pacific and central-pacific types of enso. Journal of Climate **22**(3), 615–632 (2009)
- [36] Kug, J.-S., Jin, F.-F., An, S.-I.: Two types of el niño events: cold tongue el niño and warm pool el niño. Journal of climate **22**(6), 1499–1515 (2009)
- [37] Gillett, N.P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B.D., Stone, D., Tebaldi, C.: The detection and attribution model intercomparison project (damip v1. 0) contribution to cmip6. Geoscientific Model Development **9**(10), 3685–3697 (2016)
- [38] O'Neill, B.C., Tebaldi, C., Van Vuuren, D.P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., *et al.*: The scenario model intercomparison project (scenariomip) for cmip6. Geoscientific Model Development **9**(9), 3461–3482 (2016)
- [39] Li, C., Sun, Q., Wang, J., Liang, Y., Zwiers, F.W., Zhang, X., Li, T.: Constraining projected changes in rare intense precipitation events across global land regions. Geophysical Research Letters **51**(3), 2023–105605 (2024)
- [40] Li, C., Zwiers, F.W., Zhang, X., Fischer, E.M., Du, F., Liu, J., Wang, J., Liang, Y., Li, T., Yuan, L.: Constraining the entire earth system projections for more reliable climate change adaptation planning. Science Advances **11**(9), 5346 (2025)
- [41] OpenAI: Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1>. Accessed July 2025 (2025)
- [42] OpenAI: Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini>. Accessed July 2025 (2025)
- [43] Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
- [44] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
- [45] Rayner, N.A., Parker, D.E., Horton, E., Folland, C.K., Alexander, L.V., Rowell, D., Kent, E.C., Kaplan, A.: Global analyses of sea surface temperature, sea ice, and

night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres* **108**(D14) (2003)

- [46] Morice, C.P., Kennedy, J.J., Rayner, N.A., Winn, J.P., Hogan, E., Killick, R.E., Dunn, R.J., Osborn, T.J., Jones, P.D., Simpson, I.R.: An updated assessment of near-surface temperature change from 1850: The hadcrut5 data set. *Journal of Geophysical Research: Atmospheres* **126**(3), 2019–032361 (2021)
- [47] Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., *et al.*: The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979–present). *Journal of hydrometeorology* **4**(6), 1147–1167 (2003)
- [48] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al.: Era5 monthly averaged data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) **10** (2023)
- [49] Gregory, J.M., Ingram, W.J., Palmer, M., Jones, G.S., Stott, P., Thorpe, R., Lowe, J.A., Johns, T., Williams, K.: A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical research letters* **31**(3) (2004)