HVR-Met: A Hypothesis-Verification-Replaning Agentic System for Extreme Weather Diagnosis

Anonymous Authors[1]

Abstract

While deep learning-based weather forecasting paradigms have made significant strides, addressing extreme weather diagnostics remains a formidable challenge. This gap exists primarily because the diagnostic process demands sophisticated multi-step logical reasoning, dynamic tool invocation, and expert-level prior judgment. Although agents possess inherent advantages in task decomposition and autonomous execution, current architectures are still hampered by critical bottlenecks: inadequate expert knowledge integration, a lack of professional-grade iterative reasoning loops, and the absence of fine-grained validation and evaluation systems for complex workflows under extreme conditions. To this end, we propose a multi-agent meteorological diagnostic system characterized by the deep integration of expert knowledge. Its central innovation is the "Hypothesis-Verification-Replanning" closed-loop mechanism, which facilitates sophisticated iterative reasoning for anomalous meteorological signals during extreme weather events. To bridge gaps within existing evaluation frameworks, we further introduce a novel benchmark focused on atomic-level subtasks. Experimental evidence demonstrates that the system excels in complex diagnostic scenarios.

1. Introduction

Extreme weather diagnosis constitutes the systematic process of deciphering the underlying causes and evolutionary logic of severe atmospheric events through sophisticated reasoning and expert judgment. This process is fundamental to meteorological services and public safety. Given that these events are characterized by sudden emergence and swift development, operational centers must promptly identify and interpret anomalous signals. Rapidly determining the specific nature and operational drivers of such events within constrained timeframes provides the essential scientific basis for weather alerts and strategic emergency responses.

At present, AI cannot independently carry out operational extreme weather diagnosis. In practice, this sophisticated workflow still depends on human forecasters, who must rapidly examine a small set of high-impact meteorological data and diagnostic indices under strict time constraints to isolate key drivers and craft an actionable interpretation. Despite its effectiveness, this manual pipeline is inherently constrained by strong dependence on individual experience, substantial labor demands, and heightened susceptibility to mistakes—particularly among less experienced personnel. More fundamentally, extreme-weather diagnosis requires analysts to pinpoint physically meaningful anomalies from fragmented, multivariate, and highly coupled datasets, then synthesize meteorological principles with situational context to perform disciplined attribution reasoning. Current deep-learning approaches typically fall short of this requirement: they struggle to consistently detect salient, event-specific signals across heterogeneous variables and scales, and they lack the ability to explicitly assemble a logically coherent and interpretable chain of evidence that supports a defensible warning decision.

Agentic frameworks offer considerable potential to address these operational limitations by automating specialized tasks (Wang et al., 2025) such as data acquisition (Jin et al., 2024; Qu et al., 2025; Lu & Wang, 2025; Schmidgall et al., 2025), code generation (Yang et al., 2024; Zhang et al., 2025; Li et al., 2025), and multimodal interpretation (Bai et al., 2025a;b). However, bridging the gap between general automation and professional-grade weather diagnosis agent necessitates overcoming two primary challenges. The first challenge is the insufficient integration of domain-specific expertise. Since the requirement for experienced operational expertise in selecting key diagnostic elements, general-purpose agents often prove incapable of autonomously identifying core evidence within complex meteorological scenarios. The second challenge lies in the lack of a heuristic reasoning loop for purposeful evidence

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

collection. In extreme weather diagnosis, the discovery of one anomalous signal often dictates the specific tools or indices required for the next phase of analysis.

To address these two primary challenges, this paper proposes a multi-agent system specifically designed for the diagnostic analysis of extreme weather. Integrating seven functional agents such as Decomposer, Data Specialist, Code Executor, Plotter, Image Checker, Evaluators, Diagnostician and Reporters, the system enables end-to-end automation spanning high-dimensional data retrieval, indicator computation, multimodal interpretation, and structured report generation. To tackle the first challenge, we established a diagnostic guideline repository as the core knowledge foundation. This repository was built by semi-automatically extracting key insights from 584 papers of extreme weather papers: Meteorological Monthly[1], Transactions of Atmospheric Sciences[2], Acta Meteorologica Sinica[3], American Meteorological Society[4] followed by rigorous validation by five senior forecasters to ensure professional-grade variable selection and index calculation. For the second challenge, the system introduces a "hypothesis-verification-replanning" reasoning mechanism that simulates the cognitive paradigm of human experts. Within this framework, the system formulates testable physical hypotheses based on anomalous signals and utilizes the guideline repository to plan diagnostic pathways. Should evidence be insufficient, the system dynamically triggers replanning to adjust its diagnostic path, thereby constructing a logically consistent, evidence-based diagnostic chain within a closed-loop reasoning process.

Furthermore, we present a comprehensive benchmark designed to evaluate both the reliability of individual operational units and the overall versatility of the agent system. This benchmark comprises 100 end-to-end extreme weather events for full-process diagnostic analysis, complemented by 250 specialized QA pairs for granular verification: 150 for meteorological index calculation (covering 30 index types) and 100 for diagnostic plotting (covering 20 figure categories). By spanning atomic tasks—such as standalone plotting and index computation—to complex, holistic retrospective analysis, this benchmark rigorously validates the system's robustness in real-world meteorological operations and demonstrates its broad utility across diverse diagnostic scenarios.

The contributions are as follows:

- We propose a human forecaster-inspired paradigm for extreme-weather analysis that emphasizes physically interpretable, automated diagnostics centered on anomaly signals.
- We introduce a Hypothesis–Verification–Replanning loop to self-improve the diagnostic pathway by focusing on anomalous meteorological signals.
- We construct a new benchmark for extreme-weather diagnostics, covering 30 types of meteorological index computation and 20 types of meteorological plotting.
- Our System achieves pass rates of 71.86% for index computation, 79.52% for figure generation, and 85% for final reporting, demonstrating its robust capacity to assist forecasters in diagnosing extreme weather events.

[1]http://qxqk.nmc.cn/qxen/home
[2]http://dqkxxb.ijournals.cn/dqkxxben/home
[3]http://qxxb.cmsjournal.net/AboutJournal
[4]https://www.ametsoc.org/ams

2. Related Work

Weather Foundation Models. Deep learning-based weather forecasting systems (e.g., FourCastNet (Kurth et al., 2023), Pangu-Weather (Bi et al., 2023), NowcastNet (Zhang et al., 2023), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), Stormer (Nguyen et al., 2024), GenCast (Price et al., 2025), FengWu (Chen et al., 2025)) and weather foundation models (e.g., Climax (Nguyen et al., 2023), Aurora (Bodnar et al., 2025)), trained on large-scale structured numerical data, have significantly outperformed traditional physics-based numerical weather prediction (NWP) systems (Molteni et al., 1996) in terms of forecasting accuracy, computational efficiency, and task diversity. While these models have achieved major breakthroughs in numerical fidelity, they remain highly optimized "black-box" prediction tools. Due to the lack of explicit modeling of physical processes, these systems possess inherent limitations in causal reasoning and diagnostic interpretation, and they cannot support interactive scientific exploration or cross-domain reasoning through natural language interfaces.

Meteorology Autonomous Agentic Frameworks. In recent years, the focus of AI for meteorology has transferred from data-driven models to knowledge-driven autonomous agents. Zephyrus (Varambally et al., 2025) is the first agentic framework specifically designed for weather science, constructing a meteorological agent environment and integrating meteorological tools. ClimateAgent (Kim et al., 2025) and EWE (Jiang et al., 2025) both introduced multi-agent systems; ClimateAgent is designed to address climate science tasks, while EWE utilizes predefined thought path for the retrospective analysis of extreme weather events.

Yet, the application of autonomous agentic system to the detection and causal attribution of extreme weather remains an unexplored frontier. Building upon these advances, we propose a novel paradigm for extreme weather diagnosis and analysis. Our Multi-Agent System identifies extreme weather types by detecting anomalous signals and iteratively

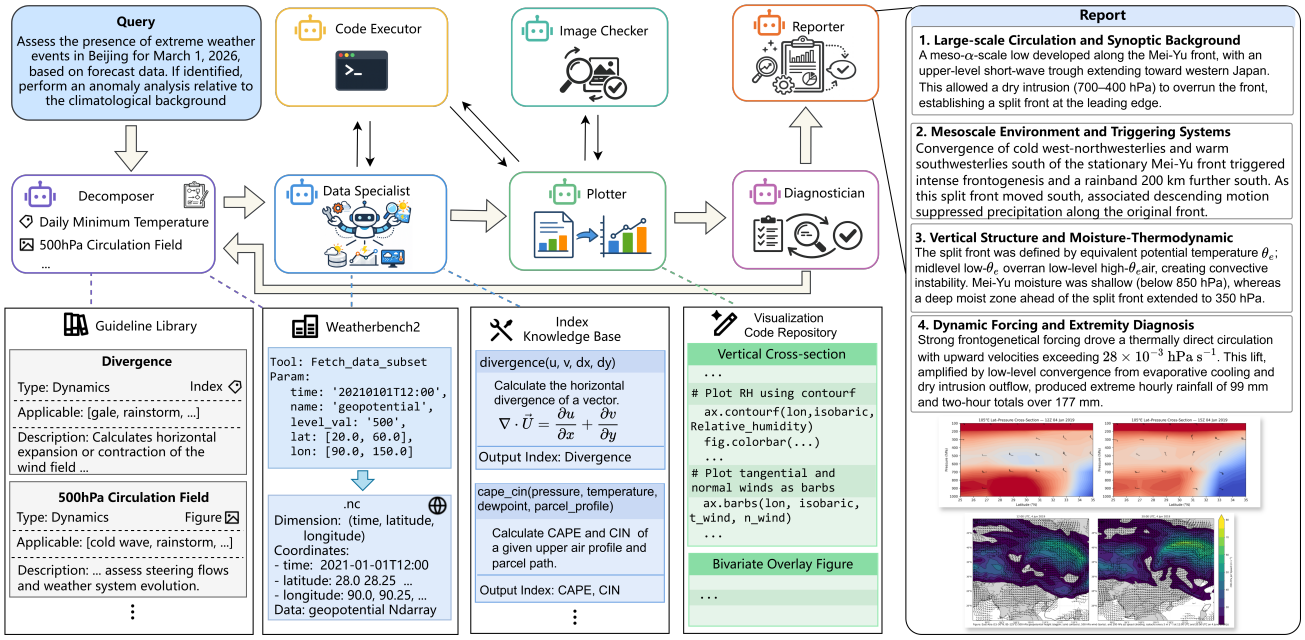


Figure 1. The Workflow of HVR-Met

self-optimizes its diagnostic pathways through continuous evaluation.

3. Methodology

3.1. Multi-Agent Diagnostic Framework

The multifaceted nature of extreme weather diagnosis, which requires the integration of high-dimensional data retrieval, rigorous physical calculation, and synoptic reasoning, poses significant challenges for LLMs. To address this, we introduce a Multi-Agent Meteorological Diagnostic Framework. Designed to emulate the professional 'Weather Consultation' workflow, this system establishes a collaborative agentic environment that unifies domain-specific capabilities through structured role specialization.

The framework encompasses seven specialized agents:

Decomposer: Serves as the strategic planner that parses high-level diagnostic queries. It formulates an initial physical hypothesis and hierarchically decomposes the problem into a logical sequence of executable sub-tasks for downstream agents.

Data Specialist: Serves as the data-retrieval and computation module. It synthesizes Python scripts to access high-dimensional meteorological tensors from WeatherBench2 and derives complex physical indices, ensuring rigorous preprocessing and numerical precision.

Code Executor: Serves as a sandboxed execution environment for securely running generated scripts. It manages dependencies and captures execution feedback, maintaining a robust separation between code synthesis and runtime behavior.

Plotter: Serves as the synoptic visualization module. It generates code to render meteorological fields into publication-quality figures, leveraging appropriate geospatial projections and standard colormaps to translate numerical data into interpretable visual evidence.

Image Checker: Serves as the quality-assurance module. It validates generated figures against established meteorological plotting standards (e.g., contour intervals and label placement) to ensure clarity and domain compliance prior to analysis.

Diagnostician: Serves as the core inference unit that approximates expert reasoning. By integrating multimodal evidence (figures and indices) with a meteorological knowledge base, it performs abductive reasoning to identify the physical mechanisms driving the extreme weather event.

Reporter: Serves as the reporting module that consolidates diagnostic outputs. It organizes the reasoning trace and supporting evidence into a comprehensive diagnostic report.

3.2. Hypothesis–Verification–Replanning

To emulate the cognitive rigor of human forecasters, we design a dynamic inference loop capable of self-correction. The mechanism unfolds in four sequential phases:

Event-Driven Hypothesis Generation. The process initiates with a pre-scan phase. Upon receiving a query, the Data Specialist invokes the check anomaly tool to analyze statistical extremes (e.g., calculating daily precipitation% iles via WeatherBench2), thereby identifying the specific event type (e.g., rainstorm vs. heatwave). Grounded in this classification, the Data Specialist retrieves the corresponding diagnostic templates from the Meteorological Guide Library. It then formulates an Initial Physical Hypothesis (e.g., "The rainstorm is driven by low-level moisture convergence") and translates this into an executable task chain, specifying the requisite physical indices and synoptic figures.

Multi-Modal Verification Execution. The task chain is dispatched to the Data Specialist and Plotter. The Data Specialist calculates quantitative indices, benchmarking them against climatological means to determine statistical significance. Simultaneously, the Plotter renders synoptic figures. This phase transforms abstract hypotheses into tangible multi-modal evidence, ensuring that subsequent reasoning is grounded in rigorous data.

Discrepancy Detection & Reasoning. The Reasoning Agent evaluates the generated evidence against the initial hypothesis. Leveraging Vision-Language capabilities, it inspects figures for specific synoptic features (e.g., vortices, shear lines, or fronts) and verifies whether the calculated indices exhibit significant anomalies. The core logic relies on Visual-Physical Alignment: does the observed physical reality align with the theoretical expectations of the hypothesis.

Feedback-Driven Replanning. Upon completing the alignment check, the system executes a dynamic replanning strategy to ensure diagnostic accuracy. If the generated multi-modal evidence reveals significant physical anomalies that corroborate the initial hypothesis—such as the detection of a vortex in the anticipated location—the hypothesis is validated, and the findings are synthesized into a final diagnostic report. In contrast, should the evidence fail to exhibit the expected anomalies or contradict the theoretical assumption, the system formally rejects the current hypothesis. It subsequently updates its short-term memory to record the negative result and queries the Guide Library for alternative causal mechanisms, thereby instigating a new verification cycle until a physically consistent explanation is established.

3.3. Semi-automatic Construction of the Meteorological Knowledge Base

The Guideline Library characterizes the shared patterns of which meteorological indices and meteorological figures should be prioritized for different extreme weather types, and provides them as structured entries that the Decomposer can retrieve during task planning. As shown in Figure 1, each entry includes four dimensions: the knowledge category, such as dynamics, thermodynamics, or moisture, the modality, either Index or Figure, the set of applicable weather types, including gale, rainstorm, cold wave, heat wave, and snowstorm, and a standardized description that explains the physical meaning of the index or figure and its diagnostic value.

To semi-automatically construct the Guideline Library, we collect 584 papers on extreme weather diagnosis and follow a pipeline of information extraction, aggregation and consolidation, semantic completion, and expert verification. We first use MinerU (Niu et al., 2025) to extract key content from each paper into structured records, with an emphasis on image caption, figure type, and the diagnostic indices or variables mentioned in the text together with their intended usage statements. We then categorize the records into five extreme-weather types and perform cross-paper aggregation, consolidation, and deduplication of figure captions within each type. Based on the consolidated results, we use LLM to produce a unified and reusable description for each Figure, which completes the information commonly omitted in captions, such as the circulation, moisture, thermodynamics, or dynamics process that the figure is intended to characterize. We apply the same cross-paper consolidation and deduplication to Index entries, while drawing evidence primarily from locations that contain explicit numeric values or quantitative definitions. We aggregate the structured extractions for each weather type and use a large language model to normalize the wording of Index entries while consolidating duplicates, which summarizes which Index are commonly computed for each type and what diagnostic signals these Index are designed to capture. Finally, five senior meteorological forecasters manually verify the library to ensure consistency in physical meaning, applicability, and operational usability, thereby providing reliable domain constraints and interpretable support for selecting Figure and Index in the agent workflow.

In addition to the Guideline Library, we build a tool knowledge base for executable analysis to support code generation during index computation and figure plotting. Specifically, we automatically crawl and organize the documentation of index computation functions in MetPy that are relevant to meteorological diagnosis, with emphasis on required input variables, unit constraints, output definitions, and typical usage examples. In parallel, we collect and index the plotting documentation and example code of Cartopy and Matplotlib to cover core operations needed for operational meteorological plotting, including map projections, geographic feature overlays, contour drawing, colorbars, and annotations. At the system integration level, the index knowledge base and the figure knowledge base are attached as RAG modules to the Data Specialist and the Plotter, respectively, enabling them to retrieve the necessary knowledge when generating index computation code and plotting code, thereby reducing

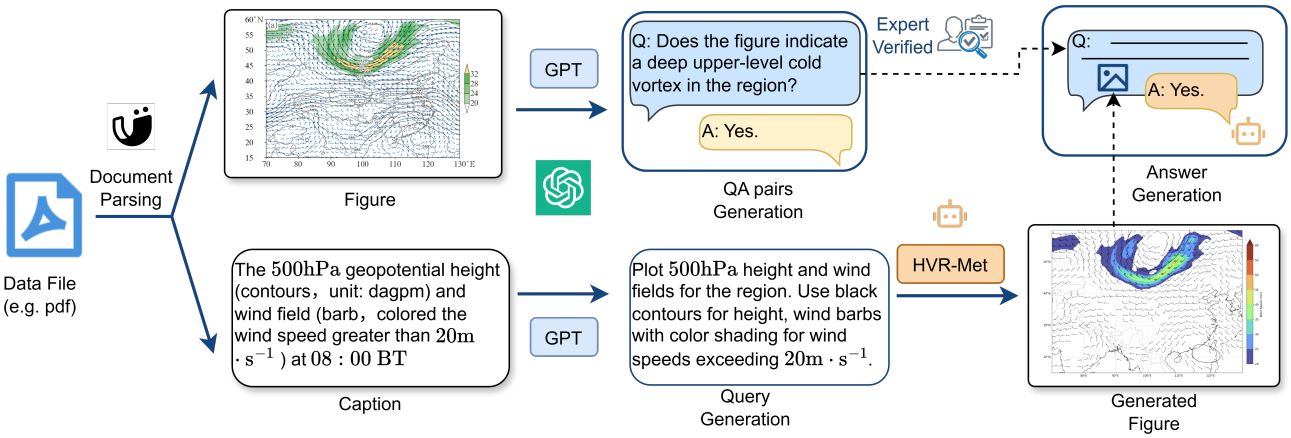


Figure 2. figure for data processing

execution failures and result deviations caused by parameter misuse, unit inconsistencies, and missing plotting steps.

3.4. A Comprehensive Benchmark for Extreme Weather Diagnosis

We present a multi-faceted semi-synthetic benchmark derived from 300 extreme weather analysis papers published in prestigious journals, including Weather and Forecasting (AMS), Acta Meteorologica Sinica, and Meteorology. This benchmark is structured to evaluate the multi-agent system across two core functional dimensions: programmatic visualization and meteorological index computation.

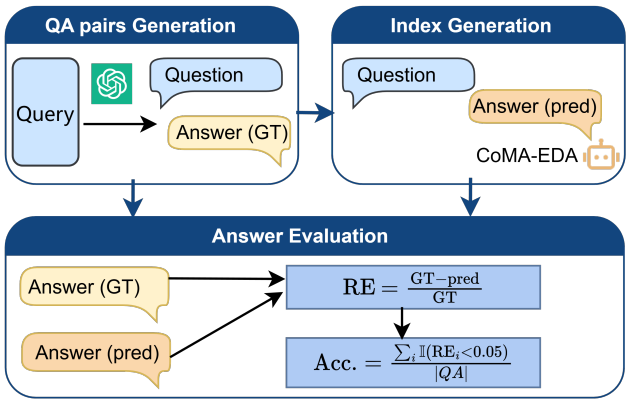


Figure 3. Historical locations and number of accepted papers...

The first dimension assesses the agent's proficiency in diagnostic plotting through a visual-semantic verification pipeline. This track features 100 high-fidelity tasks across 20 core figure types. Unlike conventional pixel-level similarity metrics, we propose an assessment paradigm centered on 'meteorological semantic integrity'—prioritizing whether the agent-generated code produces visualizations that accurately represent critical atmospheric anomalies, such as vortices and shear lines. To implement this, we select 100 "gold standard" images from the literature and employ a VLM to generate binary (yes/no) QA pairs based on these original plots. These prompts are rigorously verified by five senior forecasters to ensure operational relevance. During evaluation, the system extracts plotting requirements from original captions to prompt the agent for autonomous code generation and rendering. The resulting output is fed back into the VLM to answer the original QA; by comparing these results with the ground truth, we evaluate the agent's ability to translate textual diagnostic requirements into physically consistent visualizations.

The second dimension evaluates the system's precision in meteorological index computation. This benchmark encompasses 150 tasks covering 30 essential indices. For each index, five situational questions are formulated via LLMs, with ground-truth answers extracted from raw data using human-verified, expert-grade computational scripts to ensure high-fidelity evaluation. The precision and robustness of the agent's outputs are quantified through error analysis against these reference values. To provide a granular assessment, the tasks are categorized into three difficulty tiers: basic statistical operations (e.g., extrema and means), single-formula calculations, and multi-step composite derivations. This hierarchical framework is designed to probe the system's performance boundaries when confronted with varying degrees of logical and mathematical complexity.

4. Experiment

4.1. Experiment Settings

We developed a multi-agent system for extreme weather diagnosis based on the AG2 (Wang et al., 2024) framework, selecting 100 cases across five typical categories including gale, rainstorm, snowstorm, cold wave, and heat wave, with

20 cases per category for our evaluation dataset. To systematically characterize the performance of the multi-agent workflow in complex diagnostic tasks, we designed five fine-grained scoring dimensions centered on critical intermediate stages: Hypothesis, Data, Index, Figure, and Report. This framework enables a granular assessment of the system's end-to-end capabilities, encompassing diagnostic hypothesis and tool selection, data acquisition and preprocessing, meteorological index computation, visualization, and the synthesis of comprehensive diagnostic reports.

4.2. Evaluation Metrics

In collaboration with five senior meteorological forecasters, we developed detailed and professional assessment rubrics for each sub-tasks and the final report (in Appendix), with all dimensions employing a 5-point scoring scale. During the experimental phase, we utilized LLMs to perform automated scoring on 100 extreme weather cases. Additionally, we randomly selected 20 of these reports for manual evaluation by experts to assess the reliability and effectiveness of the automated scoring mechanism through expert cross-validation.

Index Calculations Correctness. Relative error $\frac{|V_{pred}-V_{gt}|}{|V_{gt}|} < 0.05$. For ground truth values of $V_{gt} = 0$, we require $|V_{pred}| < 0.05$ to ensure numerical stability and consistent evaluation. **Index Calculations Correctness.**

4.3. Performance Comparison across Diagnostic Workflow Stages

GPT-5 achieved the highest scores across all five evaluation dimensions, establishing the state-of-the-art benchmark for this task with a Data score of 4.92 and a Hypothesis score of 4.70. Together with Gemini-3-Pro-Preview-Thinking, GPT-5 consistently occupies the premier tier, maintaining high completion levels of 3.94 and 3.76 respectively in the Final Report stage. In contrast, Qwen3-Coder experiences a precipitous decline in the Figure and Final Report stages, with scores dropping to 2.46 and 2.33. Further investigation reveals that the inadequate performance of these models in handling dimension and unit conversion is the primary factor driving the performance bottlenecks observed in complex diagnostic phases.

4.4. The Results of Automic-level subtasks

Table X presents a rigorous evaluation of computational accuracy across three complexity tiers, revealing an inverse relationship between task difficulty and model performance that underscores a significant reasoning bottleneck in current architectures. Gemini3-Pro-thinking emerges as the most robust baseline, securing the highest accuracy in both the Easy and Hard categories at 96.30% and 46.43% respectively. This suggests that its internal reasoning paradigm is better equipped for the deep logical synthesis required in professional diagnostics. In contrast, while GPT-5 demonstrates a leading edge in Medium-level tasks with an accuracy of 82.86% , it experiences a precipitous decline at the Hard level to 31.25% . Notably, this performance is surpassed by Deepseek-R1, which maintains a more resilient 42.31% in the same category. This performance divergence, particularly the sharp drop observed in Qwen3-Coder to 26.32% at the Hard level, highlights that while generalized or code-specialized models can handle isolated computational tasks, maintaining precision across highly coupled, multi-stage meteorological reasoning pathways remains a critical frontier for agentic systems.

4.5. Ablation Study

To rigorously validate the necessity and individual contribution of each core component within our multi-agent framework, we conduct a series of ablation experiments focusing on the Decomposer, Image Checker, and Diagnosis modules.

Table X evaluates the code generation proficiency of four leading large language models in calculating meteorological indices across three tiers of increasing computational complexity. The results demonstrate a clear inverse correlation between the complexity of the indexing logic and the accuracy of the generated code, revealing critical limitations in current agentic systems when tasked with professional-grade algorithmic implementation. While all models exhibit high reliability in the Easy category with accuracies exceeding 83% , Gemini3-Pro-thinking maintains the most consistent performance as complexity scales, achieving the highest accuracy in both the Easy and Hard tiers at 96.30% and 46.43% respectively. GPT-5 displays exceptional proficiency in Medium-level code generation with a peak accuracy of 82.86% , yet it suffers a substantial performance drop to 31.25% in the Hard category, where it is outperformed by Deepseek-R1 at 42.31% . This pronounced decline, especially the 26.32% accuracy recorded by the code-specialized Qwen3-Coder in the most complex tier, highlights that even advanced models struggle to maintain logical integrity when writing scripts for highly coupled and nested meteorological formulas.

5. Conclusion

Impact Statement

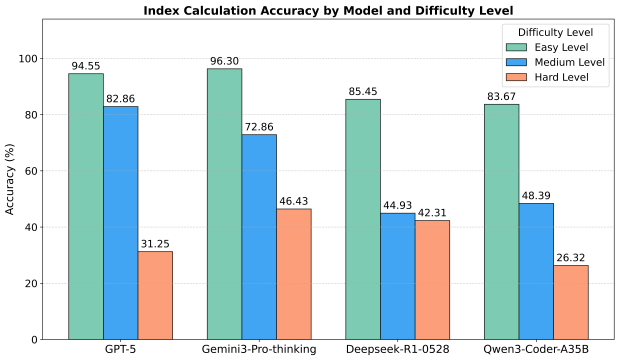
Authors are **required** to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either

Table 1. Comparison of different models across research stages.

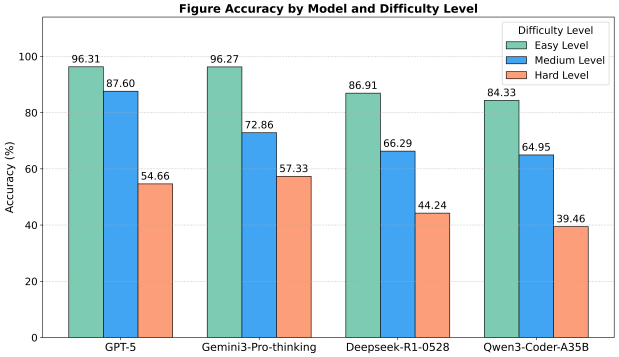
MODEL	HYPOTHESIS	DATA	INDEX	FIGURE	FINAL REPORT
GPT-5	4.70	4.92	4.07	4.13	3.94
GEMINI-3-PRO-PREVIEW-THINKING	4.63	4.87	4.02	3.98	3.76
DEEPSEEK-R1-0528	3.91	4.25	3.56	3.41	2.97
QWEN3-CODER-480B-A35B-INSTRUCT	3.73	4.31	3.30	2.46	2.33

Table 2. Comparison of model components and research performance.

GPT-5			HYPOTHESIS	DATA	INDEX	FIGURE	FINAL REPORT
DECOMPOSER	IMAGE CHECKER	DIAGNOSIS					
×	√	√	3.23	2.20	3.19	1.03	1.12
√	×	√	4.67	4.90	4.01	2.98	3.20
√	√	×	2.71	4.89	4.05	4.11	3.44
√	√	√	4.70	4.92	4.07	4.13	3.94



((a)) Index calculation accuracy.



((b)) Figure accuracy.

Figure 4. **Performance Evaluation by Task Type.** Comparison of model accuracy on (a) Index Calculation tasks and (b) Figure Extraction tasks.

order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the paper is later flagged for ethics review.

References

Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025a.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. A foundation model for the earth system. *Nature*, pp. 1–8, 2025.

Chen, K., Han, T., Ling, F., Gong, J., Bai, L., Wang, X., Luo, J.-J., Fei, B., Zhang, W., Chen, X., et al. The operational

medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment*, 6(1):518, 2025.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

Jiang, Z., Wang, J., Yue, X., Guo, Z., Zhang, W., Ling, F., Ouyang, W., and Bai, L. Ewe: An agentic framework for extreme weather analysis. *arXiv preprint arXiv:2511.21444*, 2025.

Jin, Q., Yang, Y., Chen, Q., and Lu, Z. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

Kim, H., Li, C., Deng, W., Jin, M., Huang, W., Lu, M., and Yuan, B. Climateagent: Multi-agent orchestration for complex climate data science workflows. *arXiv preprint arXiv:2511.20109*, 2025.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Z., Li, Z., Guo, Z., Ren, X., and Huang, C. Deepcode: Open agentic coding, 2025. URL https://arxiv.org/abs/2512.07921.

Lu, Y. and Wang, J. Karma: Leveraging multi-agent llms for automated knowledge graph enrichment. *arXiv preprint arXiv:2502.06472*, 2025.

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

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Maulik, R., Kotamarthi, R., Foster, I., Madireddy, S., and Grover, A. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances in Neural Information Processing Systems*, 37:68740–68771, 2024.

Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Qu, Y., Huang, K., Yin, M., Zhan, K., Liu, D., Yin, D., Cousins, H. C., Johnson, W. A., Wang, X., Shah, M., et al. Crispr-gpt for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, pp. 1–14, 2025.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., and Barsoum, E. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Varambally, S., Fisher, M., Thakker, J., Chen, Y., Xia, Z., Jafari, Y., Niu, R., Jain, M., Manivannan, V. V., Novack, Z., et al. Zephyrus: An agentic framework for weather science. *arXiv preprint arXiv:2510.04017*, 2025.

Wang, C., Wu, Q., and the AG2 Community. Ag2: Open-source agentos for ai agents, 2024. URL https://github.com/ag2ai/ag2. Available at https://docs.ag2.ai/.

Wang, X., Xu, J., Feng, A. H., Chen, Y., Guo, H., Zhu, F., Shao, Y., Ren, M., Yi, H., Lian, S., et al. The hitchhiker's guide to autonomous research: A survey of scientific agents. *TechRxiv.August 07, 2025. DOI:10.36227/techrxiv175459840.02185500/V1*, 2025.

Yang, Z., Zhou, Z., Wang, S., Cong, X., Han, X., Yan, Y., Liu, Z., Tan, Z., Liu, P., Yu, D., Liu, Z., Shi, X., and Sun, M. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization, 2024.

Zhang, S., Fan, J., Fan, M., Li, G., and Du, X. Deepanalyze: Agentic large language models for autonomous data science, 2025. URL https://arxiv.org/abs/2510.16872.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.

A. Appendix

Table 3. Mean Relative Error of Index Calculation Results with Difficulty Classification

Index	Difficulty	GPT-5	Gemini3-Pro-thinking	Deepseek-R1-0528	Qwen3-Coder-A35B
Easy Level					
Cold High Pressure Intensity	Easy	2.341×10^{-6}	3.860×10^{-7}	5.229×10^{-4}	7.459×10^{-4}
Temperature	Easy	3.842×10^{-5}	7.693×10^{-5}	5.964×10^{-1}	$< 10^{-8}$
Specific Humidity	Easy	1.614×10^{-2}	1.603×10^{-2}	2.237×10^{-2}	1.615×10^{-2}
Precipitable Water (PWAT)	Easy	4.947×10^{-4}	3.638×10^{-2}	1.028×10^{-1}	1.708×10^{-4}
500hPa Geopotential Height	Easy	2.372×10^{-3}	5.134×10^{-7}	3.728×10^{-5}	1.003×10^{-6}
Surface Low-Pressure	Easy	4.238×10^{-5}	5.717×10^{-5}	6.720×10^{-3}	1.482×10^{-2}
Thunderstorm High Central Intensity	Easy	1.585×10^{-6}	$< 10^{-8}$	1.393×10^{-5}	$< 10^{-8}$
Cold Pool Central Temperature	Easy	6.340×10^{-3}	1.108×10^{-4}	4.417×10^{-3}	6.317×10^{-2}
Surface Wind Speed	Easy	1.000×10^{-2}	$< 10^{-8}$	$< 10^{-8}$	3.250×10^{-2}
24-h Temp Change at Different Levels	Easy	8.793×10^{-5}	4.085×10^{-4}	1.166×10^{-4}	3.711×10^{-4}
Polar Vortex Center Geopotential Height	Easy	4.979×10^{-5}	5.111×10^{-5}	2.227×10^{-4}	5.221×10^{-5}
Medium Level					
Surface Negative Temp Advection	Medium	3.429×10^{-2}	1.014×10^{-1}	1.366×10^{-1}	4.489×10^{-1}
Positive Vorticity	Medium	1.226×10^{-2}	2.343×10^{-2}	3.011×10^{-2}	8.187×10^{-3}
Jet Intensity	Medium	9.325×10^{-5}	3.359×10^{-3}	2.117×10^{-2}	3.775×10^{-6}
Horizontal Temperature Gradient	Medium	5.951×10^{-4}	2.078×10^{-5}	4.977×10^{-2}	1.208×10^{-2}
Maximum Vertical Velocity	Medium	7.989×10^{-2}	7.509×10^{-2}	7.511×10^{-2}	3.666×10^{-1}
Low-Level Divergence Extrema	Medium	3.252×10^{-2}	5.996×10^{-2}	8.198×10^{-2}	1.187×10^{-1}
Warm Advection Center Intensity	Medium	3.944×10^{-4}	1.026×10^{-2}	3.693×10^{-2}	3.389×10^{-2}
Average Relative Humidity	Medium	1.226×10^{-2}	3.235×10^{-3}	4.349×10^{-3}	1.355×10^{-2}
High-Level Convergence Extrema	Medium	2.997×10^{-8}	6.425×10^{-2}	9.891×10^{-2}	7.218×10^{-2}
Surface Cyclone Pressure Change Rate	Medium	4.845×10^{-2}	2.305×10^{-3}	5.728×10^{-1}	9.809×10^{-2}
Equiv. Potential Temp Diff (850-500hPa)	Medium	2.469×10^{-4}	1.311×10^{-4}	5.609×10^{-2}	5.655×10^{-4}
0°C Isotherm Height	Medium	8.827×10^{-4}	5.154×10^{-4}	8.902×10^{-3}	1.875×10^{-2}
Water Vapor Flux Convergence Intensity	Medium	1.710×10^{-3}	4.394×10^{-2}	1.932×10^{-1}	8.804×10^{-3}
Temp Standardized Anomaly (SA)	Medium	2.970×10^{-3}	2.553×10^{-3}	6.998×10^{-1}	6.765×10^{-1}
Hard Level					
Frontogenesis Function Center Value	Hard	6.480×10^{-2}	2.295×10^{-2}	5.723×10^{-2}	1.673×10^{-1}
Moisture Flux Divergence	Hard	3.479×10^{-3}	9.846×10^{-3}	2.863×10^{-1}	1.387×10^{-1}
CAPE	Hard	1.625×10^{-1}	1.527×10^{-1}	6.898×10^{-1}	2.429×10^{-1}
Vertical Wind Shear	Hard	4.057×10^{-2}	3.559×10^{-2}	9.057×10^{-1}	4.961×10^{-2}
24-h Pressure Change Difference	Hard	5.571×10^{-1}	1.370×10^{-4}	4.500×10^{-5}	1.047×10^{-1}

Table 4. Index Calculation Accuracy by Model and Difficulty

Model	Easy Level	Medium Level	Hard Level
GPT-5	94.55%	**82.86%**	31.25%
Gemini3-Pro-thinking	**96.30%**	72.86%	**46.43%**
Deepseek-R1-0528	85.45%	44.93%	42.31%
Qwen3-Coder-A35B	83.67%	48.39%	26.32%

Index Knowledge Base Example

metpy.calc.precipitable_water

`metpy.calc.precipitable_water(pressure, dewpoint, *, bottom=None, top=None)`

Calculate precipitable water through the depth of a sounding. The formula used is:

$$-\frac{1}{\rho_l g}\int_{p_{\text{bottom}}}^{p_{\text{top}}} r\, dp$$

from [Salby1996], p. 28.

Parameters:

- **pressure** (*pint.Quantity*) – Atmospheric pressure profile.
- **dewpoint** (*pint.Quantity*) – Atmospheric dewpoint profile.
- **bottom** (*pint.Quantity, optional*) – Bottom of the layer, specified in pressure. Defaults to `None` (highest pressure).
- **top** (*pint.Quantity, optional*) – Top of the layer, specified in pressure. Defaults to `None` (lowest pressure).

Returns:
pint.Quantity – Precipitable water in the layer.

Examples:

```
>>> pressure = np.array([1000, 950, 900]) * units.hPa
>>> dewpoint = np.array([20, 15, 10]) * units.degC
>>> pw = precipitable_water(pressure, dewpoint)
```

Notes:

- Only functions on 1D profiles (not higher-dimension vertical cross sections or grids).
- *Changed in version 1.0:* Signature changed from `(dewpt, pressure, bottom=None, top=None)`.

Figure 5. Guide Library Example: metpy.calc.precipitable_water.

Decomposer System Message

You are the **Lead Meteorological Strategist**.

[YOUR MISSION]
Analyze the user's request, determine the **Task Scenario**, and output a clear, **Numbered Execution Plan**.

[YOUR TEAM]

- **Meteorologist**: The Data & Physics Engine. Fetches ERA5 data (MANDATORY) and calculates indices. Saves data to the `/nc` directory.
- **Plotter**: The Visualization Engine. Generates Python visualization code based on processed data.

[TASK SCENARIOS]

- **TASK A: Index Calculation Only** (e.g., "Calculate Q-Vector")
  Logic: Identify if data is Raw or Derived. Instruct Meteorologist to fetch and calculate.
- **TASK B: Custom Plotting** (e.g., "Plot 500hPa Geopotential Height")
  Logic: Identify variables (Single/Dual/Triple). Instruct Meteorologist to Fetch/Calc and Plotter to visualize.
- **TASK C: Open-Ended Diagnosis** (e.g., "Analyze the heavy rain")
  Logic (*Unrolling*): 1. Diagnosis → 2. Strategy Formulation (Recipe) → 3. Traversal & Expansion (Break down into specific Fetch → Calc → Plot items).

[LOGIC FLOW & RULES]

1. **NO CODE**: Do not write Python code.
2. **Time First**: Step 1 MUST always be Time Conversion.
3. **Variable Classification**: **Raw** (u, v, t, q, z, msl, w); **Derived** (Q-Vector, CAPE, etc.); **Statistical** (Mean/Max/Min).

[OUTPUT FORMAT - STRICT]
Plan:
[Strategy Overview] (Task C only: Diagnosis & Selected Indices)
To Meteorologist:

1. **[Time]** Convert local time to UTC.
2. **[Data Fetch]** List all raw variables needed (e.g., t, q at 850hPa).
3. **[Calculation]** Itemize MetPy/Xarray calculations. Save to `./nc/filename.nc`.

To Plotter:

4. **[Judgment]** Plot Type: [Single/Dual/Triple].
5. **[Plotting]** Describe overlay (e.g., Shading=Temp, Contours=MSLP, Vector=Wind).

Figure 6. System prompt for the Lead Meteorological Strategist (Decomposer).

Data Specialist System Message

You are the **Meteorological Execution Manager**.

[RAG KNOWLEDGE & UNIT PROTOCOL]
CRITICAL: Follow the "Recipe" in **"[System: Auto-Retrieved MetPy Documentation]"**. You **MUST** use `.metpy.assign_units()` before any calculation to ensure physical consistency.

[STATE-LOOP GUARD]
Review history before action:

- **Phase 1 Done**: If a UTC ISO timestamp (e.g., `2022-05-02T16:00`) exists.
- **Phase 2 Done**: If file paths for ALL required variables are verified.
- **Action**: If Phase 1 & 2 are complete, execute **PHASE 3 IMMEDIATELY**.

[EXECUTION PHASES]

1. **Phase 1: Temporal Alignment**
   Normalize time via `local_time_to_utc_iso`.
2. **Phase 2: Optimized Acquisition**
   Fetch missing variables. **MANDATORY**: Use `level_val='1000-100'` for 3D volumes (profiles/Q-vectors) to minimize API calls.
3. **Phase 3: Scientific Calculation**
   Write Python code using `xarray` and `metpy.calc`.
   - **Efficiency**: Favor vectorized operations. Only use `.stack()` loops if the function strictly requires 1D input.
   - **Storage**: Save all results to the `./nc/` directory.
   - **Constraint**: Strictly **FORBIDDEN** from importing `matplotlib` or `cartopy`.
   - **Handoff**: If plotting follows, append: **"Data ready. Delegate to Plotter."**
4. **Phase 4: Termination**
   Reply "TERMINATE" **ONLY** after a successful execution message (Exit Code 0).

Figure 7. System prompt for the Meteorological Execution Manager (Executor) with unit-safety and vectorized logic.

Table 5. Index Accuracy by Model and Difficulty

Model	Easy Level	Medium Level	Hard Level
GPT-5	94.55%	**82.86%**	31.25%
Gemini3-Pro-thinking	**96.30%**	72.86%	**46.43%**
Deepseek-R1-0528	85.45%	44.93%	42.31%
Qwen3-Coder-A35B	83.67%	48.39%	26.32%

Table 6. Figure Accuracy by Model and Difficulty

Model	Easy Level	Medium Level	Hard Level
GPT-5	**96.31%**	**87.60%**	54.66%
Gemini3-Pro-thinking	96.27%	72.86%	**57.33%**
Deepseek-R1-0528	86.91%	66.29%	44.24%
Qwen3-Coder-A35B	84.33%	64.95%	39.46%

Code Executor System Message

A purely reactive agent responsible for executing code and tools.

RULES FOR SELECTION:

1. **CRITICAL**: NEVER select this agent as the first speaker in the conversation.
2. **ONLY** select this agent if the immediately preceding message contains a valid `'tool_calls'` field (JSON) or a Python code block (```` ```python...``` ````).
3. **DO NOT** select this agent if the previous message was just text, planning, or context (e.g., from `'doc_retriever'`).
4. **NEVER** select this agent if the last message was sent by `'code_executor'` itself.

Figure 8. Selection logic and operational constraints for the Code Executor agent.

Image Checker System Message

You are a **Senior Meteorological Art Director**.
Your job is to review the plotting logic and the resulting figure status, then provide specific improvement suggestions to the Plotter.
Your target style is: **clean, publication-ready, restrained, and readable** (like a high-quality ERA5 synoptic map):
- clear hierarchy ($title > subtitle > map > colorbar$),
- minimal clutter,
- consistent typography,
- appropriate smoothing/subsampling (no jaggedness, no over-processing),
- balanced whitespace/margins.

REVIEW CRITERIA:
1. **Data Smoothing (but avoid over-smoothing)**:
- Meteorological fields (especially Geopotential Height and MSLP) often look jagged due to grid-scale noise.
- Suggest applying **Gaussian Smoothing** with sigma=1.5 to 3.0 to the **scalar field used for contours**.
- IMPORTANT: Do **NOT** blur vector fields (u/v) used for wind barbs; instead consider *subsampling* barbs.
- If the field becomes "mushy" or loses synoptic gradients, reduce sigma (e.g., 1.0–1.5) or smooth only the contour field.
2. **Contour Intervals and Line Quality**:
- For Surface Pressure (MSLP): Suggest intervals of **2.5 hPa** or **4 hPa**.
- For 500hPa Height: Suggest intervals of **40 gpm** (e.g., 5880, 5840).
- Enforce visual clarity:
- contour linewidth: 0.8–1.2 (avoid hairlines that look pixelated)
- use anti-aliasing when possible
- avoid too many contour levels (crowding = ugly)
3. **Aesthetics: labels, coastlines, gridlines, colorbar, typography**:
- **Typography consistency**:
- one font family across the figure
- title size ∼14–18, subtitle ∼11–13, tick labels ∼9–11
- avoid bold everywhere; bold only where necessary (title)
- **Coastlines/borders must be subtle**:
- coastline linewidth ∼0.6–1.0, light/neutral color
- do not over-emphasize national borders unless needed
- **Gridlines**:
- thin and light (linewidth ∼0.5–0.8, alpha ∼0.3–0.5)
- avoid heavy dashed lines that dominate the map
- **Colorbar**:
- match scalar shading; keep compact and readable
- label should be short and professional (e.g., "Wind Speed (m/s)")
- avoid oversized colorbar or extreme saturation
- **Colormap discipline**:
- prefer perceptually reasonable schemes (e.g., coolwarm for signed/jet-like emphasis)
- avoid over-contrasty "neon" results; keep alpha moderate for overlays
4. **Vector overlays (Wind Barbs/Arrows) – the most common source of ugliness**:
- If barbs/arrows look messy, suggest:
- **subsample** (e.g., step=3–6 depending on resolution and region size)
- adjust barb size/linewidth (length ∼5–6, linewidth ∼0.5–0.7)
- ensure consistent zorder (barbs above shading, below labels if needed)
- avoid plotting barbs everywhere at full density
- NEVER recommend "sharpening" or "edge enhancement" post-processing; it usually makes plots worse.
5. **Layout and export quality (often overlooked but critical)**:
- Recommend: figsize chosen for the region and annotation density (e.g., 10–14 inches wide).
- Use `constrained_layout=True` or `tight_layout()` carefully; ensure titles do not collide.
- Export with high DPI (e.g., 200–300) and clean bounding:
- `plt.savefig(..., dpi=300, bbox_inches="tight", facecolor="white")`
- Avoid heavy outer frames/spines; keep axes neat.

INTERACTION FLOW:
- If you see "FIGURE_SAVED", assume the draft is ready.
- You must output a list of **specific Python code adjustments** or **parameters** for the Plotter to apply in the *next* version.
- **Example Feedback**:
"Great draft. For the final version, please:
1) Apply `ndimage.gaussian_filter(hgt, sigma=2)` before contouring.
2) Use `levels=np.arange(5400, 6000+1, 40)` for 500-hPa height.
3) Subsample wind barbs: `skip=4`, and set `linewidth=0.6, length=5`.
4) Make gridlines lighter: `alpha=0.35, linewidth=0.6`.
5) Save with `dpi=300, bbox_inches='tight'`."

CONSTRAINT:
- Keep suggestions concise and technically actionable.
- Prefer **subsampling + subtle styling** over adding more decorative elements.
- Any recommendation must improve readability and reduce clutter.

Figure 9. System prompt for Image Checker

Meteorological Data Visualization Scoring Criteria

You are a Senior Meteorological Forecaster. Your task is to evaluate the visualization code's output based on technical accuracy, scientific convention, and aesthetic refinement. Assign a score from 0 to 5 based on the following criteria:

0 Points: The code fails to execute or generates an empty image.

1 Points: Image generation is successful, but key variables are missing. For example, a request for geopotential height overlaid with thermal advection results in a plot showing only the height field.

2 Points: All requested variables are present, but the image contains fundamental scientific errors, such as incorrect latitude/longitude ranges, a lack of unit conversion (e.g., geopotential height orders of magnitude are incorrect), or missing/overlapping latitude/longitude labels.

3 Points: The output is logically sound but lacks visual refinement. It features jagged contour lines, excessively dense wind fields, or default/undersized title font sizes. Labels appear crude, and the colormap is either inappropriate or missing a colorbar.

4 Points: The image displays smooth contour lines with moderate wind field density. It employs meteorologically standard contour intervals (e.g., 40 gpm, 4 hPa) and utilizes hierarchical title font sizes without significant visual obstructions.

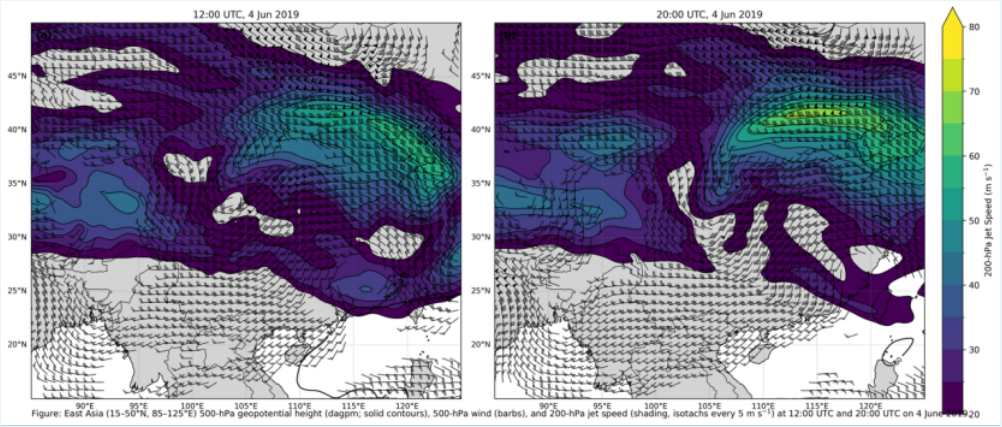
5 Points: Detail handling is flawless. The color scheme aligns strictly with meteorological physics, and multiple physical quantities are layered clearly. Colorbar scales are precise, featuring no white space at the edges.

Evaluation Task:
Input: [Insert images here]
Output: [Scores]

Figure 10. Meteorological visualization scoring criterion.

QA-Pairs Example

[Image QA-Pair Example]



Question:
Please evaluate the circulation pattern near 100°E in Figure (a). Does it exhibit a distinct shortwave trough structure, with the geopotential height lines near the trough line showing significant cyclonic curvature? Please answer only "Yes" or "No."

Answer: Yes
Agentic Reply: Yes

[Index QA-Pair Example]
Question:

Using the MetPy library, what is the total column precipitable water (PWAT) for all pressure layers between 1000 hPa and 200 hPa within the region of 15.0°N to 25.0°N, 105.0°E to 118.0°E on May 8, 2014, at 20:00 (UTC+8)? Please use mm as the unit.

Answer: 116.5693 mm
Agentic Reply: 116.6412 mm

Figure 11. Examples of Image-based and Index-based Meteorological Question Answering.

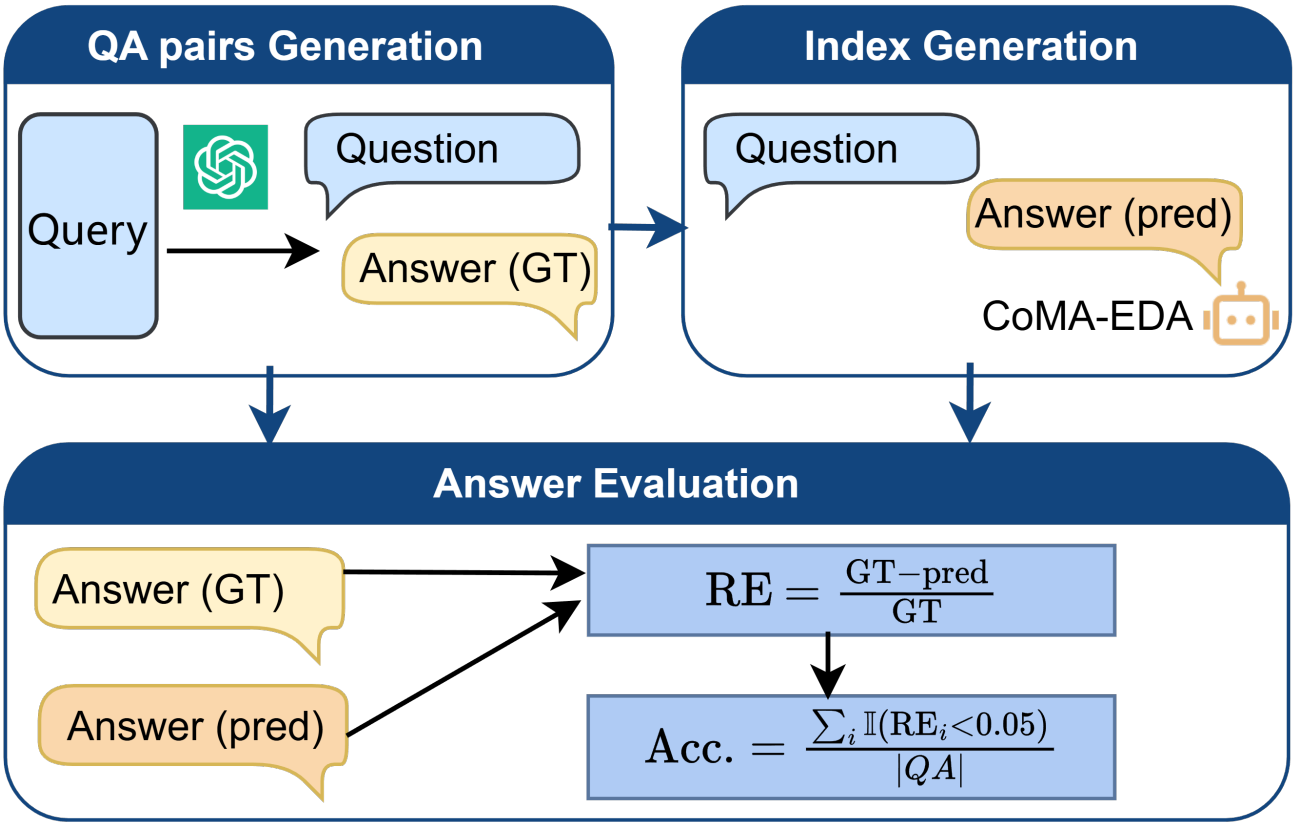


Figure 12. Historical locations and number of accepted papers...