# **AgentCaster: Reasoning-Guided Tornado Forecasting**

# **Anonymous Author(s)**

Affiliation Address email

## **Abstract**

There is a growing need to evaluate Large Language Models (LLMs) on complex, high-impact, real-world tasks to assess their true readiness as reasoning agents. To address this gap, we introduce AgentCaster, a contamination-free framework employing multimodal LLMs end-to-end for the challenging, long-horizon task of tornado forecasting. Within AgentCaster, models interpret heterogeneous spatiotemporal data from a high-resolution convection-allowing forecast archive. We assess model performance over a 40-day period featuring diverse historical data, spanning several major tornado outbreaks and including over 500 tornado reports. Each day, models query interactively from a pool of 3,625 forecast maps and 40,125 forecast soundings for a forecast horizon of 12-36 hours. Probabilistic tornado-risk polygon predictions are verified against ground truths derived from geometric comparisons across disjoint risk bands in projected coordinate space. To quantify accuracy, we propose domain-specific Tornado-Bench and Tornado-Hallucination metrics, with TornadoBench highly challenging for both LLMs and domain expert human forecasters. Notably, human experts significantly outperform state-of-the-art models, which demonstrate a strong tendency to hallucinate and overpredict risk intensity, struggle with precise geographic placement, and exhibit poor spatiotemporal reasoning in complex, dynamically evolving systems. AgentCaster aims to advance research on improving LLM agents for challenging reasoning tasks in critical domains.

## 1 Introduction

2

3

5

6

7

8

9

10 11

12

13

14

15 16

17

18

19

20

21

- LLMs have rapidly progressed from text-only pattern recognizers to general-purpose reasoning agents capable of planning, using tools, and operating in multi-turn interactions [2, 3, 24, 46, 23, 38]. As these models are increasingly envisioned for autonomous roles, evaluating their true capabilities on more challenging and higher impact problems becomes paramount [37]. This evaluation gap inhibits our understanding of both LLM limitations and progress, particularly in domains where reliable performance is critical.
- Severe convective weather represents precisely such a domain. Predicting tornadoes carries immense importance; from 2010 through 2024, tornadoes in the United States caused over USD 25 billion in property damage and claimed more than 1,200 lives [33]. Human forecasters at the NWS Storm Prediction Center (SPC) must synthesize heterogeneous high-resolution numerical weather prediction (NWP) fields, examine vertical atmospheric profiles, reason across extensive geographic areas and timeframes, and ultimately produce nested probabilistic polygons that communicate risk to emergency managers and the public [7]. However, despite decades of research, tornado forecasting remains notoriously challenging.
- To address this evaluation gap, we develop a framework that can rigorously test LLM capabilities in a real-world forecasting environment. We introduce AgentCaster, a novel, contamination-free evaluation framework that assesses multimodal LLM agents end-to-end on tornado forecasting.

As shown in Figure 2 in Appendix A, within AgentCaster, LLMs function as AI meteorologists, interactively querying a rich archive of historical, high-resolution weather forecast data. Finally, agents synthesize their findings to produce probabilistic tornado risk predictions as geospatial polygons in standard GeoJSON format, analogous to official SPC outlooks. See Appendix B for related work.

Our contributions include: (1) AgentCaster, a multimodal, interactive, and contamination-free agent framework for evaluating LLM reasoning on the challenging and real-world task of tornado 45 forecasting using daily generated high-resolution forecast data; (2) domain-specific evaluation 46 metrics based on geometric verification against ground truths; (3) a curated 40-day benchmark 47 dataset comprising 145,000 processed forecast maps, on-demand generation for 1,605,000 forecast 48 soundings, SPC outlooks for baseline comparison, and processed ground truth tornado reports; (4) 49 initial evaluation of state-of-the-art multimodal LLMs against human expert baselines; and (5) release 50 of all code and datasets to facilitate reproducibility and further research. We hope AgentCaster will 51 catalyze research on high-impact, real-world reasoning tasks and motivate progress towards agents that can meaningfully assist human experts in critical domains.

# 54 2 AgentCaster

#### 55 2.1 Framework Overview

AgentCaster is an interactive environment where an LLM agent is placed in the role of an AI meteorologist tasked with issuing a tornado risk forecast for the Continental United States (CONUS). Agents make sequential requests for meteorological data products using a defined set of tools. They begin with access to a wide array of forecast maps and can subsequently request vertical atmospheric profiles for specific locations and times. The agent must predict the probability of a tornado occurring within 25 miles of any point during a 24-hour period from 12:00 UTC on the target date to 12:00 UTC the following day, aligning with operational forecasting timelines used by human meteorologists. For all experiments reported here, we freeze a contiguous 40-day benchmark window (March 1, 2025 to April 9, 2025) to ensure fair composition and reproducibility, even though the framework is designed for live daily forecasting.

AgentCaster's design enables: (1) realistic assessment of domain expertise by requiring reasoning similar to expert human forecasters; (2) interactive exploration through deliberate tool usage to analyze heterogeneous data; and (3) contamination-free evaluation using rolling numerical weather prediction archives. Distinct from text-based or purely simulated environments, AgentCaster dynamically integrates real-world, multimodal meteorological data (including on-demand visual sounding generation triggered by agent requests) within an interactive loop, as illustrated by Figure 2. Agent-Caster is also extensible, allowing for the future inclusion and modification of different NWP models, prediction objectives, or prediction horizons.

## 74 2.2 Meteorological Data Sources

AgentCaster utilizes archived data from daily runs of the High-Resolution Rapid Refresh (HRRRv4) 75 [9] model, processed into formats suitable for multimodal LLM inputs. The HRRRv4 is the state-76 of-the-art, 3-km resolution, convection-allowing numerical weather prediction system operated by NOAA, built on the WRF-ARW dynamical core [27]. For each day, we process the 00:00 UTC 78 HRRR model run to extract and visualize all 145 available map products. These include convective 79 parameters (CAPE, CIN), wind fields (shear, helicity), moisture variables, temperature profiles, 80 and simulated radar reflectivity, among others. See Appendix K for the full list of forecast map 81 products. To access full vertical atmospheric structure near any given point, the framework provides 82 forecast soundings derived from HRRR BUFKIT data. These are generated on-demand during 83 the agent's interaction. This on-demand generation, coupled with a daily quota (defaulting to 50 84 requests), encourages: (1) targeted geographic focus; (2) efficient context window use; and (3) strategic decision-making under resource constraints.

## 87 2.3 Agent Interaction Loop

The agent operates in a multi-turn conversational loop that mirrors an iterative forecast workflow (prompts/code in Appendix H). It begins with an initial prompt specifying the task, date, and available

tools, then issues map and sounding requests as needed. Each request is processed by the backend, which returns confirmations and embeds the requested images in-line; sounding replies also report the remaining daily quota. The agent analyzes these multimodal inputs and decides what to request next, repeating a request–receive—analyze cycle until the evidence is sufficient. When confident, it submits the final tornado-risk polygons (GeoJSON), which ends the day's interaction.

# 95 3 TornadoBench and TornadoHallucination

#### 96 3.1 Ground Truth Generation

Converting discrete tornado reports into a continuous probability field requires spatial smoothing to capture the inherent uncertainty of tornado occurrences. To generate an objective verification target, we adapt and extend the Practically Perfect Forecast (PPF) methodology of [14], developing a multistep approach to construct high-resolution ground-truth risk fields. Our modified approach transforms discrete tornado observations into a continuous probability field representing a theoretically ideal probabilistic forecast, as displayed in Figure 3. See Appendix C for full details. See Appendix F for the dataset composition and summary table.

#### 3.2 TornadoBench Score and TornadoHallucination Metrics

We propose TornadoBench as the primary metric for AgentCaster. It is designed to evaluate the 105 agent's ability to accurately delineate the location, extent, and intensity of tornado risk; it addresses the 106 limitations of standard metrics by incorporating domain-specific weighting and geometric accuracy 107 across multiple probability thresholds. See Appendix D for the complete definition and formulas. 108 LLMs are known to hallucinate information [48, 15], and in a forecasting context, we define this as 109 predicting risk where none exists or predicting risk in an entirely non-overlapping location on a risk 110 day. Evaluating hallucinations is particularly important in tornado prediction, where false alarms 111 can lead to unnecessary costs and public complacency. We introduce two metrics to quantify these 112 behaviors. See Appendix E for full definitions. 113

# 4 Experiments and Evaluation

We evaluated a suite of reasoning and non-reasoning multimodal LLMs with knowledge cutoff 115 dates prior to March 1st. The human expert baseline is the first official SPC Day 1 Convective Outlook issued for the 12:00 UTC cycle, processed identically to agent predictions. All LLM agents were initialized with a detailed system prompt (see Appendix H for full prompts) outlining their 118 role as an AI meteorologist, the forecasting objective, data access tools, and the GeoJSON output 119 format requirements. The primary forecasting accuracy, hallucination metrics, and maximum risk 120 matching for the LLM configurations and the SPC baseline are presented in Table 1. Agent interaction 121 statistics and centroid distance errors are detailed in Table 2 (centroid computation described in Appendix I). The SPC baseline achieves a TornadoBench score of 18.31%, significantly outperforming all evaluated LLM agents. Among the LLM agents, performance varied, with the highest-scoring 125 models achieving TornadoBench scores below 10%. A notable challenge for several LLMs was the consistent generation of valid GeoJSON outputs. The models with the fewest valid predictions, 126 gemini-2.5-flash-preview:thinking (16 days), also had the lowest TornadoBench scores. 127

Within the GPT-5 family, increasing reasoning correlates with a monotonic drop in TornadoBench 128 (8.51%, 7.23%, 6.28%, 3.54% for gpt-5-minimal, gpt-5-low, gpt-5-medium, and gpt-5-high, re-129 spectively). This degradation occurs despite mixed shifts in hallucination severity. Furthermore, 130 claude-3.7-sonnet (non-thinking) marginally outperforms its thinking variant on TornadoBench 131 (6.79% vs. 6.64%). LLM agents exhibit a strong tendency towards hallucinations. The TornadoHallu-132 cinationHard scores for LLMs were substantially higher than SPC's, with not only more frequent 133 but also more severe hallucinations or complete misplacement of risk areas. The average centroid 134 distance errors indicate significant challenges for LLMs in accurately placing the core of the predicted 135 tornado threat, with most errors exceeding 400-500 km, compared to SPC's 182 km (overall) and 236 136 km (max risk). Agent interaction patterns varied across models. Except for one model, the number of sounding requests remained well below the daily quota of 50.

Table 1: Primary forecasting performance metrics. For TornadoHallucination metrics, lower is better. Max Risk Match shows the percentage of days the model's maximum predicted risk was Under/Match/Over the ground truth maximum risk.

Model	TornadoBench (%)	TornadoHallucination Simple	TornadoHallucination Hard	Max Risk Match (%) Under / Match / Over	
SPC (Human Expert)	18.31	0.275	0.70	5.0 / 55.0 / 40.0	
gpt-5-minimal	8.51	0.385	2.56	12.8 / 20.5 / 66.7	
gpt-5-low	7.23	0.444	1.92	11.1 / 27.8 / 61.1	
claude-3.7-sonnet	6.79	0.400	3.30	10.0 / 22.5 / 67.5	
claude-3.7-sonnet:thinking	6.64	0.359	3.10	17.9 / 23.1 / 59.0	
gpt-5-medium	6.28	0.484	2.65	9.7 / 22.6 / 67.7	
gpt-4.1	5.63	0.444	3.64	11.1 / 19.4 / 69.4	
gemini-2.5-pro-preview-03-25	4.26	0.406	4.50	15.6/21.9/62.5	
grok-4	3.85	0.538	8.85	2.6/7.7/89.7	
gpt-5-high	3.54	0.500	2.30	16.7/0.0/83.3	
o4-mini-high	3.37	0.528	5.39	11.1/13.9/75.0	
o3	3.27	0.550	5.50	10.0/7.5/82.5	
gemini-2.5-flash-preview:thinking	1.57	0.625	4.50	6.3/6.3/87.5	

Table 2: Agent interaction statistics and centroid distance errors.

Model	Prediction Days	Centroid Dist. (Avg. / Max Risk) (km)	Avg. Assistant Turns	Avg. Tool Calls	Sounding Requests (Avg. / Max)	
SPC (Human Expert)	40	182 / 236	N/A	N/A	N/A	
gpt-5-minimal	39	358 / 354	8.93	18.32	0.12 / 3	
gpt-5-low	36	417 / 469	4.00	35.58	0.05 / 1	
claude-3.7-sonnet	40	405 / 441	21.80	21.80	4.83 / 8	
claude-3.7-sonnet:thinking	39	474 / 493	21.57	21.57	4.97 / 11	
gpt-5-medium	31	398 / 447	4.45	41.27	0.05 / 1	
gpt-4.1	36	361 / 377	11.32	23.07	4.47 / 13	
gemini-2.5-pro-preview-03-25	32	494 / 561	5.55	18.38	2.23 / 5	
grok-4	39	450 / 487	5.83	24.23	4.00 / 8	
gpt-5-high	30	449 / 525	4.75	39.25	0.40 / 4	
o4-mini-high	36	583 / 623	6.58	6.55	0.12 / 1	
03	40	478 / 564	13.70	13.70	0.62 / 5	
gemini-2.5-flash-preview:thinking	16	601 / 595	7.05	32.38	2.70 / 50	

Among the three 30% risk days in our benchmark, we show March 14, 2025, the day whose SPC daily 139 TornadoBench score is closest to the top model's score. On this day, the top LLM agent achieved a 140 daily TornadoBench score of 9.45%, approaching SPC's 9.51% (Figure 1).



Figure 1: Evaluation of SPC and the top performing model on March 14, 2025. Overlapping solution regions are shaded.

#### 5 **Conclusion**

141

We introduced AgentCaster, a novel framework for evaluating multimodal LLM agents on the task of 143 tornado forecasting. Through an interactive environment utilizing high-resolution meteorological 144 data, AgentCaster assesses agentic reasoning. Our metrics, TornadoBench and TornadoHallucination, 145 applied over a 40-day period, revealed gaps between current LLM capabilities and human expert 146 performance. By establishing a challenging benchmark, we aim to drive progress toward more 147 capable and reliable AI agents while highlighting the current limitations of LLMs.

## 149 References

- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-Weather:
   A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. Nov. 3, 2022.
   DOI: 10.48550/arXiv.2211.02556. URL: http://arxiv.org/abs/2211.02556.
- 153 [2] Tom B. Brown et al. Language Models are Few-Shot Learners. July 22, 2020. DOI: 10.48550/ arXiv.2005.14165. URL: http://arxiv.org/abs/2005.14165.
- 155 [3] Yupeng Chang et al. "A Survey on Evaluation of Large Language Models". In: *ACM Trans.*156 Intell. Syst. Technol. 15.3 (Mar. 29, 2024), 39:1–39:45. ISSN: 2157-6904. DOI: 10.1145/
  157 3641289. URL: https://dl.acm.org/doi/10.1145/3641289.
- Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. *ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos*. July 2, 2024. DOI: 10.48550/arXiv.2406.19392. URL: http://arxiv.org/abs/2406.19392.
- Lin Chen et al. Are We on the Right Way for Evaluating Large Vision-Language Models? Apr. 9, 2024. DOI: 10.48550/arXiv.2403.20330. URL: http://arxiv.org/abs/2403.20330.
- [6] Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. ARC Prize 2024:
   Technical Report. Jan. 8, 2025. DOI: 10.48550/arXiv.2412.04604. URL: http://arxiv.org/abs/2412.04604.
- Stephen F. Corfidi. "The Birth and Early Years of the Storm Prediction Center". In: (Aug. 1, 1999). Section: Weather and Forecasting. ISSN: 1520-0434. URL: https://journals.ametsoc.org/view/journals/wefo/14/4/1520-0434\_1999\_014\_0507\_tbaeyo\_2\_0\_co\_2.xml.
- 170 [8] Charles A. Doswell, Steven J. Weiss, and Robert H. Johns. "Tornado forecasting: A review".

  171 In: Geophysical Monograph Series. Ed. by C. Church, D. Burgess, C. Doswell, and R. Davies172 Jones. Vol. 79. Washington, D. C.: American Geophysical Union, 1993, pp. 557–571. ISBN:
  173 978-0-87590-038-4. DOI: 10.1029/GM079p0557. URL: http://doi.wiley.com/10.
  174 1029/GM079p0557.
- David C. Dowell et al. "The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description". In: (Aug. 3, 2022). Section: Weather and Forecasting. DOI: 10.1175/WAF-D-21-0151.1. URL: https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0151.1.xml.
- Neel Guha et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. Aug. 20, 2023. DOI: 10.48550/arXiv.2308.11462. URL: http://arxiv.org/abs/2308.11462.
- Pamela L. Heinselman et al. "Warn-on-Forecast System: From Vision to Reality". In: (Dec. 22, 2023). Section: Weather and Forecasting. DOI: 10.1175/WAF-D-23-0147.1. URL: https://journals.ametsoc.org/view/journals/wefo/39/1/WAF-D-23-0147.1.xml.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. *Measuring Massive Multitask Language Understanding*. Jan. 12, 2021. DOI: 10.48550/arXiv.2009.03300. URL: http://arxiv.org/abs/2009.03300.
- Aaron J. Hill, Gregory R. Herman, and Russ S. Schumacher. "Forecasting Severe Weather with Random Forests". In: (May 1, 2020). Section: Monthly Weather Review. DOI: 10.1175/MWR-D-19-0344.1. URL: https://journals.ametsoc.org/view/journals/mwre/148/5/mwr-d-19-0344.1.xml.
- Nathan M. Hitchens, Harold E. Brooks, and Michael P. Kay. "Objective Limits on Forecasting Skill of Rare Events". In: (Apr. 1, 2013). Section: Weather and Forecasting. DOI: 10.1175/WAF-D-12-00113.1. URL: https://journals.ametsoc.org/view/journals/wefo/28/2/waf-d-12-00113\_1.xml.
- Lei Huang et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2 (Mar. 31, 2025), pp. 1–55. ISSN: 1046-8188, 1558-2868. DOI: 10.1145/3703155. URL: http://arxiv.org/abs/2311.05232.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits.

  What Disease does this Patient Have? A Large-scale Open Domain Question Answering
  Dataset from Medical Exams. Sept. 28, 2020. DOI: 10.48550/arXiv.2009.13081. URL:
  http://arxiv.org/abs/2009.13081.

- 204 [17] Ryan Lagerquist, Amy McGovern, Cameron R. Homeyer, David John Gagne Ii, and Travis
  205 Smith. "Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Pre206 diction". In: (June 24, 2020). Section: Monthly Weather Review. DOI: 10.1175/MWR-D207 19-0372.1. URL: https://journals.ametsoc.org/view/journals/mwre/148/7/
  208 mwrD190372.xml.
- 209 [18] Remi Lam et al. GraphCast: Learning skillful medium-range global weather forecasting.
  210 Aug. 4, 2023. DOI: 10.48550/arXiv.2212.12794. URL: http://arxiv.org/abs/2212.
  211 12794.
- [19] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić.
   TopViewRS: Vision-Language Models as Top-View Spatial Reasoners. June 4, 2024. DOI:
   10.48550/arXiv.2406.02537. URL: http://arxiv.org/abs/2406.02537.
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. Apr. 6, 2025. DOI: 10.48550/arXiv.2501.02189. URL: http://arxiv.org/abs/2501.02189.
- 219 [21] Percy Liang et al. *Holistic Evaluation of Language Models*. Oct. 1, 2023. DOI: 10.48550/ 220 arXiv.2211.09110. URL: http://arxiv.org/abs/2211.09110.
- 221 [22] Xiao Liu et al. *AgentBench: Evaluating LLMs as Agents*. Oct. 25, 2023. DOI: 10.48550/ 222 arXiv.2308.03688. URL: http://arxiv.org/abs/2308.03688.
- Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. *AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents*. Dec. 23, 2024. DOI: 10.48550/arXiv.2401.13178. URL: http://arxiv.org/abs/2401.13178.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. *A Comprehensive Overview of Large Language Models*. Oct. 17, 2024. DOI: 10.48550/arXiv.2307.06435. URL: http://arxiv.org/abs/2307.06435.
- Jaideep Pathak et al. FourCastNet: A Global Data-driven High-resolution Weather Model using
  Adaptive Fourier Neural Operators. Feb. 22, 2022. DOI: 10.48550/arXiv.2202.11214.
  URL: http://arxiv.org/abs/2202.11214.
- [26] Long Phan et al. Humanity's Last Exam. Apr. 19, 2025. DOI: 10.48550/arXiv.2501.14249.
   URL: http://arxiv.org/abs/2501.14249.
- Jordan G. Powers et al. "The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions". In: (Aug. 1, 2017). Section: Bulletin of the American Meteorological Society. DOI: 10.1175/BAMS-D-15-00308.1. URL: https://journals.ametsoc.org/view/journals/bams/98/8/bams-d-15-00308.1.xml.
- [28] Yujia Qin et al. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world
   APIs. Oct. 3, 2023. DOI: 10.48550/arXiv.2307.16789. URL: http://arxiv.org/abs/2307.16789.
- Alec Radford et al. Learning Transferable Visual Models From Natural Language Supervision. Feb. 26, 2021. DOI: 10.48550/arXiv.2103.00020. URL: http://arxiv.org/abs/2103.00020.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. *Identifying the Risks of LM Agents with an LM-Emulated Sandbox*. May 17, 2024. DOI: 10.48550/arXiv.2309.15817. URL: http://arxiv.org/abs/2309.15817.
- 250 [31] Karan Singhal et al. Large Language Models Encode Clinical Knowledge. Dec. 26, 2022. DOI: 10.48550/arXiv.2212.13138. URL: http://arxiv.org/abs/2212.13138.
- Aarohi Srivastava et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. June 12, 2023. DOI: 10.48550/arXiv.2206.04615. URL: http://arxiv.org/abs/2206.04615.
- 255 [33] Storm Events Database | National Centers for Environmental Information. URL: https://www.ncdc.noaa.gov/stormevents/.
- Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min zhang. Living in the Moment: Can Large Language Models Grasp Co-Temporal Reasoning? June 13, 2024. DOI: 10.48550/arXiv.2406.09072. URL: http://arxiv.org/abs/2406.09072.

- Mark S. Veillette, James M. Kurdzo, Phillip M. Stepanian, John Y. N. Cho, Siddharth Samsi, and Joseph McDonald. A Benchmark Dataset for Tornado Detection and Prediction using Full-Resolution Polarimetric Weather Radar Data. Jan. 26, 2024. DOI: 10.48550/arXiv. 2401.16437. URL: http://arxiv.org/abs/2401.16437.
- Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. *Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models*. Nov. 4, 2024. DOI: 10.48550/arXiv.2406.14852. URL: http://arxiv.org/abs/2406.14852.
- 269 [37] Lei Wang et al. "A survey on large language model based autonomous agents". In: Frontiers of Computer Science 18.6 (Mar. 22, 2024), p. 186345. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: https://doi.org/10.1007/s11704-024-40231-1.
- [38] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji.
   MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. Mar. 12,
   2024. DOI: 10.48550/arXiv.2309.10691. URL: http://arxiv.org/abs/2309.10691.
- Yubo Wang et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. Nov. 6, 2024. DOI: 10.48550/arXiv.2406.01574. URL: http://arxiv.org/abs/2406.01574.
- 278 [40] Colin White et al. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. Apr. 18, 2025. DOI: 10.48550/arXiv.2406.19314. URL: http://arxiv.org/abs/2406.19314.
- Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. ST-Think: How Multimodal Large Language Models Reason About 4D Worlds from Ego-Centric Videos. version: 2. Apr. 23, 2025. DOI: 10.48550/arXiv.2503.12542. URL: http://arxiv.org/abs/2503.12542.
- Qianqian Xie et al. FinBen: A Holistic Financial Benchmark for Large Language Models.

  June 19, 2024. DOI: 10.48550/arXiv.2402.12659. URL: http://arxiv.org/abs/2402.

  12659.
- John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. Nov. 11, 2024. DOI: 10.48550/arXiv.2405.15793. URL: http://arxiv.org/abs/2405.15793.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and
   Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). Oct. 11,
   2023. DOI: 10.48550/arXiv.2309.17421. URL: http://arxiv.org/abs/2309.17421.
- Zonghai Yao et al. MedQA-CS: Benchmarking Large Language Models Clinical Skills Using an AI-SCE Framework. Oct. 2, 2024. DOI: 10.48550/arXiv.2410.01553. URL: http://arxiv.org/abs/2410.01553.
- Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. Feb. 28, 2024. DOI: 10.48550/arXiv.2402.18013. URL: http://arxiv.org/abs/2402.18013.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. "Vision-Language Models for Vision Tasks: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (Aug. 2024), pp. 5625–5644. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2024.3369699. URL: https://ieeexplore.ieee.org/abstract/document/10445007.
- Yue Zhang et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Sept. 24, 2023. DOI: 10.48550/arXiv.2309.01219. URL: http://arxiv.org/abs/2309.01219.
- Shuyan Zhou et al. WebArena: A Realistic Web Environment for Building Autonomous Agents.

  Apr. 16, 2024. DOI: 10.48550/arXiv.2307.13854. URL: http://arxiv.org/abs/2307.

  13854.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. *MiniGPT-4:*Enhancing Vision-Language Understanding with Advanced Large Language Models. Oct. 2,
  2023. DOI: 10.48550/arXiv.2304.10592. URL: http://arxiv.org/abs/2304.10592.

# 12 A AgentCaster Simplified Overview

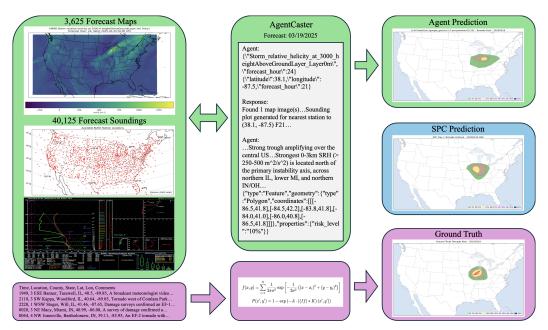


Figure 2: A simplified overview of the AgentCaster framework. LLM agents act as AI meteorologists by first requesting and analyzing forecast maps, then passing specific longitudes and latitudes which are processed to return targeted atmospheric soundings. Agents reason about severe weather dynamics and when confident, generate probabilistic tornado risk predictions as geospatial polygons. These predictions are evaluated against ground truths derived from practically perfect forecasts [14] and compared with domain expert SPC forecast baselines.

## **B** Related Work

Benchmarking LLMs and agents. Recent years have seen rapid development in benchmarks to keep pace with large language models [2], with increasingly complex reasoning assessments [12, 39, 32, 21, 6]. However, many existing benchmarks are facing saturation, with state-of-the-art models approaching or exceeding human-level performance. Some works [40] attempt to address contamination by using updated information sources, while others [26] position themselves as testing at the frontier of human knowledge. The emergence of agent frameworks has introduced new benchmarking challenges. A few approaches [22, 43] evaluate LLMs across diverse environments, and others [49, 28, 30] assess tool use in various environments.

**Multimodal reasoning.** The rise of Vision-Language Models (VLMs) [20, 47, 29, 50, 44] has spurred new approaches to evaluating visual-language integration [5]. Benchmarks for spatial reasoning [36, 19] reveals that multimodal models struggle with spatial relationships, often performing worse than text-only LLMs on spatial tasks given preference between visual and textual context. Some temporal reasoning benchmarks [34] have also been explored. Others focus on spatiotemporal understanding through videos [4] or egocentric spatiotemporal reasoning [41]; in general, evaluations show that models struggle to track changes over time, integrate spatiotemporal information, and understand causality.

**Expert domain tasks.** Specialized knowledge domains increasingly serve as benchmarks for LLMs, with notable examples in medicine [16, 31, 45], law [10], and finance [42]; furthermore, domain-specific evaluations can highlight gaps between knowledge retrieval and the nuanced reasoning required for expert-level tasks. These evaluations offer several advantages: they require deep expertise, can integrate multiple reasoning modes, and feature well-defined evaluation criteria with established human expert performance. A common limitation is that such benchmarks rely on static

question-answering or classification based on domain corpora. In contrast, AgentCaster utilizes tornado science as an expert domain but evaluates a dynamic, interactive problem-solving forecasting process.

Machine learning for weather forecasting. Weather forecasting has had significant advances through deep learning approaches. Previous work with global models [18, 25, 1] have demonstrated competitive performance with traditional NWP methods. Some experimental systems [11] update convection-allowing ensembles frequently to extend warning lead times. However, these approaches typically operate directly on gridded NWP data, maintaining a closed-loop architecture that differs fundamentally from the human forecasting process [18, 25, 1, 11, 13]. Tornado nowcasting has been explored with CNNs with some success [17, 35], but nowcasting is an entirely different process from forecasting [8]. AgentCaster is the first framework to deploy machine learning for weather forecasting through an interactive, human-like workflow.

## 348 C Ground Truth Generation

First, tornado reports from the SPC are aggregated for the relevant 24-hour forecast period (12:00 UTC to 12:00 UTC). A probability density field f(x,y) is calculated on an 80-km Lambert Conformal grid (NCEP Grid 211) using a normalized Gaussian kernel density estimator (KDE) with a smoothing parameter  $\sigma \approx 120$  km:

$$f(x,y) = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2} \left(\frac{d_n(x,y)}{\sigma}\right)^2\right]$$
 (1)

where N is the total number of tornado reports, and  $d_n(x,y)$  is the Euclidean distance from grid point (x,y) to the n-th report in the projected coordinate system. This density field  $f_{80km}$  is then bilinearly interpolated onto a finer grid with approximately 5-km spacing  $(f_{5km})$ , preserving the original projection. To align with the SPC's definition of tornado probability,  $f_{5km}$  is convolved with a uniform circular kernel of radius 40km. This integrates the probability density over the relevant neighborhood around each grid point. The result of the convolution is multiplied by the area of a 5-km grid cell  $(A_{\text{cell}})$  to yield  $\lambda(x,y)$ , the expected number of tornadoes in that neighborhood.

$$\lambda(x,y) = A_{\text{cell}} \cdot \text{Conv}(f_{5km}, \text{Disk}_{R=40km})(x,y)$$
 (2)

The ground truth probability is then calculated from this expected count via the Poisson relation.

$$P_{\text{truth}}(x,y) = 1 - e^{-\lambda(x,y)} \tag{3}$$

The continuous  $P_{\text{truth}}$  field is categorized into discrete risk levels ('0%', '2%', '5%', '10%', '15%', '30%', '45%', '60%') based on standard SPC thresholds (e.g.,  $0.02 \le P_{\text{truth}}(i) < 0.05 \rightarrow '2\%'$ ). These categorical raster areas are then converted into vector polygons; these polygons are reprojected from the Lambert Conformal grid CRS to standard geographic coordinates (WGS84) and saved as the daily ground truth file.

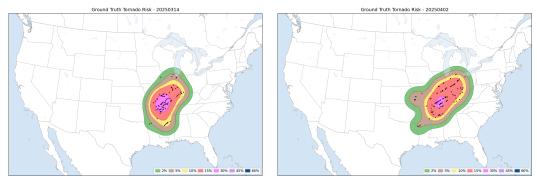


Figure 3: Days with greater than 100 tornado reports.

## 366 D TornadoBench Score

We propose TornadoBench as the primary metric for AgentCaster. It is designed to evaluate the agent's ability to accurately delineate the location, extent, and intensity of tornado risk; it addresses the limitations of standard metrics by incorporating domain-specific weighting and geometric accuracy across multiple probability thresholds. For each day d and risk category C (from 0% to 60%), we calculate the IoU between the predicted and ground truth geometries where  $GT_C$  and  $Pred_C$  are the ground truth and predicted geometries for category C. For the 0% category, we calculate the IoU of the complementary geometries. The daily TornadoBench score is then calculated as:

$$\mathrm{TB}_d = \begin{cases} 1 & \text{if } \mathrm{MaxRisk}_{\mathrm{GT},d} = 0\% \text{ and } \mathrm{MaxRisk}_{\mathrm{Pred},d} = 0\% \\ 0 & \text{if } \mathrm{MaxRisk}_{\mathrm{GT},d} = 0\% \text{ and } \mathrm{MaxRisk}_{\mathrm{Pred},d} > 0\% \\ \frac{1}{|S_d|} \sum_{C \in S_d} \frac{\mathrm{Area}(\mathrm{GT}_C \cap \mathrm{Pred}_C)}{\mathrm{Area}(\mathrm{GT}_C \cup \mathrm{Pred}_C)} & \text{if } \mathrm{MaxRisk}_{\mathrm{GT},d} > 0\% \end{cases} \tag{4}$$

where  $|S_d|$  is the number of categories in set  $S_d$ . The overall TornadoBench score is a weighted average of daily scores, where the weight for each day depends on the maximum risk level in the ground truth:

TornadoBench = 
$$\frac{\sum_{d=1}^{D} (\text{TB}_d \cdot W_d)}{\sum_{d=1}^{D} W_d}$$
 (5)

where  $W_d$  is the numerical value of the maximum risk level in the ground truth on day d (e.g., '0%'  $\to W_d = 1$ , '5%'  $\to W_d = 5$ , '30%'  $\to W_d = 30$ ).

# 379 E TornadoHallucination Metrics

TornadoHallucinationSimple measures the frequency of simple false alarms: days where the agent predicted *any* tornado risk (MaxRisk $_{Pred,d} \ge 2\%$ ) when the ground truth indicated *no* risk (MaxRisk $_{GT,d} = 0\%$ ).

TornadoHallucinationHard penalizes hallucinations based on the *magnitude* of incorrectly predicted risk. It considers two types of hallucinations: (1) any prediction of risk ( $\geq 2\%$ ) on a quiet day (GT=0%), and (2) predictions of risk ( $\geq 2\%$ ) on a risk day (GT>0%) that have zero spatial overlap with the ground truth risk areas. Each such day is assigned a penalty equal to the numerical weight of the highest risk level predicted by the agent, as defined in TornadoBench. The final TornadoHallucinationHard score is computed as the average of these daily penalties over the benchmark period.

# F Dataset Composition

The release benchmark dataset spans a continuous 40-day period from March 1, 2025, to April 9, 2025. This timeframe was selected to include a diverse range of meteorological scenarios across the CONUS, including quiet periods, marginal severe weather setups, and several significant tornado outbreak days. The overall distribution of maximum ground truth risk levels and associated tornado reports is summarized in Table 3. Detailed daily information, including the maximum ground truth risk, total tornado reports, and top affected states for each day in the benchmark period, is provided in Appendix L (Table 6). While AgentCaster is designed for live daily forecasting, for benchmarking we select an evaluation window optimized for composition.

Table 3: Distribution of maximum ground truth risk levels and associated tornado reports across the 40-day benchmark period (March 1–April 9, 2025). There were no 45% or 60% days.

Maximum Risk	Number of Days	Number of Reports
0%	22	5
2%	6	21
5%	4	44
10%	3	102
15%	2	45
30%	3	305
Total	40	522

# **G** Limitations

The evaluation period, while diverse, cannot capture the full range of meteorological conditions across multiple years. Current constraints of the benchmark include limiting sounding requests to 50 per day due to poor context handling and coherence loss with long contexts in existing models. With future models that demonstrate improved context window management and coherence, the framework could be easily extended to incorporate additional convection-allowing models or other data sources, as designed. The HRRRv4 remains the state-of-the-art convection-allowing model, providing both analyses (current conditions) and forecast maps [27]. Additionally, while we designed our interaction protocol to balance realism and reproducibility in the overall operational forecasting process, alternative approaches might better capture more specific aspects of the process. Future work could be to explore applications in related tasks, such as nowcasting or climate-scale forecasting.

# H Prompts and Code

The AgentCaster framework utilizes a structured prompting strategy to guide the LLM agent through the forecasting task. This includes an initial system prompt defining the agent's role, objectives, available tools, and evaluation criteria, followed by a first user prompt to initiate the interaction. The complete codebase for AgentCaster, including all code for data processing, agent interaction, and evaluation, as well as agent prediction GeoJSONs, is publicly available at https://github.com/agentcaster/agentcaster.

# 417 H.1 System Prompt

## **AgentCaster System Prompt**

You are AgentCaster, an expert autonomous AI meteorologist agent that issues Storm Prediction Center (SPC)-style forecasts in tornado prediction using 00z HRRR model data.

## **Objective:**

Your primary objective is to utilize HRRR forecast data to generate an SPC-style tornado risk forecast for the CONUS for the forecast day starting {date\_str} 12z to {next\_date} 12z (forecast hours 12-36 from the 00z run). This is the timeframe for which you will be making your SPC-style prediction.

# **Background & Evaluation:**

To evaluate your prediction, the ground truth is generated as follows: Observed tornado reports are used to calculate a normalized probability density field on an ~80 km grid (using a Gaussian kernel with  $\sigma \approx 120 \, \mathrm{km}$ ), which is then interpolated to a  $\sim 5 \, \mathrm{km}$  grid. This density field is convolved with a 40 km radius disk kernel to integrate the density over a neighborhood. The result is multiplied by the grid cell area to get an expected tornado count ( $\lambda$ ). Finally, this expected count is converted to a probability using  $P = 1 - e^{-\lambda}$ . This probability field is categorized using standard SPC thresholds (2%, 5%, 10%, etc.) and converted into vector polygon geometries. Your predicted risk areas (from the GeoJSON you provide) are directly compared against these ground truth geometries using vector-based geometric operations. Your final score is the average Intersection over Union (IoU) across all evaluated categories present in either your prediction or the ground truth, calculated based on the areas of the geometric intersection and union. This score ranges from 0% (no agreement) to 100% (perfect agreement). Accurate placement, spatial extent, and correct nesting of risk levels (2%, 5%, 10%, etc.) are crucial for a high score. The tornado risk probabilities you predict (e.g., 5%, 10%) represent the likelihood of a tornado occurring within 25 miles (approx. 40 km) of any point within that specific risk area during the forecast period ({date\_str} 12z to {next\_date} 12z).

# **Workflow Guidance**

- Start by calling list\_available\_map\_types to understand the data available for today.
- Then, use request\_hrrr\_map and request\_sounding (strategically, respecting the quota) to gather the information needed for your analysis.
- When confident, call submit\_tornado\_prediction with the properly formatted and nested GeoJSON output, ensuring all separate areas for each risk level are included.

# **Context & Images**

- Map and sounding images are provided as PNGs embedded directly in the conversation (base64), and they consume context.
- [context limit provided to the model]

420

419

418

#### **Autonomy**

- There is no human in the loop. Do not ask for permission or preferences.
- Decide and act yourself. If you need more evidence, request specific maps/soundings (respecting the quota and your context limit). Otherwise, proceed to call submit\_tornado\_prediction with a valid GeoJSON.
- Never ask questions like "Which would you prefer?" or "Should I proceed?" If you would ask, instead choose the action and perform it.

421

# **Tool: List Available Map Types**

list\_available\_map\_types:

Lists the available types of HRRR map plots based on the generated directories. Call this first to see what map types can be requested.

422

# **Tool: Request HRRR Map**

request\_hrrr\_map:

Requests a specific HRRR forecast map image (PNG). Provide the exact map\_type\_directory name from the list and the integer forecast\_hour (12–36).

423

# **Required Properties**

- map\_type\_directory (string): The exact directory name representing the map type. Obtain this from list\_available\_map\_types.
- forecast hour (integer): The forecast hour (e.g., 12, 18, 36) for the map.

424

# **Tool: Request Sounding**

request\_sounding:

Gets a sounding plot (PNG) for the nearest available station to a specified latitude and longitude for a specific integer forecast\_hour (12–36). Limit of {max\_soundings\_per\_day} soundings per day.

425

# **Required Properties**

- latitude (number): Target latitude in decimal degrees.
- longitude (number): Target longitude in decimal degrees.
- forecast\_hour (integer): Forecast hour (e.g., 12, 15, 24).

426

#### **Tool: Submit Tornado Prediction**

submit\_tornado\_prediction:

Call this function *only once* when you have finished analyzing all necessary maps and soundings and are ready to submit the final tornado risk prediction as GeoJSON.

**Format:** The GeoJSON must be a valid FeatureCollection string representing the tornado risk forecast. Each feature must be a polygon or multipolygon with a risk\_level key in the properties field.

- Use MultiPolygon for disjoint risk areas.
- Ensure nesting: higher risk polygons must be spatially contained within all lower risk polygons.

427

## **Required Properties**

- **prediction\_geojson** (string): Output GeoJSON FeatureCollection string as described above.

428

## 429 H.2 First User Prompt

## **First User Prompt**

Today's forecast date is {date\_str}.

You have {max\_soundings\_per\_day} sounding requests available for today.

Please start by calling list\_available\_map\_types to see the available map plots. Remember to call submit\_tornado\_prediction with your final GeoJSON prediction when you are confident with your analysis.

430

## 431 H.3 Context Limit Usage

## **Context Limit Usage (every turn)**

Token usage: The current prompt is about [prompt tokens]. The conversation so far totals about [overall tokens]. [context limit provided to the model].

432

435

## 433 I Centroid Calculation Methodology

- To complement the primary IoU-based TornadoBench score, we also calculate centroid-based metrics to quantify the geographic displacement of predicted tornado risk areas relative to the ground truth.
- These metrics capture both the central tendency of the overall risk and the core of the highest-threat
- 437 regions. All centroid calculations are performed after reprojecting geometries to a common Lambert
- 438 Conformal Conic projection (defined as TARGET\_CRS, based on NCEP Grid 211). Two primary types
- of centroids are computed for both the ground truth (GT) and the agent's prediction for each forecast
- day, as can be found in Figure 1.
- Overall risk centroid. This centroid represents the geometric center of all areas where tornado
- risk > 2%. For both GT and prediction, all individual disjoint risk polygons corresponding to risk
- levels of 2% or greater are first combined into a single geometry using a unary\_union operation.
- This results in geom\_gt\_nonzero and geom\_pred\_nonzero, respectively. The centroid of this
  - unified nonzero risk geometry is then calculated using the .centroid property of the resulting

Shapely object, yielding (x, y) coordinates in the TARGET\_CRS. This metric helps assess if the overall predicted envelope of tornado risk is geographically aligned with the observed risk envelope.

**Maximum risk centroid.** This centroid represents the geometric center of the area(s) assigned the highest specific risk level present in the GT or prediction on a given day. For the GT, the 449 maximum risk level observed on that day (e.g., "30%") is identified (current\_day\_max\_gt\_risk). 450 All disjoint polygons corresponding exclusively to this maximum risk level are combined using 451 unary\_union. The centroid of this resulting geometry (geom\_gt\_hr) is then computed. For the 452 prediction, the maximum risk level predicted by the agent (max\_risk\_pred\_level) is identified, 453 and the centroid of the union of polygons for that specific highest predicted risk (geom\_pred\_hr) is 454 calculated. This metric evaluates the agent's ability to pinpoint the core area of the most significant 455 predicted or observed tornado threat. 456

Distance calculation. Once the corresponding GT and predicted centroids (Overall Risk, Maximum Risk) are determined, the Euclidean distance between them is calculated. This distance is computed directly in the projected coordinate system (TARGET\_CRS), resulting in a value in meters. For reporting in summary tables and analyses, these distances are converted to kilometers (Table 2).

## 461 J Confidence Intervals

462

463

464

465

466

467

468

Table 4 presents the  $\pm 2\sigma$  confidence intervals for key performance metrics. These intervals were calculated using a non-parametric bootstrap procedure with 1000 iterations for each model. The confidence intervals are derived using the percentile method from the distribution of bootstrap statistics; this method captures the variability in model performance due to the specific set of daily scores available for each model and is robust to non-normally distributed data, such that we allow for asymmetric intervals. Note that models evaluated on fewer prediction days may inherently exhibit wider confidence intervals due to a smaller sample size for bootstrapping.

Table 4: Confidence intervals for performance metrics.

Model	TornadoBench (%)	TornadoHallucinationSimple	TornadoHallucinationHard
SPC (Human Expert)	[10.23, 28.34]	[0.12, 0.42]	[0.30, 1.12]
gpt-5-minimal	[4.80, 12.55]	[0.23, 0.54]	[1.58, 3.68]
gpt-5-low	[4.44, 12.32]	[0.28, 0.61]	[1.14, 2.81]
claude-3.7-sonnet	[3.51, 10.78]	[0.25, 0.53]	[1.70, 4.97]
claude-3.7-sonnet:thinking	[4.25, 10.61]	[0.21, 0.51]	[1.79, 4.56]
gpt-5-medium	[2.88, 11.21]	[0.29, 0.68]	[1.35, 4.27]
gpt-4.1	[3.58, 11.49]	[0.28, 0.61]	[2.22, 5.19]
gemini-2.5-pro-preview-03-25	[2.39, 6.94]	[0.25, 0.59]	[2.88, 6.09]
grok-4	[1.13, 7.67]	[0.38, 0.69]	[6.28, 11.67]
gpt-5-high	[1.25, 7.77]	[0.30, 0.67]	[1.39, 3.44]
o4-mini-high	[1.70, 7.28]	[0.36, 0.69]	[3.94, 6.86]
03	[1.14, 6.21]	[0.40, 0.71]	[3.84, 6.88]
gemini-2.5-flash-preview:thinking	[0.34, 14.45]	[0.38, 0.88]	[2.44, 6.69]

# 69 K Dataset Details

- 470 The complete AgentCaster benchmark dataset, including all processed HRRR map types, soundings,
- and ground truths, is publicly available for research and reproducibility. The dataset is hosted
- 472 on Hugging Face and can be accessed at https://huggingface.co/datasets/agentcaster/
- agentcaster ( $\approx 244\,\mathrm{GB}$ ).

# 474 K.1 NOAA License

## NOAA Data License

NOAA data disseminated through NODD are open to the public and can be used as desired.

NOAA makes data openly available to ensure maximum use of our data, and to spur and encourage exploration and innovation throughout the industry. NOAA requests attribution for the use or dissemination of unaltered NOAA data. However, it is not permissible to state or imply endorsement by or affiliation with NOAA. If you modify NOAA data, you may not state or imply that it is original, unaltered NOAA data.

Link: https://registry.opendata.aws/noaa-hrrr-pds.

475

## 476 K.2 List of Generated Forecast Maps

Table 5: List of the 141 map folders that were generated for the benchmark. While there are 145 available map types in the total data archive, some are organized within nested folders.

Var#	Variable Name
1	10_metre_U_wind_component_at_10_heightAboveGround
2	10_metre_V_wind_component_at_10_heightAboveGround
3	10_metre_wind_speed_at_10_heightAboveGround
4	2_metre_dewpoint_temperature_at_2_heightAboveGround
5	2_metre_relative_humidity_at_2_heightAboveGround
6	2_metre_specific_humidity_at_2_heightAboveGround
7	2_metre_temperature_at_2_heightAboveGround
8	Aerosol_optical_depth_at_0_atmosphereSingleLayer
9	Baseflow-groundwater_runoff_at_0_surface
10	Best_(4-layer)_lifted_index_at_18000_pressureFromGroundLayer_Layer0Pa
11	Boundary_layer_height_at_0_surface
12	Categorical_freezing_rain_at_0_surface
13	Categorical_ice_pellets_at_0_surface
14	Categorical_rain_at_0_surface
15	Categorical_snow_at_0_surface
16	Cloud_Forcing_Net_Solar_Flux_at_0_surface
17	Convective_available_potential_energy_at_0_heightAboveGroundLayer_
	Layer3000m
18	Convective_available_potential_energy_at_0_surface
19	Convective_available_potential_energy_at_18000
19	pressureFromGroundLayer_Layer0Pa
20	Convective_available_potential_energy_at_25500
20	pressureFromGroundLayer_Layer0Pa
21	Convective_available_potential_energy_at_9000
	pressureFromGroundLayer_Layer0Pa
22	Convective_inhibition_at_0_surface
23	Convective_inhibition_at_18000_pressureFromGroundLayer_Layer0Pa
24	Convective_inhibition_at_25500_pressureFromGroundLayer_Layer0Pa
25	Convective_inhibition_at_9000_pressureFromGroundLayer_Layer0Pa

26 27 28	Derived_radar_reflectivity_at_1000_heightAboveGround
28	
	Derived_radar_reflectivity_at_263_isothermal
	Derived_radar_reflectivity_at_4000_heightAboveGround
29	Dew_point_temperature_at_1000_isobaricInhPa
30	Dew_point_temperature_at_500_isobaricInhPa
31	Dew_point_temperature_at_700_isobaricInhPa
32	Dew_point_temperature_at_850_isobaricInhPa
33	Dew_point_temperature_at_925_isobaricInhPa
34	Downward_long-wave_radiation_flux_at_0_surface
35	Downward_short-wave_radiation_flux_at_0_surface
36	Forecast_surface_roughness_at_0_surface
37	Freezing_Rain_at_0_surface
38	Frictional_velocity_at_0_surface
39	Geometric_vertical_velocity_at_1_sigmaLayer
40	Geometric_vertical_velocity_at_700_isobaricInhPa
41	Geopotential_height_at_0_adiabaticCondensation
42	Geopotential_height_at_0_cloudBase
43	Geopotential_height_at_0_cloudCeiling
44	Geopotential_height_at_0_cloudTop
45	Geopotential_height_at_0_equilibrium
46	Geopotential_height_at_0_freeConvection
47	Geopotential_height_at_0_highestTroposphericFreezing
48	Geopotential_height_at_0_isothermZero
49	Geopotential_height_at_1000_isobaricInhPa
50	Geopotential_height_at_253_isothermal
51	Geopotential_height_at_263_isothermal
52	Geopotential_height_at_500_isobaricInhPa
53	Geopotential_height_at_700_isobaricInhPa
54	Geopotential_height_at_850_isobaricInhPa
55	Ground_heat_flux_at_0_surface
56	Hail_at_0_atmosphere
57	Hail_at_0_sigma
58	Hail_at_0_surface
59	High_cloud_cover_at_0_highCloudLayer
60	Instantaneous_surface_sensible_heat_flux_at_0_surface
61	Land-sea_mask_at_0_surface
62	Latent_heat_net_flux_at_0_surface
63	Layer_Thickness_261K-256K_Layer
64	Leaf_Area_Index_at_0_surface
65	Lightning_at_0_atmosphere
66	Low cloud cover at 0 lowCloudLayer
67	Mass_density_at_8_heightAboveGround
68	Maximum_Composite_radar_reflectivity_at_0_atmosphere
69	Medium_cloud_cover_at_0_middleCloudLayer
70	Moisture_availability_at_0_depthBelowLand
71	MSLP_(MAPS_System_Reduction)_at_0_meanSea
72	Orography_at_0_surface
73	Percent_frozen_precipitation_at_0_surface
74	Plant_canopy_surface_water_at_0_surface
75 75	Potential_temperature_at_2_heightAboveGround
75 76	Precipitable_water_at_0_atmosphereSingleLayer
70 77	Precipitation_rate_at_0_surface
78	Pressure_at_0_cloudTop
78 79	Pressure_at_0_cloud top Pressure_at_0_highestTroposphericFreezing
	Pressure_at_0_isothermZero
80	1 1055u10_at_0_150t116t1112&10

Var #	Variable Name
82	Pressure_of_level_from_which_parcel_was_lifted_at_25500
	pressureFromGroundLayer_Layer0Pa
83	Relative_humidity_at_0_highestTroposphericFreezing
84	Relative_humidity_at_0_isothermZero
85	Sea_ice_area_fraction_at_0_surface
86	Simulated_Brightness_Temperature_for_GOES_11,_Channel_3_at_0_nominalTop
87	Simulated_Brightness_Temperature_for_GOES_11,_Channel_4_at_0_nominalTop
88	Simulated_Brightness_Temperature_for_GOES_12,_Channel_3_at_0_nominalTop
89	Simulated_Brightness_Temperature_for_GOES_12,_Channel_4_at_0_nominalTop
90 91	Snow_cover_at_0_surface
91	Snow_depth_at_0_surface Storm_relative_helicity_at_1000_heightAboveGroundLayer_Layer0m
92	Storm_relative_helicity_at_1000_heightAboveGroundLayer_Layer0m
93 94	Storm_surface_runoff_at_0_surface
95	Surface_lifted_index_at_500_isobaricLayer_Layer1000hPa
96	Surface pressure at 0 surface
97	Temperature_at_0_surface
98	Temperature_at_1000_isobaricInhPa
99	Temperature at 500 isobaricInhPa
100	Temperature_at_700_isobaricInhPa
101	Temperature_at_850_isobaricInhPa
102	Temperature_at_925_isobaricInhPa
103	Total_Cloud_Cover_at_0_atmosphere
104	Total_Cloud_Cover_at_0_boundaryLayerCloudLayer
105	Total_Precipitation_at_0_surface
106	U_component_of_wind_at_1000_isobaricInhPa
107	U_component_of_wind_at_250_isobaricInhPa
108	U_component_of_wind_at_300_isobaricInhPa
109	U_component_of_wind_at_500_isobaricInhPa
110	U_component_of_wind_at_700_isobaricInhPa
111	U_component_of_wind_at_80_heightAboveGround
112	U_component_of_wind_at_850_isobaricInhPa
113	U_component_of_wind_at_925_isobaricInhPa
114	U-component_storm_motion_at_0_heightAboveGroundLayer_Layer6000m
115	Upward_long-wave_radiation_flux_at_0_nominalTop
116 117	Upward_long-wave_radiation_flux_at_0_surface Upward_short-wave_radiation_flux_at_0_nominalTop
117	Upward_short-wave_radiation_flux_at_0_nonlinarrop Upward_short-wave_radiation_flux_at_0_surface
119	V_component_of_wind_at_1000_isobaricInhPa
120	V_component_of_wind_at_250_isobaricInhPa
121	V_component_of_wind_at_300_isobaricInhPa
122	V_component_of_wind_at_500_isobaricInhPa
123	V_component_of_wind_at_700_isobaricInhPa
124	V_component_of_wind_at_80_heightAboveGround
125	V_component_of_wind_at_850_isobaricInhPa
126	V_component_of_wind_at_925_isobaricInhPa
127	V-component_storm_motion_at_0_heightAboveGroundLayer_Layer6000m
128	Vegetation_at_0_surface
129	Vegetation_Type_at_0_surface
130	Vertical_u-component_shear_at_0_heightAboveGroundLayer_Layer1000m
131	Vertical_u-component_shear_at_0_heightAboveGroundLayer_Layer6000m
132	Vertical_v-component_shear_at_0_heightAboveGroundLayer_Layer1000m
133	Vertical_v-component_shear_at_0_heightAboveGroundLayer_Layer6000m
134	Vertically-integrated_liquid_at_0_atmosphere
135	Visibility_at_0_surface
136	Visible_Beam_Downward_Solar_Flux_at_0_surface

Var #	Variable Name
137	Visible_Diffuse_Downward_Solar_Flux_at_0_surface
138	Vorticity_(relative)_at_1000_heightAboveGroundLayer_Layer0m
139	Vorticity_(relative)_at_2000_heightAboveGroundLayer_Layer0m
140	Water_equivalent_of_accumulated_snow_depth_(deprecated)_at_0_surface
141	Wind_speed_(gust)_at_0_surface

# 77 L Ground Truth Details

- Table 6 provides a day-by-day breakdown of the maximum ground truth tornado risk, the total number of observed tornado reports, and the top three states by report count for the entire 40-day benchmark
- 480 period.

Table 6: Details for all 40 days in the benchmark period (March 1 – April 9, 2025).

Date	Max Risk	Total Reports	Top 3 States (Report Count)
2025-03-01	0%	0	N/A
2025-03-02	0%	0	N/A
2025-03-03	2%	4	TX (2), OK (2)
2025-03-04	10%	26	LA (10), TX (7), OK (6)
2025-03-05	0%	2	NC (1), VA (1)
2025-03-06	0%	0	N/A
2025-03-07	0%	0	N/A
2025-03-08	0%	0	N/A
2025-03-09	0%	0	N/A
2025-03-10	0%	1	FL (1)
2025-03-11	0%	0	N/A
2025-03-12	0%	1	CA (1)
2025-03-13	0%	0	N/A
2025-03-14	30%	104	MO (41), AR (21), IL (20)
2025-03-15	30%	87	MS (48), AL (26), LA (5)
2025-03-16	5%	13	PA (8), GA (2), NC (2)
2025-03-17	0%	0	N/A
2025-03-18	0%	0	N/A
2025-03-19	15%	23	IL (15), IN (7), KY (1)
2025-03-20	0%	0	N/A
2025-03-21	0%	0	N/A
2025-03-22	0%	0	N/A
2025-03-23	2%	4	MS (4)
2025-03-24	0%	0	N/A
2025-03-25	0%	0	N/A
2025-03-26	0%	0	N/A
2025-03-27	2%	2	TX (2)
2025-03-28	2%	2	TX (1), LA (1)
2025-03-29	0%	1	OK (1)
2025-03-30	10%	51	MI (14), IN (7), KY (7)
2025-03-31	5%	9	GA (5), AL (2), LA (1)
2025-04-01	5%	9	OK (5), KS (2), CA (1)
2025-04-02	30%	114	IN (22), MO (21), IL (21)
2025-04-03	2%	5	TN (2), KY (2), AL (1)
2025-04-04	15%	22	TX (15), AR (4), MO (2)
2025-04-05	10%	25	MS (16), TN (4), AL (4)
2025-04-06	5%	13	GA (7), AL (4), MS (2)
2025-04-07	2%	4	GA (3), FL (1)

Date	Max Risk	Total Reports	Top 3 States (Report Count)
2025-04-08	0%	· ·	N/A
2025-04-09	0%		N/A

# M TornadoBench Scores

485

In the tables that follow, we first provide a concise mapping between each model's internal identifier and its short-form abbreviation (Table 7). Table 8 then presents the full set of daily TornadoBench scores for each model over the 40-day evaluation period; each row corresponds to one calendar date and dashed entries indicate days on which a model produced invalid GeoJSON output.

Table 7: Model name mapping for Table 8.

Internal Name	Abbreviation
anthropic_claude-3.7-sonnet	C3.7S
anthropic_claude-3.7-sonnet:thinking	C3.7ST
google_gemini-2.5-flash-preview:thinking	G2.5FT
google_gemini-2.5-pro-preview-03-25	G2.5P
openai_gpt-4.1	GPT4.1
openai_o3	O3
openai_o4-mini-high	O4M
openai_gpt-5-minimal	GPT5
openai_gpt-5-low	GPT5L
openai_gpt-5-medium	GPT5M
openai_gpt-5-high	GPT5H
x-ai_grok-4	G4

Table 8: Daily TornadoBench scores (rounded to the nearest percent). Dashed scores indicate invalid GeoJSONs.

Date	SPC	C3.7S	C3.7ST	G2.5FT	G2.5P	G4	GPT4.1	GPT5	GPT5H	GPT5L	GPT5M	O3	O4M
03-01	100%	0%	0%	-	-	0%	100%	0%	0%	-	0%	0%	0%
03-02	0%	100%	100%	-	0%	0%	0%	0%	0%	0%	-	0%	0%
03-03	3%	0%	0%	1%	0%	0%	1%	0%	4%	0%	0%	0%	2%
03-04	10%	3%	5%	-	0%	5%	7%	7%	5%	7%	8%	4%	10%
03-05	0%	0%	0%	-	0%	0%	0%	0%	0%	100%	0%	0%	0%
03-06	100%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%
03-07	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
03-08	0%	0%	100%	0%	0%	0%	0%	0%	-	0%	-	0%	-
03-09	0%	0%	100%	0%	-	0%	0%	0%	0%	0%	0%	0%	0%
03-10	0%	0%	0%	-	0%	0%	0%	0%	0%	100%	0%	0%	0%
03-11	100%	100%	100%	100%	100%	0%	100%	100%	-	100%	0%	0%	0%
03-12	0%	0%	100%	0%	0%	-	0%	0%	-	0%	-	0%	0%
03-13	0%	100%	0%	0%	0%	0%	-	-	0%	0%	0%	0%	0%
03-14	10%	9%	4%	-	3%	4%	-	6%	5%	7%	7%	3%	3%
03-15	18%	0%	6%	6%	3%	6%	0%	8%	7%	5%	11%	7%	4%
03-16	10%	1%	0%	0%	1%	0%	-	3%	2%	7%	-	0%	0%
03-17	100%	100%	0%	0%	-	0%	100%	100%	0%	0%	100%	0%	0%
03-18	100%	0%	0%	-	-	0%	100%	100%	0%	100%	0%	0%	0%
03-19	23%	2%	-	_	8%	4%	0%	10%	3%	-	1%	2%	0%
03-20	100%	100%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
03-21	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
03-22	0%	0%	0%	_	-	0%	0%	0%	_	0%	_	0%	0%
03-23	6%	2%	0%	-	0%	0%	0%	0%	0%	0%	0%	0%	0%
03-24	0%	0%	0%	_	0%	0%	0%	0%	0%	0%	_	0%	0%
03-25	100%	0%	0%	_	-	0%	0%	100%	_	0%	0%	0%	0%
03-26	0%	100%	100%	_	0%	0%	0%	0%	_	0%	0%	0%	0%
03-27	11%	0%	0%	_	0%	1%	0%	0%	0%	0%	0%	0%	0%
03-28	0%	0%	0%	_	-	0%	0%	0%	0%	0%	-	0%	0%
03-29	0%	0%	0%	_	0%	0%	0%	0%	_	0%	0%	0%	0%
03-30	15%	3%	2%	_	7%	3%		3%	_	2%	8%	2%	-
03-31	4%	4%	3%	_	6%	1%	0%	9%	_	10%	4%	1%	3%
04-01	2%	0%	0%	_	0%	2%	3%	5%	3%		2%	1%	-
04-02	23%	7%	1%	_	3%	8%	7%	6%	5%	6%	5%	4%	_
04-03	2%	0%	0%	5%	0%	0%	3%	0%	0%	0%	0%	0%	3%
04-04	13%	5%	6%	-	2%	2%	7%	4%	0%	4%	3%	4%	5%
04-05	7%	6%	5%	_	3%	1%	12%	12%	-	18%	13%	6%	7%
04-06	26%	6%	7%	3%	6%	3%	19%	4%	13%	5%	-	3%	1%
04-07	4%	0%	0%	-	-	2%	3%	0%	7%	-	_	0%	0%
04-08	100%	0%	0%	_	100%	0%	0%	100%	0%	100%	100%	0%	0%
04-09	100%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	0%	100%

# N LLM Reasoning Details

Table 9: Reasoning capabilities of evaluated models. Parentheses indicate maximum length of reasoning (tokens) if relevant.

Provider	Model Name	Reasoning
Anthropic	claude-3.7-sonnet	No
Anthropic	claude-3.7-sonnet:thinking	Yes (32k)
Google	gemini-2.5-flash-preview: thinking	Yes (25k)
Google	gemini-2.5-pro-preview-03-25	Yes
OpenAI	gpt-4.1	No
OpenAI	03	Yes
OpenAI	o4-mini-high	Yes
OpenAI	gpt-5-minimal	Yes
OpenAI	gpt-5-low	Yes
OpenAI	gpt-5-medium	Yes
OpenAI	gpt-5-high	Yes
xĀI	grok-4	Yes

# **488 O Experiment Resource Costs**

Evaluations in this study were performed using commercial LLM APIs, including OpenAI gpt-4.1, o3, o4-mini-high, and the GPT-5 family (gpt-5-minimal, gpt-5-low, gpt-5-medium, gpt-5-high); Anthropic claude-3.7-sonnet and claude-3.7-sonnet: thinking; Google gemini-2.5-pro-preview-03-25 and gemini-2.5-flash-preview: thinking; and xAI grok-4. The total cost of API calls for model-based evaluation was approximately \$500.

# **NeurIPS Paper Checklist**

#### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Please see Section 1.

## Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the
  contributions made in the paper and important assumptions and limitations. A No or
  NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals
  are not attained by the paper.

#### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: Yes

Justification: Please see Appendix G.

#### Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was
  only tested on a few datasets or with a few runs. In general, empirical results often
  depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

#### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have any theoretical results.

#### Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and crossreferenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if
  they appear in the supplemental material, the authors are encouraged to provide a short
  proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

## 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We released our code on Github and our dataset on Hugging Face; we thoroughly described our methodology and dataset. Please see Section L and Appendices H, K.

#### Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We released our code on Github and our dataset on Hugging Face. Please see Appendices H, K for information and links.

#### Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be
  possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not
  including code, unless this is central to the contribution (e.g., for a new open-source
  benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

#### 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We did not perform any model training; we fully specify our evaluation protocol including the 40-day test period (Section ??), the quota and usage of tools (Section H), dataset composition and access (Appendix K), and all evaluation metrics and procedures (Section 3).

## Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

#### 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please see Appendix J.

#### Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error
  of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how
  they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please see Appendix O.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our study uses open-licensed meteorological data (under the NOAA Data License; Appendix K), involves no human subjects or sensitive personal data, and aims to improve public safety through better tornado forecasting.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a
  deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We outline the potential to improve tornado warning accuracy and public safety by augmenting human forecasters (Section 1), and we also highlight risks such as AI-driven false alarms leading to public complacency (Sections 3, 4).

#### Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal
  impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

## Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do
  not require this, but we encourage authors to take this into account and make a best
  faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all authors. For licenses, please see Section K.

#### Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
  - If assets are released, the license, copyright information, and terms of use in the
    package should be provided. For popular datasets, paperswithcode.com/datasets
    has curated licenses for some datasets. Their licensing guide can help determine the
    license of a dataset.
  - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
  - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

#### 13. New assets

752

753

754

755

756

757

758

759

760

761

762

763

764 765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We released our code on Github and our dataset on Hugging Face. Please see Appendices H, K for information and links.

#### Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human participants or crowdsourcing were involved.

## Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

# 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human participants or crowdsourcing were involved.

#### Guidelines:

 The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions
  and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
  guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

## Answer: [Yes]

Justification: The core methodology deploys pretrained LLMs as interactive AI meteorologist agents; their prompts, tool definitions, and interaction loop are detailed in Section 2 and Appendix H.

#### Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.