HAZARDCAST: BENCHMARKING LLM AGENTS FOR FORECAST CONSTRUCTION ACROSS FIVE NATURAL HAZARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-hazard forecasting requires turning heterogeneous environmental observations into prediction procedures that respect task-specific targets, horizons, units, and information constraints. Large language model (LLM) agents can inspect data, write code, fit models, and revise workflows, but evaluating them requires more than measuring final prediction error. We introduce HazardCast, a benchmark of 24 forecasting questions spanning tropical cyclones, wildfire, heat, drought, and floods. Each question provides prepared environmental data and an explicit prediction contract while leaving feature construction and forecasting methodology to the agent. Evaluation combines native forecast losses, performance relative to a fitted local model, and checks of numerical and forecast-time validity. Across a 24-question comparison, selected agent workflows frequently improve on local models, but gains are uneven. Higher configured reasoning effort does not yield a uniform advantage, paired training conditions show substantial sensitivity to population design, and run-level audits identify concrete output transformations, prediction restrictions, and delivery failures. HazardCast therefore evaluates not only whether an agent produces accurate forecasts, but whether the procedure producing them is scientifically valid.

1 INTRODUCTION

Forecasting natural hazards requires connecting environmental observations to predictive procedures that match a specific scientific question. A cyclone-track forecast, a next-day wildfire map, a seasonal low-flow probability, and a six-hour river-stage forecast differ in representation, forecast horizon, temporal alignment, and verification. LLM agents offer a possible way to automate the construction of such forecasting workflows because they can inspect data, write and execute code, fit models, and iterate using validation feedback (Chan et al., 2025; Chen et al., 2025; Starace et al., 2025). This raises a central question: *can LLM agents construct effective and scientifically valid forecasting workflows across heterogeneous natural-hazard tasks?*

This automation could be particularly valuable in natural-hazard forecasting, where constructing a forecasting workflow requires substantial task-specific choices in data preparation, model selection, temporal alignment, and verification. These choices often need to be reconfigured across regions, forecast horizons, and scientific questions. An agent that can reliably assemble these components could make forecasting more scalable and adaptable, reducing the need to engineer a separate pipeline for every task. At the same time, this flexibility is useful only if the resulting forecasting procedure remains scientifically valid.

The challenge is that strong predictive performance does not by itself show that an agent has correctly solved the forecasting task. In HazardCast, each question specifies not only a target variable, but also a forecast horizon, output representation, units, case structure, and verification rule. An agent must translate these requirements into a complete modeling workflow: selecting and transforming the supplied inputs, constructing an appropriate predictor, producing outputs in the required form, and maintaining the correct correspondence between predictions and evaluation cases. Errors at any of these stages can produce forecasts that appear numerically competitive while no longer matching the task that was specified. We therefore evaluate not only the numerical quality of an agent's forecasts,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but also whether the procedure it constructs faithfully implements the intended forecasting problem. This distinction is especially important in a benchmark spanning multiple hazard families, where the meaning of a valid prediction differs substantially across trajectory, regression, classification, probability, and spatial forecasting tasks.

Existing work provides two complementary foundations. Forecast benchmarks such as WeatherBench 2 standardize verification for a defined forecasting setting (Rasp et al., 2024), while agent benchmarks evaluate whether language-model systems can construct predictive or scientific workflows through computation (Chan et al., 2025; Chen et al., 2025; Bragg et al., 2026). Forecast benchmarks generally assume a predefined forecasting setting, whereas general agent benchmarks do not test whether agent-constructed procedures satisfy task-specific forecast-time and scientific constraints across heterogeneous hazards. HazardCast connects these perspectives. We fix the scientific target, forecast horizon, permissible evidence, output schema, and evaluation rule, while allowing the agent to choose and implement its forecasting method.

We introduce **HazardCast**, a collection of **24 prediction questions across five hazard families**: tropical cyclones, wildfire and associated environmental conditions, heat and heat-related impacts, drought, and floods. Each question links source-derived environmental data to an explicit target, forecast horizon, chronological split, output specification, native loss, and reference forecast. Published studies inform task and input design, while comparisons use methods actually evaluated on the benchmark's cases. The benchmark evaluates forecast construction from prepared data; it does not require agents to retrieve and reconstruct raw environmental archives.

Our empirical study asks whether agents can construct effective and valid forecasts, when that capability holds, and why it breaks. We first compare overall performance across agents, then examine robustness to reasoning effort and forecast horizon. We next assess sensitivity to population design, distinguishing training selection from the cases used for evaluation. Finally, we connect prediction errors and procedural failures to recorded agent decisions. The analyses build a cumulative account of capability and its limitations while retaining their distinct comparison conditions.

Our contributions follow the construction, evaluation, and analysis of the benchmark.

Contribution 1. Benchmark construction. We assemble 24 forecasting questions across five hazard families, linking heterogeneous environmental inputs to explicit targets, horizons, temporal partitions, and output contracts (Section 2). Agents choose and implement forecasting methods within these task definitions.

Contribution 2. Evaluation protocol. We connect frozen numerical forecasts to task-specific errors, matched reference comparisons, and separate checks of numerical and forecast-time validity (Section 3). The protocol distinguishes the supplied input context from the scored population and records differences in data access.

Contribution 3. Empirical diagnosis. We conduct a systematic empirical analysis of agent-constructed forecasting workflows across capability, robustness, population sensitivity, and failure mechanisms (Section 4). We summarize four main findings:

- ***Finding 1.*** *Selected agent workflows frequently outperform fitted local models across heterogeneous hazard tasks, but gains are uneven and strong numerical performance does not necessarily imply a scientifically valid forecast.*
- ***Finding 2.*** *Higher reasoning effort does not yield a uniform performance advantage, while forecast horizon can affect absolute error and local-relative performance differently.*
- ***Finding 3.*** *Paired experiments show substantial sensitivity to training-population design, with balanced-training workflows outperforming their natural-training counterparts across all seven fixed-test comparisons.*
- ***Finding 4.*** *Run-level audits reveal heterogeneous failure mechanisms, including harmful output transformations, restrictive prediction rules, delivery and feedback failures, and forecast-time information-boundary qualifications.*

2 HAZARDCAST BENCHMARK

HazardCast turns heterogeneous hazard records into 24 executable forecasting questions. We describe the task coverage and forecast units, the source-derived inputs supplied to agents, and the temporal partitions and sampling rules used to construct evaluation cases. Section 3 then defines how agent procedures and their predictions are assessed.

2.1 TASK COVERAGE

HazardCast contains 24 questions organized into five hazard families (Table 1). A question defines a prediction target and forecasting setting; a case identifies an entity and forecast origin; and a lead defines the interval to the target. A question can contain multiple leads or many spatial units without becoming multiple benchmark tasks. Spatial pixels, station–lead terms, and repeated issue times are evaluation units rather than independent hazard events.

Table 1: HazardCast task coverage.

Family	Questions	Example targets
Tropical cyclone	5	Center position, maximum wind, wind radii, extratropical transition, regional occurrence
Wildfire and environment	6	Active-fire pixels, area growth, final size class, radiative energy, new discovery, PM2.5
Heat and impacts	4	Temperature field, apparent temperature, heatstroke transports, daily minimum temperature
Drought	5	Drought category, SPI12, seasonal low flow, flash-drought onset, vegetation stress
Flood	4	Discharge, river stage, threshold crossing, newly inundated pixels
Total	24	Regression, classification, event probability, and spatial prediction

The tasks span horizons from six-hour river stage to seasonal low flow. Their semantics remain task-specific: for example, W+3 predicts total site-level PM2.5 rather than a wildfire-attributable component, and F3 predicts a threshold-crossing event rather than continuous river stage. Coverage is concentrated in the United States, with additional cyclone basins, Canadian fire records, and Japanese heat-health data; HazardCast is therefore not a representative sample of global hazard exposure.

2.2 PREPARED INPUTS, TEMPORAL ALIGNMENT, AND POPULATIONS

Each task connects a forecasting question to source-derived observations, candidate predictors, and a verifiable future outcome. Supplied information can include recent hazard state, environmental conditions, spatial context, terrain, or exposure variables. Feature dictionaries specify units, ordering, missingness, and temporal support. Agents decide how to transform the permitted inputs and which estimator to fit.

Each task also specifies a forecast origin, target time or target interval, and eligible predictor history. Correct temporal alignment is part of the forecasting problem: a prediction for the right valid time is not necessarily valid if it uses a later issue or later observation. Train, validation, and test periods follow task-specific chronological partitions, and final-fitting permissions are recorded separately from access to validation feedback.

Eight questions use balanced, stratified, or matched evaluation selections to increase representation of less frequent outcomes. Training selection and scoring selection are treated as distinct design choices. Within any reported comparison, agent and reference losses are computed on the same cases and weights.

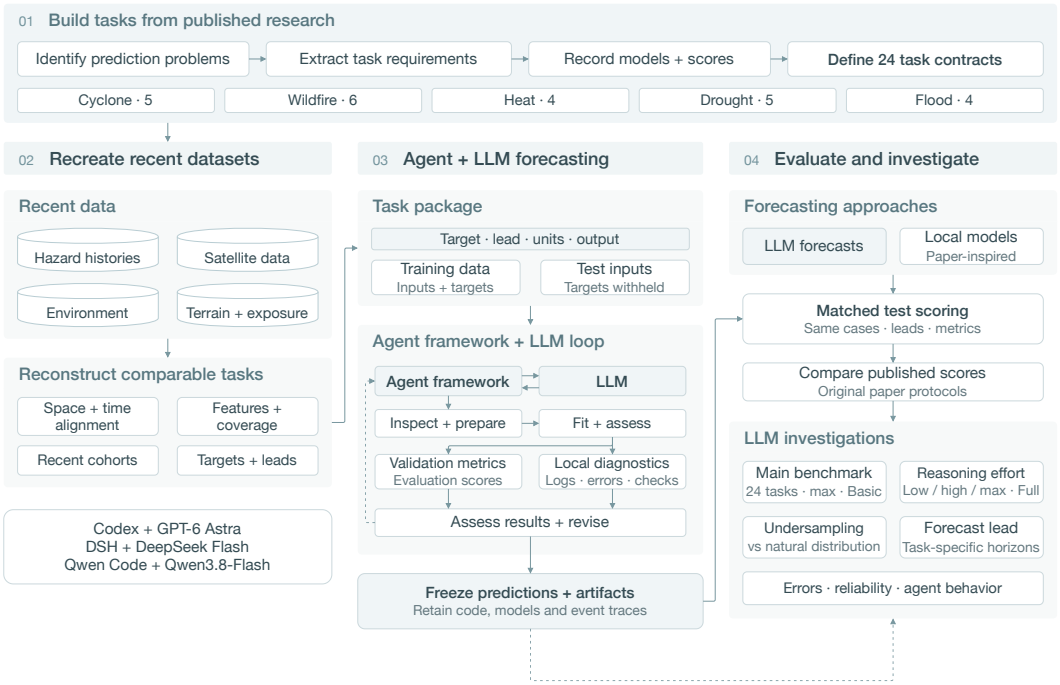

Figure 1: HazardCast study workflow. Literature-grounded task contracts and recent data support agent-built forecasters. Validation metrics and local execution diagnostics guide revision while test targets remain withheld. Frozen predictions are compared with baselines on matched cases, and published scores retain their original-protocol context. The diagram summarizes the 24-question experimental allocation used in the study; DSH denotes DeepSeek Harness.

3 Evaluation Protocol

3.1 Agentic Forecasting Pipeline

Each episode begins with a task specification, prepared data, an output schema, and a development policy. The agent inspects the permitted inputs, writes and executes code, constructs a predictor, and revises its method using whatever validation feedback is available under that policy. The final procedure consists of prediction code and fitted artifacts together with frozen numerical predictions and a method description.

Validation and test serve different purposes. Validation can inform model selection when allowed; test outcomes remain outside the development loop. Frozen test predictions are not subsequently replaced by diagnostic interventions reported later in the paper.

Our comparisons use multiple agent configurations. RQ1 compares Qwen at xhigh effort, DeepSeek Flash at Max effort with Basic capabilities, and GPT-6 Astra at High effort. RQ2 varies DeepSeek Low, High, and Max effort under Full capabilities. Basic provides local execution and aggregate validation feedback; Full additionally exposes the fixed scientific-document collection and specialist forecasting services. The primary comparison contains 24 questions. Full configuration and provenance details are left to Appendix C.4.

3.2 Forecast Error and Cross-task Normalization

We first measure forecast error against the observed targets using each question's native loss, including position error, MAE, RMSE, Brier loss, and misclassification rate. This preserves the physical or probabilistic meaning of each prediction. Appendix D specifies the metrics, evaluation cases, spatial weights, and decision rules.

To compare heterogeneous tasks, we normalize each agent loss against the fitted local model on the same target, cases, weights, and lead. For question q and lead h, let $L_{q,h}$ and $L_{q,h}^{\mathrm{local}} > 0$ denote the agent and local-model losses, and let $\mathcal{H}_q$ be the set of scored leads. We define

$$S_{q,h} = \frac{100}{1 + \left(L_{q,h}/L_{q,h}^{\mathrm{local}}\right)^2}, \qquad S_q = \frac{1}{|\mathcal{H}_q|} \sum_{h \in \mathcal{H}_q} S_{q,h}. \tag{1}$$

The fitted local model scores 50 and zero error scores 100. Scores above 50 indicate lower error than the fitted local model; the score is not an accuracy percentage. Multi-lead tasks are normalized lead by lead before averaging. Persistence, climatology, and other simple forecasts are reported separately where appropriate.

Historical batch-input results that permit broader cross-case context are retained with explicit qualifications. Past-only history, same-time spatial context, and label-free adaptation are not automatically classified as future-label leakage; the relevant question is whether they are permitted by the task and comparison policy.

3.3 Submission Validity

We separately check completion, numerical validity, and forecast-time validity. Numerical checks verify output shape, units, finite values, probability bounds, and case alignment. Forecast-time checks inspect whether the submitted procedure satisfies the task's specified information boundary. These checks are evaluated separately from predictive loss; Appendix B.4 gives the full criteria and reproducibility boundaries.

Historical batch-input results that permit broader cross-case context are retained with explicit qualifications. Past-only history, same-time spatial context, and label-free adaptation are not automatically classified as future-label leakage; the relevant question is whether they are permitted by the task and comparison policy.

4 Experiments

Can LLM agents construct reliable forecasting workflows across heterogeneous natural hazards? We first measure the capability of the selected agents (RQ1), then examine its robustness to reasoning effort and forecast horizon (RQ2). We next test how population design affects the conclusions (RQ3), before diagnosing concrete failure mechanisms (RQ4). This progression connects observed performance to its conditions and limitations; it does not treat the comparisons as a causal chain.

4.1 Experimental Setup

RQ1 compares Qwen at xhigh effort, DeepSeek Flash at Max effort with Basic capabilities, and GPT-6 Astra at High effort, using the selected outcomes in the combined table. Basic provides local code execution and aggregate validation feedback. RQ2a varies DeepSeek Low, High, and Max effort within Full, which additionally provides scientific-document retrieval and specialist forecasting services; RQ2b examines forecast leads. All scores use Equation 1: the matched local model scores 50 and zero error scores 100.

The benchmark and current comparison cover the same 24 questions. We retain one selected outcome per condition, including configuration-matched restricted-input and rebuilt-objective replacements. RQ3 uses seven paired training conditions; RQ4 audits the saved procedures. Model access, fitting rules, and result qualifications are detailed in Appendix C.4; these comparisons do not assume identical agent protocols.

4.2 RQ1: How Effectively Do Agents Construct Forecasts Across Hazards?

Selected agent workflows frequently improve on fitted local models, but gains are uneven and not all numerically strong results support a valid forecast interpretation. The full 24-task score table is provided in Appendix C.1; Table 2 summarizes the cross-task results.

Table 2: Cross-task summary for RQ1 on the same 24 questions. Mean scores use Equation 1; the last column counts scores strictly above 50.

Workflow	Mean score (24 tasks)	Above local (tasks)
Qwen xhigh	53.15	21/24
DeepSeek Max Basic	52.97	19/24
Astra High	55.53	22/24

The benchmark exposes substantial task heterogeneity. On F2 river stage, Astra and Qwen obtain normalized scores of 78.94 and 71.73, respectively, while all three workflows remain close to the local-model anchor on W+2 new-fire discovery. No workflow leads every task.

These means do not isolate underlying model quality. The displayed workflows differ in runner, capability setting, realized information access, and verification status. DeepSeek Basic C1 and C2 use later-issued forecasts, W+1 uses later-event values, several Qwen replacements remain provisional, and some imported procedures use different cross-case context. We therefore interpret the cross-task scores as properties of the selected forecasting workflows, not as a protocol-equivalent ranking of language models.

Summarized Answer to RQ1. Selected workflows improve on the local model on 21/24 tasks for Qwen, 19/24 for DeepSeek Max Basic, and 22/24 for Astra. Gains vary across hazards, and numerical performance alone does not establish forecast-time validity or a protocol-equivalent model ranking.

4.3 RQ2: How Robust Is Performance to Effort and Forecast Horizon?

We examine robustness along two separate axes: the agent's configured reasoning effort and the forecasting task's lead time. The first changes how a workflow is constructed; the second changes the prediction problem it must solve. They are analyzed separately, without estimating an effort-by-horizon interaction. Figure 2 summarizes both analyses.

RQ2a: Reasoning effort. Max effort wins the most tasks, but higher effort does not consistently improve scores. We compare only DeepSeek Low, High, and Max with Full capabilities across all 24 active tasks, using the selected completed runs and configuration-matched replacements. Scores use the local-model anchor of 50, with zero error at 100.

Low has the highest mean score (52.67), followed by Max (52.06) and High (51.99), yet Max achieves the best score on 11 tasks, compared with eight for Low and five for High. Max improves on Low on 13 tasks and performs worse on 11. Individual profiles illustrate this heterogeneity: C3 improves from 51.46 to 53.08 to 54.02 as effort increases, whereas C+2 falls from 56.97 at Low to 50.50 at High and 50.87 at Max. F2 instead favors High (70.84), above both Low (52.73) and Max (44.94). Thus, neither the mean nor the task wins establish one effort level as uniformly superior.

RQ2b: Forecast lead time. Longer forecast leads usually increase absolute forecast error, but relative performance against the fitted local model can improve when the local model degrades faster. The two examples in Figure 2 illustrate this contrast: for C3, Astra's wind-radius MAE increases from 26.51 to 41.42 nmi between 24 and 72 hours, while its local-relative score rises from 53.35 to 59.48. Similarly, DeepSeek Max Basic's F1 discharge RMSE increases from 513.68 to 857.89 cfs between one and fourteen days, yet its score improves from 50.24 to 54.13. In both cases, the local model deteriorates faster. Restricted D1 Max Basic instead remains near 50 across all three horizons. Thus, the effect of lead time depends on both the task and the comparison model. Full per-lead profiles for seven questions are reported in Appendix C.3.

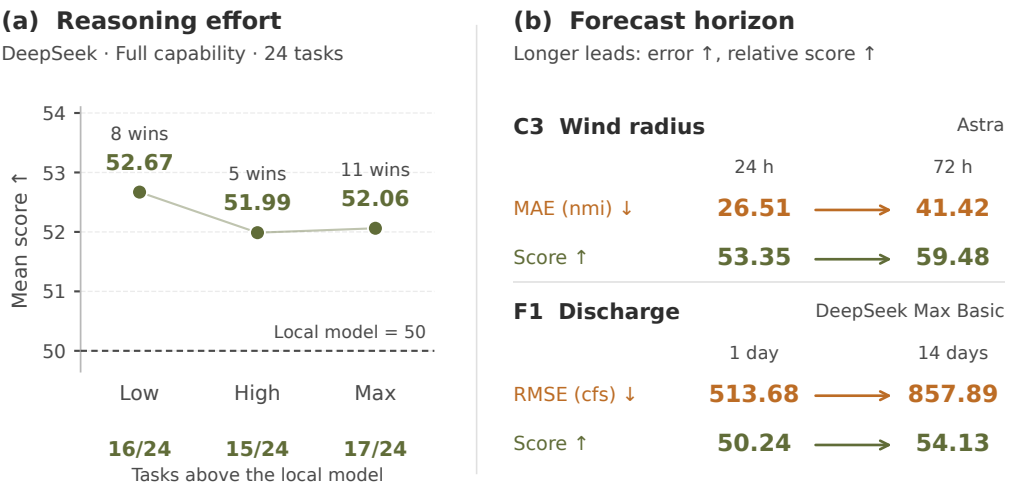

Figure 2: Robustness to reasoning effort and forecast horizon. (a) Higher configured effort does not monotonically improve the selected DeepSeek Full workflows: Low has the highest mean score, while Max wins the most tasks and exceeds the local model on the most tasks. (b) Longer horizons can increase absolute forecast error while improving performance relative to the fitted local model when that reference deteriorates faster. Full task-level effort results and seven-task lead profiles appear in Appendices C.2 and C.3.

Summarized Answer to RQ2. More reasoning effort is not uniformly better: Low has the highest mean score (52.67), while Max wins the most tasks (11/24). Longer horizons can increase absolute error yet improve the local-relative score when the reference deteriorates faster.

4.4 RQ3: How sensitive is performance to training-population design?

Population choices materially change the observed performance of agent workflows. We distinguish the training population from the evaluation population. The available paired runs test training-population sensitivity on fixed test cases: balanced-training workflows outperform natural-training workflows in all seven observed comparisons. Table 3 compares DeepSeek Max Basic under the two training conditions, using identical test cases within each task. Six tasks use balanced evaluation cohorts; F1 retains its natural test population. Scores are anchored at 50 for the local model and 100 for zero error.

Table 3: Training-population comparison on fixed test cases. Scores use Equation 1; Δ is natural minus balanced, computed before rounding.

Task	Balanced	Natural	Δ
W1	57.01	17.26	-39.74
W3	42.50	24.46	-18.04
D+1	46.77	17.22	-29.55
W+2	50.27	33.55	-16.72
F3	48.32	32.76	-15.56
F1	52.26	49.20	-3.06
F+1	49.11	28.15	-20.96

The largest difference occurs on W1: balanced training scores 57.01 versus 17.26 for natural training, a 39.74-point advantage. D+1 shows a similarly large gap of 29.55 points (46.77 versus 17.22). The remaining balanced-cohort tasks favor balanced training by 15.56–20.96 points: F3, W+2, W3, and F+1. By contrast, F1 differs by only 3.06 points (52.26 versus 49.20). Balanced-training workflows

exceed the local reference on three tasks, whereas natural-training workflows fall below it on all seven.

The advantage is consistent in direction but varies substantially in size. Thus, performance under balanced learning does not directly characterize workflows trained on natural event frequencies. Evaluation cases are fixed within each pair, so the comparison reflects differences between the paired training conditions rather than changes in the test population. The fixed evaluation cases also make the direction and magnitude of the observed score differences directly comparable within each task. Across all seven comparisons, balanced-training workflows score higher, although the size of that advantage remains strongly task dependent. The smaller F1 gap is a cross-task observation.

Summarized Answer to RQ3. Balanced-training workflows score higher in all seven paired comparisons on fixed test cases, by 3.06–39.74 points. The result demonstrates sensitivity to the paired training conditions. The consistency of this advantage across all seven tasks underscores that agent-constructed forecasting performance is highly sensitive to the paired training conditions.

4.5 RQ4: Why Do Agent-Constructed Forecasts Fail?

Aggregate scores hide distinct failure modes. We audit the saved procedures associated with the selected results and distinguish isolated component effects, verified structural restrictions, information-boundary qualifications, and broader hypotheses. Table 4 summarizes the mechanisms; Appendix C.5 reports the run-level evidence and its limitations.

Table 4: Selected run-level failure mechanisms. Component effects are measured on fixed cases; broader explanations remain qualified.

Mechanism	Evidence from selected runs
Output transformations	Removing W3 class weights reduces error by 16.9% for Basic and 12.6% for High; removing C+1 logit offsets reduces Brier loss by 5.9% for Low and 7.3% for High.
Restricted predictions	D1 Astra and Low never predict worsening; D3 High assigns one probability per year, eliminating within-year site discrimination.
Development transfer	Validation-fitted recalibration does not reliably transfer; a fixed scalar mapping improves some outcomes but worsens more.
Delivery behavior	F+1 High and Max exactly return the supplied specialist predictions, explaining both scores of 50.
Feedback failures	Retained broker records contain 281 validation attempts with no numerical responses across 16 selected DeepSeek runs.

The clearest component-level example is W3. Basic and High apply final class weights that favor the small-fire class. Removing only those recorded final weights—without changing the fitted estimator, training data, or 630 scored test cases—reduces misclassification error from 0.4413 to 0.3667 for Basic and to 0.3857 for High. Medium-class recall increases at the same time. The intervention identifies a harmful decision mapping on these cases; it is not a new benchmark submission and does not explain all remaining error.

Structural restrictions explain other failures. In D1, Astra and DeepSeek Low implement fixed-decrement decisions and therefore miss all worsening cases at every scored lead. D3 High produces only one probability per test year, yielding no within-year discrimination. In F+1, High and Max exactly reproduce the supplied specialist predictions over all delivered pixels, so their score of 50 reflects a fallback path rather than two independently learned predictors converging to identical skill.

Broader residual patterns identify hypotheses rather than unique causes. Several workflows under-predict the strongest cyclone-intensity cases; large wildfire-growth events account for most absolute error on W2; and W+2 exhibits weak ranking across multiple recipes. These observations motivate

explanations involving tail representation, feature handling, or missing predictive signal, but the available evidence does not separate those possibilities cleanly.

Summarized Answer to RQ4. Forecasting failures have distinct mechanisms: harmful output mappings, restrictive prediction rules, and delivery or feedback failures. W3 provides an isolated component effect, while D1 and F+1 expose verified structural or delivery restrictions. Other residual patterns remain hypotheses rather than uniquely identified causes.

5 Related Work

Agents as model builders. MLE-bench evaluates data preparation, model training, and experimentation in machine-learning competitions (Chan et al., 2025). ScienceAgentBench evaluates executable scientific tasks derived from peer-reviewed studies (Chen et al., 2025), while PaperBench extends evaluation to paper-scale research replication (Starace et al., 2025). AstaBench emphasizes reproducible scientific-agent environments and the importance of controlling tool access and resource differences (Bragg et al., 2026). Within atmospheric science, AtmosSci-Bench evaluates large language models on domain-specific knowledge and reasoning tasks (Li et al., 2026), while ClimateAgent studies multi-agent orchestration for complex climate data science workflows (Kim et al., 2025). HazardCast complements these efforts by focusing on a particular scientific output of agent-constructed workflows: forecasts that must align with future environmental observations under explicit forecast-time information constraints.

Hazard forecasting and verification. Prior work illustrates the diversity of hazard-specific prediction settings, including tropical-cyclone tracks (Wang et al., 2023), next-day wildfire prediction (Huot et al., 2022), heatwave prediction from station networks (Li et al., 2023), weekly drought-category prediction (Cao et al., 2023), and flood-inundation forecasting (Sun et al., 2023). WeatherBench 2 provides a standardized evaluation framework for global medium-range weather forecasting (Rasp et al., 2024). HazardCast differs in evaluating whether agents can construct forecasting procedures across heterogeneous hazard targets while satisfying task-specific units, leads, output forms, and forecast-time information boundaries. The cited hazard studies motivate scientific problem formulations; the benchmark does not claim to reproduce their full architectures, data coverage, or original reported scores.

6 Discussion and Limitations

HazardCast shows why forecast-construction agents require evaluation at two levels. At the prediction level, selected workflows often improve on fitted local models, but performance varies strongly by task, effort setting, forecast horizon, and training population. The study has several limitations. First, the reported configurations use one selected workflow per condition and therefore do not estimate stochastic run-to-run variation. Large numbers of pixels or forecast rows do not substitute for independent agent repetitions or event-level uncertainty analysis. Second, the compared agent systems differ in runner, capability set, and realized information access, so RQ1 should not be interpreted as a model-only causal comparison. Third, some results remain qualified by provisional contracts, missing replay artifacts, or timing concerns.

7 Conclusion

HazardCast evaluates whether LLM agents can construct effective and scientifically valid forecasting workflows across 24 questions and five natural-hazard families. Selected agent workflows frequently improve on fitted local models, but gains are uneven and some apparent successes fail forecast-time validity checks. Higher reasoning effort does not produce a uniform advantage, longer horizons separate absolute error from relative skill, paired training conditions reveal substantial population sensitivity, and run-level audits identify concrete output, prediction, delivery, and information-boundary failures. Reliable evaluation therefore requires both forecast verification and examination of how each forecast was constructed.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, to assist with four aspects of this work: polishing and restructuring the writing; retrieving relevant literature and identifying candidate related work; research ideation and execution; and drafting sections of the manuscript. Research assistance included refining benchmark questions and experimental designs, developing and debugging code for data preparation, evaluation, and analysis, inspecting run artifacts, and exploring interpretations of the results. All AI-assisted work was subsequently reviewed and verified by the authors, with generated code, analyses, interpretations, tables, figures, and manuscript text checked and revised as appropriate before inclusion. The authors retained final responsibility for all methodological decisions, scientific claims, and reported results.

REFERENCES

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2510.21652. URL https://arxiv.org/abs/2510.21652v2.

Junjun Cao, Fu Guan, Xiang Zhang, et al. Multiple Markov chains for categorial drought prediction on the U.S. Drought Monitor at weekly scale. *Journal of Applied Meteorology and Climatology*, 62(10):1415–1435, 2023. doi: 10.1175/JAMC-D-23-0061.1. URL https://journals.ametsoc.org/view/journals/apme/62/10/JAMC-D-23-0061.1.xml.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.07095. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7e3767db483c942b883eb4f8cfb74e31-Abstract-Conference.html.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.05080. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning data set to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. doi: 10.1109/TGRS.2022.3192974. URL https://arxiv.org/abs/2112.02447.

Hyeonjae Kim, Chenyue Li, Wen Deng, Mengxi Jin, Wen Huang, Mengqian Lu, and Binhang Yuan. Climateagent: Multi-agent orchestration for complex climate data science workflows. *arXiv preprint arXiv:2511.20109*, 2025.

Chenyue Li, Wen Deng, Mengqian Lu, and Binhang Yuan. Atmossci-bench: evaluating the recent advance of large language model for atmospheric science. *Advances in Neural Information Processing Systems*, 38, 2026.

Peiyuan Li, Yin Yu, Daning Huang, Zhi-Hua Wang, and Ashish Sharma. Regional heatwave prediction using graph neural network and weather station data. *Geophysical Research Letters*, 50 (7):e2023GL103405, 2023. doi: 10.1029/2023GL103405. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103405.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019. URL https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023MS004019.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research, 2025. URL https://arxiv.org/abs/2504.01848.

Alexander Y. Sun, Zhi Li, Wonhyun Lee, Qixing Huang, Bridget R. Scanlon, and Clint Dawson. Rapid flood inundation forecast using Fourier neural operator, 2023. URL https://arxiv.org/abs/2307.16090. Preprint; also presented at the ICCV 2023 AIHADR workshop.

Liang Wang, Bingcheng Wan, Shaohui Zhou, Haofei Sun, and Zhiqiu Gao. Forecasting tropical cyclone tracks in the northwestern Pacific based on a deep-learning model. *Geoscientific Model Development*, 16:2167–2179, 2023. doi: 10.5194/gmd-16-2167-2023. URL https://gmd.copernicus.org/articles/16/2167/2023/.

A TASK DEFINITIONS AND CHRONOLOGICAL PARTITIONS

The benchmark contains the 24 questions listed below. Each question is counted once, even when it has multiple leads, quadrants, pixels, or scoring blocks. The experimental tables report the selected completed outcomes for these tasks.

Table 5: Complete 24-question target, horizon, metric, and reference inventory. F+1 has an observation-defined interval; W3 predicts final size rather than a fixed-time increment.

ID and target	Horizon	Native loss	Simple reference
C1: cyclone center	24/48/72 h	Mean great-circle error, km	Constant motion
C2: maximum wind	24/48/72 h	MAE, kt	Intensity persistence
C3: quadrant wind radii	24/48/72 h	MAE, nautical miles	Issue-time radii
C+1: extratropical transition	Within 48 h	Brier	Balanced training climatology
C+2: regional cyclone occurrence	Next 7 d	Brier	Region-month training climatology
W1: active-fire pixels	Next day	Brier	Current fire mask
W2: additional burned area	Next day	MAE, km^2	Previous-day area growth
W3: final fire-size class	Final size from discovery-time information	Misclassification rate	Constant size class
W+1: radiative energy	Next day	RMSE, MJ	Issue-day energy
W+2: new fire discovery	Next day, after 7 discovery-free days	Brier	Balanced training climatology
W+3: site PM2.5	Next day	RMSE, $\mu g\, m^{-3}$	Previous-day concentration
H1: temperature field	1/3/7 d	Latitude-weighted RMSE, °C	Temperature persistence
H3: apparent temperature	1/2/3 d	RMSE, °C	Daily apparent temperature
H+1: heatstroke transports	Next day	RMSE, count	Count persistence in current packet
H+2: daily minimum temperature	Next reported daily period	RMSE, °C	Last completed daily minimum
D1: USDM category	1/2/4 wk	Misclassification rate	Current category
D2: precipitation deficit	Next month	RMSE, SPI12	Current SPI12
D3: seasonal low flow	July mean from February flow	Brier	Archived native reference; training prevalence for research
D+1: flash-drought onset	Next 5-day period	Brier	Balanced training climatology
D+2: vegetation stress	Next week	RMSE, percentage points	Current stressed area
F1: discharge	1/7/14 d	RMSE, cubic feet/s	Last discharge
F2: river stage	6 h	RMSE, feet	Current stage
F3: flood-threshold crossing	Within 6 h	Brier	Balanced training climatology
F+1: new inundation	Next target satellite observation	Brier	Balanced training climatology

Reference definitions are versioned with the scoring endpoint. H+1's current expanded packet uses lagged-count persistence, whereas its archived results used a validation-frozen naive reference. W3's balanced native endpoint uses a fixed expected reference error of $2/3$; its natural-population research endpoint uses deterministic small-class predictions on 87,241 test incidents. D3 preserves its archived reference for native scoring and uses training-only event prevalence for research scoring. These reference losses must not be interchanged when computing skill.

Chronological periods are recorded in Table 6. They specify the data partitions, not a universal permission to refit on training plus validation. Final-fitting permissions belong to the corresponding run protocol. The US concentration of many tasks and their differing temporal coverage limit geographic and period-level generalization.

Table 6: Task-specific chronological partitions. C3 has no eligible 2016–2017 cases in the inspected source period; H+1 lacks 2018; F2 does not have continuous complete coverage.

ID	Region or observation setting	Training / validation / test
C1	Western North Pacific	2016–2023 / 2024 / 2025
C2	Atlantic	2016–2022 / 2023 / 2024–2025
C3	Atlantic	Eligible 2018–2022 / 2023 / 2024–2025
C+1	Global main-track tropical storms	Storm cohorts beginning 2016–2022 / 2023 / 2024
C+2	Australian, South Indian, South Pacific sectors	Seasons starting 2016–2021 / 2022 / Nov 2023–Apr 2024
W1	US wildfire imagery	2020–2021 / 2022 / 2024
W2	Canadian fire records	2016–2022 / 2023 / 2024
W3	Contiguous US incidents	2016–2022 / 2023 / 2024
W+1	US incident-associated fire components	2024 / separate 2024 development partition / 2025
W+2	60 training-selected US cells	2016–2022 / 2023 / 2024
W+3	California air-quality monitors	2016–2022 / 2023 / 2024
H1	Contiguous US summer grid	2016–2022 summers / 2023 summer / 2024–2025 summers
H3	Houston, Chicago, Phoenix stations	2021–2023 / 2024 / 2025
H+1	47 Japanese prefectures	2016–2017 and 2019–2021 / 2022 / 2023–2024
H+2	US GSOD stations	2016–2023 / 2024 / 2025 through Aug 27
D1	24 US monitoring sites	2016–2023 / 2024 / 2025
D2	344 US climate divisions	Eligible issues Jan 2017–Nov 2023 / 2024 / 2025
D3	US CAMELS/USGS gauges	2016–2021 / 2022 / 2023–2025
D+1	16 US NASA POWER sites	2016–2022 / 2023 / 2024–2025
D+2	20 US states	2016–2023 / 2024 / 2025
F1	US multi-gauge discharge network	2016–2023 / 2024 / 2025
F2	Three USGS gauges	Available 2016–2023, excluding 2019 / 2024 / 2025
F3	Six USGS/NWPS gauges	2016–2022 / 2023 / 2024–2025
F+1	Contiguous US satellite-water maps	Sep–Oct 2024 / Nov 2024 / 2025

B Data construction, sampling, and verification details

B.1 Data representations and source adaptations

Each task is represented by its question, source and feature descriptions, development and evaluation partitions, prediction schema, and scorer. These records distinguish variables that are supplied from those actually used by a fitted procedure. IDs used to align outputs are not automatically scientific predictors: admissibility depends on the active task and run policy. Saved code, model descriptions, and prediction artifacts identify the procedure to which a result belongs.

The following examples illustrate the relationship between scientific motivation and the implemented data. C1 represents recent storm motion through an eight-step, eleven-channel history and provides atmospheric, sea-surface, and availability fields. C+2 adds lagged climate-index information to regional occurrence predictors; the supplied construction uses previous-day RMM components and a four-month-lagged Trans-Niño Index. W+3 combines concentration history and meteorology with satellite fire detections and radiative-power summaries, without defining the target as smoke-attributable PM2.5. H+1 combines weather information with population exposure for a count outcome. D1 supplies current and previous weekly drought categories. These are candidate inputs, not an assertion that every agent uses every field.

For the F+1 **CONUS42** variant, 42 maps are partitioned into 16 training, seven validation, and 19 test maps. The 72 predictors include 15 original water-map features, 36 terrain/rainfall features, and 21 modeled-ponding summaries. National Water Model ponding at native 250 m resolution is assigned to the 30 m output grid; this assignment neither creates 30 m physical information nor converts modeled ponding into observed depth. The current scoring contract selects 104,612 test pixels from predictions covering 8,416,206 test-grid rows. The earlier archived F+1 result used a different 20-map, 213,700-pixel scoring cohort. Its score is not transferred to CONUS42. Version and mask identity are part of any result comparison.

B.2 Sampling populations

Outcome-dependent selection is defined by the evaluator and must remain distinct from an agent's modeling choices. Table 7 lists the eight questions with balanced or matched native scoring populations. Selection increases the representation of particular outcomes while defining a different population from the full eligible data. Labels and label-dependent scoring masks remain outside the agent's permitted test information.

Table 7: Native balanced or matched evaluation designs. Training selection, scoring selection, and full-population probability analysis are separate quantities.

Question	Selection rule	Interpretation
C+1	Transition/non-transition balance	Selected transition windows
W1	Active-fire/non-fire pixel balance	Pixel score with complete map context retained
W3	Equal small/medium/large classes	Final-size classification under balanced classes
W+2	Discovery/control matching within cell, year, and quarter	Selected discovery opportunities
D1	Improving/unchanged/worsening balance within split and lead	Predict future category; sampling strata describe change
D+1	Onset/non-onset balance	Selected flash-drought windows
F3	Same-gauge controls matched on threshold margin and stage slopes	Selected threshold-crossing opportunities
F+1	Still-dry controls among initially dry valid pixels, stratified by nearby water	Versioned spatial scoring mask

W1's archived native score uses 3,286 balanced pixels from 30 maps, whereas its natural-population analysis uses 120,287 valid pixels. Balanced-cohort and natural-population losses must not be treated as interchangeable measurements. F1 uses equal development sampling across five flow regimes but retains its full eligible test population.

B.3 Losses, aggregation, and comparison validity

Regression questions retain physical units through mean absolute error or root mean squared error. C1 uses great-circle distance between predicted and observed coordinates. Binary probability tasks use Brier loss, $N^{-1}\sum_i (p_i - y_i)^2$. Categorical tasks use the misclassification rate. H1 additionally weights spatial squared errors by cosine latitude before taking the square root, and C3 pools its eligible quadrant rows within each lead. The native scorer determines these weights; they are not replaced by equal event weights after the fact.

Equation 1 uses the matched, frozen local-model loss for each lead. The denominator must use the same metric, case set, spatial weighting, and target representation as the agent loss. Each lead is normalized separately before equal averaging; the ratio of averaged losses is not substituted. The local model therefore scores 50 at each lead and after task aggregation. Zero agent loss scores 100. If the local loss is zero or unavailable, the normalized score is undefined; we retain the native error and report the affected comparison separately rather than adding an arbitrary constant.

For comparison with a simple forecast, we may separately report relative skill $100(1 - L/B)$, where $B > 0$ is that forecast's loss on the same cases. This is a percentage reduction in the stated loss, not the normalized benchmark score. For RMSE it measures reduction in RMSE rather than mean squared error. This reference-relative skill is distinct from the local-model-normalized score in Equation 1.

Cross-model summaries report the shared task set and its size; protocol and eligibility qualifications remain explicit. A missing output, an ineligible procedure, and a completed but unscored submission are different statuses. The run-level appendix distinguishes procedural failures from numerical outcomes. Because multiple rows can belong to the same storm, map, site, gauge, or time block, row counts alone do not establish independent replication; uncertainty calculations must respect that dependence. Repeated agent episodes address a different source of variation from resampling observed events.

B.4 FORECAST-TIME VALIDITY AND REPRODUCIBILITY BOUNDARIES

Forecast-time checking distinguishes an observation's valid time, its publication time, and the procedure's forecast origin. A row issued tomorrow may contain information useful for tomorrow's target but cannot establish a forecast made today. The same distinction applies to centered indices, revised archives, and finalized reanalysis. The study records permissible supplied-data use under the active protocol and separately audits whether a procedure supports an original-issue forecast interpretation. Validation feedback is development information and cannot be substituted for held-out evaluation.

Retained artifacts support inspection of inputs, schemas, training choices, predictions, native losses, references, and procedural checks. Exact environment and source versions remain relevant to fresh-run reproduction, especially for API model aliases and externally stored datasets. These records do not by themselves establish a clean-room rebuild, historical product-release latency, or permission to redistribute every upstream dataset. The present construction evaluates retrospective forecasting workflows on prepared data.

C SUPPLEMENTARY RESULTS AND DIAGNOSTICS

The tables report the selected completed outcomes used in the main analyses. B, L, H, and M denote DeepSeek Max Basic, Low Full, High Full, and Max Full; A denotes Astra. Scores use the local-model anchor in Equation 1. Native losses, input-access qualifications, and run-level diagnostics are reported separately. Intermediate queue snapshots and superseded score inventories are omitted.

C.1 OVERALL-RESULT PROVENANCE

Table 8: Overall performance from the current combined table. Scores are normalized to local model = 50 and zero error = 100. Superscripts retain provenance and validity qualifications described below; values are not filtered into a protocol-equivalent ranking.

Task	Qwen xhigh	DeepSeek Max Basic	Astra High
C1	52.78	53.55[t]	52.58
C2	53.07[c]	65.74[t]	52.11
C3	54.58	51.63	56.64
C+1	55.41[c]	50.35	54.30
C+2	52.08[c]	50.49	51.05
W1	51.33	57.01[b]	64.68
W2	51.28[r]	51.66	51.51
W3	51.70	42.50	51.92
W+1	50.02	49.85[t]	50.06
W+2	50.18[r]	50.27[r]	50.54
W+3	51.23	50.88	48.93
H1	50.36	51.41	52.56
H3	51.61	52.24	51.60
H+1	57.97[c]	64.45	64.20
H+2	53.41[c]	52.89	55.90
D1	49.76	50.20[r]	50.00
D2	49.88[r]	59.55	59.67
D3	55.95[c]	52.81	56.01
D+1	46.81	46.77	50.29
D+2	59.60[r]	64.42[r]	62.91
F1	52.75	52.26	54.19
F2	71.73[c]	52.94	78.94
F3	50.53	48.32[b]	50.67
F+1	51.73	49.11	61.48
Mean (24 displayed tasks)	53.15	52.97	55.53

r: restricted input; b: rebuilt balanced objective; t: known later-issued information use; c: cross-case history, same-time context, or unlabeled adaptation requiring a matched access policy.

Table 8 reports the current combined-table selection on 24 active tasks, for Qwen xhigh, DeepSeek Max Basic, and Astra High. Each task score is the equal-lead average in Equation 1; overall means weight tasks equally. Qwen W1/W3/D1/F3/F+1 retain provisional scoring-contract qualifications in the run-level audit.

C.2 UPDATED FULL-EFFORT COMPARISON

The effort comparison contains 72 completed outcomes across 24 tasks and three Full settings. Each task score averages the lead-specific values in Equation 1; Table 9 weights tasks equally.

C.3 COMPLETE PER-LEAD PROFILES

Table 11 reports seven questions, five agent configurations, and three leads per configuration (105 agent–lead measurements), with 21 local-model losses. D1 uses restricted inputs in all four DeepSeek settings; Astra retains its recorded protocol.

Table 9: DeepSeek reasoning-effort comparison with Full capabilities on 24 common tasks. Scores are averaged over leads within a task, then equally across tasks. Winner counts use unrounded scores; each task has one winner.

Summary	Low Full	High Full	Max Full
Mean score	52.67	51.99	52.06
Median score	51.55	50.39	50.86
Best score (tasks)	8	5	11
Above local model (tasks)	16	15	17

Table 10: Complete DeepSeek Full-effort scores used in RQ2a. Bold identifies the highest unrounded score per task.

Task	Low Full	High Full	Max Full
C1	**49.77**	44.76	46.79
C2	**52.69**	50.11	49.45
C3	51.46	53.08	**54.02**
C+1	43.57	43.06	**47.96**
C+2	**56.97**	50.50	50.87
W1	51.63	**59.13**	52.13
W2	51.64	51.04	**51.92**
W3	49.38	42.50	**50.84**
W+1	49.80	49.95	**50.22**
W+2	**50.44**	49.55	50.33
W+3	51.06	**51.50**	50.32
H1	**72.79**	49.81	50.38
H3	51.45	57.29	**60.30**
H+1	64.66	**65.94**	65.02
H+2	**55.36**	52.29	52.11
D1	50.00	**50.20**	48.92
D2	49.79	50.52	**59.05**
D3	51.88	50.28	**53.58**
D+1	48.73	48.57	**50.88**
D+2	**57.81**	57.21	56.74
F1	52.41	52.73	**52.78**
F2	52.73	**70.84**	44.94
F3	46.37	46.85	**49.92**
F+1	**51.67**	50.00	50.00

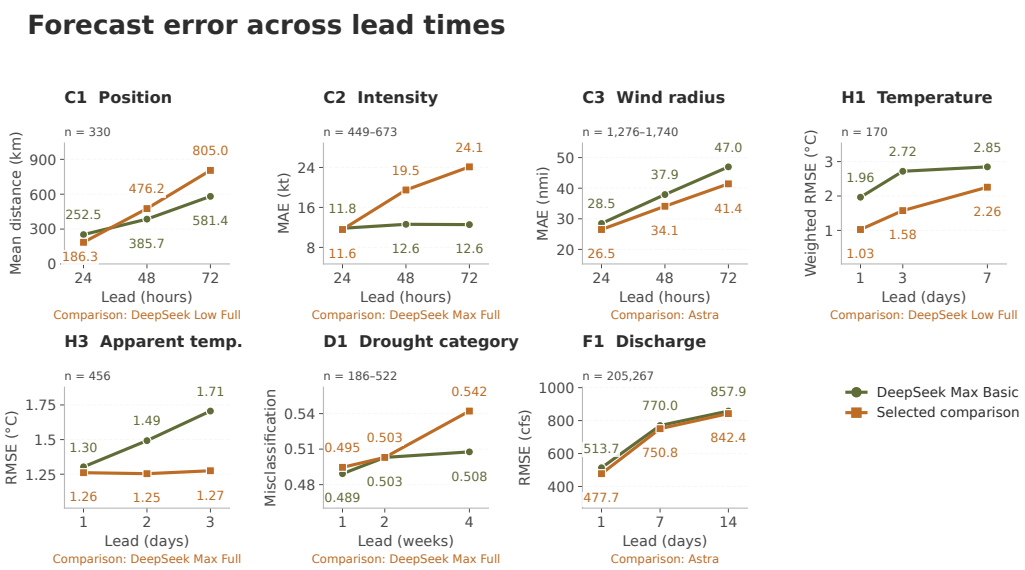

Figure 3: Forecast errors across lead times for seven questions. Green denotes DeepSeek Max Basic; the orange comparator is named beneath each plot. Comparators are selected for profile contrast, not model ranking. Units and axis ranges are task-specific; case counts may vary across leads. C1/C2 Basic, H1 Low, and H3 High/Max retain information-timing qualifications described in Appendix C.5.

Leads are 24/48/72 hours for C1–C3, 1/3/7 days for H1, 1/2/3 days for H3, 1/2/4 weeks for D1, and 1/7/14 days for F1. Case counts are respectively 330/330/330, 673/557/449, 1,740/1,528/1,276, 170/170/170, 456/456/456, 186/336/522, and 205,267/205,267/205,267. These are scoring units, not independent events; cohorts can differ by lead. Losses are mean distance (km), MAE (kt), MAE (nmi), latitude-weighted RMSE (°C), RMSE (°C), misclassification, and RMSE (cfs), respectively. Scores use the corresponding lead-specific local denominator in Equation 1.

Figure 3 retains B and selects, among A/L/H/M, the comparator with the largest mean absolute difference in local-normalized error across leads. This illustrates profile contrast, not model ranking.

Table 11: Complete RQ2b endpoint profiles. Triples follow increasing lead. Errors are lower-is-better; local-relative scores are higher-is-better.

Task	Method	Native error	Local-relative score
C1	Local	188.04 / 467.31 / 801.31	50.00 / 50.00 / 50.00
C1	A	173.11 / 449.07 / 775.82	54.13 / 51.99 / 51.62
C1	B	252.50 / 385.73 / 581.42	35.68 / 59.48 / 65.51
C1	L	186.29 / 476.25 / 804.96	50.47 / 49.05 / 49.77
C1	H	234.02 / 503.84 / 820.98	39.23 / 46.24 / 48.79
C1	M	207.43 / 502.79 / 819.17	45.11 / 46.35 / 48.90
C2	Local	11.90 / 19.13 / 23.14	50.00 / 50.00 / 50.00
C2	A	10.90 / 18.82 / 22.63	54.39 / 50.82 / 51.11
C2	B	11.83 / 12.62 / 12.56	50.31 / 69.68 / 77.24
C2	L	11.60 / 18.29 / 21.12	51.26 / 52.25 / 54.54
C2	H	11.89 / 19.32 / 22.77	50.03 / 49.51 / 50.79
C2	M	11.58 / 19.49 / 24.12	51.36 / 49.07 / 47.92
C3	Local	28.34 / 39.33 / 50.18	50.00 / 50.00 / 50.00
C3	A	26.51 / 34.10 / 41.42	53.35 / 57.09 / 59.48
C3	B	28.48 / 37.92 / 46.97	49.77 / 51.82 / 53.30
C3	L	28.72 / 38.76 / 46.00	49.33 / 50.72 / 54.34
C3	H	27.20 / 36.59 / 46.70	52.05 / 53.60 / 53.59
C3	M	28.10 / 36.53 / 42.73	50.42 / 53.68 / 57.97
H1	Local	1.99 / 2.80 / 2.98	50.00 / 50.00 / 50.00
H1	A	1.83 / 2.69 / 2.89	54.21 / 52.03 / 51.43
H1	B	1.96 / 2.72 / 2.85	50.57 / 51.44 / 52.21
H1	L	1.03 / 1.58 / 2.26	78.94 / 75.93 / 63.50
H1	H	2.02 / 2.79 / 2.96	49.13 / 50.13 / 50.19
H1	M	1.95 / 2.78 / 2.97	50.84 / 50.29 / 50.02
H3	Local	1.37 / 1.56 / 1.77	50.00 / 50.00 / 50.00
H3	A	1.31 / 1.51 / 1.74	52.36 / 51.53 / 50.91
H3	B	1.30 / 1.49 / 1.71	52.63 / 52.21 / 51.87
H3	L	1.33 / 1.51 / 1.73	51.46 / 51.71 / 51.17
H3	H	1.33 / 1.35 / 1.35	51.48 / 57.29 / 63.10
H3	M	1.26 / 1.25 / 1.27	54.27 / 60.75 / 65.87
D1	Local	0.4892 / 0.5089 / 0.5077	50.00 / 50.00 / 50.00
D1	A	0.4892 / 0.5089 / 0.5077	50.00 / 50.00 / 50.00
D1	B	0.4892 / 0.5030 / 0.5077	50.00 / 50.59 / 50.00
D1	L	0.4892 / 0.5089 / 0.5077	50.00 / 50.00 / 50.00
D1	H	0.4892 / 0.5030 / 0.5077	50.00 / 50.59 / 50.00
D1	M	0.4946 / 0.5030 / 0.5421	49.45 / 50.59 / 46.72
F1	Local	516.14 / 808.14 / 931.93	50.00 / 50.00 / 50.00
F1	A	477.66 / 750.85 / 842.41	53.87 / 53.67 / 55.03
F1	B	513.68 / 770.02 / 857.89	50.24 / 52.41 / 54.13
F1	L	502.10 / 766.85 / 873.75	51.38 / 52.62 / 53.22
F1	H	499.64 / 765.11 / 863.23	51.62 / 52.73 / 53.82
F1	M	505.97 / 762.26 / 853.01	51.00 / 52.92 / 54.41

C.4 AGENT CONFIGURATION AND REPORTING DETAILS

DeepSeek Flash uses the provider alias `deepseek-flash` and DeepSeek Harness commit `46a7f68b0922371ce7144b668b90e377d8e799f4`; the alias is not an immutable checkpoint. Basic supplies local execution and aggregate validation feedback. Full additionally exposes the fixed scientific-document collection and specialist forecasting services. Agent web access is disabled. Each episode allows 60 active minutes, two numerical threads, 8 GiB memory, and 2,000 tokens per command output; API recovery intervals are excluded from active time.

GPT-6 Astra uses High reasoning effort, local execution and installed numerical libraries, 60 active minutes, two numerical threads, and a 2,000-token command-output limit. Its recorded development-label access and fitting rules are retained. Qwen uses the configuration metadata supplied with its imported artifacts. Nominal resource limits do not make these systems protocol-equivalent.

The current selection has 24 tasks and 72 Full-effort outcomes. RQ3 contains seven paired training conditions on fixed test cases; F1 retains its natural test population. RQ2b contains seven multi-lead questions with endpoint-specific cohorts. Runs are matched by task, effort, and capability, not chosen by test score. One selected workflow per condition does not estimate stochastic run-to-run variation. Diagnostic interventions do not alter frozen benchmark submissions.

Verification checks prediction hashes, recorded entrypoints and transforms, stored feedback, and retained replay records. It establishes artifact identity and numerical consistency, not a clean-room retraining of every estimator or operational information availability. The run-level audit below states the quantitative coverage and remaining qualifications.

C.5 EVERY SELECTED OUTCOME

Table 12 retains every audited outcome. A/B/L/H/M denote Astra and the four DeepSeek settings; Qw denotes Qwen, with p indicating a provisional scoring contract. The Qwen W+3 row marked "archived diagnoses the superseded procedure (score 98.90); it is not the current RQ1 result (51.23), whose saved predictions were unavailable for this residual audit. Each row reports the strongest supported finding on the scored cases; multi-lead tail diagnostics use the longest lead unless stated otherwise. Native units are defined in Appendix D.

Evidence codes are I: isolated output-component effect; S: verified prediction or delivery restriction; H: source-supported hypothesis; U: unresolved signal-versus-model limitation; and Q: information-boundary qualification. Probability bias is mean prediction minus prevalence. Tail diagnostics condition on observed-outcome deciles: error shares use squared error for RMSE and absolute error for MAE. These conditional patterns do not alone establish a unique cause.

Table 12: Run-level diagnosis for all 144 selected outcomes. Scores use local model = 50. I/S/H/U/Q denote the evidence codes defined above. Native-error units follow Appendix D.

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
C1	B	53.55	24/48/72 h: 252/386/581 km	Later-issued short-lead substitutions; apparent skill qualified (Q).
C1	L	49.77	24/48/72 h: 186/476/805 km	Missing-environment handling and feature transfer (H).
C1	H	44.76	24/48/72 h: 234/504/821 km	Missing-environment handling and feature transfer (H).
C1	M	46.79	24/48/72 h: 207/503/819 km	Missing-environment handling and feature transfer (H).
C1	A	52.58	24/48/72 h: 173/449/776 km	Missing-environment handling and feature transfer (H).
C1	Qw	52.78	24/48/72 h: 173/447/771 km	Missing-environment handling and feature transfer (H).
C2	B	65.74	Upper-decile bias -22; error share 22%	Later-issued short-lead substitutions; apparent skill qualified (Q).
C2	L	52.69	Upper-decile bias -64.5; error share 37%	Underprediction of intense storms; tail representation (H).

Table 12: Run-level diagnosis (continued).

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
C2	H	50.11	Upper-decile bias -63.5; error share 34%	Underprediction of intense storms; tail representation (H).
C2	M	49.45	Upper-decile bias -64.6; error share 33%	Underprediction of intense storms; tail representation (H).
C2	A	52.11	Upper-decile bias -65.9; error share 36%	Underprediction of intense storms; tail representation (H).
C2	Qw	53.07	Upper-decile bias -64; error share 36%	Underprediction of intense storms; tail representation (H).
C3	B	51.63	Upper-decile bias -119; error share 26%	Under-expansion of large wind fields (H).
C3	L	51.46	Upper-decile bias -87.1; error share 20%	Under-expansion of large wind fields (H).
C3	H	53.08	Upper-decile bias -90.4; error share 21%	Under-expansion of large wind fields (H).
C3	M	54.02	Upper-decile bias -85; error share 21%	Under-expansion of large wind fields (H).
C3	A	56.64	Upper-decile bias -89.5; error share 22%	Under-expansion of large wind fields (H).
C3	Qw	54.58	Upper-decile bias -80.4; error share 19%	Under-expansion of large wind fields (H).
C+1	B	50.35	AUROC 0.949; probability bias -0.031	Storm/cohort transfer and probability mapping (H).
C+1	L	43.57	AUROC 0.943; probability bias -0.048	Remove final mapping: loss 0.1019 to 0.0959 (I).
C+1	H	43.06	AUROC 0.958; probability bias -0.078	Remove final mapping: loss 0.1030 to 0.0954 (I).
C+1	M	47.96	AUROC 0.959; probability bias -0.053	Storm/cohort transfer and probability mapping (H).
C+1	A	54.30	AUROC 0.966; probability bias -0.041	Storm/cohort transfer and probability mapping (H).
C+1	Qw	55.41	AUROC 0.965; probability bias -0.027	Storm/cohort transfer and probability mapping (H).
C+2	B	50.49	AUROC 0.805; probability bias +0.071	Seasonal transfer/calibration with only 75 cases (H).
C+2	L	56.97	AUROC 0.844; probability bias +0.003	Seasonal transfer/calibration with only 75 cases (H).
C+2	H	50.50	AUROC 0.811; probability bias +0.082	Seasonal transfer/calibration with only 75 cases (H).
C+2	M	50.87	AUROC 0.815; probability bias +0.069	Seasonal transfer/calibration with only 75 cases (H).
C+2	A	51.05	AUROC 0.824; probability bias +0.071	Seasonal transfer/calibration with only 75 cases (H).
C+2	Qw	52.08	AUROC 0.830; probability bias +0.075	Seasonal transfer/calibration with only 75 cases (H).
W1	B	57.01	AUROC 0.814; probability bias +0.005	Spatial feature choice and transfer across fires (H).
W1	L	51.63	AUROC 0.797; probability bias -0.068	Spatial feature choice and transfer across fires (H).
W1	H	59.13	AUROC 0.831; probability bias -0.027	Spatial feature choice and transfer across fires (H).
W1	M	52.13	AUROC 0.818; probability bias -0.002	Spatial feature choice and transfer across fires (H).
W1	A	64.68	AUROC 0.869; probability bias -0.005	Spatial feature choice and transfer across fires (H).
W1	Qw^{p}	51.33	AUROC 0.807; probability bias -0.091	Spatial feature choice and transfer across fires (H).
W2	B	51.66	Upper-decile bias -20.8; error share 77%	Rare growth bursts missed by median-oriented models (H).

Table 12: Run-level diagnosis (continued).

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
W2	L	51.64	Upper-decile bias -21.8; error share 79%	Rare growth bursts missed by median-oriented models (H).
W2	H	51.04	Upper-decile bias -21.4; error share 78%	Rare growth bursts missed by median-oriented models (H).
W2	M	51.92	Upper-decile bias -20.2; error share 77%	Rare growth bursts missed by median-oriented models (H).
W2	A	51.51	Upper-decile bias -18.9; error share 75%	Rare growth bursts missed by median-oriented models (H).
W2	Qw	51.28	Upper-decile bias -20.2; error share 77%	Rare growth bursts missed by median-oriented models (H).
W3	B	42.50	Medium-class recall 36.2%	Undo class weights: error 0.441 to 0.367 (I).
W3	L	49.38	Medium-class recall 38.6%	Medium-class separation and decision weights (H).
W3	H	42.50	Medium-class recall 18.1%	Undo class weights: error 0.441 to 0.386 (I).
W3	M	50.84	Medium-class recall 40.5%	Medium-class separation and decision weights (H).
W3	A	51.92	Medium-class recall 45.2%	Medium-class separation and decision weights (H).
W3	Qw^{P}	51.70	Medium-class recall 44.3%	Medium-class separation and decision weights (H).
W+1	B	49.85	Upper-decile bias -1.07×10^8; error share 99%	Later-event adjustment/target copies; apparent skill qualified (Q).
W+1	L	49.80	Upper-decile bias -1.19×10^8; error share 100%	Sparse event support and restricted extrapolation (H).
W+1	H	49.95	Upper-decile bias -1.14×10^8; error share 100%	Later-event observed-value override; apparent skill qualified (Q).
W+1	M	50.22	Upper-decile bias -1.1×10^8; error share 100%	Sparse event support and restricted extrapolation (H).
W+1	A	50.06	Upper-decile bias -1.02×10^8; error share 99%	Sparse event support and restricted extrapolation (H).
W+1	Qw	50.02	Upper-decile bias -1.04×10^8; error share 99%	Sparse event support and restricted extrapolation (H).
W+2	B	50.27	AUROC 0.605; probability bias -0.001	Weak ranking; missing signal versus model capacity unresolved (U).
W+2	L	50.44	AUROC 0.608; probability bias +0.001	Weak ranking; missing signal versus model capacity unresolved (U).
W+2	H	49.55	AUROC 0.600; probability bias +0.007	Weak ranking; missing signal versus model capacity unresolved (U).
W+2	M	50.33	AUROC 0.604; probability bias +0.011	Weak ranking; missing signal versus model capacity unresolved (U).
W+2	A	50.54	AUROC 0.605; probability bias -0.006	Weak ranking; missing signal versus model capacity unresolved (U).
W+2	Qw	50.18	AUROC 0.597; probability bias +0.004	Weak ranking; missing signal versus model capacity unresolved (U).
W+3	B	50.88	Upper-decile bias -3.92; error share 57%	Pollution-peak errors; model/feature transfer (H).
W+3	L	51.06	Upper-decile bias -4.27; error share 59%	Pollution-peak errors; model/feature transfer (H).
W+3	H	51.50	Upper-decile bias -4.71; error share 61%	Pollution-peak errors; model/feature transfer (H).
W+3	M	50.32	Upper-decile bias -4.63; error share 63%	Pollution-peak errors; model/feature transfer (H).
W+3	A	48.93	Upper-decile bias -4.37; error share 62%	Pollution-peak errors; model/feature transfer (H).
W+3	Qw	98.90	Upper-decile bias -0.0971; error share 75%	Direct future-row PM2.5 copying; apparent skill qualified (Q).
H1	B	51.41	1/3/7 d RMSE 1.96/2.72/2.85	Spatial/temporal representation remains a candidate (H).

Table 12: Run-level diagnosis (continued).

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
H1	L	72.79	1/3/7 d RMSE 1.03/1.58/2.26	Smoothing across the sample/date axis; apparent skill qualified (Q).
H1	H	49.81	1/3/7 d RMSE 2.02/2.79/2.96	Spatial/temporal representation remains a candidate (H).
H1	M	50.38	1/3/7 d RMSE 1.95/2.78/2.97	Spatial/temporal representation remains a candidate (H).
H1	A	52.56	1/3/7 d RMSE 1.83/2.69/2.89	Spatial/temporal representation remains a candidate (H).
H1	Qw	50.36	1/3/7 d RMSE 2.02/2.78/2.88	Spatial/temporal representation remains a candidate (H).
H3	B	52.24	Upper-decile bias -0.565; error share 5%	Station/date correction and representation (H).
H3	L	51.45	Upper-decile bias -0.85; error share 6%	Station/date correction and representation (H).
H3	H	57.29	Upper-decile bias -0.738; error share 7%	Pooling forecasts across issue dates; apparent skill qualified (Q).
H3	M	60.30	Upper-decile bias -0.848; error share 10%	Cross-issue pooling and two-sided smoothing; apparent skill qualified (Q).
H3	A	51.60	Upper-decile bias -0.983; error share 6%	Station/date correction and representation (H).
H3	Qw	51.61	Upper-decile bias -0.735; error share 5%	Station/date correction and representation (H).
H+1	B	64.45	Upper-decile bias -3.03; error share 69%	Count/exposure modeling and seasonal transfer (H).
H+1	L	64.66	Upper-decile bias -8.45; error share 73%	Count/exposure modeling and seasonal transfer (H).
H+1	H	65.94	Upper-decile bias -4.7; error share 71%	Count/exposure modeling and seasonal transfer (H).
H+1	M	65.02	Upper-decile bias -5.4; error share 71%	Count/exposure modeling and seasonal transfer (H).
H+1	A	64.20	Upper-decile bias -4.54; error share 74%	Count/exposure modeling and seasonal transfer (H).
H+1	Qw	57.97	Upper-decile bias -2.14; error share 74%	Count/exposure modeling and seasonal transfer (H).
H+2	B	52.89	Upper-decile bias -0.743; error share 5%	Spatial feature design and local temperature residuals (H).
H+2	L	55.36	Upper-decile bias -0.616; error share 5%	Spatial feature design and local temperature residuals (H).
H+2	H	52.29	Upper-decile bias -0.745; error share 5%	Spatial feature design and local temperature residuals (H).
H+2	M	52.11	Upper-decile bias -0.695; error share 5%	Spatial feature design and local temperature residuals (H).
H+2	A	55.90	Upper-decile bias -0.626; error share 5%	Spatial feature design and local temperature residuals (H).
H+2	Qw	53.41	Upper-decile bias -0.746; error share 5%	Spatial feature design and local temperature residuals (H).
D1	B	50.20	Worsening recall 56.5/57.1/0.0%	Short-lead recall recovered; 4-week worsening still missed (S).
D1	L	50.00	Worsening recall 0.0/0.0/0.0%	Fixed decrement forces zero worsening recall (S).
D1	H	50.20	Worsening recall 56.5/57.1/0.0%	Short-lead recall recovered; 4-week worsening still missed (S).
D1	M	48.92	Worsening recall 6.5/57.1/52.9%	Transition-rule recall/accuracy tradeoff (S).
D1	A	50.00	Worsening recall 0.0/0.0/0.0%	Fixed decrement forces zero worsening recall (S).
D1	Qw^{p}	49.76	Worsening recall 0.0/1.8/2.3%	Near-decrement decisions miss most worsening; provisional (S).
D2	B	59.55	RMSE 0.291; dry-decile bias +0.139	Lag-history access differs; attribution is confounded (H).

Table 12: Run-level diagnosis (continued).

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
D2	L	49.79	RMSE 0.355; dry-decile bias +0.194	Lag-history access differs; attribution is confounded (H).
D2	H	50.52	RMSE 0.350; dry-decile bias +0.226	Lag-history access differs; attribution is confounded (H).
D2	M	59.05	RMSE 0.294; dry-decile bias +0.129	Lag-history access differs; attribution is confounded (H).
D2	A	59.67	RMSE 0.291; dry-decile bias +0.126	Lag-history access differs; attribution is confounded (H).
D2	Qw	49.88	RMSE 0.354; dry-decile bias +0.199	Lag-history access differs; attribution is confounded (H).
D3	B	52.81	AUROC 0.718; probability bias +0.010	Site-risk discrimination and probability spread (H).
D3	L	51.88	AUROC 0.711; probability bias +0.011	Site-risk discrimination and probability spread (H).
D3	H	50.28	AUROC 0.512; probability bias -0.011	One risk per year; within-year AUROC 0.5 (S).
D3	M	53.58	AUROC 0.717; probability bias -0.002	Site-risk discrimination and probability spread (H).
D3	A	56.01	AUROC 0.746; probability bias -0.024	Site-risk discrimination and probability spread (H).
D3	Qw	55.95	AUROC 0.718; probability bias +0.014	Site-risk discrimination and probability spread (H).
D+1	B	46.77	AUROC 0.713; probability bias +0.007	Small-sample ranking and calibration transfer (H).
D+1	L	48.73	AUROC 0.750; probability bias +0.012	Small-sample ranking and calibration transfer (H).
D+1	H	48.57	AUROC 0.749; probability bias +0.011	Small-sample ranking and calibration transfer (H).
D+1	M	50.88	AUROC 0.716; probability bias +0.009	Small-sample ranking and calibration transfer (H).
D+1	A	50.29	AUROC 0.711; probability bias +0.016	Small-sample ranking and calibration transfer (H).
D+1	Qw	46.81	AUROC 0.656; probability bias +0.029	Small-sample ranking and calibration transfer (H).
D+2	B	64.42	Upper-decile bias -0.899; error share 12%	Residual features/model selection; no old copy diagnosis (H).
D+2	L	57.81	Upper-decile bias -1.4; error share 14%	Residual features/model selection; no old copy diagnosis (H).
D+2	H	57.21	Upper-decile bias -1.49; error share 15%	Residual features/model selection; no old copy diagnosis (H).
D+2	M	56.74	Upper-decile bias -1.59; error share 14%	Residual features/model selection; no old copy diagnosis (H).
D+2	A	62.91	Upper-decile bias -0.858; error share 12%	Residual features/model selection; no old copy diagnosis (H).
D+2	Qw	59.60	Upper-decile bias -1.48; error share 18%	Residual features/model selection; no old copy diagnosis (H).
F1	B	52.26	Upper-decile bias -434; error share 97%	Large-flow errors; scale/log/blend choices (H).
F1	L	52.41	Upper-decile bias -808; error share 98%	Large-flow errors; scale/log/blend choices (H).
F1	H	52.73	Upper-decile bias -641; error share 97%	Large-flow errors; scale/log/blend choices (H).
F1	M	52.78	Upper-decile bias -740; error share 98%	Large-flow errors; scale/log/blend choices (H).
F1	A	54.19	Upper-decile bias -425; error share 93%	Large-flow errors; scale/log/blend choices (H).
F1	Qw	52.75	Upper-decile bias -703; error share 98%	Large-flow errors; scale/log/blend choices (H).

Table 12: Run-level diagnosis (continued).

Task	Run	Score	Measured error pattern	Attribution or remaining hypothesis
F2	B	52.94	Upper-decile bias -0.865; error share 70%	History representation and residual-model choice (H).
F2	L	52.73	Upper-decile bias -1.41; error share 79%	History representation and residual-model choice (H).
F2	H	70.84	Upper-decile bias -0.723; error share 64%	History representation and residual-model choice (H).
F2	M	44.94	Upper-decile bias -1.75; error share 72%	History representation and residual-model choice (H).
F2	A	78.94	Upper-decile bias -0.372; error share 69%	History representation and residual-model choice (H).
F2	Qw	71.73	Upper-decile bias -0.481; error share 66%	History representation and residual-model choice (H).
F3	B	48.32	AUROC 0.864; probability bias -0.034	Threshold discrimination and unavailable feedback (H).
F3	L	46.37	AUROC 0.852; probability bias -0.041	Remove final mapping: loss 0.1623 to 0.1591 (I).
F3	H	46.85	AUROC 0.841; probability bias -0.027	Remove final mapping: loss 0.1607 to 0.1570 (I).
F3	M	49.92	AUROC 0.871; probability bias -0.041	Threshold discrimination and unavailable feedback (H).
F3	A	50.67	AUROC 0.895; probability bias -0.056	Event discrimination and physical-feature transfer (H).
F3	Qw^P	50.53	AUROC 0.887; probability bias -0.055	Event discrimination and physical-feature transfer (H).
F+1	B	49.11	AUROC 0.806; probability bias -0.023	Map transfer/calibration; DeepSeek feedback unavailable (H).
F+1	L	51.67	AUROC 0.826; probability bias -0.084	Map transfer/calibration; DeepSeek feedback unavailable (H).
F+1	H	50.00	AUROC 0.809; probability bias -0.056	Exact specialist fallback explains score 50 (S).
F+1	M	50.00	AUROC 0.809; probability bias -0.056	Exact specialist fallback explains score 50 (S).
F+1	A	61.48	AUROC 0.871; probability bias -0.025	Map transfer and probability calibration (H).
F+1	Qw^P	51.73	AUROC 0.836; probability bias -0.067	Map transfer and probability calibration (H).

D QUESTION-SPECIFIC ERROR COMPUTATION

This appendix defines the loss between a submitted forecast and its observed target for all 24 questions. Lower loss is better. These losses are distinct from the local-model-normalized benchmark score and from the signed-error diagnostics used in RQ4. The same definitions apply to validation and test wherever the corresponding endpoint is available.

D.1 COMMON SCORING CONVENTIONS

For question q and lead h, let $\mathcal{I}_{q,h}$ be the evaluator-defined set of eligible scored units and $N_{q,h} = |\mathcal{I}_{q,h}|$. A scored unit may be a storm issue, a station–date, a gauge–date, an incident, or a pixel, as specified below. Predictions are aligned to targets by the question's case, time, lead, and spatial identifiers. The scoring mask is applied after alignment. Invalid submitted values are rejected rather than silently dropped; missing or invalid targets are excluded according to the fixed eligibility definition. A valid forecast is required for every requested output row, including rows outside a private scoring mask.

For scalar targets y_i and forecasts $\hat{y}_i$, the regression losses are

$$L_{\mathrm{MAE}} = \frac{1}{N} \sum_{i \in \mathcal{I}} |\hat{y}_i - y_i|, \tag{2}$$

$$L_{\mathrm{RMSE}} = \left[\frac{1}{N} \sum_{i \in \mathcal{I}} (\hat{y}_i - y_i)^2 \right]^{1/2}. \tag{3}$$

Errors are computed in the target's reported physical units, after any model-specific inverse transformation. RMSE takes the square root after pooling squared errors; it is not the mean of separate station-, event-, or map-level RMSEs.

For a binary target $y_i \in \{0, 1\}$ and submitted event probability $p_i \in [0, 1]$, the Brier loss is

$$L_{\mathrm{Brier}} = \frac{1}{N} \sum_{i \in \mathcal{I}} (p_i - y_i)^2. \tag{4}$$

Probabilities are not thresholded for this loss. For categorical outcomes, $\hat{c}_i = \arg\max_k p_{ik}$ and

$$L_{\mathrm{class}} = \frac{1}{N} \sum_{i \in \mathcal{I}} \mathbf{1}[\hat{c}_i \neq y_i]. \tag{5}$$

Class probabilities follow the supplied class order and sum to one; ties select the first class in that order. Brier and classification losses are dimensionless. Unless explicitly stated otherwise, eligible units receive equal weight within a lead. Pooling rows does not give each storm, station, gauge, or map equal total weight.

D.2 TROPICAL CYCLONES

C1: cyclone center position. At each of 24, 48, and 72 hours, compare the predicted and observed latitude–longitude pair for each storm issue using great-circle distance. With latitudes $\phi_i, \hat{\phi}_i$ and longitudes $\lambda_i, \hat{\lambda}_i$ in radians,

$$a_i = \sin^2\left(\frac{\hat{\phi}_i - \phi_i}{2}\right) + \cos\phi_i \cos\hat{\phi}_i \sin^2\left(\frac{\hat{\lambda}_i - \lambda_i}{2}\right), \tag{6}$$

$$L_{\mathrm{C1},h} = \frac{1}{N_h} \sum_{i \in \mathcal{I}_h} 2R \arcsin\Big(\sqrt{\mathrm{clip}(a_i, 0, 1)}\Big), \qquad R = 6371.0088\ \mathrm{km}. \tag{7}$$

Coordinates are supplied in degrees and converted to radians for this calculation. The loss is mean distance in kilometres, not coordinate-wise MAE or squared distance.

C2: maximum wind. Compute MAE between predicted and observed maximum wind speed, in knots, separately at 24, 48, and 72 hours. Each eligible storm issue contributes one absolute error at each scored lead.

C3: quadrant wind radii. Compute MAE in nautical miles for the 34-knot wind radius at 24, 48, and 72 hours. Each eligible storm-issue–quadrant row contributes one error; all eligible quadrants are pooled within a lead. The loss is not an error in wind speed or in the average radius of a storm.

C+1: extratropical transition. Compute Brier loss for transition within 48 hours on the fixed transition/non-transition scoring cohort. Each selected storm window contributes one probability error. The native constant-probability reference is $p = 0.5$, with loss 0.25.

C+2: regional cyclone occurrence. Compute Brier loss for occurrence within the next seven days, pooling the eligible region–issue records. Each probability is compared with the binary occurrence target for the same region and forecast window; this is not a track-distance metric.

D.3 Wildfire and associated environmental targets

W1: next-day active-fire pixels. Compute Brier loss at eligible pixels against the next-day active-fire indicator. The native endpoint uses the fixed balanced fire/non-fire mask; the natural-population analysis uses all eligible pixels. Pixels are pooled across maps before taking the mean, so maps with more scored pixels contribute more terms.

W2: additional burned area. Compute MAE in square kilometres between predicted and observed next-day additional burned area for each eligible incident–day. The target is an area increment, not cumulative final area. Average absolute errors over the selected rows without an additional per-incident averaging step.

W3: final fire-size class. Convert the three submitted probabilities, ordered small, medium, and large, to an argmax class and compute the misclassification rate over incidents. A one-class and a two-class mistake both count as one error. The native mask balances the three classes; the natural-population analysis uses the full eligible incident cohort. Class boundaries and target labels are fixed during benchmark preparation, not chosen from the evaluated predictions.

W+1: fire radiative energy. Compute RMSE in megajoules for next-day radiative energy over eligible incident-component–day records. Squared energy errors are pooled before taking the square root; the loss is neither a percentage error nor an error in radiative power.

W+2: new fire discovery. Compute Brier loss for a recorded discovery on the next day in a cell with seven preceding discovery-free days. The scoring population comprises the matched discovery/control opportunities described in Appendix B.2; it does not include every cell–day indiscriminately.

W+3: site PM2.5. Compute RMSE in $\mu\mathrm{g\,m^{-3}}$ for next-day total PM2.5 concentration over eligible monitor–day records. The target is the observed site concentration, not an estimated wildfire-attributable component.

D.4 Heat and heat-related impacts

H1: temperature fields. For each lead of 1, 3, and 7 days, compute latitude-weighted RMSE in degrees Celsius over the target fields. Let t index eligible forecast issues and j, k index latitude and longitude, with $w_j = \cos\phi_j$. The implemented loss is

$$L_{\mathrm{H1},h} = \left[\frac{1}{N_h} \sum_{t \in \mathcal{I}_h} \frac{\sum_{j,k} w_j (\hat{T}_{tjk,h} - T_{tjk,h})^2}{\sum_{j,k} w_j} \right]^{1/2} . \tag{8}$$

The weights act along latitude, not the forecast-issue axis. Spatial weighted mean squared error is averaged across issues before taking a single square root.

H3: apparent temperature. Compute RMSE in degrees Celsius over station–issue records, separately at 1, 2, and 3 days. Compare with the prepared apparent-temperature target, rather than air temperature alone. Pool squared errors across the eligible stations and dates within each lead.

H+1: heatstroke transports. Compute RMSE in transport counts for the next day over eligible prefecture–day records. Continuous count predictions are compared directly with the observed count; the metric does not divide by population or average prefecture-specific RMSEs.

H+2: daily minimum temperature. Compute RMSE in degrees Celsius for the next reported daily observation period, pooling eligible station–date records. The target is the reported daily minimum; the observation period is not replaced by an assumed universal 24-hour interval.

D.5 Drought

D1: drought category. Compute the misclassification rate for the six-category USDM target, encoded 0–5, separately at 1, 2, and 4 weeks. The scoring mask balances improving, unchanged, and worsening transitions within split and lead, but the predicted label remains the future category. Errors are not weighted by category distance, and the metric is not three-way transition classification.

D2: precipitation deficit. Compute RMSE between predicted and observed next-month SPI12 over eligible climate-division–issue records. SPI12 is a standardized index, so the error is reported in index units rather than millimetres of precipitation.

D3: seasonal low flow. Compute Brier loss for the July low-flow event from February information over eligible gauge–year records. The event label is fixed by the prepared task definition; this loss does not compare predicted and observed continuous discharge. The native reference is available only for the test split in the current contract. Research scoring uses training-only event prevalence as its reference.

D+1: flash-drought onset. Compute Brier loss for onset during the next five-day period over the fixed balanced onset/non-onset site–issue cohort. Each selected onset opportunity contributes one error, independent of how many daily observations support its predictors.

D+2: vegetation stress. Compute RMSE for next-week regional stressed-area percentage over eligible state–week records. Predictions and targets use percentage points on the same scale; this is not a relative percentage error.

D.6 Floods

F1: river discharge. Compute RMSE in cubic feet per second over eligible gauge–issue records separately at 1, 7, and 14 days. Test scoring retains the full eligible population. The score pools squared errors across gauges and issues before the square root; it is not gauge-normalized RMSE or an equal-gauge average.

F2: river stage. Compute RMSE in feet between predicted and observed stage six hours after each eligible gauge issue. Compare with the target for the same gauge and target time; threshold exceedance is evaluated separately in F3.

F3: flood-threshold crossing. Compute Brier loss for crossing the task's gauge-specific flood threshold within six hours. Each selected issue window contributes one binary-event probability error. The evaluator uses the fixed matched opportunity cohort, including same-gauge controls selected by threshold margin and stage slopes; individual measurements inside the six-hour window are not separate scoring terms.

F+1: newly inundated pixels. Compute Brier loss for becoming inundated at the next target satellite observation among initially dry, valid pixels. In the CONUS42 variant, the private scoring mask contains 104,612 test pixels from the 19 test maps. Predictions are required for all supplied grid rows, but only masked pixels enter the loss. Errors are pooled over pixels, not averaged equally over maps. Existing water pixels are outside the newly inundated target population.

D.7 From question losses to scores

The losses above enter Equation 1 using the fitted local model's loss on the same cases, weights, metric, and lead. The local model scores 50 and zero error scores 100. Transform each lead's loss before taking the equal-lead task mean; cross-task summaries then weight the stated common task set equally. Native errors with different units are not pooled, and a zero local loss leaves the ratio undefined. Simple references and historical scores are distinct from this normalization. Lead-time results retain the endpoint-specific cohorts documented in Appendix C.3.