

EGU26-14350, updated on 17 Oct 2026

<https://doi.org/10.5194/egusphere-egu26-14350>

EGU General Assembly 2026

© Author(s) 2026. This work is distributed under the Creative Commons Attribution 4.0 License.



Augmenting Large Language Models with Climate Indicator Computation for Next-Generation Climate Services

Jacopo Grassi^{1,3}, Dmitrii Pantiukhin², Ivan Kuznetsov², Nikolay Koldunov², Massimo Dragan³, and Jost von Hardenberg¹

¹Polytechnic University of Turin, Department of Environment, Land and Infrastructure Engineering, Italy
(jacopo.grassi@polito.it)

²Alfred Wegener Institute, Bremerhaven, Germany

³WSP in Italy, Turin, Italy

Large Language Models (LLMs) and agentic AI are increasingly explored as interfaces for geoscience information, risk communication, and decision support in natural hazards and disaster management. However, most LLM-based assistants remain limited in quantitative reasoning and often lack traceability, reproducibility, and robust uncertainty communication. Here we present XCLIM-AI, an agentic system that couples LLM-based interpretation with deterministic computation of climate indicators through the open-source xclim library. XCLIM-AI can compute >200 standardized climate indices from CMIP6 HighResMIP projection ensembles, enabling responses that combine narrative explanations with transparent, auditable quantitative outputs (e.g., heatwave metrics, drought duration, extreme precipitation indices) and explicit provenance of assumptions and processing steps.

A key aspect of this work is the integration of XCLIM-AI within ClimSight, a multi-agent platform for localized climate information. In the integrated architecture, general-purpose agents handle retrieval and reasoning over scientific and contextual information, while XCLIM-AI performs on-demand, tool-based computation of indicators requested by the user query.

We evaluate four system configurations: (1) a plain LLM baseline, (2) XCLIM-AI, (3) ClimSight, and (4) an integrated ClimSight-XCLIM architecture, using a hybrid assessment protocol that combines scalable LLM-as-a-judge scoring with blinded human expert evaluation. Performance is assessed across four criteria central to climate- and hazard-relevant services: relevance, credibility, uncertainty communication, and actionability. Results show systematic gains over the baseline, with the strongest improvements in actionability and uncertainty reporting when indicator computation is available and properly integrated. We also observe that simply increasing contextual information does not automatically increase perceived credibility, highlighting the importance of traceable quantitative evidence and evaluation protocols tailored to operational trust. We conclude by discussing implications for the reliable adoption of agentic AI in geosciences and hazard-facing workflows, and by outlining a generalizable evaluation framework for tool-augmented LLM systems.

