

<https://doi.org/10.1038/s44168-025-00300-y>

Transforming climate services with LLMs and multi-source data integration



Ivan Kuznetsov¹ ✉, Antonia Anna Jost¹, Dmitrii Pantiukhin¹, Boris Shapkin¹, Thomas Jung^{1,2} & Nikolay Koldunov¹

Integrating Large Language Models (LLMs) with climate model data, scientific literature, and unstructured text enables a new generation of climate information systems that deliver accurate, localized, and context-aware insights. Our primary objective is to develop and evaluate ClimSight, a scalable platform that turns complex heterogeneous data into actionable information. We augment LLMs with Retrieval Augmented Generation, a method that retrieves relevant climate models and reports at query time to ground responses. An agent-based architecture orchestrates specialized modules that route and process user queries with task-specific tools. Real-world evaluations compare multiple LLM configurations and analyze trade-offs between speed, cost, and accuracy. Results show improved scalability and precision in climate assessments, democratizing access to localized information. This paradigm shift equips stakeholders in agriculture, urban planning, disaster management, and policy with effective tools for forward planning and risk management.

Climate change is increasingly affecting societies worldwide, leading to a growing need for accurate and localized climate assessments. Traditional climate information systems often struggle to provide tailored insights due to limitations in data granularity, interpretability, and contextual relevance. Moreover, the widespread impacts of climate change mean that people across the globe are facing more climate-related questions regarding future planning - ranging from agricultural decisions to urban infrastructure adaptation. This demand necessitates scalable solutions capable of addressing climate-related inquiries at both local and global levels.

Recent advances in artificial intelligence, particularly in Large Language Models (LLMs), have opened new possibilities for enhancing climate services. LLMs, like those developed by OpenAI and others¹⁻³, are capable of understanding complex language patterns, synthesizing information from diverse sources, and generating human-like text. However, standard LLMs face significant challenges when interpreting specialized climate data, mainly due to their general-purpose design and lack of domain-specific training⁴. Several AI-driven climate information platforms have emerged in recent years. For example, ChatClimate⁵ grounds a GPT-4 model's responses in the IPCC Sixth Assessment Report to ensure up-to-date, scientifically accurate answers, and ClimateQ&A⁶ leverages an LLM with a knowledge base from IPCC/IPBES reports. Other efforts include ClimateGPT⁷, which uses climate-specific LLMs trained on interdisciplinary literature, and ChatNetZero⁸, a retrieval-augmented chatbot for climate pledges. Additionally, The Climate Explorer⁹, while not AI-based, provides tools and information to help build climate resilience. These

systems represent important advances in climate information delivery, yet they often rely on specific document corpora or narrow use cases, which hinder their ability to provide accurate, context-aware climate assessments at scale.

To meet these challenges, there is a pressing need for climate information systems that are not only accurate and localized but also scalable to address the increasing volume of climate-related questions worldwide. ClimSight¹⁰ aims to bridge this gap by leveraging advanced LLMs and integrating diverse data sources, including unstructured text and domain-specific databases, ensuring comprehensive, context-aware, and scalable climate assessments.

ClimSight integrates advanced LLMs with high-resolution geographical and climate data, providing localized climate insights tailored to specific user needs and activities. The system leverages Retrieval-Augmented Generation (RAG) techniques^{11,12} and an agent-based architecture¹³⁻¹⁵ to overcome the limitations of general-purpose models. By incorporating domain-specific knowledge and high-resolution spatial data¹⁶⁻¹⁸, ClimSight delivers accurate, reliable, and reproducible climate assessments at a granular, location-specific level.

The developed approach distinguishes itself from existing climate information systems through several key advancements:

- Integration of LLMs: The proposed methodology leverages state-of-the-art LLMs to interpret complex climate-related queries, synthesizing information from diverse data sources, including scientific reports, IPCC documents, and geographical databases. This allows the system

¹Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. ²Institute of Environmental Physics, University of Bremen, Bremen, Germany. ✉e-mail: Ivan.Kuznetsov@awi.de

to provide nuanced, context-aware insights that go beyond traditional numerical models.

- **Multi-Source Data Integration:** Unlike conventional systems that rely solely on structured climate data, our platform integrates information from multiple sources, including unstructured text, domain-specific literature, and online databases. This approach ensures comprehensive climate assessments tailored to specific user needs.
- **RAG System:** ClimSight enhances LLMs' contextual understanding by retrieving relevant knowledge from external sources, ensuring evidence-based and contextually accurate answers. This approach bridges the gap between static datasets and dynamic, context-aware insights.
- **Agent-Based Architecture:** By employing a modular, agent-based design, the framework ensures scalability, flexibility, and improved system efficiency. Specialized agents handle distinct tasks, such as data retrieval, contextual understanding, and result synthesis, leading to more accurate and coherent outputs.
- **Practical Use Cases and Real-World Applications:** The solution is validated through practical examples, such as assessing climate risks for specific agricultural activities and urban planning scenarios. These use cases demonstrate its potential to support climate-informed decision-making across various sectors.

The target audience includes researchers, providers of climate services, policymakers, agricultural planners, urban developers, and other stakeholders who require detailed climate information. ClimSight is designed to democratize access to climate data, empowering users with actionable insights relevant to their specific contexts.

This paper provides a comprehensive overview of the system's architecture, methodologies, and evaluation framework. It demonstrates the platform's effectiveness through real-world examples and comparative analyses, highlighting its contribution to climate services. By addressing the limitations of traditional systems and leveraging state-of-the-art AI

techniques, ClimSight contributes to the evolving landscape of digital climate twins and next-generation climate information systems.

Results

This section presents the performance outcomes of ClimSight, including usage examples with two distinct models and their corresponding visualizations.

As an example of the concept's capabilities, we tested its response to a question previously explored in ref. 10: "What are the potential risks and benefits of climate change for growing wheat?" at a location in Morocco (31.6912N, 8.1098W). The answer was generated using GPT-4o and is shown in Box 1. To illustrate how the response was created, Fig. 1 presents a visual representation of the process. This diagram shows the different steps involved, including retrieving relevant information, formatting the data, and reasoning through the response. The process relies on several specialized components, each performing a specific role (In parentheses, the names of the components are indicated in Fig. 1):

- **Location Requests (not in LangGraph):**
Retrieves addresses and land/water status from coordinates
- **Doorman (intro_agent):**
Introductory interface of the system
- **API & Local Data Requests (zero_rag_agent):**
Retrieves location-specific data from various datasets.
- **IPCC (ipcc_rag_agent):**
Retrieves climate-related information from IPCC reports.
- **Reports (general_rag_agent):**
Pulls additional context from broader climate literature.
- **Climate model data (data_agent):**
Derive data from climate models output
- **Smart agent (smart_agent):**
Retrieve and synthesize information from multiple heterogeneous data sources.
- **Combine (combine_agent):**

Box 1 | ClimSight with gpt-4o: question about wheat near by Marrakech

The potential risks and benefits of growing wheat in Saâda, Morocco, under the influence of climate change can be evaluated with regard to the prevailing and projected climatic and environmental conditions.

Risks:

1. **Temperature Increases:** The data indicates a significant rise in temperatures over the decades. For instance, the average temperature in August is projected to rise from 35.11 °C in the 2020s to 36.77 °C in the 2040s. Such temperatures exceed the optimal range for wheat (15 °C to 23 °C) and approach the absolute maximum temperature threshold of 27 °C for wheat cultivation. This increase can lead to heat stress, negatively impacting wheat yields during critical growth phases.
2. **Precipitation Deficits:** The annual precipitation levels are well below the optimal requirement for wheat, which ranges from 750 mm to 900 mm. The highest monthly precipitation is expected to be just 10.05 mm in December by the 2040s, indicating a severe shortfall and potential for drought stress. Wheat requires consistent moisture, especially during germination and early growth stages, which this region may not be able to provide without substantial irrigation.
3. **Flood Risks:** Historical data shows that the region has experienced floods (in 1995, 1996, 1997, and 2008), which could pose additional risks such as soil erosion or waterlogged conditions that are detrimental to wheat.

Benefits:

1. **Extended Growing Season:** The rise in temperatures might allow for a longer growing season, which could be beneficial if managed correctly with appropriate wheat varieties that are heat-tolerant and have shorter growing cycles.
2. **CO₂ Fertilization Effect:** Increased atmospheric CO₂ levels can enhance photosynthesis in C₃ plants like wheat, potentially offsetting some yield losses due to heat stress. However, this benefit is contingent upon sufficient water availability, which remains a significant constraint in this region.
3. **Increased Wind Speed:** With slight increases in wind speeds over the years, although generally not beneficial for wheat pollination (as it is predominantly self-pollinating), it may aid in reducing humidity and preventing certain fungal diseases that thrive in stagnant, moist air conditions.

In summary, while the projected climatic changes in Saâda, Morocco, present several risks to wheat cultivation, particularly concerning temperature increases and inadequate precipitation, there are potential benefits that could be exploited with careful management. Adaptation strategies such as selecting heat-tolerant wheat varieties, improving irrigation infrastructure, and implementing soil moisture conservation practices will be crucial to mitigating the adverse effects of climate change on wheat production in this region.

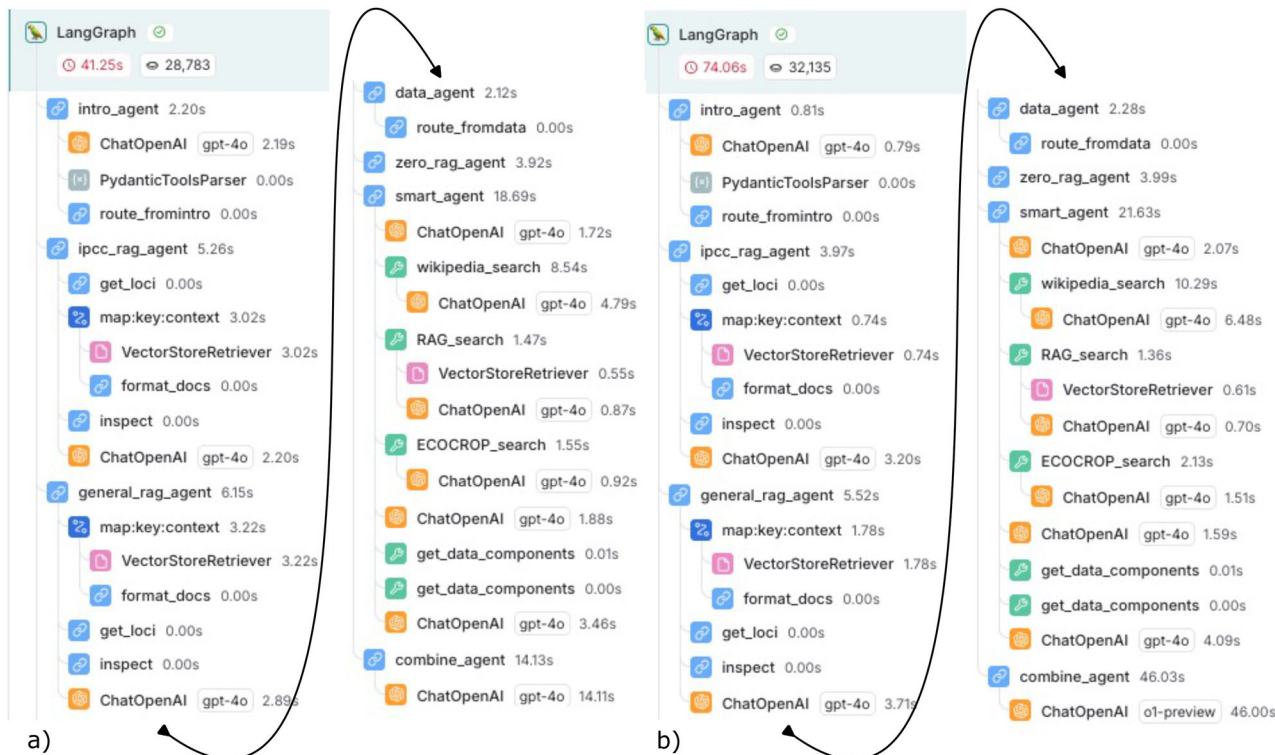


Fig. 1 | ClimSight LangGraph query. Visualization of a LangGraph schema for a question about wheat near by Marrakech using GPT-4o (a) and o1-preview (b).

Synthesizes information to generate a well-structured response.

These components operate partly sequentially and partly in parallel, processing the request in 41.25 seconds at a cost of about 8 cents, handling approximately 28,000 tokens (units of text processed by the AI). To compare performance, we repeated the same query but used a different model, o1-preview, for the final step (Box 2). This adjustment significantly increased the response time to 74 seconds, mainly because the final processing step took 46 seconds instead of 14 seconds. The cost also rose to 39 cents.

Both answers use the same input climate model data and basic information but differ in depth and detail. Answer 1 (gpt-4o) provides a concise evaluation, focusing mainly on temperature increases, precipitation deficits, and flood risks, with only brief mentions of adaptation strategies. It emphasizes the direct impact of temperature and moisture on wheat without addressing broader factors like pests or soil degradation. In contrast, Answer 2 (o1-preview) offers a more comprehensive analysis, incorporating decade-specific climate projections and detailed risk factors, such as increased pest pressure and soil degradation, as well as specific adaptation strategies like drought-tolerant varieties and efficient irrigation methods. Answer 2 also includes additional contextual information, such as the region’s soil type and land use, providing a broader perspective on the challenges and opportunities for wheat cultivation. Overall, Answer 2 is more thorough, offering a fuller picture of potential climate change impacts and adaptive responses for wheat farming in this region.

Despite differences in responses caused by using different models - referred to here as 4o and o1 - in the final step, the answers are fundamentally similar. The similarity comes from the nearly identical input data. However, since data from the database is first processed by the LLM before reaching ‘Combine’, variations arise in the input to ‘Combine’. The analysis potential of model o1 is greater. However, using model o1 requires more financial resources and time. It can be concluded that model o1 is justified for final decision-making. For quick queries, model 4o is sufficient. It provides a preliminary impression and a first approximation.

We also conducted experiments with an open Gemma 7B model¹². This model was significantly reduced to 7 billion parameters. Such a size makes it usable on a personal laptop. However, in our tests, the results were

considerably more modest compared to those of model 4o. However, using such small models can be justified in intermediate steps. This approach significantly speeds up queries and reduces costs.

In the current version, we used high-resolution data averaged for the 2020s, compared to the previous version¹⁰, which relied on lower-resolution CMIP6 simulations. While absolute temperature and precipitation values differ slightly between the versions, both indicate significantly higher temperatures than the critical thresholds for wheat growth. In the older version, these thresholds were derived from pre-trained datasets, whereas the newer version employs precise values from the ECOCROP database⁹, enhancing accuracy. Furthermore, the updated analysis focuses on short-term climate impacts (2020-2049), in contrast to the longer-term period (2070-2100) considered previously, resulting in a more actionable timeframe. The newer version also introduces additional plots to visualize the data, providing clearer insights into key climate trends. Finally, the updated analysis integrates nuanced risk factors, such as increases in wind speed, and aligns with Morocco’s climate adaptation policies, presenting a more comprehensive and practical assessment. A ‘References’ tab was introduced, allowing users to trace the sources of information utilized in generating the report.

Discussion

The proposed framework demonstrates significant potential in providing accurate and localized climate insights by integrating Large Language Models (LLMs) with diverse data sources, including text, scientific literature, and climate data. The results highlight ClimSight’s capability to synthesize complex information into actionable, context-aware assessments. This is particularly relevant for decision-makers in sectors such as agriculture, urban planning, and disaster management, where localized climate information is crucial for informed planning and risk management. We acknowledge that our study presents a conceptual framework at this stage and requires comprehensive validation through user studies, particularly given the involvement of large language models in generating climate assessments.

The evaluation shows that the developed approach enhances the precision and relevance of climate assessments while maintaining scalability, enabling it to address the growing volume of climate-related questions

Box 2 | ClimSight with o1-preview: question about wheat near by Marrakech

Climate change is poised to significantly impact the cultivation of wheat in Saâda, Morocco (latitude 31.6912, longitude -8.1098). Situated at an elevation of 384 meters above sea level with Luvisols soil - known for their fertility and suitability for agriculture - the region currently utilizes its land primarily for orchards. However, the prospect of growing wheat here faces both challenges and opportunities due to shifting climatic conditions.

Temperature trends and Implications

Atmospheric model simulations indicate a clear upward trend in temperatures for Saâda:

- **2020–2029:** Monthly average temperatures range from 8.36 °C in January to 35.11 °C in August.
- **2030–2039:** Temperatures rise slightly, with January averaging 8.76 °C and August 36.12 °C.
- **2040–2049:** Further increases project January at 9.41 °C and August at 36.77 °C.

Wheat thrives in optimal temperatures between 15 °C and 23 °C (ECOCROP database). Temperatures above 27 °C approach the absolute maximum for wheat growth. The projected increases, particularly in the hottest months, suggest that wheat crops may experience heat stress, especially during critical growth periods like grain filling. Elevated temperatures can accelerate plant development unnaturally, reducing the time for grain maturation and potentially leading to lower yields.

Precipitation patterns and water availability

The region's precipitation is notably low and is expected to fluctuate:

- **2020–2029:** Total monthly precipitation varies from 0.33 mm in August to 8.4 mm in February.
- **2030–2039:** A significant decrease is projected in February, dropping to 2.66 mm.
- **2040–2049:** February precipitation further declines to a mere 0.16 mm, while December sees an increase to 10.05 mm.

Wheat requires annual rainfall between 750 mm and 900 mm for optimal growth (ECOCROP database). The current and projected precipitation levels in Saâda are well below these requirements. This substantial deficit indicates a high risk of drought stress, making wheat cultivation heavily reliant on irrigation. Moreover, decreasing rainfall during crucial planting seasons could exacerbate water scarcity, increasing the dependency on irrigation systems and elevating production costs.

Potential Risks

1. **Heat stress on Crops:** Rising temperatures may surpass wheat's tolerable limits, especially during summer, affecting photosynthesis and leading to reduced grain quality and yields.
2. **Water Scarcity:** Lower precipitation combined with higher evaporation rates can diminish water availability, stressing both rainfed and irrigated wheat systems.
3. **Increased Pest and Disease Pressure:** Warmer temperatures may alter the lifecycle and prevalence of pests and diseases harmful to wheat, potentially increasing infestation risks.
4. **Extreme Weather Events:** Although the historical data mentions floods in 1995, 1996, 1997, and 2008, climate change could lead to more frequent and severe events, threatening crop stability.

5. **Soil Degradation:** Higher temperatures and erratic rainfall can lead to soil moisture depletion, affecting nutrient uptake and soil structure.

Potential Benefits

1. **CO₂ Fertilization Effect:** Elevated atmospheric CO₂ levels can enhance photosynthesis in C₃ plants like wheat, potentially improving biomass and grain yields. However, this benefit may be limited by nutrient and water availability.
2. **Extended Growing Seasons:** Warmer temperatures might allow for longer growing seasons or the possibility of double cropping, provided that water resources are sufficient.

Adaptation Strategies

To navigate these challenges and harness potential benefits, several strategies can be employed:

- **Drought-Tolerant Varieties:** Developing or adopting wheat varieties bred for heat and drought tolerance can mitigate some of the adverse effects.
- **Efficient Irrigation Techniques:** Implementing water-saving irrigation methods such as drip irrigation can optimize water use and reduce wastage.
- **Soil Conservation Practices:** Techniques like mulching, cover cropping, and conservation tillage can improve soil moisture retention and reduce erosion.
- **Adjusted Planting Schedules:** Shifting planting dates to cooler periods can help avoid peak temperature stresses.
- **Integrated Pest Management:** Monitoring and controlling pest populations proactively can reduce the risk of infestations exacerbated by climate change.

Conclusion

The cultivation of wheat in Saâda, Morocco, faces considerable risks due to climate change, primarily from increasing temperatures and decreasing precipitation during key growth periods. These factors threaten to exacerbate heat and drought stresses, potentially leading to reduced yields and increased reliance on irrigation.

While there are potential benefits such as the CO₂ fertilization effect and possibly extended growing seasons, these are contingent upon effective management of water resources and other agronomic practices. Proactive adaptation strategies are essential to mitigate risks and sustain wheat production in this region.

References

- **ECOCROP Database:** Provides essential crop requirements, indicating optimal temperatures of 15 °C to 23 °C and annual rainfall needs of 750 mm to 900 mm for wheat.
- **Atmospheric Model Simulations (2020-2049):** Projected climate data for Saâda, showing increases in average temperatures and fluctuations in precipitation.
- **Smart Data Extractor Agent Analysis:** Highlights the discrepancies between wheat's requirements and the projected climatic conditions in Saâda.

worldwide. By leveraging Retrieval-Augmented Generation (RAG) and an agent-based architecture, this methodology effectively overcomes the limitations of traditional climate information systems, providing a more nuanced understanding of localized climate impacts. However, the evaluation dataset currently consists of only 30 QA pairs, primarily used for initial system tuning. Future work will expand this dataset to several

thousand QA pairs, enabling more comprehensive system evaluation and supporting the fine-tuning of open-source models. However, the current evaluation relies solely on LLM-based scoring, which, while scalable, lacks external expert validation. This may limit the robustness of the conclusions, especially in high-stakes or policy-relevant contexts. To address this, future work will incorporate expert reviews, structured usability studies, and

domain-specific user feedback to assess ClimSight's outputs. Particular attention will be given to collecting targeted evaluations from users in agriculture, urban planning, and policy sectors to guide iterative improvements and validate system credibility. Implementing these stakeholder-driven studies will require a dedicated mixed-methods effort that lies beyond the present project's resources, so this paper intentionally confines itself to an automated, scalable evaluation as a transparent baseline for future user-centered validation.

While this version focuses on demonstrating a scalable automated assessment pipeline, we acknowledge the need for more rigorous empirical grounding. This includes future integration of observational data and probabilistic climate model ensembles to capture uncertainty, as well as broader real-world use cases beyond those presented here.

A key strength of the proposed framework lies in its integration of LLMs with multiple data sources, including unstructured text, which allows for a holistic view of climate information. This approach enables the platform to contextualize data according to specific user needs, enhancing the relevance and usability of the outputs. The agent-based architecture further contributes to system efficiency and scalability, ensuring flexible integration of future data sources and model updates.

However, the implemented methodology also faces limitations, particularly in its dependency on external APIs and data availability. The quality and reliability of outputs are influenced by the accuracy of input data and the limitations of LLMs, including their sensitivity to prompt design and potential biases. These potential biases include hallucinations, where the LLM generates plausible but incorrect information not derived from the provided data, and instances where numerical values are drawn from the model's internal parameters rather than validated external datasets. Additionally, integrating unstructured text and diverse data sources introduces challenges in maintaining consistency and verifiability of the assessments. Moreover, the current implementation presents climate model data without explicit uncertainty quantification. Further development should also focus on incorporating explicit uncertainty quantifications, confidence intervals for projected climate trends, and recalculated biases from climate models to enhance the interpretability and reliability of ClimSight outputs.

Traditional systems typically rely on static datasets and numerical models with limited contextual adaptation. In contrast, ClimSight dynamically retrieves and synthesizes relevant information, offering evidence-based insights tailored to user-specific queries. Compared to other AI-driven climate platforms, the agent-based architecture enhances modularity and flexibility, allowing for targeted data retrieval and contextual understanding. This modular approach not only optimizes performance but also enables the system to adapt to rapidly evolving climate science and user needs.

The growing impacts of climate change globally necessitate scalable climate information systems capable of addressing diverse, localized questions. The presented concept contributes to this need by democratizing access to localized climate information, empowering stakeholders ranging from researchers and policymakers to communities and businesses.

Future developments will focus on enhancing real-time data integration and expanding the application of this approach across different climate-sensitive sectors. Additionally, refining the evaluation framework to include cross-validation with observed data will improve the reliability and consistency of the assessments. Efforts to minimize risks associated with false information will include developing additional verification mechanisms. For instance, agent prompts will explicitly highlight discrepancies when different datasets provide contradictory information. Additionally, future plans involve creating "second-eye" agents to double-check statistical calculations performed by ClimSight and dedicated fact-checking agents to further reduce misinformation risks. Integrating more advanced LLMs and expanding the diversity of data sources, including social and economic data, will further enhance the system's contextual awareness and usability.

The current reliance on the OpenAI API poses potential cost challenges with increased usage. Nevertheless, rapid advancements in LLM technologies are continuously reducing these costs. For example, in 2024-

2025, GPT-4o usage costs decreased significantly, and the introduction of GPT-4.1 Nano further reduced query expenses to a fraction of the earlier costs. ClimSight's flexible architecture, built upon the LangChain framework, allows seamless integration and switching between various LLMs, including open-source alternatives capable of running locally. Future research and development efforts will prioritize extending support for these open-source models, which will further mitigate costs and enhance the scalability and accessibility of the platform.

This novel methodology represents a paradigm shift in how climate data is synthesized and utilized, bridging the gap between scientific complexity and practical decision making. As climate change continues to present complex challenges, scalable and adaptable solutions like the proposed system are essential for supporting informed future planning and effective risk management.

Methods

This section outlines the methodology and system architecture behind ClimSight. Additionally, the evaluation subsection provides a comparative analysis of the system's responses across multiple queries, supported by relevant statistical data.

ClimSight: concept and architecture

ClimSight leverages LLMs to generate climate assessments tailored to user-defined locations and activities. The initial concept, introduced by¹⁰, demonstrated that supplementing LLM prompts with localized climate data improves the accuracy and relevance of climate information. In this work, we extend this approach by integrating additional environmental datasets, high-resolution climate model outputs, and contextual knowledge from climate reports and scientific literature.

To manage the increasing data volume and complexity, ClimSight now follows a modular architecture, with specialized components handling specific tasks (Fig. 2). The system follows these key steps:

1. User Input—The user provides a question and geographic coordinates.
2. Location Processing—ClimSight determines the location details and contextual information.
3. Data Retrieval—The system queries online and pre-downloaded climate databases, extracts relevant climate model outputs, and retrieves climate reports.
4. Smart Analysis—A specialized agent interprets data based on the user's request (e.g., applying climate thresholds relevant to a specific crop).
5. Information Filtering—The LLM processes and refines relevant content from databases and reports.
6. Final Report Generation—A "combine agent" synthesizes findings into a user-friendly climate impact assessment.

To evaluate the system's reliability, ClimSight includes an evaluation module that systematically assesses the quality of its responses (see Evaluation section). ClimSight is primarily developed in Python, with a Streamlit-based web interface and a command-line version. It uses LangChain and LangGraph frameworks²⁰ for managing agent interactions and LLM calls. The modular design enables easy integration of different LLMs, and while OpenAI models are currently used, preliminary tests have been conducted with the open-source Gemma 7B model. The system's modularity allows seamless adaptation to different applications and continuous updates through a CI pipeline. ClimSight is open-source (BSD 3-Clause License) and available for public use, modification, and expansion. Below, we describe the core components in more detail.

Core components

User input. The user submits a question along with geographic coordinates (longitude and latitude).

Location requests. This module determines the location's address (via the Nominatim API²¹) and whether it is on land or water (using Natural Earth datasets²²).

Fig. 2 | ClimSight schematic. Dashed lines indicate conditional calls. Solid arrows represent the mandatory path from one agent to another. Numbers show the key steps mentioned in the section “ClimSight: Concept and Architecture”.

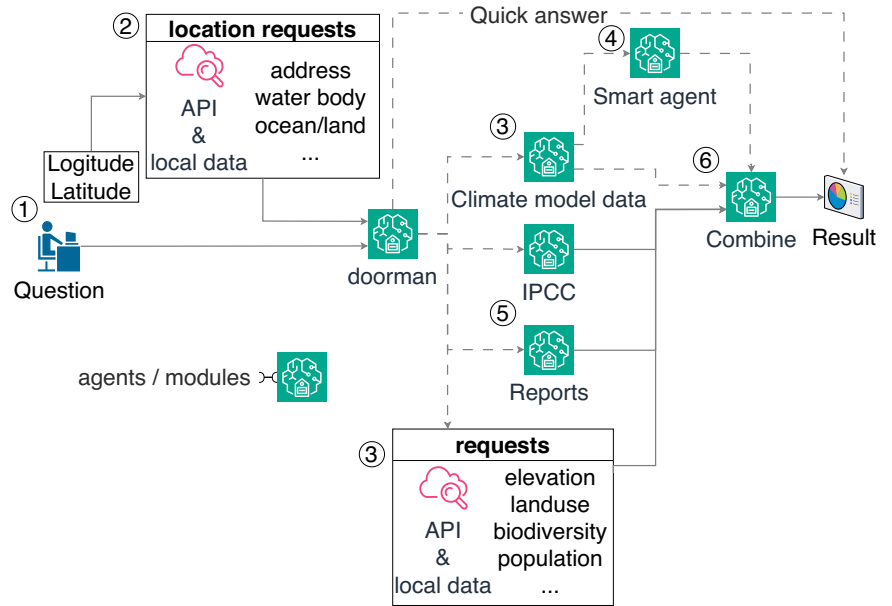
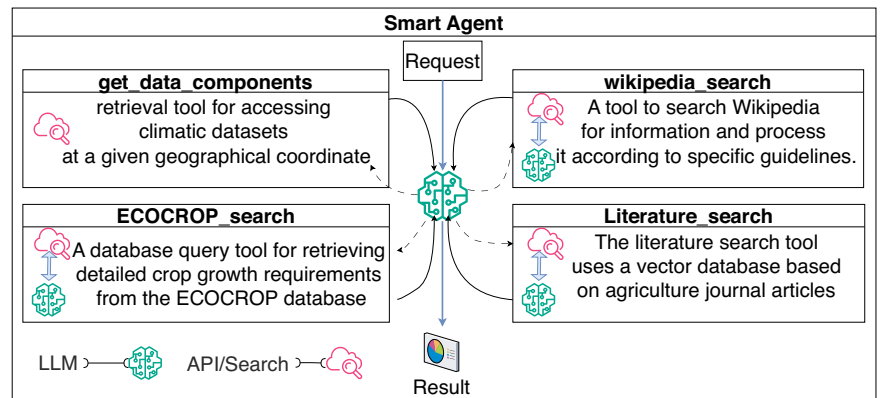


Fig. 3 | Smart agent schematic. Dashed lines indicate conditional calls. Solid arrows represent the mandatory path.



Query filtering. The “Doorman” serves as an initial filter, evaluating whether the user’s query is relevant to climate-informed decision-making. If so, it extracts preliminary contextual information before passing the request to specialized components.

Data retrieval. ClimSight collects location-specific environmental and climate information from multiple sources:

- Natural hazards - EM-DAT / GDIS datasets (1960-2018)^{23,24}
- Population data - UN World Population Prospects (2022)²⁵
- Biodiversity records - GBIF API²⁶
- Elevation data - Open Topo Data API^{27,28}
- Land use - Overpass API (OpenStreetMap)²⁹
- Soil types - ISRIC SoilGrids API³⁰
- Coastline proximity - Natural Earth datasets²²

Climate reports and literature. Two modules provide contextual climate knowledge,

1. IPCC Reports Module—Focuses on the IPCC 6th Assessment Report, ensuring high credibility.
2. Scientific Reports Module—Includes a broader set of WMO reports, German climate-health reports (RKI), and others.

By keeping these databases separate, ClimSight ensures users receive high-quality assessments based on both authoritative global perspectives and

localized insights. If no relevant information is found, these sources are omitted from the final output.

Climate model data processing. ClimSight analyzes four primary climate variables: 2m air temperature, precipitation, wind speed, and wind direction. It incorporates:

- Coarse-resolution climate model data (100 km) from AWI-CM (CMIP6), covering past (1985-2004) and future (2070-2100) climate scenarios.
- High-resolution simulations from nextGEMS (9 km), coupled with the FESOM ocean model, spanning 2020-2049^{31,32}.

Smart agent analysis. The Smart agent (Fig. 3) functions as a specialized analytical component that evaluates climate-relevant parameters based on user queries. This LLM-powered component synthesizes information from multiple data sources to provide contextually appropriate climate assessments. The agent operates through four primary information retrieval mechanisms:

- Encyclopedic Knowledge: Conducts targeted Wikipedia searches to establish baseline environmental parameters and thresholds for specific activities or crops.
- Scientific Literature Integration: Queries a vector database of agricultural and climate research papers to identify climate sensitivity factors and environmental constraints.

- Agricultural Requirements Database: For crop-specific queries, consults the ECOCROP database¹⁹ to determine optimal growing conditions and climate thresholds.
- Climate Data Retrieval: Accesses pre-processed climate model outputs for the specified geographical coordinates, extracting temperature, precipitation, and wind parameters relevant to the query.

The agent employs a two-phase workflow. first determining which environmental parameters are relevant to the user’s query through knowledge base consultation, then retrieving and analyzing the specific climate data for those parameters.

Final report generation and accessibility. After data retrieval and analysis, ClimSight synthesizes findings into a user-friendly climate assessment report. The system can be accessed via a web browser or terminal, requiring an OpenAI API key for operation. Essential datasets are preloaded, while additional data is fetched dynamically via API requests. This structured approach ensures ClimSight remains flexible, scalable, and accessible for climate-informed decision-making across various applications.

Evaluation

We employed a modified version of the evaluation framework proposed by Liu et al.³³ to assess the quality of responses generated by ClimSight. This framework adopts a reference-based approach, in which the generated answers are compared against predefined reference answers using LLMs (see Fig. 4). Consistent with³³, we incorporated chain-of-thought (CoT) reasoning by instructing the LLM to evaluate responses across multiple criteria: Completeness, Accuracy, Relevance, Clarity, and Coherence. Each answer is assigned a score from 1 to 5 for each criterion, and these scores are compiled to produce an overall assessment of response quality. Detailed explanations for each evaluation criterion are provided in the

Supplementary Note section “Prompt for evaluation”. Our evaluation prompt guides the LLM through a structured process:

- Understand the Question Context
- Analyze the Provided Answers
- Compare Answers for Completeness
- Assess Accuracy and Relevance
- Evaluate Clarity and Coherence
- Rate the Answer
- Provide Feedback and Suggestions

By leveraging LLMs with CoT reasoning, this evaluation framework provides a systematic analysis of ClimSight’s responses. As demonstrated by³³, such an approach correlates strongly with human judgments. This evaluator is actively employed in application development to assess modifications including the addition or removal of components, and to identify specific issues related to modernization, such as changes in the sequence of components calls.

In the context of evaluating system performance, we used questions derived from IPCC reports to verify the system’s ability to address general inquiries and to test the effectiveness of its RAG. We also compiled a set of 30 question-answer pairs generated by the o1 model using ClimSight, which we verified based on our expertise in climate research and climate assessments. As an example, one of the questions was, “Will expanding wheat fields in this region ensure consistent harvests over the next decade?” supplemented with geographic coordinates in Spain. Questions for evaluation were grouped into several thematic categories: “General Climate Risks,” “Business-Specific Concerns,” “Agriculture and Food Production,” “Renewable Energy and Infrastructure,” “Urban and Community Planning,” “Coastal and Water-Related Risks,” “Community and Ecosystem Impacts,” “Energy and Supply Chain Resilience,” “Industry-Specific Impacts,” and “Long-Term Sustainability.” Within each category, questions were divided into “general” and “specific” types; for example, a general question included “How will climate change impact agricultural yields at this location?” while a specific question was “Can we continue growing grapes at our vineyard located here under future climate scenarios?” The locations associated with each question were randomly selected globally, ensuring each point lies on land. To evaluate how our system’s responses depend on different LLM models, we used ClimSight to answer the same set of questions, while varying the final model.

In experiments involving gpt-4o and higher-level models, we employed gpt-4o for textual RAG and changed only the model in the final step “Combine” since the o1 family does not integrate tools well in the current LangChain environment. The comparative results (see Table 1) indicate that using more advanced models for combining information, as expected, yields higher scores. We do not show a direct comparison with o1 itself alone because it tends to self-score at maximum (5), diminishing its utility for direct comparison. The last row in Table 1 (“Pure gpt-4o, no ClimSight”) provides a control experiment in which gpt-4o was used without ClimSight. Although gpt-4o typically

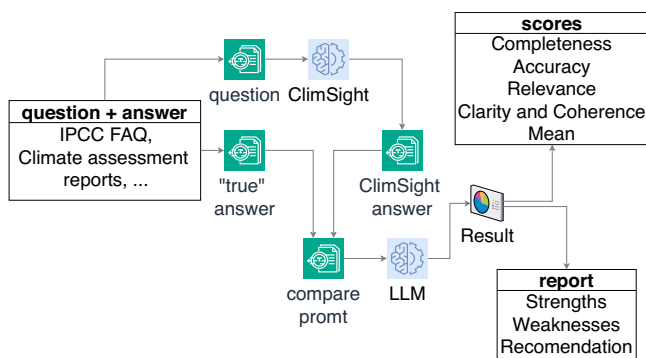


Fig. 4 | Schematic of the evaluation module.

Table 1 | Results of the evaluation using different LLM models, with and without ClimSight and the ‘smart’ agent, based on 30 pairs of questions and answers, as well as MMLU (Massive Multitask Language Understanding) and GPQA (General-Purpose Question Answering) benchmarks

Model	Completeness	Accuracy	Relevance	Clarity and coherence	Mean	MMLU 0-shot CoT	GPQA
gpt-3.5-turbo-0125	2.4 (0.5)	2.9 (0.5)	2.8 (0.5)	4.0 (0.3)	3.0 (0.4)	69.8	30.8
gpt-4o-mini	2.9 (0.5)	3.3 (0.6)	3.5 (0.6)	4.2 (0.4)	3.5 (0.4)	82.0	40.2
gpt-4o	3.2 (0.6)	3.7 (0.7)	3.6 (0.7)	4.4 (0.5)	3.8 (0.5)	88.7	53.6
o1-mini	4.0 (0.7)	4.0 (0.8)	4.3 (0.6)	4.9 (0.3)	4.3 (0.5)	85.2	60.0
gpt-4o without “Smart agent”	2.9 (0.4)	3.3 (0.6)	3.4 (0.6)	4.2 (0.4)	3.5 (0.4)		
Pure gpt-4o no ClimSight	2.2 (0.4)	2.9 (0.4)	2.9 (0.7)	4.1 (0.3)	3.0 (0.3)		

Mean values are reported with standard deviation in parentheses.

recognized the geographic context of a question often pinpointing a city or region its answers lacked the additional data that ClimSight provides. As a result, the overall performance of a pure gpt-4o system approached that of gpt-3.5 paired with ClimSight. Notably, the gpt-4o-only responses often contained vague statements or very broad temperature ranges (e.g., “average temperatures may increase by 1–5 °C”), which offer limited value for a detailed climate assessment.

By comparison, simpler based models like Gemma 7B were generally unable to link geographic coordinates to relevant geographic attributes and therefore performed significantly worse in similar tests (results not shown). Furthermore, when we supplemented our system with the “Smart agent”, the average response quality (mean score) increased from 3.475 to 3.76, largely because of improvements in Completeness and Accuracy. This effect was most pronounced in questions related to crops, where the system could leverage information from an ECOCROP database tied to the specific plant in question.

In practice, this evaluation methodology helps identify how much the system’s responses improve after modifications, such as adding or removing components, and supports quantitative assessments when the underlying LLM is changed. As shown in Table 1, the o1 models tend to be more computationally expensive and slower compared to gpt-4o. For more efficient deployment, one could use ClimSight with gpt-4o or even gpt-3.5 to handle preliminary inquiries, reserving o1 for final, complex queries where its enhanced capabilities justify the additional cost.

In addition, we observe that GPQA (A Graduate-Level Google-Proof Q&A Benchmark) has an exceptionally high correlation with our own evaluation metrics (Table 1), on the order of 0.97, while the correlation with MMLU (Massive Multitask Language Understanding Benchmark) is 0.8. This result seems intuitive when looking at o1-mini and GPT-4o: although o1-mini’s MMLU score (85.2) is slightly lower than GPT-4o’s (88.7), its GPQA score (60.0) exceeds GPT-4o’s (53.6), reflecting o1-mini’s stronger performance on our custom metrics (4.29 vs. 3.76 mean score). In other words, o1-mini appears to offer more robust domain-specific reasoning and completeness, likely due to a more STEM-focused training that better matches the specialized tasks in ClimSight. Given this near-perfect correlation between GPQA and our custom metrics, we consider GPQA a crucial indicator when selecting LLMs for ClimSight.

In conclusion, our evaluation framework, adapted from Liu et al.³³, demonstrates how integrating ClimSight and specific components can enhance response quality and provide measurable improvements in various criteria. Moreover, these findings clearly illustrate ClimSight’s superiority in handling climate-specific inquiries, as it consistently provides more specialized, accurate and contextual responses than LLMs alone. Including domain specific databases and task-oriented agents, the developed approach delivers deeper insights that are critical for robust climate assessment. Nevertheless, explicitly verifying the accuracy of specific numerical values, such as seasonal mean temperatures or precipitation, remains a challenge that requires additional verification mechanisms, currently under development in an updated version of ClimSight. This approach is particularly valuable during system development, allowing for systematic assessments of how each component contributes to overall performance.

Data availability

The ClimSight code used in this publication is available at <https://github.com/CliDyn/climsight> (release v1.0.0) and can also be found on Zenodo: 10.5281/zenodo.14945362, along with evaluation results. Reference to data is provided within the manuscript or the supplementary information files. The datasets are also listed in the ClimSight YAML configuration file and can be downloaded using the corresponding ClimSight script.

Code availability

The ClimSight code used in this publication is available at <https://github.com/CliDyn/climsight> (release v1.0.0) and can also be found on Zenodo: <https://doi.org/10.5281/zenodo.14945362>, along with evaluation results.

Received: 11 March 2025; Accepted: 18 September 2025;

Published online: 25 October 2025

References

- OpenAI et al. GPT-4 Technical Report. Preprint at <http://arxiv.org/abs/2303.08774> (2024).
- OpenAI. Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>.
- Gemini Team et al. Gemini: a family of highly capable multimodal models. Preprint at <https://arxiv.org/abs/2312.11805> (2024).
- Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
- Vaghefi, S. A. et al. Chatclimate: Grounding conversational AI in climate science. *Commun. Earth Environ.* **4**, 480 (2023).
- Calzada, N. D. L., Costa, T. A. D., Blangero, A. & Chesneau, N. ClimateQ&A: Bridging the gap between climate scientists and the general public. In *Proceedings of the Tackling Climate Change with Machine Learning Workshop at ICLR 2024* (Vienna, Austria, 2024).
- Thulke, D. et al. Climategpt: Towards AI synthesizing interdisciplinary research on climate change. Preprint at <https://arxiv.org/abs/2401.09646> (2024).
- Hsu, A., Laney, M., Zhang, J., Many, D. & Farczadi, L. Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)* (eds Stambach, D. et al.) 82–92 (Association for Computational Linguistics, 2024). <https://aclanthology.org/2024.climate-nlp-1.6/>.
- U.S. Federal Government. U.s. climate resilience toolkit. (2025). (Accessed 5 Aug 2025).
- Koldunov, N. & Jung, T. Local climate services for all, courtesy of large language models. *Commun. Earth Environ.* **5**, 13 (2024).
- Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
- Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models, Version Number 1. Preprint at <https://arxiv.org/abs/2302.13971> (2023).
- Millidge, B. Scaffolded LLMs as natural language computers. <http://www.beren.io/2023-04-11-Scaffolded-LLMs-natural-language-computers/>.
- Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (eds Follmer, S., Han, J., Steimle, J. & Henry Riche, N.) 1–22 (Association for Computing Machinery, New York, NY, USA, 2023). <https://doi.org/10.1145/3586183.3606763>.
- Pantiukhin, D., Shapkin, B., Kuznetsov, I., Jost, A. A. & Koldunov, N. Accelerating earth science discovery via multi-agent LLM systems. Preprint at <https://arxiv.org/abs/2503.05854> (2025).
- Rackow, T. et al. Multi-year simulations at kilometre scale with the integrated forecasting system coupled to FESOM2.5 and NEMOV3.4. *Geosci. Model Dev.* **18**, 33–69 (2025).
- Hohenegger, C. et al. Icon-sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geosci. Model Dev.* **16**, 779–811 (2023).
- Moon, J.-Y. et al. Earth’s future climate and its variability simulated at 9 km global resolution. *EGU Sphere* **2024**, 1–46 (2024).
- FAO & IASA. *Global Agro-Ecological Zones (GAEZ v4)—Data Portal User’s Guide* 1st edn (FAO and IASA, 2021).
- LangChain, Original-date: 2022-10-17T02:58:36Z. <https://github.com/langchain-ai/langchain> (2022).
- Nominatim. <https://nominatim.org/>.
- Natural Earth. Free vector and raster map data. <https://www.naturalearthdata.com>.

23. Rosvold, E. L. & Buhaug, H. GDIS, a global dataset of geocoded disaster locations **8**, 61. <https://www.nature.com/articles/s41597-021-00846-6>.
 24. Rosvold, E. & Buhaug, H. Geocoded Disasters (GDIS) Dataset.
 25. United Nations, Department of Economic and Social Affairs, Population Division. World population prospects 2024, online edition (2024). <https://population.un.org/wpp/downloads?folder=Standard%20Projections&group=CSV%20format>.
 26. GBIF.org. Occurrence Search. <https://www.gbif.org/>.
 27. Open topo data. <https://www.opentopodata.org/> (Accessed 11 Mar 2025).
 28. Amante, C. & Eakins, B. W. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24, National Geophysical Data Center, NOAA <https://repository.library.noaa.gov/view/noaa/1163> (2009) (Accessed 11 Mar 2025).
 29. OpenStreetMap Wiki. Overpass api — openstreetmap wiki https://wiki.openstreetmap.org/w/index.php?title=Overpass_API&oldid=2809610 (2025).
 30. Poggio, L. et al. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217–240 (2021).
 31. Moon, J.-Y. et al. Earth's future climate and its variability simulated at 9 km global resolution. <https://egusphere.copernicus.org/preprints/2024/egusphere-2024-2491/>.
 32. Rackow, T. et al. Multi-year simulations at kilometre scale with the integrated forecasting system coupled to fesom2.5/nemov3.4. *EGUsphere* **2024**, 1–59 (2024).
 33. Liu, Y. et al. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522 (Association for Computational Linguistics, 2023). <https://aclanthology.org/2023.emnlp-main.153>.
- German Research Foundation, project no. 274762653). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

I.K., A.A.J., D.P., B.S., T.J., and N.K. co-developed the idea and wrote the manuscript. I.K., A.A.J., D.P., B.S., and N.K. implemented the code.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s44168-025-00300-y>.

Correspondence and requests for materials should be addressed to Ivan Kuznetsov.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

Acknowledgements

This work was supported by the European Union's Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts, implementing part of this Initiative with funding by the European Union. We also thank the Helmholtz Association and the Federal Ministry of Education and Research (BMBF) for supporting the DataHub Initiative of the Research Field Earth and Environment, which contributed to this study. Additionally, this work is supported by project S1: Diagnosis and Metrics in Climate Models of the Collaborative Research Centre TRR 181 "Energy Transfer in Atmosphere and Ocean," funded by the Deutsche Forschungsgemeinschaft (DFG,