METEOSAGE: AN OPERATIONAL WEATHER AGENT CONNECTING HETEROGENEOUS OBSERVATIONS TO FORECASTS AND SERVICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Operational weather forecasting requires synthesizing heterogeneous observations and specialized forecast products to assess evolving weather hazards and deliver timely public services. Although AI has advanced individual forecasting tasks, integrating these capabilities into a coherent, traceable workflow remains challenging because of heterogeneous data semantics, conflicting evidence, and context-dependent forecasting strategies. We introduce `MeteoSage`, an agentic framework connecting observations, analysis, forecasting, and weather services through three complementary components: (i) *Meteorological Agentic Environment* provides unified access to heterogeneous data and controlled tool execution; (ii) *Event-Centered Evidence Workspace* links evolving assessments and unresolved discrepancies to the underlying observations to guide conflict assessment and resolution; and (iii) *Recursive Hierarchical Memory* distills eligible task records and feedback into reusable, context-specific guidance. We further introduce WxFlowBench, a benchmark built from multi-source observations, operational forecast products, and real-world meteorological reports. Its five tracks cover evidence-grounded analysis (T0), nowcasting over 0–3 h (T1), forecasting from 3 h to 15 d (T2), meteorological report generation (T3), and public weather services and emergency support (T4). Comprehensive evaluation on WxFlowBench shows that `MeteoSage` strong performance across all tasks. Further analysis identifies distinct failure modes involving forecast-model selection and sensor-modality mismatches, highlighting the challenging nature of our benchmark. *Dataset and code:* `https://github.com/useliz/MeteoSage`.

1 INTRODUCTION

Weather forecasting is the first line of defense against rapidly evolving weather hazards, supporting risk warning, emergency response, and public-safety decision-making (Majumdar et al., 2021). For high-impact events such as severe convection, heavy rainfall, floods, and tropical cyclones, the key challenge is not merely detecting their occurrence, but anticipating their evolution, continuously assessing changing risks, and determining when warnings should be escalated (Reichstein et al., 2025). This requires forecasters to integrate heterogeneous observations and meteorological products into continuously updated assessments of the atmospheric state and emerging hazards. Failure to recognize rapid intensification in time can turn a manageable event into a severe disaster.

Recent advances in artificial intelligence have improved individual components of the forecasting workflow, including nowcasting (Jun et al.; Chen et al., 2026), AI-based numerical weather prediction (Lam et al., 2023; Bi et al., 2023), ensemble forecasting (Alet et al., 2026; Price et al., 2025), extreme-weather detection (Racah et al., 2017; Ran et al., 2025), and meteorological report generation (Zheng et al., 2026; Camps-Valls et al., 2025). Yet operational forecasting still relies heavily on human forecasters, who must repeatedly synthesize radar, satellite, station, analysis-field, and model products, making the process labor-intensive, experience-dependent, and prone to delayed or inconsistent interpretation (Stuart et al., 2022). Large language models (LLMs) have recently enabled agentic frameworks for weather and climate science, with systems such as Zephyrus supporting rapid analysis of extreme events and climate impacts (Varambally et al., 2026; Tang et al., 2026; Ni et al., 2026; Li et al., 2026). Recent weather agents have advanced scientific analysis, forecast-

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.

ing, and extreme-weather early warning. However, jointly supporting heterogeneous observations, multiple forecasting horizons, and downstream weather services remains underexplored. These advances motivate a complementary question: ***can an agent coordinate heterogeneous meteorological evidence and forecasting tools across lead times to produce traceable analyses, forecasts, and task-appropriate service reports like a human forecaster?***

Agent harnesses provide a feasible foundation for building such a system. Within this framework, LLMs can serve as state coordinators that interpret forecasting tasks, select specialized tools, and integrate validated conclusions. With this division of responsibilities, deterministic meteorological computations and task-specific forecasting models produce quantitative outputs. The agent then coordinates their use and synthesizes the resulting evidence into meteorological assessments, reports, and task-appropriate services.

However, organizing this observation-to-service workflow requires more than stacking tools and presents three challenges. **Heterogeneous Data Semantics (C1):** Gridded fields, station time series, radar volume scans, satellite imagery, model ensembles, and documentary records differ in units, spatial and temporal definitions, and evidential roles. Rigid integration rules can overlook these differences, leading to invalid comparisons and misleading conclusions. **Non-equivalent Evidence Conflicts (C2):** Meteorological evidence is often incomplete or conflicting, but different forms of disagreement arise from different mechanisms. Station–grid discrepancies may reflect sampling differences, radar artifacts indicate quality issues, and ensemble disagreement may reflect forecast uncertainty. **Context-Dependent Forecasting Strategies (C3):** Effective strategies vary with region, season, hazard type, forecast horizon, and event evolution, limiting the applicability of fixed workflows and uniform source-selection rules.

To address these challenges, we introduce `MeteoSage`, an agentic framework that connects observations to analysis, forecasting, and weather services through three complementary components. Its *Meteorological Agentic Environment* supports the compatible use of heterogeneous data (**C1**) by providing common access to data and scientific tools while preserving units, temporal definitions, and provenance. The *Event-Centered Evidence Workspace* supports the assessment of conflicting evidence (**C2**) by linking evolving weather assessments to observations, computed results, and their sources. These links help the controller investigate discrepancies, select further checks, and retain unresolved uncertainty in reports. The *Recursive Hierarchical Memory* supports context-dependent strategy selection (**C3**) by connecting working memory for the active case with experience memory distilled from completed tasks. The agent retrieves guidance according to the forecasting context. Feedback from task execution refines the experience that informs subsequent forecasts, creating a recursive update cycle without updating controller parameters.

Our main contributions are summarized as follows:

•**An agentic framework for weather forecasting.** We introduce `MeteoSage`, the first agentic framework that unifies multi-source meteorological observations, multi-horizon forecasting, and public-service and emergency support. Its coordinates three components: (i) *Meteorological Agentic Environment*, (ii) *Event-Centered Evidence Workspace*, and (iii) *Recursive Hierarchical Memory*.

• **A comprehensive benchmark for operational weather agents.** We curate WxFlowBench, a five-track benchmark spanning the operational forecasting workflow: meteorological analysis (T0), radar nowcasting (T1), multi-horizon forecasting (T2), meteorological report generation (T3), and public weather services and emergency support (T4).

• Comprehensive evaluation shows that `MeteoSage` achieves encouraging benchmark results, demonstrating its potential to assist forecasters in operational forecasting workflows. The identified failure modes involving forecast-model selection, sensor semantics, and forecast-pipeline composition provide concrete directions for the design of future weather agents.

2 Related Work

2.1 Machine Learning for Meteorology

Machine learning has advanced multiple components of weather application. StormInsight and StormMind model convective evolution for nowcasting (Jun et al.; Chen et al., 2026), while GraphCast

and Pangu-Weather provide efficient global forecasting (Lam et al., 2023; Bi et al., 2023). GenCast and WeatherNext Cyclones extend prediction to probabilistic atmospheric evolution and tropical cyclone forecasting (Price et al., 2025; Alet et al., 2026). ExtremeWeather and HR-Extreme support extreme-event detection (Racah et al., 2017; Ran et al., 2025), whereas WeatherSyn and CLLMate connect meteorological data with reports and event narratives (Zheng et al., 2026; Li et al., 2025). However, these works primarily address individual prediction, diagnosis, or communication tasks, leaving their integration into a observation-to-service workflow underexplored.

2.2 Agentic Frameworks for Meteorology

Meteorological agents combine language-based reasoning with scientific tools. Zephyrus supports weather-data analysis, forecasting, and simulation (Varambally et al., 2026), while ClimateAgent automate climate-analysis workflows (Li et al., 2026). EWE and HVR-Met emphasize knowledge-guided extreme-weather diagnosis (Jiang et al., 2025; Tang et al., 2026), and TerraAgent coordinates heterogeneous Earth-system computation (Nguyen et al., 2026). SIREN further established an experience-based end-to-end warning chain. (Ni et al., 2026). Nevertheless, existing frameworks only provide limited support for a unified operational workflow which integrates radar, satellite, and station observations with forecast products for weather service.

2.3 Multimodal Weather Benchmarks

Recent meteorological benchmarks increasingly combine multimodal evidence. ClimateIQA, WeatherQA, and CLLMate pair meteorological fields or imagery with text for anomaly interpretation, and event forecasting (Chen et al., 2025; Ma et al., 2024; Li et al., 2025). Terra combines time series with geographical imagery and descriptive text (Chen et al., 2024). Climate-Agent-Bench-85 evaluates climate-analysis reports (Li et al., 2026). TerraBench integrates cross-domain scientific tasks with numerical evaluation, while SIREN-Bench assesses individual warning procedures and connected warning chains (Nguyen et al., 2026; Ni et al., 2026). OpenEarthAgent and Earth-Agent further evaluate structured geospatial reasoning (Shabbir et al., 2026; Feng et al., 2026). However, existing benchmarks do not jointly cover observation-grounded analysis, forecast skill across lead times, and public-service support within a unified operational forecasting workflow.

3 MeteoSage: An Agentic Framework for Operational Weather Forecasting

To construct a comprehensive forecasting workflow, we introduce `MeteoSage`, an operations-inspired agentic framework composed of three main components: (i) *Meteorological Agentic Environment*, (ii) *Event-Centered Evidence Workspace*, and (iii) *Recursive Hierarchical Memory*

3.1 Meteorological Agentic Environment

MeteoSage provides an executable environment for accessing meteorological evidence, performing scientific computations, and generating forecast products, see Figure 1. The environment combines multi-source data, composable meteorological tools, and a controlled runtime. Common data and provenance contracts allow results to pass between tools while preserving their physical meaning and connection to the underlying sources.

3.1.1 Multi-source meteorological data.

- **Gridded atmospheric fields** comprise reanalysis, numerical analysis, and forecast products that describe atmospheric states and their predicted evolution, supporting physical diagnosis and forecast initialization or comparison.
- **Radar observations** capture precipitation structures and their temporal evolution, supporting storm tracking and observation-driven nowcasting.
- **Satellite observations** provide spatial information on cloud systems, including brightness-temperature patterns and changes, complementing radar and atmospheric fields.

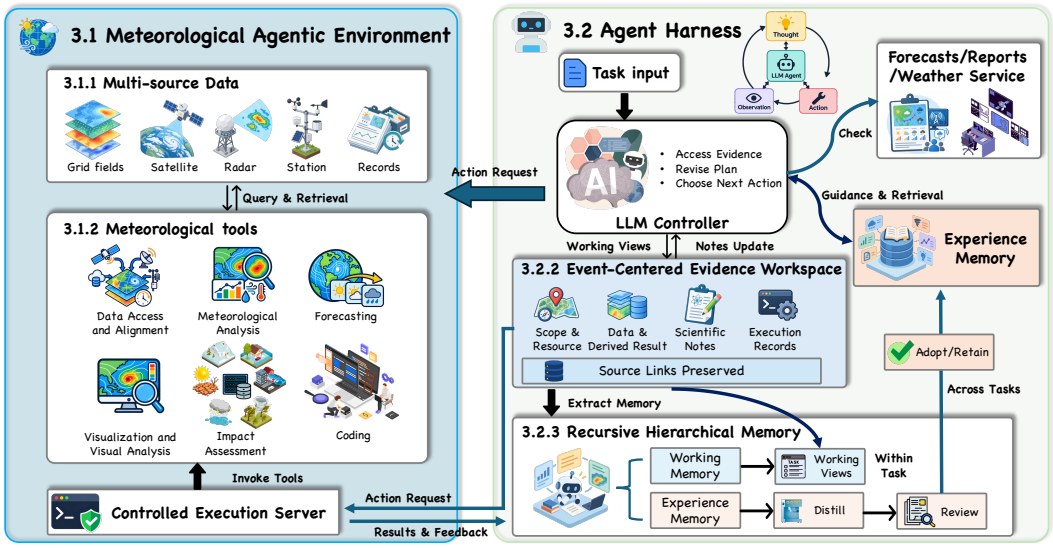

Figure 1: The Framework of `MeteoSage`.

- **Station observations** provide local time series of surface conditions for temporal analysis and comparisons with gridded estimates.
- **Meteorological historical records** provide documented weather assessments, event evolution, and reported impacts, supporting historical comparison, impact analysis, and evidence-grounded reporting. Access is restricted to documents permitted by the task.

The environment provides a common interface for discovering and retrieving data from different sources, while preserving their variable definitions, units, spatial and temporal support, quality information, and provenance, allowing Heterogeneous meteorological data to be compared under compatible physical and spatiotemporal definitions.

3.1.2 METEOROLOGICAL TOOLS.

The environment provides six groups of reusable tools for meteorological analysis,forecasting and service workflows tasks in WxFlowBench:

- **Data access and alignment.** *Source discovery and retrieval* provide access to task-permitted observations, reanalysis, and forecast products. *Spatial processing* supports subsetting, region masking, station–grid matching, and grid alignment. *Temporal and variable alignment* reconciles time windows and compatible variable definitions and units, while preserving missing-data masks, coverage, and source information.
- **Meteorological analysis.** *Statistical analysis* computes regional statistics, temporal accumulations, changes, extremes, and threshold-exceedance extent or duration, including comparisons with specified climatological references when available. *Physical diagnostics* computes dynamical, thermodynamic, and moisture-related quantities using scientific libraries such as MetPy. *Event and cross-source analysis* identifies and tracks meteorological features and compares signals from different sources under compatible physical and spatiotemporal definitions.
- **Forecasting.**
  The environment uses optical-flow extrapolation and SimVP for observation-driven nowcasting (Pulkkinen et al., 2019; Gao et al., 2022) and GraphCast and Pangu-Weather via model-specific atmospheric forecasting adapters (Lam et al., 2023; Bi et al., 2023). Other operational forecast products are retrieved via their source APIs.
- **Visualization and visual analysis.** *Visualization* produces georeferenced maps, time-series plots, and comparison figures from retrieved or computed data. *Visual interpretation* invokes vision-language models (Bai et al., 2023) to examine spatial structures and their evolution in the resulting images, returning qualitative interpretations to the controller.

- **Impact assessment.** Tools assemble hazard-intensity information, local exposure data, and task-permitted historical cases. The LLM controller interprets these inputs to characterize potential impacts on assets and populations. The resulting assessments retain supporting sources and uncertainty to inform warning assessments and emergency-response support.
- **Coding.** *CodeRun* provides agents with programmable scientific computing tools to handle analysis requirements not directly covered by predefined meteorological tools.

MeteoSage combines these tools according to the task, reusing intermediate outputs across computations. For example, a precipitation forecast can undergo regional exceedance analysis using diagnostic tools or *CodeRun*, followed by visualization.

3.1.3 CONTROLLED EXECUTION.

The execution layer runs meteorological tools and Python code. Generated code runs in an isolated workspace with restricted access to data and computational resources. Its outputs undergo separate format checks before being registered as reusable *artifacts*. Each operation returns result summaries and artifact references, or error feedback to guide the next action.

3.2 METEOSAGE AGENT HARNESS

The agent harness organizes the task into cycles of action and feedback (Yao et al., 2023), maintains the evidence and current assessment, and supplies relevant experience from previous tasks.

3.2.1 OBSERVATION-DRIVEN FORMULATION

Given a question q, the agent iteratively gathers evidence to produce an output $\hat{y}$. At agent interaction step t, the controller uses the task state S_t, available actions $\mathcal{A}_t$, and retrieved experience R_t:

$$\begin{aligned} c_t &= \mathrm{Project}(q, S_t, \mathcal{A}_t, R_t), \quad (a_t, \Delta n_t) = \phi_\theta(c_t), \\ o_t &= \mathrm{Process}(a_t), \qquad\qquad\qquad S_{t+1} = \mathrm{Update}(S_t, \Delta n_t, a_t, o_t). \end{aligned} \tag{1}$$

Here, $\mathrm{Project}$ constructs the working view c_t, from which the LLM controller ϕ_θ selects an action $a_t \in \mathcal{A}_t$ and an optional note update Δn_t. After valid notes are saved, $\mathrm{Process}$ executes the requested operation or checks a submitted answer. Update records the notes, action, and returned execution result or feedback message o_t, supplying the state for the next decision.

Submission checks validate output structure, artifact references, source eligibility, and explicit delivery conditions. Rejected submissions return feedback for revision within the same task execution. The task ends when a submission is accepted or a declared termination condition is reached.

3.2.2 EVENT-CENTERED EVIDENCE WORKSPACE

To support assessments based on incomplete or conflicting meteorological evidence, we maintain an event-centered workspace that separates source data and computed results from provisional interpretations. The LLM controller proposes explanations, selects further checks, and revises its assessment. This division enables traceable reasoning.

Task state and evidence tracking. We distinguish four parts of the task state:

$$S_t = (U_t, D_t, N_t, L_t), \tag{2}$$

where U_t contains task scope, available inputs, and resource information; D_t stores observations, derived artifacts, and their metadata and source links; N_t contains scientific working notes; and L_t records actions, parameters, results, and errors. The runtime updates task-execution information and logs, while the controller revises provisional findings, unresolved discrepancies, and planned checks in N_t. The controller is instructed to retain the values, units, spatial and temporal scope, and evidence references needed to interpret each finding.

Revised notes enter subsequent working views c_t (Eq. 1) alongside the original question, task information from U_t, recent records, the latest observation, available actions, and retrieved advice when

present, guiding action selection. Large arrays, long documents, and the complete execution history remain in storage, represented in c_t by concise descriptions and identifiers. The controller can retrieve the underlying evidence to reassess earlier findings.

Conflict analysis and revision. When evidence conflicts, the controller uses retained metadata and source links to select further checks. For example, a station–grid rainfall discrepancy prompts checks of variable definitions, units, accumulation intervals, and spatial support. If the permitted data support a compatible comparison, the controller invokes tools to align the data and recompute the difference; otherwise, it records the comparison as *indeterminate*. A discrepancy alone does not establish that either source is faulty. The controller updates findings, unresolved questions, and planned checks in N_t to guide subsequent actions, retaining unresolved discrepancies and relevant qualifications in the final output.

3.2.3 Recursive Hierarchical Memory

To adapt forecasting strategies under different evidence and task conditions, we adopt a *Recursive Hierarchical Memory* that couples current-case working memory with recursively refined experience. The unit of reuse is a methodological recommendation together with its applicability and exclusion conditions: these are assessed together during experience updates and preserved together during retrieval.

Memory organization. Inspired by cognitive theories (Baddeley, 2000; 2003), we connect memory across two timescales: working memory for the active case and experience memory across tasks. Working memory carries task-specific information across interaction steps, drawing on the workspace state S_t. Experience memory stores reusable recommendations, their conditions of use, and source information linking them to the records from which they were distilled.

Task-specific retrieval. The original question initiates retrieval from an experience version fixed for the task, and the agent may issue more specific queries during execution. Source restrictions first define eligible entries; text-based retrieval narrows large candidate sets, and an LLM selects and orders guidance using its applicability and exclusion conditions. Entries with known scope, input, or definition conflicts are excluded, and selection may be empty. The controller decides whether to use the supplied advice with current-task evidence; retrieved experience remains methodological guidance.

Recursive experience update. Experience evolves recursively from the outcomes of tasks guided by earlier versions. Between tasks, an LLM distills eligible completed-task records and requirement-level feedback into proposed changes to the guidance and its conditions of use. The records retain errors together with subsequent corrections, while feedback distinguishes requirements that were met, failed, or insufficiently supported. An LLM reviewer assesses each recommendation, its stated scope, and its supporting reason against these records and feedback, with limited correction and rechecking where needed. Rejected and unresolved proposals are not incorporated. Accepted changes are assembled into the next experience version.

Let E_k denote the experience collection at update round k, and let $\mathcal{T}_k$ and $\mathcal{F}_k$ contain the task records and feedback permitted for learning. The process is

$$E_k \xrightarrow{\text{Run}} (\mathcal{T}_k, \mathcal{F}_k) \xrightarrow{\text{Update}} E_{k+1} \xrightarrow{\text{Run}} (\mathcal{T}_{k+1}, \mathcal{F}_{k+1}), \tag{3}$$

where Update comprises distillation, record-grounded review, any permitted repair, and assembly. The resulting experience guides subsequent tasks, whose records and feedback support further refinement, while controller parameters θ, tools, and the control loop remain fixed. Appendix A.2 details the representation, feedback, update procedure, and retrieval.

4 WxFlowBench: A Benchmark for Weather Workflows

WxFlowBench evaluates evidence-grounded analysis, forecasting, reporting, and weather-service tasks through five complementary tracks, as shown in Figure 2. T0 includes globally scoped meteorological analysis, while T1–T4 primarily use cases from France and the United States, where the required observations, forecast products, and reporting sources can be linked. Tasks cover different forecast horizons under fixed information cutoffs. Operational relevance refers to their data products, analytical objectives, and service responsibilities. Appendix A.3–A.8 provide more details.

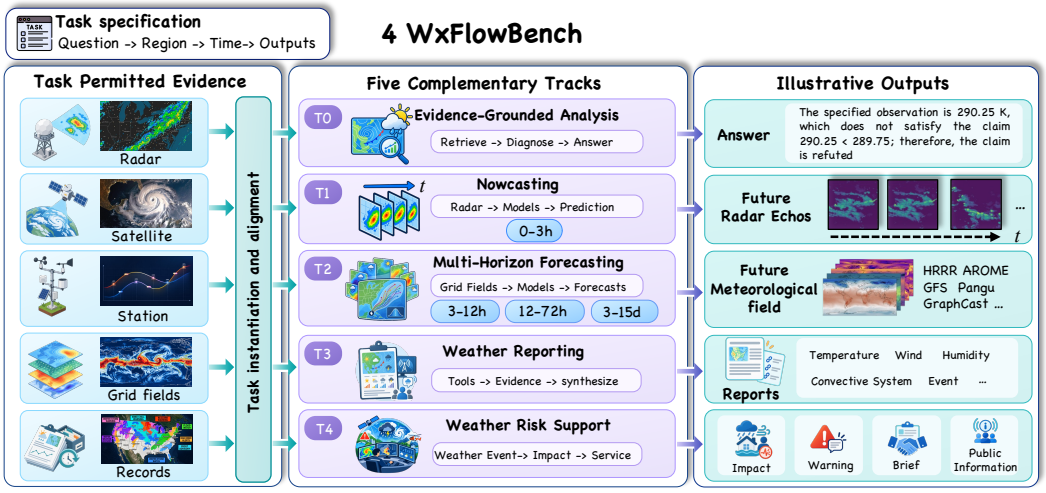

Figure 2: Benchmark.

4.1 Task Design and Coverage

T0: Evidence-Grounded Analysis evaluates data retrieval, deterministic statistics, climatological extremes, physical diagnostics, cross-source consistency, and claim verification. Tasks require answers in predefined formats, supported by traceable evidence.

T1: Nowcasting evaluates radar-based nowcasting over 0–3 h through forecast generation, targeted queries, source selection, and forecast comparison and arbitration.

T2: Multi-Horizon Forecasting evaluates Forecast, multimodel selection and ranking, threshold-exceedance probabilities, and hazard localization over very-short-range (3–12 h), short-range (12–72 h), and medium-range (3–15 d) horizons.

T3: Weather Reporting T3 evaluates the synthesis of meteorological evidence into traceable reports on current conditions, nowcasts, forecasts, integrated assessments, or retrospective events. The report's geographic scope, time period, and analytical focus are specified by the task.

T4: Public Weather Services and Emergency Support evaluates the translation of meteorological assessments into recipient-specific impact information, warning recommendations, partner-support briefs, and public communications, within the responsibilities of meteorological agencies.

4.2 Data sources.

WxFlowBench draws on radar, satellite, and station observations; atmospheric reanalysis; numerical and AI-based forecast products, including ensembles; and meteorological event records and operational reports. Detailed source descriptions and references are provided in Appendix A.3.1.

4.3 Benchmark Construction

WxFlowBench binds each task to a specified region, temporal scope, information cutoff, permitted evidence, and required output. Source alignment reconciles coordinates, units, and temporal support while preserving the original metadata, distinct time roles, and missing-value semantics. Reference targets are constructed independently of evaluated-agent outputs according to each track's output requirements. The quality-assurance protocol combines programmatic checks with technical and meteorological review to assess task clarity, scientific correctness, and consistency between task requirements, evidence, and reference targets. The reference construction and reviewer procedures are detailed in Appendix A.3.

5 EVALUATION

5.1 EVALUATION METRICS

We evaluate scientific output quality and task execution across five tracks (Detailed metric definitions are provided in Appendix A.9.).

T0: Answer correctness. We report Accuracy (Acc) as the mean of task-specific binary Hit Scores (HS). A hit requires all answer, evidence, and artifact requirements to be satisfied.

T1/T2: Forecast quality. Forecast fields and T1 derived queries are evaluated using CRPS. Source selection and ranking are assessed using the selected forecast's CRPS, Selection Regret, and Best-hit Rate. T2 probability and localization tasks use the Brier Score and FSS, respectively.

T3/T4: Report and service quality. We report Factual Consistency (FC) as the proportion of substantive checkable claims supported by the adjudication evidence, and Report Quality (RQ) as the proportion of applicable content and usability criteria satisfied. Groundedness (GRD) and BLEU-1 provide complementary diagnostics of support from delivered evidence and lexical agreement with the reference.

Cross-track execution metrics. Completion Rate measures the proportion of started executions producing valid final submissions. Tool-call Validity (TCV) measures compliance with declared tool-interface, invocation, and access requirements, averaged over applicable runs.

5.2 LLM BACKBONES

To validate comprehensively `MeteoSage`, we evaluate across five representative LLMs categorized into two tiers: (1) **Frontier**: *GPT-5.6 sol* (OpenAI, 2026a) for state-of-the-art reasoning and planning; and (2) **Lightweight**: *DeepSeek v4.1 Flash* (DeepSeek-AI, 2026), *GLM-5.3 Flash* (GLM-5-Team et al., 2026), *Gemini 3.8 Flash* (Google DeepMind, 2026), and *Qwen3.8-Flash* (Qwen Team, 2026) to assess practicality and resilience in latency-sensitive, resource-constrained settings. Additionally, we evaluate a **Vanilla baseline** using *DeepSeek v4.1 Flash* without any agentic harness. For all reasoning-capable models, the thinking effort is uniformly set to medium.

5.3 OVERALL PERFORMANCE

Table 1 compares `MeteoSage` across LLM controllers with a vanilla DeepSeek v4.1 Flash baseline, highlighting three findings.

The agentic workflow improves performance with the same backbone. Using DeepSeek v4.1 Flash, `MeteoSage` raises T0 Accuracy from 18.12% to 98.01% and improves Completion by 50.90–85.68 percentage points across T0–T2. T1 Best-hit Rate increases from 16.00% to 59.00%, while T2 Best-hit Rate doubles from 10.00% to 20.00%. These gains highlight the value of integrating meteorological evidence, specialized tools, and stateful execution to support accurate analysis, task delivery, and forecast-source decisions.

The benefits extend to reports and weather services. With the same DeepSeek v4.1 Flash backbone, `MeteoSage` raises the reported RQ from 6.96% to 52.53% on T3 and from 4.07% to 29.65% on T4. Gemini also achieves higher RQ summaries than the vanilla baseline on both tracks, and these two agent configurations attain Completion Rates of 96.67–99.33% across T3 and T4. The improvements extend beyond numerical analysis to synthesizing heterogeneous evidence into task-appropriate reports and service outputs.

Different tasks favor different controllers. GPT-5.6 Sol leads T0 Accuracy at 98.19% and achieves the lowest reported T2 pressure-field CRPS of 109.62 Pa. GLM-5.3 Flash leads T1 forecast generation at 7.99 dBZ, while Gemini-3.8 Flash achieves the highest T1 Best-hit Rate of 92.00% and the highest TCV on T3 and T4. DeepSeek obtains the highest reported T2 FSS of 0.8017. These complementary strengths show that controller advantages depend on the forecasting stage and requested product within the common framework.

More results are provided in Appendix A.1.3–A.1.7

Table 1: **Selected results across T0–T4 in WxFlowBench.**

Track	Metric	GPT-5.6 Sol	DeepSeek v4.1 Flash	Gemini-3.8 Flash	GLM-5.3 Flash	Qwen-3.8 Flash	Vanilla Baseline
T0	Accuracy (%) ↑	**98.19**	98.01	97.10	95.11	84.06	18.12
	Completion (%) ↑	**100.00**	98.91	99.82	96.56	91.85	48.01
	TCV (%) ↑	95.07	86.02	**99.09**	86.24	88.20	N/A
T1	Best-hit Rate (%) ↑	52.00	59.00	**92.00**	75.00	47.00	16.00
	Field CRPS (dBZ) ↓	21.09	15.60	10.31	**7.99**	13.39	N/A
	Completion (%) ↑	**100.00**	99.41	**100.00**	**100.00**	92.94	30.59
	TCV (%) ↑	94.66	95.33	**99.43**	97.72	94.27	N/A
T2	Best-hit Rate (%) ↑	**20.00**	**20.00**	**20.00**	15.00	5.00	10.00
	Field CRPS (Pa) ↓	**109.62**	113.12	117.99	112.02	181.84	N/A
	FSS ($\times 10^{-2}$) ↑	79.45	**80.17**	77.31	79.38	76.93	N/A
	Completion (%) ↑	**100.00**	**100.00**	**100.00**	97.14	67.19	14.32
	TCV (%) ↑	93.88	94.38	92.73	**97.84**	88.77	N/A
T3	FC (%) ↑	31.02	**35.45**	27.85	7.86	21.52	13.29
	RQ (%) ↑	17.09	**52.53**	44.94	11.43	13.93	6.96
	GRD (%) ↑	91.76	**95.89**	86.38	86.50	76.13	0.00
	BLEU-1 ↑	**36.03**	30.75	27.48	0.13	20.59	2.99
	Completion (%) ↑	94.67	97.33	99.33	27.33	94.67	**100.00**
	TCV (%) ↑	72.35	85.24	**98.20**	85.64	95.97	N/A
T4	FC (%) ↑	**38.38**	36.63	29.65	21.88	30.81	13.95
	RQ (%) ↑	11.05	**29.65**	19.77	9.38	9.30	4.07
	GRD (%) ↑	94.94	**95.86**	89.05	76.39	86.82	5.32
	BLEU-1 ↑	**34.31**	27.72	30.75	0.54	10.33	15.05
	Completion (%) ↑	96.00	96.67	98.00	36.00	89.33	**100.00**
	TCV (%) ↑	73.25	81.57	**97.45**	88.01	96.18	N/A

5.4 Additional Analysis

Beyond overall performance, we also analyze representative failure cases (Appendix A.1.2) and conduct tool-category ablations (Appendix A.1.1) to characterize agent behavior and assess the contribution of specialized tools to meteorological service tasks. These analyses collectively provide further support for the design of `MeteoSage`, explain its limitations, and offer practical insights for developing more capable and reliable meteorological agents.

6 Conclusion

Our agentic framework, `MeteoSage`, connects heterogeneous meteorological observations to analysis, forecasting, and weather services through a stateful, evidence-grounded workflow. Its meteorological environment and agent harness coordinate specialized tools, maintain source-linked intermediate results, and support iterative assessment. We also introduce WxFlowBench, a five-track benchmark spanning meteorological analysis, radar nowcasting, multi-horizon forecasting, meteorological reporting, and public weather services and emergency support. Experiments with multiple LLM controllers demonstrate the value of organizing model capabilities into executable meteorological workflows. With DeepSeek v4.1 Flash as the backbone, `MeteoSage` raises T0 Accuracy by 79.89 percentage points (from 18.12% to 98.01%) and T2 Completion by 85.68 percentage points (from 14.32% to 100.00%). These same-backbone gains highlight the contribution of the integrated workflow to both correct scientific answers and successful task delivery. Together, `MeteoSage` and WxFlowBench provide an extensible foundation for developing and evaluating weather agents that turn heterogeneous evidence into traceable analyses, forecasts, and task-appropriate services.

AI USE STATEMENT

In this study, generative AI tools were used to help with writing the manuscript and developing the research, including refining the language, optimizing the phrasing, structuring the text, formatting it in LaTeX, and generating and debugging partial code. The authors made all determinations regarding the research design, experimental setup, metric definitions, and scientific conclusions. Furthermore, this study involves large language model (LLM)-driven agents and LLM-assisted evaluation. Using such models is an integral part of the methodology investigated and evaluated in this paper. Their configurations and evaluation procedures are detailed in the main text and appendices, which are distinct from the aforementioned writing assistance.

ETHICS STATEMENT

This work does not involve experiments on human subjects or the collection of personally identifiable information. The benchmark is constructed from meteorological observations, forecasts, reports, and related operational products obtained from established data sources. Any released benchmark artifacts and source references should be used in accordance with the corresponding data access and licensing requirements.

REPRODUCIBILITY STATEMENT

We have made substantial efforts to support the reproducibility of this work. The paper and appendix provide detailed descriptions of the benchmark design, task construction, data sources, evaluation protocols, metric definitions, and implementation settings required to reproduce the reported experiments. We also provide links in the paper to the project resources. Subject to the access and redistribution policies of the underlying data providers, the benchmark dataset will be publicly released, including the task instances and associated metadata needed for evaluation. The source code for benchmark construction, evaluation, and experimental execution will also be publicly released. These resources are intended to enable the independent reproduction of the benchmark, evaluation procedure, and reported results, as well as the future evaluation of additional models and agents under the same protocol.

REFERENCES

Ferran Alet, Tom R Andersson, Ilan Price, Stratis Markou, Andrew El-Kadi, Dominic Masters, Amy Li, Samier Merchant, Natalie Williams, Gregory Thornton, et al. Operational tropical cyclone forecasting with ai. *Nature*, pp. 1–3, 2026.

Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839, 2003.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature communications*, 16(1):1919, 2025.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5322–5333, 2025.

Jun Chen, Minghui Qiu, Lin Chen, Yan Fang, Shuxin Zhong, Binghong Chen, and Kaishun Wu. Stormmind: Disentangled layerwise modeling for convective weather systems. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2145–2154, 2026.

Wei Chen, Xixuan Hao, Yuankai Wu, and Yuxuan Liang. Terra: A multimodal spatio-temporal dataset spanning the earth. *Advances in Neural Information Processing Systems*, 37:66329–66356, 2024.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

David C Dowell, Curtis R Alexander, Eric P James, Stephen S Weygandt, Stanley G Benjamin, Geoffrey S Manikin, Benjamin T Blake, John M Brown, Joseph B Olson, Ming Hu, et al. The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, LEI BAI, Conghui He, and Weijia Li. Earth-agent: Unlocking the full landscape of earth observation with agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dkIXAbWuxO.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3160–3170. IEEE, 2022.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan

Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google DeepMind. Gemini 3.8 flash best for tackling complex agentic tasks at scale. https://deepmind.google/models/gemini/flash/, 2026.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

IEM. National weather service text product archive: Area forecast discussions. Iowa State University. URL https://mesonet.agron.iastate.edu/wx/afos/.

Zhe Jiang, Jiong Wang, Xiaoyu Yue, Zijie Guo, Wenlong Zhang, Fenghua Ling, Wanli Ouyang, and Lei Bai. Ewe: An agentic framework for extreme weather analysis. *arXiv preprint arXiv:2511.21444*, 2025.

CHEN Jun, Yan Fang, Minghui Qiu, Yueran Qiu, CHEN Lin, Shuxin Zhong, Yu Zhang, and Kaishun Wu. Storminsight: Hierarchical environmental forcing and vertical coupling for weather system evolution. In *Forty-third International Conference on Machine Learning*.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Gwennaëlle Larvor and Lea Berthomier. Meteonet: An open reference weather dataset for ai by météo-france. In *American Meteorological Society Meeting Abstracts*, volume 101, pp. 1–ii, 2021.

Chenyue Li, Hyeonjae Kim, Wen Deng, Mengxi Jin, HUANG Wen, Mengqian Lu, and Binhang Yuan. Climateagent: Multi-agent orchestration for complex climate data science workflows. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=XLWvXNumGa.

Haobo Li, Zhaowei Wang, Jiachen Wang, YueYa Wang, Alexis Kai Hon Lau, and Huamin Qu. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17547–17573, 2025.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

Sharanya J Majumdar, Juanzhen Sun, Brian Golding, Paul Joe, Jimy Dudhia, Olivier Caumont, Krushna Chandra Gouda, Peter Steinle, Béatrice Vincendon, Jianjie Wang, et al. Multiscale forecasting of high-impact weather: Current status and future challenges. *Bulletin of the American Meteorological Society*, 102(3):E635–E659, 2021.

Météo-France. Vigilance météorologique. Météo-France. URL https://www.data.gouv.fr/datasets/vigilance-meteorologique-archivee.

NCEI. Storm events database. National Centers for Environmental Information. URL https://www.ncei.noaa.gov/stormevents/. NCEI Dataset C00510.

NCEI. Global forecast system (gfs) 0.5 degree, 2006. URL https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00634. NCEI Data Set Identifier 6182.

Dat Tien Nguyen, Thao Nguyen, Fadillah Adamsyah Maani, Huy M Le, Muhammad Umer Sheikh, Numan Saeed, Muhammad Haris Khan, and Salman Khan. Terrabench: Can agents reason over heterogeneous earth-system data? *arXiv preprint arXiv:2606.13148*, 2026.

Hang Ni, Weijia Zhang, Fan Liu, Mengqian Lu, and Hao Liu. Siren: Towards end-to-end extreme-weather early warning with experience-grounded llm agents. *arXiv preprint arXiv:2607.24588*, 2026.

NOAA NHC. National hurricane center data archive. NOAA/NWS/National Hurricane Center. URL https://www.nhc.noaa.gov/data/.

NOAA SPC. Storm prediction center product and event archives. NOAA/NWS/NCEP Storm Prediction Center. URL https://www.spc.noaa.gov/archive/.

NOAA WPC. Weather prediction center product archives. NOAA/NWS/NCEP Weather Prediction Center. URL https://www.wpc.ncep.noaa.gov/archives/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026a.

OpenAI. Gpt-6 astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, September 2026b.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Seppo Pulkkinen, Daniele Nerini, Andrés A Pérez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems*, 30, 2017.

Nian Ran, Peng Xiao, Yue Wang, Wesley Shi, Jianxin Lin, Qi Meng, and Richard Allmendinger. Hr-extreme: a high-resolution dataset for extreme weather forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 19823–19839, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. 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.

Markus Reichstein, Vitus Benson, Jan Blunk, Gustau Camps-Valls, Felix Creutzig, Carina J Fearnley, Boran Han, Kai Kornhuber, Nasim Rahaman, Bernhard Schölkopf, et al. Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, 16(1):2564, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Yann Seity, Pierre Brousseau, Sophie Malardel, G Hello, P Bénard, F Bouttier, Ch Lac, and V Masson. The arome-france convective-scale operational model. *Monthly Weather Review*, 139(3): 976–991, 2011.

Akashah Shabbir, Muhammad Umer Sheikh, Muhammad Akhtar Munir, Hiyam Debary, Mustansar Fiaz, Muhammad Zaigham Zaheer, Paolo Fraccaro, Fahad Shahbaz Khan, Muhammad Haris Khan, Xiao Xiang Zhu, et al. Openearthagent: A unified framework for tool-augmented geospatial agents. *arXiv preprint arXiv:2602.17665*, 2026.

Neil A Stuart, Gail Hartfield, David M Schultz, Katie Wilson, Gregory West, Robert Hoffman, Gary Lackmann, Harold Brooks, Paul Roebber, Teresa Bals-Elsholz, et al. The evolving role of humans in weather prediction and communication. *Bulletin of the American Meteorological Society*, 103 (8):E1720–E1746, 2022.

Richard Swinbank, Masayuki Kyouda, Piers Buchanan, Lizzie Froude, Thomas M Hamill, Tim D Hewson, Julia H Keller, Mio Matsueda, John Methven, Florian Pappenberger, et al. The tigge project and its achievements. *Bulletin of the American Meteorological Society*, 97(1):49–67, 2016.

Shuo Tang, Jiadong Zhang, Gengxian Zhou, Qizhao Jin, Qinxuan Wang, Yi Hu, Ning Hu, Hongchang Ren, Lingli He, Shiming Xiang, et al. Hvr-met: A hypothesis-verification-replanning agentic system for extreme weather diagnosis. *arXiv preprint arXiv:2603.01121*, 2026.

Sumanth Varambally, Marshall Fisher, Jas Thakker, Yiwei Chen, Zhirui Xia, Yasaman Jafari, Ruijia Niu, Manas Jain, Veeramakali Vignesh Manivannan, Zachary Novack, et al. Zephyrus: An agentic framework for weather science. In *International Conference on Learning Representations*, volume 2026, pp. 156654–156704, 2026.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in neural information processing systems*, 33:22009–22019, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jian Zhang, Kenneth Howard, Carrie Langston, Brian Kaney, Youcun Qi, Lin Tang, Heather Grams, Yadong Wang, Stephen Cocks, Steven Martinaitis, et al. Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(4):621–638, 2016.

Zinan Zheng, Yang Liu, Nuo Chen, Juepeng Zheng, Hong Cheng, and Jia Li. Weathersyn: An instruction tuning mllm for weather forecasting report generation. *arXiv preprint arXiv:2605.07522*, 2026.

A Appendix

A.1 Extra Experiment Results

Table 2: **Tool ablations on T3 and T4.** All configurations use DeepSeek v4.1 Flash.

Track	Configuration	FC (%) ↑	RQ (%) ↑	GRD (%) ↑	BLEU-1 ↑	Completion (%) ↑	TCV (%) ↑
T3	**Full (`MeteoSage`)**	35.45	52.53	95.89	30.75	97.33	85.24
	`w/o Analysis`	24.18	43.17	86.68	21.45	94.00	86.12
	`w/o Forecast`	24.81	29.75	89.88	31.04	100.00	85.11
	`w/o IA`	21.14	44.56	95.01	30.63	96.00	87.33
	`w/o Coding`	22.28	48.10	86.98	29.39	92.67	85.54
	`w/o memory`	34.64	46.00	97.14	24.28	96.67	85.28
	`Vanilla`	13.29	6.96	0.00	2.99	100.00	N/A
T4	**Full (`MeteoSage`)**	36.63	29.65	95.86	27.72	96.67	81.57
	`w/o Analysis`	21.77	22.35	87.56	20.31	96.67	83.50
	`w/o Forecast`	27.79	27.91	83.11	30.10	98.00	83.87
	`w/o IA`	21.28	29.65	87.35	29.34	96.67	85.32
	`w/o Coding`	23.61	26.04	84.62	29.28	92.67	85.17
	`w/o memory`	33.33	28.57	96.54	27.61	96.67	81.89
	`Vanilla`	13.95	4.07	5.32	15.05	100.00	N/A

A.1.1 Ablation Study

To assess the contribution of each tool and design, we conduct comprehensive ablation study:

- `w/o Analysis`: Remove *Meteorological analysis & Visualization and visual analysis* tools to assess the contribution of specialized meteorological knowledge and the role of visual information processing.
- `w/o Forecast`: Remove *Forecasting* tools to evaluate the influence of forecaster.
- `w/o IA`: Remove *Impact assessment* tools to evaluate its contribution to event impact analysis.
- `w/o Coding`: Remove *Coding* tools to assess the contribution of code-based computation.
- `Vanilla`: Remove agentic harness to evaluate their collective contribution.

The results yield several key observations (Table 2):

The integrated workflow improves report and service quality. Relative to the same-backbone vanilla baseline, the full configuration increases the reported RQ midpoint from 6.96% to 52.53% on T3 and from 4.07% to 29.65% on T4, corresponding to gains of 45.57 and 25.58 percentage points. Although Vanilla achieves 100.00% Completion on both tracks, its substantially lower report-quality summaries highlight the contribution of the integrated workflow to substantive content rather than text delivery alone.

Analysis and coding support complementary aspects of report construction. Jointly removing meteorological and visual analysis reduces RQ by 9.36 percentage points on T3 and 7.30 points on T4, the largest T4 RQ decrease among the four tool-ablation variants. Removing Coding reduces FC by 13.17 and 13.02 percentage points on T3 and T4, respectively, and lowers Completion to 92.67% on both tracks. These patterns associate diagnostic and visual support with report and service quality, while programmable computation supports both factual consistency and successful task delivery.

Impact assessment contributes most clearly to factual consistency. Removing Impact assessment produces the largest FC decreases among the four tool-ablation variants: 14.31 percentage points on T3 and 15.35 points on T4. On T3, RQ also decreases from 52.53% to 44.56%. On T4, the RQ

midpoint remains at 29.65%, while diagnostic GRD decreases from 95.86% to 87.35%. The reported pattern therefore associates impact-assessment tools primarily with the factual and evidential quality of outputs, rather than a uniform increase in service-requirement satisfaction.

Forecasting tools have the largest effect on T3 report quality. Removing Forecasting reduces T3 RQ from 52.53% to 29.75%, a decrease of 22.78 percentage points—the largest among the four tool-ablation variants. Completion nevertheless reaches 100.00%, locating the difference in the quality of submitted reports. On T4, RQ decreases by 1.74 percentage points, while diagnostic GRD decreases by 12.75 points. These task-dependent changes support specialized forecasting capabilities as an important component of meteorological reporting and evidence-supported service generation.

Memory supports report content. Removing memory weakens report quality and lexical agreement on T3, with smaller declines in factual consistency and report quality on T4. Completion remains largely stable. Groundedness and tool-call validity do not consistently favor the full configuration, suggesting that memory mainly helps the agent organize task-relevant content rather than deliver a final response.

A.1.2 Failure Pattern Analysis

A.1.2.1 T0 Failure Analysis

T0 failure case: Repeated inspection without a final spatial answer

Question (restated). Using the supplied snow-depth field and climatological reference, identify the native grid cells in the specified region where snow depth is at or below the corresponding climatological 5th-percentile threshold at 00:00 UTC on 29 June 2017.

Agent outcome (Gemini-3.8 Flash). No final task answer was produced after 53 interaction rounds and 58 model calls. The agent accessed the data artifacts and executed code, but repeatedly returned to tool descriptions and software files to determine how to represent the spatial answer. The run terminated during an unsuccessful repair of a tool request.

Required output. An identifiable set of native grid cells satisfying the prescribed snow-depth comparison, with the required spatial information and supporting evidence. The requested result is a spatial selection, rather than a regional mean or a general description of snow conditions.

Observed bottleneck. Repeated inspection did not resolve the transition from intermediate work to the required spatial output. Even in the final interaction rounds, the agent continued searching for output-representation definitions instead of submitting the requested result. The observed failure concerns task completion rather than a verified numerical mismatch.

Design lesson. Progress tracking should distinguish obtaining the data, performing the prescribed comparison, and delivering the result. Recovery should target the unresolved step while preserving completed work; successful intermediate operations alone do not establish completion of the scientific task.

We examine selected T0 failures through the relationship between the meteorological question, the available evidence, and the final answer. The cases distinguish errors in judging what the evidence supports from difficulties in carrying an analysis through to a verifiable result.

Finding a source does not establish its suitability. In a task concerning vertical velocity at 500 hPa, DeepSeek v4.1 Flash returned a specific ERA5 source, whereas the reference assessment required withholding that identification because of conflicting source semantics. The agent's working notes had already proposed checking the variable and unit, but the final answer did not reflect the restriction required by the reference. The important distinction is between knowing where a record comes from and establishing that it represents the quantity requested by the task. A source can be traceable while remaining unsuitable for a particular scientific use.

Uncertainty must be interpreted relative to the question. Evidence limitations affect different questions differently. In a specific-humidity task, GLM-5.3 Flash recorded the requested valid time from the supplied metadata but withheld the answer on data-integrity grounds. The task required the timestamp, rather than an analysis of the humidity values, and the retained diagnosis identified the abstention as unnecessary. A temperature-comparison task illustrates a different distinction: Gemini-3.8 Flash recognized that one input described mean sea-level pressure rather than 2 m temperature, but treated this as an inability to complete the task instead of returning the requested

indeterminate comparison. Here, the agent recognized the physical incompatibility. The error was in expressing its consequence. Together, these cases show why an agent must distinguish an unanswerable measurement request from an answerable question about the limits of the evidence.

More interaction does not necessarily advance the analysis. In a snow-depth task, the requested result was the set of grid cells at or below a specified climatological threshold. Gemini-3.8 Flash repeatedly inspected data, tool descriptions, and software files to determine how to represent the spatial answer. The run reached 53 rounds of interaction without a valid final submission. The observable difficulty was completing and delivering the analysis, rather than a demonstrated error in the snow-depth calculation. This pattern highlights the need to distinguish progress on the scientific question from activity spent navigating the execution interface. Recovery should address the specific obstacle while preserving the requested variable, region, time, and completed work.

Potential Avenues for Improvement. These cases motivate tracking what each finding depends on, which checks remain open, and what remains to be delivered. Evidence storage and execution records provide the basis for this tracking. The controller must connect them to decisions about further analysis, justified uncertainty, and final submission.

A.1.2.2 T1 Failure Analysis: Exploration and Forecast Assessment

We examine candidate exploration, forecast interpretation, and source attribution in GPT-5.6 Sol's T1 trajectories. The audited snapshot comprises 48 scored forecast-generation runs and 38 scored source-selection runs, with recorded losses verified against the submitted forecast artifacts.

Stopping after obtaining a valid forecast limits candidate exploration. Of the 48 forecast-generation runs, 41 retrieved only one forecast, and 46 ultimately submitted optical-flow predictions. The resulting mean CRPS of 21.09 dBZ closely matches the 22.29 dBZ obtained by always selecting optical flow on the same tasks. The agent thus frequently ended the workflow before exploring alternative forecast fields, placing the bottleneck at the exploration and stopping stage. This behavior may reflect task guidance that emphasizes obtaining a compliant forecast and submitting it correctly, making submission validity more salient than forecast-quality optimization. Optical flow also appeared first in every audited request catalog, suggesting a possible presentation-order effect.

Apparent physical plausibility can mislead forecast-quality judgments. In source-selection tasks, 37 of 38 runs retrieved all three candidates. The agent selected optical flow in 27 runs and SimVP in seven, whereas retrospective verification identified SimVP as a minimum-loss candidate in 34 runs. In a representative case, the agent preferred persistence because its echoes appeared more localized, while judging SimVP's widespread weak echoes implausible. Subsequent verification favored SimVP, with a CRPS of 4.05 dBZ compared with 17.90 dBZ for persistence. Further inspection identified 332 of persistence's 442 scored grid cells as carrying the prescribed -10 dBZ fill value. Its apparent sparsity therefore reflected the filling convention as well as the forecast content. This case illustrates how an intuitively plausible echo pattern can become an unreliable indicator of future accuracy when coverage and data-processing conventions are overlooked.

Potential Avenues for Improvement. These findings suggest that expected forecast quality should be made explicit in task guidance and that candidate comparison should be encouraged before termination when suitable alternatives are available. Diagnostic summaries could include source identities, coverage, and filling information, as well as the computed quantities. There could also be explicit checks to ensure that the selected candidate matches the forecast being assessed. In operational weather forecasting, selecting a model is an empirical decision with no definitive answer. Consequently, considerations extend beyond the performance of models from the previous forecast cycle. They also include trade-offs regarding model performance under varying weather conditions and the preferences of the LLM acting as the controller. Additional independent development tasks help further evaluate model selection strategies.

A.1.2.3 T2 Failure Analysis: Candidate Tracking and Error Recovery

We analyzed 20 Qwen-3.8 Flash runs on the T2 multi-model selection and ranking task. We focused on failure modes where the model struggled to provide a valid response. Each run involved between 11 and 51 prediction candidates. All runs successfully retrieved the full pool of candidates and attempted programmatic analysis. Eighteen runs attempted to submit ranking results, but only two

produced valid, scorable answers. The primary failure modes included errors when reading prediction arrays, failing to maintain correspondence between candidates and their associated prediction fields, and issues when attempting to rectify rejected submissions.

Loss of candidate identification information between analysis and submission phases. Each candidate corresponds to a specific prediction model, initialization cycle, and configuration. This association must be maintained throughout the ranking process. In 13 runs, the agent submitted duplicate field references and failed to maintain this necessary correspondence. In one case involving 11 candidates, the initial code correctly identified the reference for each prediction. Subsequently, however, the agent replaced these references with 11 copies recorded in a single tool call and repeatedly submitted invalid results. This approach confused the retrieval of a set with the specific identities of the predictions to be ranked. As a result, the final answer could not link the ranked candidates to their supporting prediction fields.

Data reading errors led to invalid recovery loops. We sampled 14 runs for analysis and found 71 "file not found" errors. The generated code repeatedly constructed file paths itself rather than using the provided execution interface paths. In one instance, all 35 prediction fields were available; however, the initial reader employed by the agent was incompatible with the NumPy array format. However, the agent failed to recognize the incompatibility as the root cause and instead attempted to alter file locations rather than correcting the reader. This resulted in repeated attempts to access nonexistent directories. That run concluded after 23 code execution attempts without producing a valid ranking. This process illustrates a scenario of recovery failure. A known format mismatch failed to guide the agent's subsequent actions effectively, causing it to shift its focus from comparing predictions to repeatedly attempting to access files.

Truncated feedback hinders targeted corrections. The submission interface only retained the first 600 characters of the validation information. For long ranked lists, the first 600 characters included only the submitted identifiers. Explanations regarding list duplicates or size errors appeared in the discarded portion. Thirteen runs encountered this type of feedback, and none produced a valid final submission. The missing explanations reduced the available information for targeted fixes. Similarly, recovery capability at the controller level is critical. In another instance, the agent identified a need for 27 distinct field references, but mistakenly believed that only one existed, even though all 27 had been published. Thus, the observed recovery failures involved the completeness of the error feedback and the controller's comprehension capabilities.

Precipitation magnitude was mistaken for forecast quality.

In an 11-candidate precipitation task, Qwen interpreted an array dimension of length 24 as an hourly time axis and summed over it without verifying its meaning in the metadata. Although the task specified a 24-hour accumulation window, this did not clarify the meaning of that array dimension. The agent then ranked candidates by descending precipitation totals, treating larger amounts as evidence of better forecasts. This procedure substituted precipitation magnitude for expected forecast accuracy. This case underscores the importance of verifying the physical meaning of derived quantities and evaluating candidates based on the desired forecast quality.

Potential Avenues for Improvement. Therefore, when selecting forecasting models, it is necessary to present and utilize the existing index of candidates more clearly. This index should link each prediction's identifying information, file location, and evidentiary references. Providing examples of how to read specific formats and outlining fixes for identified errors can reduce redundant trial-and-error efforts while preserving existing work. Feedback on validation could be improved by placing details regarding reasons for failure, expected versus actual list sizes, and affected entries before the lengthy list of identifiers. This would mitigate issues caused by truncation. Additionally, task guidelines should clearly distinguish between expected prediction accuracy and the magnitude of predicted values while explicitly specifying array dimensions and cumulative time windows.

A.1.3 T0 Performance Analysis

Table 3 extends the main evaluation with Kimi K3 and Qwen-3.7 Plus. The comparison highlights two complementary factors in T0 performance: the execution structure provided by the agentic harness and the capabilities of the underlying controller.

Table 3: **T0 evaluation results.**

Metric	GPT-5.6 Sol	DeepSeek v4.1 Flash	Gemini-3.8 Flash	Kimi K3	GLM-5.3 Flash	Qwen-3.8 Flash	Qwen-3.7 Plus	Vanilla Baseline
Accuracy ↑	**98.19**	98.01	97.10	97.10	95.11	84.06	75.36	18.12
Completion Rate ↑	**100.00**	98.91	99.82	98.73	96.56	91.85	89.49	48.01
TCV ↑	95.07	86.02	**99.09**	91.07	86.24	88.20	81.73	N/A

Table 4: **T1 evaluation results.**

Metric	GPT-5.6 Sol	DeepSeek v4.1 Flash	Gemini-3.8 Flash	GLM-5.3 Flash	Qwen-3.8 Flash	Vanilla Baseline
Field CRPS (dBZ) ↓	21.09	15.60	10.31	**7.99**	13.39	N/A
Query CRPS ($\times 10^{-2}$) ↓	**7.63**	9.10	9.18	8.59	8.69	N/A
Selected CRPS (dBZ) ↓	16.78	15.52	**8.11**	14.83	11.94	16.28
Selection Regret (dBZ) ↓	10.28	9.01	**1.60**	8.33	5.31	9.78
Best-hit Rate (%) ↑	52.00	59.00	**92.00**	75.00	47.00	16.00
Completion (%) ↑	**100.00**	99.41	**100.00**	**100.00**	92.94	30.59
TCV (%) ↑	94.66	95.33	**99.43**	97.72	94.27	N/A

The harness connects model capabilities to executable meteorological analysis. The vanilla multimodal baseline achieves 18.12% Accuracy and 48.01% Completion Rate, whereas the seven configurations within `MeteoSage` reach 75.36–98.19% Accuracy and 89.49–100.00% Completion Rate. The contrast highlights the demands of analyzing heterogeneous weather data: interpreting the question must be followed by retrieving the relevant observations, respecting their units and spatial and temporal support, performing the required computations, and returning an evidence-supported answer. `MeteoSage` organizes these operations through scientific tools, persistent task state, and execution feedback. The results support an agentic harness as an important complement to general multimodal reasoning, connecting model capabilities to the data-processing and verification steps required for successful task delivery. Successful T0 analysis depends not only on the underlying model, but also on how its capabilities are connected to meteorological evidence and executable operations.

The agentic harness strengthens performance with the same backbone. Using DeepSeek v4.1 Flash in both configurations, `MeteoSage` achieves 98.01% Accuracy and 98.91% Completion Rate, compared with 18.12% and 48.01% for the vanilla baseline. These correspond to gains of 79.89 and 50.90 %, respectively.This structure supports the transition from interpreting a question to performing the required analysis and delivering a verifiable answer. The gains in both metrics demonstrate the value of the integrated agentic workflow for heterogeneous weather-data tasks with the same underlying language model.

Controller upgrades strengthen performance within the same framework. GPT-5.6 Sol achieves the highest observed Accuracy of 98.19% with a 100.00% Completion Rate, illustrating the performance attainable with a flagship controller. DeepSeek v4.1 Flash and Kimi K3 also achieve high accuracies of 98.01% and 97.10%, respectively. The Qwen comparison further illustrates the benefit of controller upgrades: Qwen-3.8 Flash reaches 84.06% Accuracy, exceeding Qwen-3.7 Plus by 8.70 percentage points. Its Completion Rate increases from 89.49% to 91.85%, and TCV from 81.73% to 88.20%, showing improvements in both answer correctness and execution. Together, these results indicate that `MeteoSage` can benefit from advances in its underlying models while retaining a common meteorological toolset and workflow structure. The harness and controller therefore play complementary roles: the former organizes executable analysis, while improvements in the latter can translate into stronger end-to-end performance.

A.1.4 T1 Performance Analysis

Table 4 summarizes T1 performance across forecast generation, regional queries, source selection, and execution reliability. We highlight three findings.

The agentic workflow improves task completion and source selection. Across five controller configurations, `MeteoSage` achieves Completion Rates of 92.94–100.00% and TCV scores of 94.27–99.43%. Using DeepSeek v4.1 Flash as the backbone, `MeteoSage` achieves 99.41% Completion and a Best-hit Rate of 59.00%, compared with 30.59% and 16.00%, respectively, for the vanilla baseline—gains of 68.82 and 43.00 percentage points. These improvements highlight the value of connecting radar observations, forecasting tools, and evidence-based analysis within an executable workflow. The advantage extends beyond producing valid submissions to selecting forecast products suited to the requested region and lead time.

Forecast generation and source selection favor different controller configurations. GLM-5.3 Flash achieves the lowest reported Field CRPS of 7.99 dBZ, whereas Gemini-3.8 Flash leads source selection with the highest Best-hit Rate of 92.00% and the lowest Selection Regret of 1.60 dBZ. Gemini's results combine frequent selection of the minimum-loss candidate with a small average excess loss relative to the best available forecast. These complementary measures capture both how often the agent selects an optimal candidate and the consequences of its selection errors. The differing rankings distinguish obtaining accurate nowcast fields from identifying the best product within a fixed candidate pool, supporting separate evaluation of these two capabilities.

GPT-5.6 Sol leads regional-query performance. On targeted-query tasks, GPT-5.6 Sol achieves the lowest reported Query CRPS of 7.63×10^{-2}, approximately 11.2% below the next-lowest score of 8.59×10^{-2} achieved by GLM-5.3 Flash. The remaining configurations obtain scores between 8.69×10^{-2} and 9.18×10^{-2}. This result identifies prediction of the requested regional quantities as a particular strength of the GPT-based configuration. Across T1, the leading configurations are therefore task-dependent: GPT-5.6 Sol performs best on regional queries, GLM-5.3 Flash on forecast generation, and Gemini-3.8 Flash on source selection.

A.1.5 T2 Performance Analysis

Table 5: **T2 evaluation results.** CRPS and selection regret are reported separately by physical unit.

Metric	GPT-5.6 Sol	DeepSeek v4.1 Flash	Gemini-3.8 Flash	GLM-5.3 Flash	Qwen-3.8 Flash	Vanilla Baseline
Field CRPS (K) ↓	1.74	1.90	**1.67**	1.80	2.05	N/A
Field CRPS (mm) ↓	**4.47**	**4.47**	5.36	5.22	8.67	N/A
Field CRPS ($\mathrm{m\,s^{-1}}$) ↓	**0.68**	0.85	1.08	0.81	1.13	N/A
Field CRPS (Pa) ↓	**109.62**	113.12	117.99	112.02	181.84	N/A
Field CRPS ($\mathrm{m^2\,s^{-2}}$) ↓	156.60	**155.79**	155.82	168.94	369.67	N/A
Selection Regret (K) ↓	**0.62**	0.98	0.93	0.93	N/A	2.99
Selection Regret (mm) ↓	0.16	0.17	0.36	0.14	**0.13**	0.33
Selection Regret ($\mathrm{m\,s^{-1}}$) ↓	0.11	**0.05**	0.07	0.11	N/A	0.30
Selection Regret (Pa) ↓	60.02	41.21	**21.93**	41.21	N/A	119.20
Selection Regret ($\mathrm{m^2\,s^{-2}}$) ↓	579.66	572.78	**392.20**	784.39	N/A	N/A
Best-hit Rate (%) ↑	**20.00**	**20.00**	**20.00**	15.00	5.00	10.00
FSS ($\times 10^{-2}$) ↑	79.45	**80.17**	77.31	79.38	76.93	N/A
Brier Score ($\times 10^{-2}$) ↓	2.06	2.05	**1.98**	2.10	3.18	N/A
Completion (%) ↑	**100.00**	**100.00**	**100.00**	<u>97.14</u>	67.19	14.32
TCV (%) ↑	93.88	94.38	92.73	**97.84**	88.77	N/A

Table 5 summarizes T2 results for forecasting, source selection, event probability, spatial localization, and task execution.

The workflow improves task completion. With DeepSeek v4.1 Flash, `MeteoSage` achieves 100.00% Completion, compared with 14.32% for the vanilla baseline, a gain of 85.68 percentage points. Its Best-hit Rate also rises from 10.00% to 20.00%. GPT-5.6 Sol and Gemini-3.8 Flash likewise achieve 100.00% Completion, while GLM-5.3 Flash reaches 97.14%.

Forecast accuracy varies by variable. GPT-5.6 Sol has the lowest Field CRPS for wind ($0.68\,\mathrm{m\,s^{-1}}$) and pressure ($109.62\,\mathrm{Pa}$), and ties DeepSeek for precipitation ($4.47\,\mathrm{mm}$). Gemini leads temperature forecasting ($1.67\,\mathrm{K}$), while DeepSeek has the lowest geopotential CRPS ($155.79\,\mathrm{m^2\,s^{-2}}$).

Other T2 metrics show complementary strengths. DeepSeek achieves the highest FSS (0.8017), and Gemini has the lowest Brier Score (0.0198). GPT, DeepSeek, and Gemini each attain a 20.00% Best-hit Rate, although their selection regrets differ by variable: GPT is lowest for temperature, DeepSeek for wind, and Gemini for pressure and geopotential.

A.1.6 T3 Performance Analysis

On T3, `MeteoSage` substantially improves report quality over the vanilla DeepSeek v4.1 Flash baseline: DeepSeek achieves an RQ of 52.53%, compared with 6.96% for Vanilla, while Gemini reaches 44.94%. Although Vanilla attains 100.00% Completion, this mainly reflects its ability to directly return a textual response without performing tool-mediated retrieval, computation, or evidence integration; its substantially lower FC, RQ, and GRD show that successful delivery does not imply completion of the required meteorological analysis. Overall, the results indicate that the agentic workflow primarily improves the substantive quality and grounding of generated reports rather than merely their deliverability.

A.1.7 T4 Performance Analysis

T4 shows a similar pattern for public-weather and emergency-support tasks. DeepSeek reaches an RQ of 29.65%, compared with 4.07% for the vanilla baseline, while Gemini achieves 19.77%. The relatively high Completion of several configurations therefore does not by itself indicate successful service generation: T4 requires the agent to connect meteorological evidence with impacts, recipient-specific information, and permitted recommendations. These results suggest that the agentic workflow is particularly important for translating weather evidence into complete and task-appropriate service outputs.

The GLM-5.3 Flash exhibited frequent API timeouts during the T3/T4 experiments, resulting in substantial run-to-run instability and markedly lower completion rates on these tracks. Therefore, its T3/T4 results should be interpreted in light of this service-level instability. We attempted to modify the source to resolve this issue, but there was no improvement.

A.2 Details of Recursive Hierarchical Memory

A.2.1 Memory Organization and Experience Representation

Working memory carries task-specific information across interaction steps through the workspace state S_t, as described in Section 3.2.2. It preserves the information needed to continue the current analysis, while experience memory stores reusable guidance distilled from completed tasks.

Each experience entry contains a method recommendation, conditions under which it should be used, conditions under which it should not be used, and source information. For example, guidance on reporting the fraction of a region where precipitation exceeds a threshold may recommend stating whether the denominator is the area with valid observations or the full target area. The entry informs how the agent performs or reports the calculation; the numerical result remains grounded in current-task data. Applicability conditions allow the controller to judge whether the recommendation fits the available inputs and the requested analysis.

A.2.2 Task Records and Outcome Feedback

Offline experience updates use completed task records and explicit feedback on their outcomes. The records retain the original question, executed actions, key results, final output, and both errors and

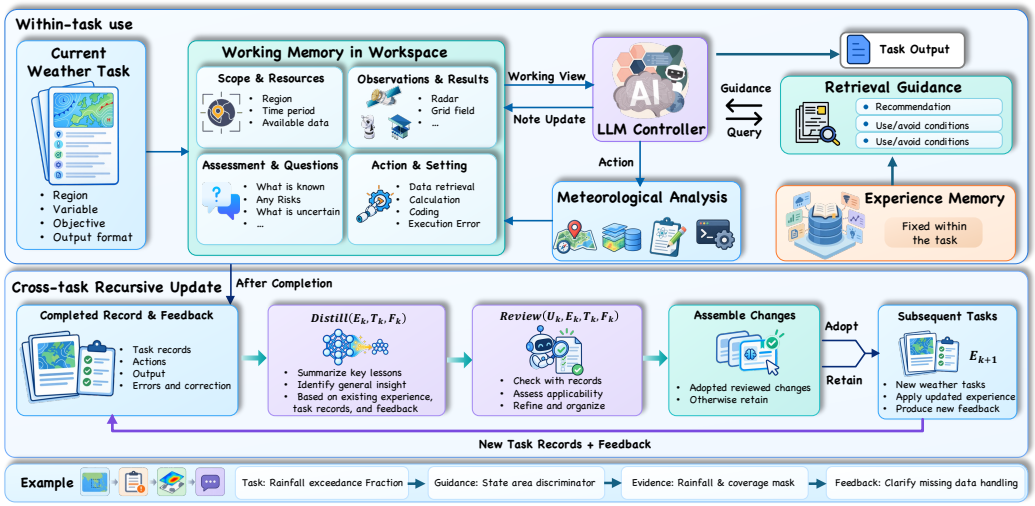

Figure 3: **Recursive Hierarchical Memory across two timescales. Top:** the controller uses the current-case workspace and retrieves recommendations with explicit applicability and exclusion conditions from an experience version fixed during the task. **Bottom:** eligible completed-task records and outcome feedback support offline updates through experience distillation, review, repair when needed, and assembly of accepted changes. Updated experience guides subsequent tasks without changing controller parameters. The rainfall example illustrates reusable guidance on the area denominator and missing-data treatment when reporting an exceedance fraction.

subsequent corrections. Keeping errors together with their corrections allows experience extraction to account for how the task was ultimately resolved.

Feedback identifies requirements that were met, failed, or insufficiently supported. Numerical judgments are grounded in reproducible calculations or reference information. At update round k, $\mathcal{T}_k$ contains the task records permitted for learning, and $\mathcal{F}_k$ contains their associated feedback. The Run notation in Eq. 3 includes task execution and the collection of these records and outcome assessments.

A.2.3 Experience Distillation, Review, and Retention

Experience retention follows a proposal, review, repair, and assembly pipeline:

$$\begin{aligned} \mathcal{U}_k &= \mathrm{Distill}(E_k, \mathcal{T}_k, \mathcal{F}_k), \\ \mathcal{U}_k^{\mathrm{acc}} &= \mathrm{ReviewRepair}(\mathcal{U}_k; E_k, \mathcal{T}_k, \mathcal{F}_k), \\ E_{k+1} &= \mathrm{Assemble}(E_k, \mathcal{U}_k^{\mathrm{acc}}). \end{aligned} \tag{4}$$

Here, $\mathcal{U}_k$ contains proposed experience changes, and $\mathcal{U}_k^{\mathrm{acc}}$ contains those accepted after review, any permitted repair, and the applicable programmatic checks.

The reviewer assesses the guidance, applicability conditions, negative conditions, and supporting reason against the disclosed task records. It assigns accept, revise, reject, or uncertain. Revised proposals are checked again using their corrected text; rejected and unresolved proposals are not incorporated. Acceptance indicates support within the stated scope, not scientific certification or independently measured performance improvement.

Programmatic checks enforce the change schema, valid references to disclosed material, permitted update operations, and inherited source-isolation metadata. They do not replace the LLM's semantic assessment. Experience retention is therefore not a separate held-out-performance selection between complete memory versions.

A.2.4 Task-Specific Retrieval and Context Management

Retrieval begins with the original question and may be repeated with more specific queries during execution. Programmatic source restrictions determine the eligible candidate set. In the docu-

mented selection configuration, sets containing more than 32 entries are narrowed by text-encoder retrieval (Reimers & Gurevych, 2019) ; an LLM selection call then chooses a subset and its delivery order. The selection may be empty.

Selection preserves each entry's guidance, applicability, and negative conditions. An entry with a known scope, input, or definition conflict is excluded rather than made applicable by deleting its conditions. Delivery limits and duplicate controls bound context usage. The controller subsequently decides whether to use the supplied advice with current-task evidence. The experience version remains fixed within each task execution.

A.3 Details of WxFlowBench

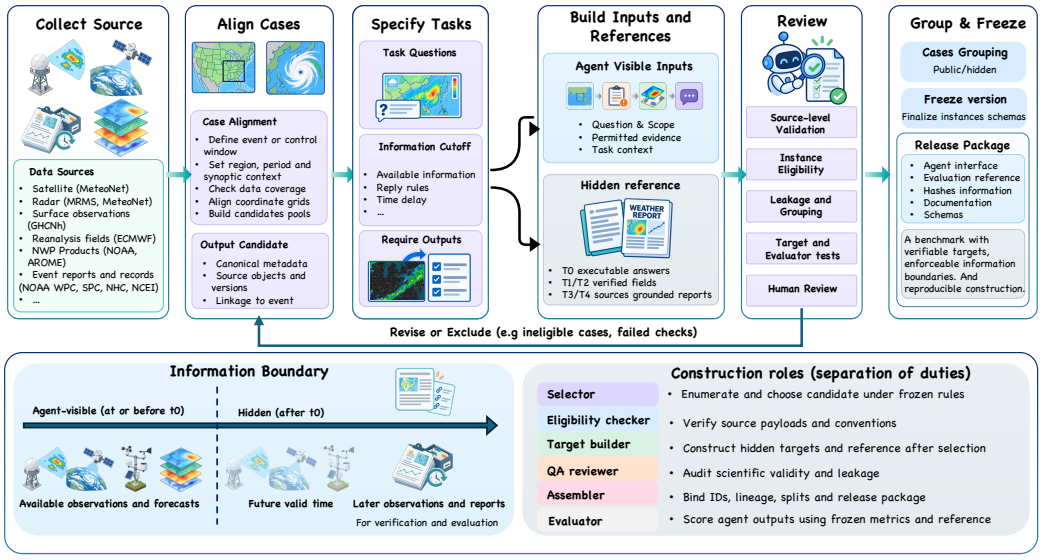

Figure 4: **WxFlowBench construction pipeline and information boundary.** Source collection, case alignment, and task specification define the evidence and output requirements for each instance. Agent-visible inputs and independently constructed reference targets are prepared separately, reviewed, and packaged with case grouping and version information. The lower timeline distinguishes information available by the cutoff t_0 under the declared replay policy from subsequent observations reserved for verification; eligible forecasts may predict times after t_0. Retrospective tasks use their specified review cutoff.

A.3.1 Data sources

- *Observations and reanalysis*: We use French radar, station, and satellite data from MeteoNet (Larvor & Berthomier, 2021); U.S. MRMS radar products (Zhang et al., 2016), station/METAR records, and SEVIR satellite imagery (Veillette et al., 2020); and ERA5 reanalysis accessed through WeatherBench 2 (Hersbach et al., 2020; Rasp et al., 2024).
- *Forecast products*: Forecast candidates include GFS (NCEI, 2006), HRRR (Dowell et al., 2022), AROME/ARPEGE (Seity et al., 2011), TIGGE ensembles (Swinbank et al., 2016), Pangu-Weather, and GraphCast (Bi et al., 2023; Lam et al., 2023).
- *Event records and operational reports*: Sources include NCEI Storm Events records (NCEI) and Météo-France Vigilance maps and follow-up bulletins (Météo-France). U.S. reports comprise NWS Area Forecast Discussions from 13 forecast offices and local warnings obtained through the Iowa Environmental Mesonet (IEM), SPC convective outlooks, WPC precipitation discussions, and NHC discussions and advisories (NOAA SPC; NOAA WPC; NOAA NHC).

A.3.2 Event Coverage

The T3/T4 construction protocol organizes weather coverage into five core families and additional challenge categories (Table 6). Core families define mandatory coverage, while challenge cate-

gories broaden the range of reporting and service scenarios for which qualified evidence supports the requested task. T3 includes controls for quiet weather and subthreshold near misses to evaluate whether agents accurately describe non-hazardous conditions. T4 focuses on hazard-related impact assessments and services.

Table 6: **Weather categories in the T3/T4 construction protocol.** Five Core families define mandatory hazard coverage, while Challenge categories extend coverage subject to source qualification and task validity.

Category	Description
Core weather families	
Heavy Precipitation	Intense or sustained rainfall with substantial accumulation.
Convective Hazards	Thunderstorms and associated severe-weather phenomena.
Non-convective Wind	Strong winds not directly generated by thunderstorms.
Temperature Extremes	Hazardous heat or cold, including heat waves, cold spells, and freezing conditions.
Frozen Precipitation and Snow	Selected frozen-precipitation and snow events with qualified source support.
Challenge weather categories	
Flood and Inundation	Flash, riverine, urban, or coastal flooding and inundation.
Visibility and Airborne Hazards	Fog, dust, sand, or smoke affecting visibility or atmospheric conditions.
Fire Weather	Meteorological conditions conducive to fire ignition or rapid spread.
Marine Weather and Sea State	Hazardous marine winds, waves, and related sea conditions.
Slow-onset Hydroclimate	Persistent precipitation deficits or prolonged dry conditions, including drought.

Coverage requirements. The T3 allocation additionally includes 15 non-hazard controls covering routine quiet weather and subthreshold near misses. T4 is entirely hazard-event based.

A.3.3 Task instantiation and alignment.

For T0, predefined question templates are instantiated with specific regions, time windows, and meteorological variables using the relevant observation and reanalysis sources. For T1/T2, each task specifies an observation history, forecast reference time, and target lead times, which constrain eligible observations and forecast products. T3 tasks are derived from official meteorological analyses, forecast discussions, and post-event reports, preserving their geographic scope, forecast horizon, and analytical focus. T4 extends selected cases into public-service and emergency-support scenarios.

Because the same event may be described with different timestamps, geographic identifiers, or hazard labels, sources are aligned to a shared event schema while retaining their original metadata. Reports are linked through event identifiers, update histories, and spatiotemporal consistency. Observation times, report issue times, forecast initialization times, and forecast valid times remain distinct. Coordinates, units, temporal support, and missing-value conventions are reconciled for comparison while preserving each product's physical meaning.

A.3.4 Reference Target Construction

Reference targets are constructed independently of the outputs of the evaluated agent. Each target is linked to its source records, scientific scope, and construction procedure. The form of each target follows the task at hand, which may be a numerical or categorical answer, a future verification field, a derived quantity, or a reference report.

- **T0: Independently verified answers.**
  Numerical and diagnostic references are produced using independent executable calculations. Retrieval and metadata answers are established from the specified source records, while categorical judgments follow predefined comparison and evidence-sufficiency rules. For evidence-modified variant questions, the expected answer is reassessed under the modified evidence conditions. These variants retain the underlying question while changing specified evidence conditions, such

as source availability, coverage, or reliability. The underlying facts may remain unchanged, but the evidence may no longer support the original determination. Therefore, reference construction distinguishes a supported answer from a task-permitted statement of insufficient evidence. Each instance specifies its answer type, required fields, units, and matching criteria.

- **T1: Radar observations and derived targets.**
  Source selection is evaluated using candidate losses against the same future radar targets. Comparison references are instead computed from the specified forecast artifacts using the task's fixed region, target time, threshold, and comparison operator.
- **T2: Variable-specific verification fields.** Across forecast horizons, primary verification uses ERA5 fields from WeatherBench 2 for atmospheric states and observation-based rainfall products for precipitation, as detailed in Appendix A.6.3. Forecasts and reference fields are matched by variable, valid time, and temporal support.
- **T3/T4: Source-constrained reference reports.** Each reference is a project-authored English report based on one adequate primary official report and supporting material added where necessary. The construction sources are fixed and include the selected reports, task requirements, and source-linked summaries of permitted observations and forecasts. Translation, reorganization, and condensation may improve readability, but must not introduce unsupported event facts. Operational references respect the task's information cutoff. T3 retrospective references instead use their specified review cutoff. T4 references additionally preserve the supplied warning state, service responsibilities, and distinction between potential impacts and realized losses.

Sources, versions, temporal coverage, and relevant processing steps are retained for reference verification and revision. Metric definitions and matching rules are provided in Appendix A.9.

A.3.5 Information and Learning Boundaries

Evidence eligibility. Valid time identifies the weather period represented by the data, whereas availability time determines when an object may be exposed under the task's declared replay policy. An observation is eligible only if its observation time or support end and its policy-defined availability are both no later than the as-of cutoff, and the source-access rules are satisfied. Forecasts may have future valid times when their cycle and required inputs are eligible by the cutoff.

The T1–T4 archive-replay configuration assumes zero publication delay: observations are nominally available at their observation time or support end, and forecasts at initialization or reference time. This assumption permits T1's cutoff to equal t_0, the final radar frame's nominal valid time; actual scans after t_0 remain excluded. T3-R uses its declared review cutoff. This retrospective setting evaluates agent workflows under declared availability assumptions rather than reconstructing historical operational latencies. Source issue times remain distinct from valid times; archive and retrieval timestamps serve as provenance metadata.

Learning eligibility. For static hidden-test evaluation, experience is constructed only from eligible public development records and their released feedback. Validation feedback may guide pre-test prompt and evaluator calibration, but validation records and feedback are excluded from experience distillation. Each evaluation condition's experience snapshot is frozen before hidden testing and remains read-only throughout it. Hidden-test outputs, execution records, and evaluator feedback are reserved for evaluation, with feedback released after all preregistered conditions finish. Linked tasks, including cross-track variants and T4 chain nodes, inherit the same statistical-group split and learning restrictions. Independent hidden executions do not share task-derived state. Within declared T4 chains, agent-generated upstream reports are passed as task inputs, while experience remains fixed and hidden references remain evaluator-only.

A.3.6 Reviewer Procedures

The review protocol checks whether each task is clearly specified, supported by permitted evidence, and paired with an appropriate reference target. It combines programmatic validation with human assessment of meteorological meaning and reporting requirements.

Reviewer responsibilities.

- **Technical reviewers** who are graduate students specializing in artificial intelligence, check consistency among the question, source records, reference target, and output requirements. Their checks cover variable identity, units, coordinates, time windows, missing-data treatment, and evidence availability. For example, a task requesting a three-hour rainfall total must not be paired with a reference calculated over 24 hours. For numerical tasks, reviewers inspect the independent reference calculation and any required artifacts.
- **Meteorological reviewers** who are experts from meteorological agencies, assess whether the question and reference interpretation are scientifically justified. They examine the suitability of diagnostics, the comparability of sources, and whether conclusions are reasonable. For report and service tasks, they additionally check important omissions, temporal consistency, faithful translation, and the assigned operational responsibilities. For example, a reference public message must preserve the given warning state, and a potential impact must not be presented as an observed loss.

Numerical-task review.

- **T0**: The pipeline combines programmatic reference checks with a stratified human audit across its capability families. Selected base tasks receive two independent reviews. Cases involving naturally faulty evidence, historical information cutoffs, and active compound evidence interventions receive additional targeted coverage. For evidence-modified variant questions, reviewers reassess answerability and the expected answer after the evidence change. Disagreements in the independent T0 reviews are referred to a third reviewer.
- **T1/T2**: Programmatic checks verify the required inputs, forecast and target alignment, spatial masks, and metric settings. Human review should complement these checks by examining whether the assigned question is compatible with the available forecast candidates and verification data. For example, a ranking task requires comparable candidates at the requested valid time, while a precipitation task requires the correct accumulation interval.

Report and service review. The release protocol requires one human review of every T3/T4 reference report. Reviewers check source support, scientific interpretation, required content, temporal consistency, and the responsibilities of the service type. An LLM critique assists reference preparation but does not replace this review. For T4 chains, reviewers also check that downstream tasks receive the context required for their service role.

A.4 T0: Question Taxonomy and Composition

Table 7: Data types, sources, and their uses in T0.

Data type	Source	Use in T0
Reanalysis fields	WeatherBench 2	Retrieve atmospheric variables, compute regional statistics, and derive physical diagnostics such as wind shear and relative vorticity.
Station observations	MeteoNet, GHCNh	Query local weather measurements, analyze temporal changes, and compare station observations with matched gridded estimates.
Radar and Satellite	MeteoNet, MRMS	Locate scans and interpret reflectivity metadata, compare strong-echo coverage, and quantify observed echo displacement over time.
Forecast products	MeteoNet ARPEGE	Identify forecast cycles and valid times, interpret accumulation intervals, and compare archived forecasts with observations at matching locations and times.
Event records	NOAA Storm Events Database	Retrieve documented event types, locations, and times, and verify claims against the corresponding event records.

A.4.1 SCOPE AND TASK STRUCTURE.

T0 evaluates evidence-grounded meteorological analysis, covering data grounding and retrieval, deterministic statistics, climatology and extremes, physical diagnostics, cross-source consistency, and atomic claim verification. The data sources used in T0 include satellites, radar, automatic weather stations, reanalysis data, forecast products, and weather reports, as in Table 7. It includes analysis of historical weather states and verification of event-related facts. Each task specifies a question, meteorological variables, a spatial domain, a valid time or temporal window, and an information cutoff (*as-of* time). Within these boundaries, agents choose their retrieval and analysis steps without following a prescribed tool sequence. Tasks span reanalysis fields, station measurements, satellite and radar observations, archived forecasts, climatological references, and documentary records. They are classified according to the capability required to answer the question, not the data source. T0 comprises 552 base task instances organized into six capability families (Table 8).

A.4.2 TASK FAMILIES AND QUESTION SUBTYPES

Inspired by the executable weather-science tasks in ZephyrusBench (Varambally et al., 2026), T0 combines verifiable meteorological questions with explicit tests of how agents identify, compare, and assess heterogeneous evidence. Alongside reanalysis-based retrieval, statistics, and anomaly analysis, it incorporates station observations, archived operational forecasts, and radar and satellite catalog and indexing tasks. We make source and time interpretation, physical diagnostics, cross-source comparability, and evidence sufficiency explicit evaluation targets, including cases where the available evidence does not permit a determination. The questions are organized into **six capability families** according to the ability being assessed.

- **F1: Grounding and Retrieval.** F1 tests whether an agent can identify the requested data and interpret their spatial, temporal, and physical meaning. Questions cover source identity and evidence role, point values, named regions, pressure levels and vertical profiles, valid times and accumulation windows, and units, coordinates, and source versions. Radar and satellite questions extend this coverage to product or channel identification, archive-object lookup, temporal and spatial indexing, and metadata queries. These observation-specific questions assess data access and interpretation, not weather-pattern recognition in imagery.
- **F2: Deterministic Statistics.** F2 tests whether an agent can compute a reproducible summary of a specified dataset. Questions ask for area-weighted averages, maxima and quantiles, where or when maxima occur, and where or for how long a threshold condition holds. Further subtypes examine temporal trends and changes, differences between regions, and statistics calculated only where a condition on a second variable is satisfied. Spatial weighting, time windows, missing-data treatment, and rules for tied values are specified so that each answer can be checked against an independent calculation.
- **F3: Climatology, Extremes, and Compounds.** F3 assesses weather conditions relative to a predefined seasonal climate reference or an absolute extreme threshold. Questions quantify departures from climatological quantile thresholds, express anomalies relative to the climatological standard deviation, and determine whether an extreme criterion is met. They also measure how long extreme conditions persist, what fraction of a region they cover, and where conditions on two variables hold simultaneously. Therefore, this family evaluates unusualness, persistence, and spatial extent rather than magnitude alone. Its thresholds are the criteria used for benchmark analysis, not official warning levels.
- **F4: Physical Diagnostics.** F4 tests whether an agent can derive or verify meteorological quantities using physical relationships. For example, spatial changes in wind quantify flow rotation or spreading, while humidity and wind together determine moisture transport and its convergence. Other subtypes examine wind differences between pressure levels, geostrophic or ageostrophic wind, temperature lapse rate or static stability, and pressure or geopotential gradients. Each question specifies the required diagnostic, input variables and levels, and output units, making the resulting physical assessment numerically verifiable.
- **F5: Cross-source Consistency and Representativeness.** F5 tests whether an agent can distinguish genuine disagreement from a comparison that the available data do not support. Questions compare stations with grids, different gridded products, and observations with reanalyses or forecasts at matching valid times. Additional subtypes examine whether differences in resolution or

Table 8: Question subtypes and instance counts across the six T0 capability families).

Capability families	Subtypes	Instances
F1: Grounding and Retrieval (108 instances)	Source, product, channel, and evidence-role identification	16
	Point lookup	16
	Named-region identification	12
	Vertical-level and profile retrieval	11
	Valid-time and accumulation interpretation	12
	Unit, coordinate, and version identification	14
	Source-object resolution	6
	Temporal-index lookup	8
	Spatial-index lookup	7
	Native data schema, shape, and missing-value interpretation	6
F2: Deterministic Statistics (101 instances)	Aggregation	15
	Extrema and quantiles	16
	Extrema location or occurrence time	17
	Threshold geometry and duration	14
	Trend and change	14
	Conditional and spatial comparisons	17
	Observed strong-radar-echo coverage comparison	4
	Observed radar-echo centroid displacement	4
F3: Climatology, Extremes, and Compounds (101 instances)	Departure from climatological quantiles	19
	Standardized anomalies	18
	Absolute or climatological extremes	16
	Persistence	19
	Spatial extent	16
	Two-variable compound conditions	13
F4: Physical Diagnostics (98 instances)	Relative vorticity or horizontal divergence	18
	Vertical wind shear	14
	Geostrophic or ageostrophic wind	16
	Moisture transport or moisture-flux convergence	16
	Lapse rate or static stability	19
	Pressure or geopotential gradients	15
F5: Cross-source Consistency and Representativeness (101 instances)	Station–grid comparison	14
	Grid–grid alignment	17
	Observation–reanalysis comparison	21
	Observation–forecast comparison at the same valid time	15
	Resolution and representativeness assessment	13
	Disagreement without source corruption	21
F6: Evidence Sufficiency and Atomic Claim Verification (43 instances)	Quantitative scientific claims	20
	Source and temporal metadata claims	16
	NOAA documentary cross-checks	7
Total		**552**

the spatial area represented by each value make the comparison unreliable, and whether observed differences fall within an expected range. The outcomes are `agreement`, `disagreement`, and `indeterminate`. A difference between sources is not, by itself, evidence that either source is faulty.

- **F6: Evidence Sufficiency and Atomic Claim Verification.** F6 tests whether an agent can judge a single checkable statement using the permitted evidence. The three subtypes concern quantitative weather claims, source and time metadata, and event facts documented in NOAA Storm Events records. Unlike F1, which asks the agent to retrieve information, F6 supplies a statement whose validity must be assessed. Answers distinguish `supported`, `refuted`, and `insufficient`, preserving the difference between evidence that contradicts a statement and evidence that cannot establish whether it is true.

A.4.3 T0 Construction Pipeline

T0 construction proceeds from predefined task specifications and archived meteorological data to independently verified questions and controlled evidence variants. Sampling and acceptance rules are fixed before final generation.

Fix the construction specification. An initial pilot checks the sampler, reference programs, question templates, and output schemas; pilot instances are excluded from the benchmark. Before final generation, we fix the source snapshots, task subtypes, region definitions, scientific conventions, sampling seeds, and allocation across task families, sources, and splits. Scientific conventions include units, accumulation windows, climatological references, and diagnostic definitions, ensuring that generated questions have unambiguous meanings.

Sample and screen candidate instances. Candidates combine a question subtype with a data source, location or region, time period, and required variables. Sampling covers two complementary groups. For calendar-based candidates, dates and regions are selected before inspecting weather values, retaining ordinary conditions and simple outcomes. Challenge candidates use predefined diagnostics to cover extremes, coverage boundaries, source disagreements, and more demanding analyses. Eligibility checks then verify source availability, variable identity, pressure levels, spatial and temporal coverage, and missing-data conditions. Documented archive faults are retained only in designated cases with appropriate answer requirements, rather than treated as valid measurements.

Establish independent reference answers. Reference programs read the source data through a code path independent of the agent's analysis tools. They compute numerical targets or verify metadata and documentary judgments, recording the relevant sources, calculation conventions, and required artifacts. Reference validation includes tests on arrays with known answers, checks of units and coordinates, and repeated execution to verify reproducibility. For answerable tasks, the reference package identifies at least one permitted evidence path that supports the answer; evidence-sufficiency tasks instead establish which judgment the available evidence supports. Candidates passing these checks are selected under the predefined allocations and duplicate and split constraints. The reference procedure establishes correctness without requiring agents to follow the same tool sequence.

Render and check the questions. Each selected specification is expressed as a natural-language question using a controlled template: a direct request, an operational context, or a comparison. The wording must preserve the specified variables, region, time period, calculation, and output requirements. For example, a question asking for a three-hour rainfall total must retain that interval in both its wording and reference calculation. Automatic consistency checks and human review assess this agreement. Wording alternatives do not count as additional base tasks.

Construct evidence-modified variants. Selected base tasks receive variants that retain the exact question text while changing a declared evidence condition, such as source availability, coverage, variable metadata, or data integrity. Unrelated task conditions remain unchanged, and the expected answer is reassessed under the modified evidence. Depending on the intervention, a justified response may retain the original answer, use alternative evidence, revise the conclusion, or report insufficient evidence. Naturally occurring archive problems and deliberately introduced modifications are recorded separately. Every variant retains its link to the base task and inherits its data split.

Finalize the dataset. The final audit checks agreement among questions, sources, reference answers, output requirements, and variant definitions. Technical and meteorological reviewers examine a stratified sample, with additional review of designated fault, historical-cutoff, and compound-intervention cases. Disagreements are adjudicated before acceptance. Related tasks remain grouped across splits, and hidden reference answers and intervention annotations are excluded from agent inputs. The construction record retains software and source versions, sampling seeds, selection and rejection reasons, and review decisions, allowing each instance to be traced back to its source data and generation procedure.

Table 9: Answer types in T0. Each task specifies its required fields and artifacts.

Answer type	**Typical content**
Numerical scalar	A regional mean, climatological anomaly, or physical diagnostic.
Structured record	A source identity, archive index, valid time, or requested metadata.
Profile or time series	Values associated with specified vertical levels or timestamps.
Location or spatial set	An extreme-value location or the set of locations satisfying a specified condition.
Categorical judgment	A source category, cross-source consistency judgment, or claim-verification outcome.

A.4.4 Illustrative Examples.

T0 answers range from factual and metadata records to numerical results and evidence-based judgments as shown in Table 9. To provide a more intuitive understanding of the question types, we present one example from each capability family as a case card. Each card identifies the question and assessed capability, summarizes the input data and their spatial and temporal scope, and explains the constraints and evidence underlying the reference answer. Questions are restated for readability rather than reproduced as complete machine-readable prompts. Their purpose is to make clear what the task asks, why the answer is supported, and which capability distinguishes it from a simpler lookup. The examples below illustrate these distinctions using actual task instances: satellite frame retrieval returns an index and timestamp; regional aggregation, climatological anomaly analysis, and physical diagnosis return scalars with units; cross-source comparison and claim verification return categorical judgments.

Example 1 (F1): Retrieving a satellite frame

Question and capability
For the MeteoNet IR108 satellite archive covering northwestern France in 2017, which stored frame is the latest one at or before 25 March 2017, 10:30 UTC? Return its zero-based index and exact valid time. *Capability:* Temporal grounding and structured retrieval. The task tests whether the agent locates the correct observation in an archive, not whether it interprets cloud imagery.

Data and scope
MeteoNet satellite observations, IR108 channel, NW domain, 2017 archive. The required evidence is the source identity and its native time-coordinate array; image-pixel analysis is not required.

Constraints and answer basis
Search within the specified archive and select the stored frame with the greatest valid time not exceeding the requested time. Do not interpolate, select a later frame, or switch to another archive. The reference lookup identifies 10:00 UTC at zero-based index 2000 as the latest eligible frame.

Reference answer: Status: `available`; frame index: `2000`; valid time: `2017-03-25T10:00:00Z`.

Example 2 (F2): Computing a regional mean

Question and capability
What is the cosine-latitude-weighted mean boundary-layer height over the specified Arctic region at 06:00 UTC on 25 February 2010? Answer in metres.
Capability: Deterministic spatial aggregation. This task involves combining grid cell values according to a specified weighting rule rather than retrieving a single value.

Data and scope
WeatherBench 2 ERA5 boundary-layer height for `macro-arctic`, at one valid time. The source has a 1.5° latitude–longitude grid and 6-hourly temporal sampling. The frozen regional mask selects 3,585 grid cells.

Constraints and answer basis
Apply the prescribed regional mask without changing its boundary or silently filling missing values. The reference calculation uses

$$\overline{h} = \frac{\sum_{i \in R} h_i \cos \phi_i}{\sum_{i \in R} \cos \phi_i},$$

where h_i is boundary-layer height, ϕ_i is latitude, and R is the selected set of grid cells. An unweighted grid-cell average is not the requested statistic.

Reference answer: $\boxed{363.3220\ \mathrm{m}}$ (numerical scalar).

Example 3 (F3): Measuring a climatological anomaly

Question and capability
For `tile-15deg-12` (105°–120°E, 15°–30°N) at 18:00 UTC on 13 March 2006, what is the regional 2 m temperature anomaly relative to the specified climatological 99th-percentile reference? Use the high-tail sign convention and answer in kelvin.
Capability: Reference-based anomaly calculation. Unlike a regional mean, this answer requires interpreting the weather state relative to a prescribed climatological reference.

Data and scope
WeatherBench 2 ERA5 2 m temperature and the frozen climatological reference for the task. The instantaneous temperature field is sampled on a 1.5° grid with 6-hourly temporal sampling; 110 grid cells fall within the selected region.

Constraints and answer basis
Use the specified 0.99 quantile and high-tail convention: temperature minus reference threshold. Preserve the regional mask and the prescribed reference calculation. The output is a temperature difference, not a percentile rank, a standardized anomaly, or an operational warning level.

Reference answer: $\boxed{-11.3216\ \mathrm{K}}$ (numerical scalar).
The negative sign indicates a regional diagnostic below the specified high-tail reference under the task's sign convention.

Example 4 (F4): Deriving regional relative vorticity

Question and capability
What is the regional mean relative vorticity at 925 hPa over `tile-15deg-03` (60°–45°W, 45°–30°S) at 18:00 UTC on 30 July 2008? Use the prescribed cosine-latitude weighting and answer in s^{-1}.
Capability: Physical diagnosis from primitive variables. Rather than retrieving either input wind component, the agent must derive a wind-field diagnostic.

Data and scope
WeatherBench 2 ERA5 zonal and meridional wind components (u and v), in $\mathrm{m\,s}^{-1}$, at 925 hPa. The source has a 1.5° grid and 6-hourly temporal sampling. The regional mask selects 110 grid cells for each component.

Constraints and answer basis
The reference calculation forms a relative-vorticity field from the native-grid wind components before applying the regional mask. It then computes

$$\overline{\zeta} = \frac{\sum_{i \in R} \zeta_i \cos \phi_i}{\sum_{i \in R} \cos \phi_i},$$

where ζ_i is the diagnosed relative vorticity and R contains the 110 selected cells. All selected diagnostic values are finite in the reference record.

Reference answer: $\boxed{-5.315641 \times 10^{-6}\ \mathrm{s}^{-1}}$ (numerical scalar).

Example 5 (F5): Comparing an observation and forecast

Question and capability
Do the MeteoNet station observation and archived ARPEGE forecast agree on 2 m temperature at `point-fixed-22` at 06:00 UTC on 12 July 2017, under the specified 5 K agreement tolerance?
Capability: Cross-source comparison with spatial and temporal support checks. This is a same-valid-time comparison, not a request to generate a future forecast.

Data and scope
One MeteoNet station observation and one collocated ARPEGE forecast-grid value, both expressed in kelvin. The station is at approximately 46.81°N, 1.60°W. MeteoNet stations have nominal 6-minute sampling; the archived ARPEGE product uses a 0.1° grid. The selected grid cell is approximately 3.56 km from the station.

Constraints and answer basis
Require exact valid-time matching, a collocation distance no greater than 8 km, at least one valid pair, and the prescribed support and agreement fractions of at least 0.8. The reference record contains one valid pair with full support. Its absolute temperature difference is approximately 1.1029 K, below the 5 K tolerance. Numerical proximity alone would not justify agreement if the support checks failed.

Reference answer: `agreement` (cross-source judgment).

Example 6 (F6): Verifying a quantitative claim

Question and capability
Does the specified MeteoNet observation support the claim that 2 m temperature at `point-fixed-24` at 06:00 UTC on 22 August 2016 was below 289.75 K?
Capability: Atomic claim verification. The task asks whether a particular proposition is supported, rather than simply requesting the temperature value.

Data and scope
One MeteoNet station observation of 2 m temperature at the specified station and exact valid time. The source is available within the task cutoff, and the comparison uses temperature in kelvin.

Constraints and answer basis
Use the specified observation, station, valid time, unit, and strict "less than" comparison. The reference record gives a temperature of 290.25 K; therefore, $290.25 < 289.75$ is false. Since the claim is contradicted by the eligible evidence, the appropriate judgment is refutation rather than insufficient evidence.

Reference answer: `refuted` (claim-verification judgment).

A.5 T1: Question Taxonomy and Composition

A.5.1 Scope and Task Structure

Evaluation scope. T1 evaluates observation-driven nowcasting over the next 0–3 h through forecast generation, targeted queries, source selection, and comparison of forecast products. Forecasts, derived queries, and source choices are verified against subsequent radar observations. Comparison tasks instead check whether the agent correctly extracts and compares quantities from the specified forecast evidence.

Task specification and temporal protocol. Each task is associated with a reference time t_0, a spatial domain, a radar observation history, and a specified output. The observation history contains 13 frames at 5-minute intervals from $t_0 - 60$,min to t_0. The principal forecast lead times are +60, +120, and +180 min. Under the benchmark's setting, t_0 is both the information cutoff and the nominal valid time of the final input frame. Actual observations acquired after this cutoff are excluded from agent inputs.

Radar data and spatial configuration. Radar data are drawn from MeteoNet in France and MRMS in the United States. Each case uses a local azimuthal equidistant grid with 1 km spacing, covering a 384×384,km input domain. Verification is restricted to the central 256×256,km domain, leaving surrounding observations available to capture weather systems moving toward the target region. Reflectivity is expressed in dBZ and accompanied by a validity mask; missing measurements remain distinct from valid non-detection.

Candidates for the nowcasting model. The candidate set comprises radar persistence, optical-flow extrapolation, and SimVP-based prediction. Persistence retains the latest observed field, optical-flow extrapolation projects estimated echo motion, and SimVP predicts evolution from the observation sequence. All candidates use the same radar history. Tasks specify the required result while allowing agents to choose their retrieval, forecasting, and analysis steps.

A.5.2 Task Families and Output Types

T1 contains four task Families, distinguished by the decision or product requested from the agent. Table 10 summarizes their Task families, assessed capabilities, and target allocation. The same weather case can support different families: producing a forecast field, answering a regional question, and selecting a forecast candidate assess different aspects of the nowcasting workflow.

- **Forecast generation.** This family asks the agent to produce a nowcast using eligible forecasting tools. Questions request reflectivity fields prediction. Field outputs retain their spatial coordinates, valid times, and missing-data information.
- **Targeted query.** This family asks the agent to extract a task-relevant quantity from the predictive information. Questions including threshold-exceedance values within a specified region or expected peak intensity.
- **Source selection.** The agent selects one candidate from the task's frozen set and binds its identifier to the corresponding forecast evidence. Selection may use the current observations, candidate forecasts, the agent's prior knowledge, and optional historical performance verified before t_0. Its meteorological quality is determined by the selected forecast's loss against the task's subsequent observations.
- **Forecast-evidence comparison.** This family assesses whether the agent correctly reads and compares specified forecast products. For example, the agent reports each forecast's predicted area fraction with reflectivity of at least 30 dBZ over the same region and valid time, their signed difference in %, and their ordering, including ties. The evaluator recomputes these quantities from the linked forecast fields. Each forecast's accuracy against subsequent observations is evaluated separately.

Table 10: T1 task families, assessed capabilities, and target allocation.

Task family	Capability assessed	Tasks
Forecast generation	Use supported forecasting tools to produce accurate radar nowcasts for the requested region and lead times.	100
Targeted query	Obtain forecasts that accurately predict a specified regional quantity under the prescribed spatial, temporal, and threshold definitions.	100
Source selection	Select the candidate expected to provide the most accurate forecast for the current task using information available at the decision time.	80
Forecast Evidence Comparison	Extract and compare task-defined quantities from forecast products, including their numerical differences and ordering.	60
Total		**340**

A.5.3 T1 Construction Pipeline

T1 construction pairs radar-based weather cases with questions that assess different nowcasting capabilities. Each case specifies a reference time, spatial domain, radar history, and candidate forecast definitions, and supports two tasks from different families. This track contains 170 cases and 340 tasks, equally divided between France and the United States (Table 11).

Establish candidate pools. For each country, we construct a fixed pool of 255 candidates, three times the target of 85 cases. The pool contains 102 event-associated candidates and 153 candidates sampled from the time–location inventory. Event candidates are linked to documented weather events; the remaining candidates broaden coverage beyond explicitly selected events and are not necessarily calm or precipitation-free. Candidate ordering follows predefined event-linking rules and sampling seeds, without using forecast errors or evaluated-system performance.

Assign questions and required data. Before final case selection, each candidate receives two tasks from different families. Each family has two semantic templates, instantiated with a region, forecast lead, threshold, event definition, and output requirement as applicable. These choices determine the observations and verification data needed for that candidate and must fit the declared capabilities of the forecast products. Each template has two or three controlled wording variants; alternative phrasings preserve the same task identity and do not increase the dataset count.

Verify data and select cases. Candidates are checked in the predefined order for the required radar history, valid timestamps, units, coordinates, spatial coverage, and missing-data masks, together with the required future verification support. These checks establish data eligibility without using forecast skill or future weather intensity to favor particular cases. Selection retains 85 cases per country: 34 event-associated cases and 51 cases from the time–location pool. Ineligible candidates are handled through the predefined selection procedure, with rejection reasons recorded. Cases are not replaced because an evaluated system performs poorly on them.

Construct forecast and verification packages. For each selected case, persistence, optical-flow extrapolation, and SimVP candidates share the same 13-frame radar history covering the preceding hour. Forecast outputs are aligned to leads of 60, 120, and 180 minutes and retain their member identities, weights, and validity masks. Threshold-exceedance fields are derived from the available members; a deterministic single-member forecast yields a binary exceedance field. Future radar observations provide the hidden verification targets. Verification regions, thresholds, spatial alignment, and scoring masks are fixed before evaluated-system outputs are inspected.

Group related cases and finalize splits. The public development subset contains 18 cases and 36 tasks; the hidden subset contains 152 cases and 304 tasks. Both tasks from a case remain in the same split. Cases sharing a weather event, forecast run, overlapping verification data, or a common base instance are also grouped together. These grouping constraints apply during case selection and split assignment, preventing related instances from being separated merely to satisfy allocation targets. The same groups are used for confidence-interval estimation to account for dependence among tasks.

Table 11: T1 composition by country and sampling stratum. Each base weather case supports two tasks from different families.

Country	Event cases	Natural cases	Total cases	Tasks
France	34	51	85	170
United States	34	51	85	170
Total	**68**	**102**	**170**	**340**

A.5.4 Illustrative Examples

The following templates illustrate the questions and required submissions for the four families. Each task specifies the region, information cutoff t_0, lead times relative to t_0, permitted tools, candidate set where applicable, query parameters, and output requirements.

- **Forecast generation.** *Using the permitted radar history and a registered nowcasting tool, obtain radar-reflectivity forecasts for the target domain at +60, +120, and +180 min. Submit the resulting forecast artifact and its evidence reference.* The tool may return a cached forecast matching the registered request or execute a supported configuration. Field quality is evaluated against subsequent radar observations.
- **Targeted query.** *For this U.S. reflectivity task, forecast the fraction of the specified region with reflectivity of at least 30 dBZ at +120 min. Submit the underlying forecast artifact, its evidence reference, and the prescribed region, lead time, threshold, and area-fraction operator.* The evaluator extracts forecast and observed fractions using physical area weights on the same observation-valid support. The requested quantity is an area fraction, not the probability of a regional event.
- **Source selection.** *Which candidate from the task's frozen set should be used for the +180 min reflectivity forecast over the target region, given the recent radar evolution? Submit the candidate identifier and its forecast evidence reference.* The selected forecast is evaluated using CRPS against subsequent radar observations for the specified region and lead time.
- **Forecast-evidence comparison.** *For forecasts A and B, what fraction of the specified region has predicted reflectivity of at least 30 dBZ at the target time? Report both fractions, their signed difference (A minus B) in %, and which fraction is larger, or whether they are equal. Include the evidence references for both forecasts.* The evaluator checks these values and relations by recomputing them from the specified forecast artifacts. Each forecast's quality against subsequent observations is reported separately.

A.6 T2: Question Taxonomy and Composition

A.6.1 Scope and Task Structure

Evaluation scope. T2 evaluates four task families from more than 3 h to 15 days ahead: forecast generation, multi-model selection and ranking, threshold-exceedance probability, and spatial extent and hazard localization. Whereas T1 focuses on near-term radar-echo evolution, T2 addresses gridded atmospheric forecasts across multiple variables, initialization cycles, and forecast horizons.

Task specification and temporal protocol. Each task specifies a decision time t_0, a target region, a variable, a forecast horizon, and the required output. The three horizon groups are very-short-range (T2-VS), short-range (T2-S), and medium-range (T2-M), as summarized in Table 12. The benchmark horizon is measured from the decision time. Each candidate separately retains its initialization time and model lead time, so forecasts from different cycles can be compared at the same target valid time.

Forecast variables and spatial configuration. Tasks cover 2 m temperature, 10 m wind components or derived wind speed, mean sea-level pressure, 500 hPa geopotential height, and precipitation accumulation. Each case provides a $550\,\mathrm{km} \times 550\,\mathrm{km}$ outer context domain and a central $450\,\mathrm{km} \times 450\,\mathrm{km}$ verification domain.

Candidates for the forecasting model. Candidate products include GFS, TIGGE ensemble forecasts, Pangu-Weather, and GraphCast, together with regional products such as AROME/ARPEGE in France and HRRR in the United States. Their eligibility depends on the archived product version, initialization cycle, variable, spatial coverage, and supported lead time. The candidate set can therefore differ across cases and horizons; a regional model is not assumed to provide every medium-range target.

Table 12: Forecast horizons and temporal support in T2.

Group	Horizon	Target offsets from t_0
T2-VS	$(3, 12]$ h	+6, +12 h
T2-S	$(12, 72]$ h	+24, +48, +72 h
T2-M	$(72, 360]$ h	+120, +168, +240, +360 h

A.6.2 Task Families and Output Types

T2 contains four task families with 96 instances each (Table 13), which assess complementary forecasting capabilities: generating forecast fields, selecting among competing forecast sources, estimating the probability of a defined meteorological event, and localizing its spatial extent.

- **Multi-horizon forecast generation.** This family evaluates whether the agent can use legally available gridded meteorological states and registered forecasting tools to produce future atmospheric fields over the requested region and forecast horizon. Unlike T1, where the predictive input is a recent radar sequence, T2 operates on operational or numerical-weather- prediction fields associated with a specified initialization or reference time. The task fixes the target variable, spatial domain, initialization time, forecast valid time, and required lead or accumulation window. The applicable forecasting systems depend on the requested horizon.
- **Multi-model selection and ranking.**
  The agent submits a complete ordering of the frozen candidate pool from expected best to worst.
- **Threshold-exceedance probability.** This family asks the agent to quantify the likelihood or regional coverage of a specified meteorological condition. Questions request a gridpoint probability field or a regional quantity derived from threshold exceedance. The task fixes the variable, comparison direction, threshold, region, and valid time or accumulation window. A regional exceedance fraction, its expected value across members, and the probability that this fraction exceeds a specified level are different quantities and must not be used interchangeably. Outputs retain the event definition and spatial support needed to interpret the reported value.
- **Spatial extent and localization.** This family evaluates whether the agent can identify where a specified hazardous condition is expected and represent its spatial extent. The task fixes the meteorological event definition, threshold, region, forecast horizon, and spatial support. The agent submits a spatial prediction product, such as a hazard-probability field or a task-defined representation of the predicted location and extent. The output must preserve the coordinates, event definition, and spatial support needed for verification. Spatial skill is evaluated against the subsequently observed hazard field at pre-specified physical scales.

A.6.3 T2 Construction Pipeline

T2 construction pairs weather cases with tasks spanning very-short-, short-, and medium-range forecasting. Each case fixes the decision time, spatial domain, required atmospheric context, and candidate source definitions. This track contains 144 cases and 384 task, equally divided between France and the United States (Table 14).

Establish candidate pools. For each country, we construct a fixed pool of 216 candidates, three times the target of 72 cases. The pool comprises 87 event-associated candidates and 129 candidates sampled from the time–location inventory. Event candidates are linked to documented weather

Table 13: T2 task families, assessed capabilities, and target allocation.

Task family	Capability assessed	Tasks
Multi-model forecast generation	Generate gridded weather forecasts from legally available initialization fields using forecasting systems compatible with the requested variable and forecast horizon.	96
Multi-model selection and ranking	Select and rank competing forecast sources for the specified weather target using information available at the decision time.	96
Threshold-exceedance probability	Quantify the probability of a precisely defined meteorological event over the requested spatial and temporal support.	96
Spatial extent and hazard localization	Predict where a defined hazardous condition will occur and represent its spatial extent and spatial structure.	96
Total		**384**

events, while time–location sampling provides broader coverage without requiring calm or hazard-free conditions. Candidate ordering follows predefined sampling and event-linking rules, without using evaluated-system errors or scores.

Assign tasks and forecast horizons. Before final case selection, each candidate receives two tasks from different non-generation families. Task specifications define the variable, region, forecast horizon, temporal support, and required output. For example, a precipitation question specifies both the forecast valid time and the accumulation interval. Each family has two semantic templates and two or three controlled wording variants per template; alternative phrasings do not create additional instances. After the 72 final cases in each country are fixed, 48 cases are deterministically assigned one additional forecast-generation task. They therefore do not introduce a new weather case, alter the root's existing task support, or trigger root resampling. Within each country, each of these three non-generation families contributes 20 very-short-range, 16 short-range, and 12 medium-range tasks.

Verify inputs and select cases. Eligibility checks establish the required atmospheric context and the exact verification support needed by both assigned tasks. The common input includes 2 m temperature, 10 m wind components, mean sea-level pressure, and 500,hPa geopotential height from the latest compatible GFS cycle within the preceding 24 hours. These fields support regional analysis and forecast selection; forecast execution additionally requires the selected model's complete, task-eligible initialization inputs. Other forecast candidates are checked against their archived variables, initialization times, and supported lead times. Unavailable optional products remain explicitly marked as unavailable; every model and variable need not be present in every case. Selection follows the predefined order and grouping constraints, retaining 29 event-associated and 43 time–location cases per country. Rejection reasons are recorded, and cases are not replaced because of an evaluated system's performance.

Construct variable-specific verification targets. Atmospheric state forecasts are verified against ERA5 from WeatherBench 2 on the common 0.25° grid. Precipitation targets use quality-controlled MeteoNet direct rainfall in France and MRMS gauge-corrected precipitation in the United States. MeteoNet rainfall is distinct from the radar reflectivity used in T1. Valid-time HRRR analyses may provide supplementary very-short-range state verification, with their shared forecast-system provenance identified. Supplementary native-grid results are reported separately from common-grid results. Forecasts and references are matched by variable, units, valid time, and spatial and temporal support. A precipitation target ending at time t with duration Δ represents $(t - \Delta, t]$. Where forecasts provide run-total accumulations, interval totals are obtained by differencing within the same initialization cycle before spatial remapping and thresholding. All required observation intervals must be present: missing periods are not treated as zero rainfall, and MRMS and Stage IV are not mixed hour by hour to complete a verification window. Thresholds, masks, and alignment rules are fixed before evaluated-system outputs are inspected.

Group related roots and finalize splits. The public development subsetcontains 14 cases and the hidden subset contains 130 cases. After adding forecast-generation views, these roots support 38 public and 346 hidden tasks, respectively. Equivalently, each country contains 7 public cases with

19 tasks and 65 hidden cases with 173 tasks. Cases linked by shared weather events, forecast runs, overlapping verification data, or cross-track ancestry are kept together. These constraints apply during selection and split assignment, instead of treating all tasks as independent weather cases.

Table 14: T2 composition by country and sampling stratum. Each country contains 72 roots and 192 TaskViews; 48 cases support three Tasks and the remaining 24 support two.

Country	Event cases	Natural cases	Total cases	Tasks
France	29	43	72	192
United States	29	43	72	192
Total	**58**	**86**	**144**	**384**

A.6.4 Illustrative Examples

The following schematic templates illustrate the four task families. Actual instances specify the region, threshold, candidate products, and output contract. The examples do not provide numerical reference answers.

- **Forecast generation.** *Using the permitted initialization data and forecasting tools, produce the requested atmospheric field over the target region at the specified valid time. Submit the forecast artifact with its variable, units, coordinates, and temporal support.* The submitted field is scored against the task's designated verification product.
- **Multi-model selection and ranking.** *Rank the eligible forecasts of 2 m temperature over the target region at $t_0 + 120$,h, accounting for their initialization times, coverage, and available predictive evidence.*

  The expected output is an ordered list of candidate identifiers. Candidates must represent the same variable and target valid time, even when their initialization cycles and model lead times differ. The reference ordering is derived from their subsequent verification losses.
- **Threshold-exceedance probability.** *For the 24,h period ending at $t_0 + 72$,h, estimate the probability that accumulated precipitation exceeds the specified threshold at each location in the target region.*

  The expected output is a georeferenced probability field with the accumulation window and threshold retained. A forecast of instantaneous precipitation or a probability for a different accumulation interval does not answer the question.
- **Spatial extent and hazard localization.** *At $t_0 + 12$,h, identify the location and extent of the region where predicted 10 m wind speed exceeds the specified threshold, using the task-defined spatial rule.*

  The expected output identifies the affected spatial set or meteorological object in geographic coordinates. Wind speed is derived consistently from the horizontal components, and the spatial result retains the valid time and threshold used to define it. A regional maximum alone does not describe the requested extent.

A.7 T3: Question Taxonomy and Composition

A.7.1 Scope and Required Output

T3 evaluates whether an agent can turn meteorological observations, forecast products, and computed diagnostics into a coherent weather report. Tasks specify the reporting region, information cutoff, report type, relevant time periods, and required content. The agent retrieves permitted information, performs any necessary analysis, and produces a complete English report. For example, a heavy-rainfall report may combine recent radar evolution, station rainfall totals, and forecast precipitation to describe the current situation and its expected development. The main output is the report itself. Data files, figures, and additional evidence references are required only when specified by the task. Reports need not follow identical headings or reproduce the reference wording, but they must satisfy the requested scientific scope and preserve relevant qualifications.

A.7.2 Report Types and Task Requirements

T3 defines five report types (Table 15), distinguished by their temporal focus and analytical objective.

- **T3-A: Current-weather analysis.** Reports characterize weather conditions within the specified analysis window using available observations and analysis products. They describe the location, intensity, and recent evolution of relevant weather features.
- **T3-N: Nowcasting report.** Reports describe expected conditions over the next 0–3 h by combining recent observations with short-term predictions. They assess the anticipated movement, intensification, or weakening of weather features, including arrival times where supported.
- **T3-F: Forecast report.** Reports summarize expected weather within one or more specified intervals: $(3, 12]$ h, $(12, 72]$ h, or $(3, 15]$ days. They synthesize relevant forecast products and diagnostics to describe the timing, location, and magnitude of anticipated conditions.
- **T3-X: Integrated report.** Reports integrate current analysis, nowcasting (0–3 h), and at least one later forecast interval into a coherent account of weather evolution. For example, a report may connect an observed rainband with its expected near-term movement and the larger-scale conditions that could sustain rainfall into the following day.
- **T3-R: Retrospective event analysis.** Reports reconstruct an event's chronology, spatial evolution, and meteorological characteristics using information available by a specified review cutoff. Where permitted, structured post-event records may supplement the reconstruction. Findings established after the event must not be presented as information available to forecasters during it.

Across all types, reports must distinguish observations from forecasts and preserve the time windows of reported quantities, avoiding double counting of accumulations. Their spatial and temporal detail and treatment of uncertainty must be supported by the available evidence.

Table 15: Report types and task allocation in T3. Non-hazard controls are included in the instance counts.

Type	Reporting objective	Temporal scope	Count	Controls
T3-A	Current-weather analysis	Current conditions	30	5
T3-N	Nowcasting report	Next $(0, 3]$ h	20	3
T3-F	Forecast report	At least one interval beyond 3 h, up to 15 days	40	5
T3-X	Integrated analysis and forecast	Current conditions, nowcasting, and a subsequent forecast interval	30	2
T3-R	Retrospective event analysis	Event history through a specified review cutoff	30	0
Total			**150**	**15**

A.7.3 Evidence Sources and Temporal Boundaries

Information available to the agent. Each task provides access to a defined collection of permitted sources instead of requiring a fixed sequence of tool calls. These sources may include radar and satellite observations, station measurements, atmospheric analyses, and forecast products. For example, radar can show the extent and movement of echoes, stations can establish local rainfall totals, and a forecast field can describe expected precipitation during the next day. Their different roles determine which statements they can support.

Information cutoff. Operational reports use a specified issue-time cutoff. An observation must describe a period ending no later than that cutoff and satisfy the source's availability rules. For example, a report issued at 12:00 UTC cannot use a station measurement collected at 13:00 UTC. It may use a forecast issued before 12:00 UTC that predicts conditions at 18:00 UTC, because a forecast's future valid time does not make it future observational evidence. Retrospective tasks

instead use a review cutoff. An event review prepared the following day may therefore use permitted observations collected after the original operational issue time, provided they were available by the review cutoff.

Reference reports and historical documents. Official reports used to construct a reference answer are separated from the sources accessible to the agent. Same-event reasoning reports, warnings, public products, selected retrospective narratives, and normalized reference reports remain hidden. This prevents the reporting task from becoming retrieval or paraphrasing of its target answer. Permitted raw observations from the same event remain available.

Reports from unrelated historical events may be provided as background . For example, a previous storm report may illustrate a reporting approach, but its rainfall totals cannot support statements about the current event. Retrospective tasks may access permitted structured post-event observations and impact records, while the selected reference narrative remains hidden. Reference construction is detailed in Appendix A.3.4.

A.7.4 Task Construction and Dataset Composition

Construction begins with a documented event or a non-hazard control window. The report type, region, cutoff, reporting windows, and content requirements are defined against the available source support. Each operational task must have a legal evidence path relevant to its primary weather family. Its cutoff is anchored to an actual product issue time, whereas T3-R uses a separate review cutoff. The reporting region cannot be derived from the event's future realized footprint.

A reference report is prepared from a closed set of selected source reports, task requirements, and frozen source-linked summaries. Normalization may translate, reorganize, and condense this material, but cannot add unsupported event facts. Mandatory report content is restricted to task-relevant findings supported by permitted evidence under the task's cutoff and access rules, at the required spatial and temporal detail and level of certainty. Official-report details beyond this support are excluded from mandatory requirements. Task-relevant gaps established by the permitted evidence are reflected in the required qualifications. Independent LLM critique and human sign-off are required before formal release. Report assessment follows Appendix A.9.

Report and horizon allocation. The construction protocol specifies 150 T3 instances across France and the United States. The 40 T3-F tasks are allocated by their longest evaluated horizon: 12 very-short-range, 18 short-range, and 10 medium-range reports. All 30 T3-X tasks include current analysis and nowcasting; their maximum integrated coverage extends through 12 h, 72 h, and 15 days for 10, 14, and 6 instances, respectively. A report covering several periods is counted once. T3-N and T3-X are restricted to France 2018 and United States 2019 cases because the first archive year is excluded from scored future-nowcast use. Other temporal and regional coverage depends on the qualified source inventory.

Weather coverage and non-hazard controls. Sampling covers five core weather families shared with T4: heavy precipitation, convective hazards, non-convective wind, temperature extremes, and selected frozen-precipitation and snow events. The allocation requires at least ten instances per core family, with no primary family exceeding 30% of hazard instances. It also covers all four meteorological seasons, with no season exceeding 40% of instances. The broader event taxonomy appears in Table 6. The allocation includes 15 non-hazard controls, distributed across report types as shown in Table 15. These include at least five routine quiet-weather cases and at least five subthreshold near-miss cases. They test whether the agent can describe ordinary or subthreshold conditions without exaggerating risk. Insufficient evidence, however, does not establish that hazardous conditions are absent.

Group-preserving splits. The protocol allocates 50 instances to training/development, 25 to validation, and 75 to hidden testing. All tasks within the same statistical group, including issue-time variants and linked T4 tasks, are assigned to the same split. Multiple reports from one event count as separate tasks, not independent weather cases. The split is event-group-disjoint.

A.7.5 Illustrative Examples

The case below illustrates a T3-A task. It presents an excerpt of the actual question, a summary of the permitted sources, and a short answer illustration linked to supporting observations. Evidence

labels E1–E4 show how its claims are supported. The case highlights the distinction between current observations, earlier context, and forecast background, while showing how missing measurements limit what a report can conclude.

T3-A: Current wind analysis—6 March 2017

Task. `instance-brittany-current-a-v1`. As-of: 6 March 2017, 12:05 UTC. Current-analysis window: [11:05,12:05) UTC. Spatial scope: native observation points within the specified Brittany retrieval rectangle, not complete administrative coverage.

Question (actual excerpt). *"Write an English standalone current-weather analysis of ordinary wind as of 2017-03-06T12:05:00Z. Assess current conditions only within [2017-03-06T11:05:00Z,2017-03-06T12:05:00Z) at native points inside the supplied retrieval rectangle. Describe spatial contrasts, recent evolution and uncertainty. . . . "*

Permitted sources. MeteoNet NW station observations in [06:05,12:05) UTC, with earlier observations used as background. AROME and ARPEGE native 10-m wind fields from the daily 00 UTC cycle, valid at 06–12 UTC, provide forecast background only. Station `ff` and model `ws` represent ordinary wind, not gusts. Same-event reference reports remain hidden.

Example answer (evidence-based reconstruction, not the frozen reference). Within the current window, sampled ordinary-wind speeds ranged from 8.200 to 13.000 $\mathrm{m\,s^{-1}}$ at station 22092001 and from 6.100 to 8.800 $\mathrm{m\,s^{-1}}$ at station 22005003 [E1–E2]. At station 22092001, the 12:00 UTC value of 8.200 $\mathrm{m\,s^{-1}}$ was lower than the 06:06 UTC value of 12.100 $\mathrm{m\,s^{-1}}$ [E3], but this comparison does not establish monotonic or region-wide weakening. Station 22147006 has ten current records with missing wind-speed values, so its wind conditions remain unknown [E4]. These observations describe sampled station conditions, not gusts or continuous extrema. Forecast fields cannot replace missing observations or establish an exact 12:05 UTC observation. Physical quality control and spatial representativeness remain unverified.

Supporting evidence. All station records below originate from `NW_Ground_Stations_2017.csv`. Times are UTC on 6 March 2017.

E1: Current sampled range. Station 22092001: minimum 8.200 $\mathrm{m\,s^{-1}}$ at 12:00 (CSV record 3886554); maximum 13.000 $\mathrm{m\,s^{-1}}$ at 11:42 (record 3885810).

E2: Contrast with another station. Station 22005003: minimum 6.100 $\mathrm{m\,s^{-1}}$ at 11:36 (record 3885559); maximum 8.800 $\mathrm{m\,s^{-1}}$ at 11:24 (record 3885061).

E3: Earlier observation. Station 22092001 reported 12.100 $\mathrm{m\,s^{-1}}$ at 06:06 (record 3871777), outside the current-analysis window.

E4: Missing values. Station 22147006 has ten current rows at six-minute labels from 11:06 through 12:00, all with `ff=null`. The records exist, but the wind-speed values are unknown.

A.8 T4: Question Taxonomy and Composition

A.8.1 Scope and Common Impact Analysis

T4 evaluates whether an agent can translate meteorological assessments into impact information and task-appropriate weather services. Each task specifies a region, an information cutoff, relevant forecast periods, a service type, and the operational context needed for the requested output. For example, a forecast of strong winds describes a weather hazard; relating those winds to exposed transport routes and communicating the relevant time window to a transport operator constitutes a service task.

All T4 outputs contain a common impact analysis, referred to as *ImpactBasis*. It describes the hazard and its evolution, the relevant area and time window, severity or probability, uncertainty, potentially affected objects, and the implications of any task-specified thresholds. Here, affected objects may include populations, roads, water-management infrastructure, or electricity networks. Exposure identifies which objects lie within the potentially affected area, while sensitivity describes task-provided conditions that may make them susceptible to the hazard.

Impact-assessment tasks deliver this common analysis as their main output. The other service types add a warning recommendation, partner-support brief, or public communication. T4 therefore extends T3's meteorological reporting through explicit service objectives and responsibilities.

A.8.2 Service Types and Task Requirements

T4 contains four service types (Table 16). Their differences concern the requested decision, available operational information, and permitted responsibility.

- **T4-I: Impact assessment.** These tasks ask which objects may be affected by the weather, during which periods, and through what plausible mechanisms. The agent combines meteorological evidence with the supplied exposure and sensitivity information. For example, forecast heavy rainfall and task-provided information about drainage-sensitive locations may support a qualified assessment of possible local disruption. The output retains the assumptions and uncertainties behind the potential impacts and does not present them as realized losses.
- **T4-W: Warning support.** These tasks ask the agent to recommend an operation within a specified warning context. The current warning state describes the status already in effect for the relevant area and product. Together with thresholds, jurisdiction, and the permitted operations, it constrains the recommendation. For example, a task may ask whether the available evidence supports maintaining or adjusting a given warning state. The output is a simulated recommendation with an evidential basis, not a claim that an official warning has been issued or changed.
- **T4-S: Partner support.** These tasks request meteorological support for one recipient: emergency management, transportation, water or flood management, or energy and utilities. The brief identifies weather conditions relevant to that recipient's area, assets, and task-provided operational triggers. For example, a transport brief may highlight when forecast winds approach a supplied route-specific threshold. It provides information and reminders within the permitted scope rather than issuing a mandatory closure or directing the recipient's final action.
- **T4-P: Public communication.** These tasks ask the agent to communicate supplied hazard facts and a given warning state to a specified audience through a specified channel. The message preserves the affected area, timing, uncertainty, and permitted guidance while adapting its wording and detail. For example, a short public message may explain when strong winds are expected and where to obtain official updates. Unlike T4-W, this task does not ask the agent to reconsider the warning state or introduce new advice beyond that allowed by the task.

Each service output must contain the information required for its own use. A public message that omits the affected area is not made complete merely because that area appears in the accompanying impact analysis. Scientific statements across the complete output must remain consistent. Evaluation criteria are detailed in Appendix A.9.

Table 16: Service types and principal outputs in T4. All outputs retain a common impact analysis.

Type	Service objective	Additional task information	Principal output
T4-I	Impact assessment	Exposed objects, qualitative sensitivity, and any applicable thresholds.	Common impact analysis.
T4-W	Warning support	Current warning state, jurisdiction, thresholds, and permitted operations.	Warning-operation recommendation.
T4-S	Partner support	Recipient, service area or assets, operational triggers, and permitted advisory scope.	Recipient-specific support brief.
T4-P	Public service	Given warning state, approved hazard facts, audience, channel, and permitted guidance.	Public-facing weather message.

A.8.3 Evidence and Operational Context

Meteorological information. T4 uses permitted observations, analyses, and forecast products under the information-cutoff rules described for T3. Radar sequences may establish recent storm evolution, station measurements may describe local conditions, and eligible forecasts may characterize subsequent weather. These sources remain necessary even when the task provides detailed information about exposed objects. Knowing that a region contains roads or electricity assets does not establish that hazardous weather will affect them.

Service-specific information. Tasks supplement meteorological sources with public or project-authored context appropriate to the service. This may describe exposed object classes, qualitative sensitivity, the current warning state, or a recipient's operational trigger. For example, a supplied trigger can define when a department requires a weather update; it does not imply that the department has already taken action.

The information differs by service type. T4-I provides the context needed to assess potential impacts. T4-W adds warning rules and permitted operations. T4-S specifies one recipient and its support requirements. T4-P supplies the facts, status, and guidance approved for the requested communication. Public communication must preserve these supplied facts rather than treating them as an invitation to revise the underlying warning decision.

Hidden information. Same-event reference reports, subsequent observations, realized losses, final warning operations, and partner actions are excluded from the agent's evidence. These outcomes cannot be used to justify an operational recommendation retrospectively. Legitimate task inputs, such as the warning state already in effect at the cutoff, remain available. Historical examples may provide permitted background, but cannot establish facts about the current event.

A.8.4 Standalone Tasks and Service Chains

T4 includes standalone tasks and linked service chains. In a standalone task, the agent develops the required impact analysis and service output from its own permitted inputs. A chain makes the information handoff explicit: an upstream T4-I task produces an impact analysis that is passed to one or more downstream service tasks.

The allocation contains four chains:

$$\mathrm{I} \rightarrow \mathrm{W},\ \mathrm{I} \quad \rightarrow \mathrm{S},\ \mathrm{I} \rightarrow \mathrm{P},\ \mathrm{I} \quad \rightarrow \mathrm{W}, \mathrm{S}, \mathrm{P}, \tag{5}$$

where I, W, S, and P denote the four T4 service types. All nodes within a chain share the same information cutoff. The arrows represent transfer of an assessment between tasks, rather than the arrival of new weather observations.

Each downstream task receives the actual agent-generated upstream output, its status, and its own task requirements and permitted sources. It does not receive a replacement reference impact analysis. If the upstream output is missing or invalid, that condition is passed onward rather than repaired with hidden target content. In the branching chain, W, S, and P execute independently without access to one another's outputs.

This arrangement evaluates whether service products remain supported and appropriately qualified when they depend on an earlier agent assessment. It also preserves the consequences of upstream omissions or errors, which would be concealed if every downstream task received an ideal reference analysis.

A.8.5 Task Construction and Dataset Composition

Task construction combines a weather event and reporting time with a service type, region, forecast scope, and the operational information needed for that service. The requested responsibility must be supported by both the meteorological sources and the supplied service context. For example, a warning-operation task requires a defined warning state and applicable rules; weather data alone do not specify which operation the agent is authorized to recommend.

Reference outputs are prepared within the source boundaries described in Appendix A.3.4. They preserve meteorological qualifications and the responsibilities of the service type. The release protocol requires checks of source support, time consistency, potential-impact language, and compliance with the assigned role.

Standalone and chain allocation. T4 contains 150 task instances: 140 standalone tasks, with 35 per service type, and ten nodes from the four chains in Eq. equation 5. Table 17 gives the resulting totals. Chain nodes are included in the 150 instances; the chains themselves are not counted as additional tasks.

Weather and temporal coverage. T4 covers hazard-event cases across France and the United States, using the weather categories in Table 6. Unlike T3, it does not include non-hazard controls.

Within each service type, at least five instances include an evaluated 0–3,h component, at least five include short- or medium-range content beyond 12,h, and at least three connect current conditions with future evolution. These coverage requirements may overlap within an instance. Tasks with evaluated nowcasting use France 2018 and United States 2019 cases, following the radar allocation.

The design targets coverage across all four meteorological seasons and the five core hazard families. Country balance is pursued subject to source quality and task validity, rather than assumed to be exactly equal.

Evidence limitations and grouping. At most five instances may be deliberately designed to require a qualified or indeterminate assessment because the permitted evidence cannot support full certainty. Such limitations are established before agent execution. For example, a task may support describing a weather hazard while lacking the object-specific information needed to assess a particular consequence. This is distinct from an agent failing to retrieve available information in an otherwise answerable task.

The split allocation is 50 development, 25 validation, and 75 hidden-evaluation instances. All nodes of a chain and other tasks linked to the same event remain in the same split. Grouping also preserves event reuse across T3 and T4. These relationships are retained for uncertainty estimation; 150 task instances do not imply 150 independent weather events.

Table 17: T4 task allocation. The four chains contribute ten task nodes already included in the total.

Service type	Standalone	Chain nodes	Total
T4-I	35	4	39
T4-W	35	2	37
T4-S	35	2	37
T4-P	35	2	37
Total	**140**	**10**	**150**

A.8.6 Illustrative Examples

The following cases illustrate the four types of T4 tasks. The Brittany wind cases are separate standalone I and S tasks; the June convection cases are W and P nodes of the same registered chain. Each case pairs a question excerpt with task-specific information and output requirements. The S case additionally includes a retained reference excerpt and supporting observations.

T4-I: Potential wind impacts—6 March 2017

Task and scope. `instance-brittany-current-i-v1`. As-of: 6 March 2017, 12:05 UTC. Current window: [11:05,12:05) UTC. Standalone analysis of native points in the Brittany retrieval rectangle, not complete administrative coverage.

Service request (registered excerpt). *"Write an English standalone potential-impact interpretation of ordinary wind as of 2017-03-06T12:05:00Z. … Discuss conditional discomfort or control difficulty, not structural failure, observed road conditions, accidents, closures or instructions to act. …"*

Permitted sources and exposure context. MeteoNet NW station observations over [06:05,12:05) UTC and AROME/ARPEGE 10-m wind forecasts initialized at 00 UTC, valid at 06–12 UTC. Earlier observations and all model fields provide background rather than additional current observations. Project-authored exposure assumptions concern outdoor activity participants and road users; their actual exposure, road conditions and responses are unknown. No numerical threshold is supplied for this I task.

Required output. Complete the six `ImpactBasis` fields, connecting current ordinary-wind conditions and recent sampled evolution to conditional discomfort or control difficulty for the assumed objects. Preserve point-level support, missing values and unverified observation quality. Do not infer gusts, realized damage, official severity or actions already taken. No prior Agent report is supplied.

T4-W: Internal warning coordination—9 June 2018

Task and scope. `instance-june-warning-fanout-v1`. As-of: 9 June 2018, 15:41 UTC. Current window: [14:41,15:41) UTC. The task concerns Ille-et-Vilaine, Mayenne and Sarthe and is a downstream W node in the registered June chain.

Service request (registered excerpt). *"Write an English current-state internal warning coordination recommendation for Ille-et-Vilaine, Mayenne and Sarthe as of 9 June 2018 15:41 UTC. . . . "*

Registered inputs and review rule. The task calls for the actual I report supplied for that execution, the node's own legal weather queries, the supplied authority state, and an explicitly synthetic internal review rule involving 40 dBZ and two adjacent five-minute scans. The rule is applied separately to each department within the current window. The upstream report is fallible data, not an instruction or a substitute for checking consequential claims.

Required output. Provide a self-contained Warning Operation Card and all six `ImpactBasis` fields. Identify the supporting scan times and native-centre evidence, or distinguish an excluded criterion from unresolved missing support. Explain the rule interpretation and any recommendation using only the permitted operations: `monitor`, `convene`, and `maintain`. Do not infer surface rainfall, persistence of the same storm cell, official warning severity or future weather from the review rule. The recommendation is not a real issuance or executed action.

T4-S: Transport-support briefing—6 March 2017

Task and scope. `instance-service2017-march6-s-v1`. As-of: 6 March 2017, 12:05 UTC. Current window: [11:05,12:05) UTC. Standalone support for a simulated transportation coordination desk, restricted to native station points in the Brittany retrieval rectangle. No actual route or asset inventory is supplied.

Service request (registered excerpt). *"Write an English standalone service briefing for the simulated transportation partner as of 2017-03-06T12:05:00Z. . . . Apply the exact synthetic station screening rule across the tool-advertised background station universe. Report the three sample-predicate categories, their spatial contrasts, missing support and representative source-linked values; summarize station groups rather than asserting region-wide conditions. . . . "*

Permitted sources and service rule. MeteoNet NW station observations over [06:05,12:05) UTC; AROME/ARPEGE daily 00 UTC forecasts valid at 06–12 UTC as background only; and project-authored exposure, service-area, screening and reminder resources. The synthetic rule compares current ordinary-wind `ff` with 10 $\mathrm{m\,s^{-1}}$. Any finite current value reaching the threshold establishes `observed_condition_met`, even when other samples are missing. A below-checklist result requires finite values below the threshold at all ten prescribed six-minute timestamps from 11:06 through 12:00, with no other current exceedance; otherwise, support is insufficient. This is not an actual partner SOP or a safety threshold.

Required output. Report the station-group screening results, spatial contrasts, recent sampled evolution and uncertainty in a complete `ImpactBasis` and a self-contained S briefing. Only ask the partner to evaluate its own procedures, check local conditions and maintain weather awareness; do not order closures or claim that a final action has been taken.

Reference excerpt. *"For the assumed outdoor participants and road users, wind could cause discomfort or control difficulty conditional on actual exposure; this screening does not establish accidents, closures or damage. Recipient: simulated transportation desk; service scope: only native station points in the supplied rectangle. Ask the partner to evaluate its own procedures, check local conditions and maintain weather awareness while addressing information gaps. . . . "*

Screening results and supporting evidence. The retained screening result classifies 49 stations: 33 as `observed_condition_met`, 9 as `prescribed_sample_set_below`, and 7 as `insufficient_samples`. Representative records from `NW_Ground_Stations_2017.csv` illustrate these outcomes:

E1: Observed threshold attainment. Station 22092001 reported 13.000 $\mathrm{m\,s^{-1}}$ at 11:42 UTC (CSV record 3885810). Its later 12:00 value of 8.200 $\mathrm{m\,s^{-1}}$ (record 3886554) does not undo the earlier current-window exceedance.

E2: Below the prescribed checklist. Station 22005003 has all ten prescribed current values finite, with a sampled range of 6.100–8.800 $\mathrm{m\,s^{-1}}$. Its maximum occurs at 11:24 UTC (record 3885061). This establishes the sample-based category, not continuous safety.

E3: Insufficient support. Station 22147006 has ten current rows, all with `ff=null`. Present records with missing values cannot establish either wind speed or a below-threshold classification.

The impact relation in the reference is supported by the task's explicitly synthetic exposure and sensitivity assumptions, not by observed accidents or losses. Its preparation reminders are restricted by the supplied reminder-scope resource.

T4-P: Public weather notice—9 June 2018

Task and scope. `instance-june-public-fanout-v1`. As-of: 9 June 2018, 15:41 UTC. Audience: the public in Ille-et-Vilaine, Mayenne and Sarthe. Channel: a historical public web notice. This P task is a sibling of the W task in the registered June chain, not a subsequent stage that receives the W output.

Service request (registered excerpt). *"Write an English historical public web notice for Ille-et-Vilaine, Mayenne and Sarthe as of 9 June 2018 15:41 UTC. … Communicate only supplied authority state, attributed approved hazard facts and approved safety guidance. … "*

Registered inputs and communication boundary. The actual I report supplied for the execution, the node's own legal weather queries, and the task-supplied authority state, approved hazard facts and approved guidance. No warning colour or expiry is supplied; neither may be inferred from the headings of safety-guidance material.

Required output. Produce all six `ImpactBasis` fields and a self-contained public notice naming the area, information cutoff, supplied state, attributed facts and approved guidance. Verify consequential weather interpretations rather than treating the upstream I report as authoritative. Do not reassess the warning level, add unapproved advice, infer realized damage or forecast future evolution. The notice is a historical communication task, not a live warning issuance.

A.9 Details of Evaluation Protocol

The five tracks produce different output types, from evidence-grounded answers and forecasts to meteorological and service reports. We therefore use task-specific quality metrics, complemented by common execution metrics and evidence diagnostics (Table 18). T0 evaluates answer correctness; T1/T2 evaluate forecast products, source choices, and evidence-based comparisons; and T3/T4 evaluate factual consistency and task-adequate reporting.

Table 18: Evaluation across tracks. Task-specific quality metrics are complemented by common execution metrics and diagnostic results.

Track	Evaluation object	Task-specific quality	Execution and diagnostics
T0	Evidence-grounded answers	Accuracy defined by instance-level Hit Scores	Completion, TCV
T1	Radar forecast fields, derived forecast queries, source selection, and forecast evidenc comparison.	Field CRPS, Query CRPS, Selected CRPS, Selection Regret, Best-hit Rate	Completion, TCV
T2	Multi-horizon forecast fields, source selection, event probability, spatial localization	Field CRPS, Selection Regret, Best-hit Rate, FSS, Brier score	Completion, TCV
T3	Meteorological reports	Factual-consistency and report-quality	Completion, TCV, groundedness, BLEU-1
T4	Weather-service and decision-support reports	Factual-consistency and report-quality	Completion, TCV, groundedness, BLEU-1

A.9.1 Common Execution Metrics

Across all tracks, we report Completion Rate and Tool-call Validity (TCV) to assess whether agents produce valid final submissions and invoke the available tools correctly.

Completion Rate. A run is completed when a final output passes the task's submission checks before the declared termination condition. These checks cover the required output structure, nonempty mandatory content, artifact references, and source eligibility.

Let $c_i \in \{\text{pass}, \text{fail}, \text{unresolved}\}$ denote the completion decision. It is *pass* when an accepted submission is recorded, and *fail* when the complete execution record establishes termination without one. Incomplete records that prevent this determination yield *unresolved*. Submission repairs do not create additional task executions, and all attempts remain in the execution record.

$$\text{Completion} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[c_i = \text{pass}]\,. \tag{6}$$

Started runs without valid outputs remain in the denominator.

Tool-call Validity. For run i, let A_i be the number of agent-issued tool attempts and V_i the number satisfying the declared tool identity, argument schema, invocation preconditions, and access constraints. For the set $\mathcal{I}_{\text{call}}$ of runs with at least one attempt and complete attempt logs,

$$v_i = \frac{V_i}{A_i}, \qquad \text{TCV} = \frac{100}{|\mathcal{I}_{\text{call}}|} \sum_{i \in \mathcal{I}_{\text{call}}} v_i. \tag{7}$$

All agent-issued attempts count, including malformed requests, unknown tools, rejected arguments, pagination, and retries. Backend-internal operations do not count as additional attempts. A valid request may return no data or encounter a service failure.

Trustworthy zero-call runs are marked *NA*; incomplete or success-only attempt logs are unresolved. If potentially applicable runs have unresolved logging, the available-run mean is provisional. An empty applicable set gives *NA*.

A.9.2 T0: Evidence-Grounded Analysis Evaluation

T0 requires task-specific answer matching because its outputs include numerical, categorical, temporal, spatial, set-valued, and evidence-sufficiency answers. We define correctness using an instance-level Hit Score and report its mean as Accuracy.

Hit Score and Accuracy. For instance i, let $\mathcal{K}_i$ denote the set of mandatory evaluation requirements fixed by its task and answer contract. These requirements specify the expected answer type and, where applicable, the target variable, spatial and temporal support, units, numerical tolerance, required structured fields, evidence conditions, and artifact requirements. The instance-level Hit Score is

$$\text{HS}_i = \mathbf{1}\left[\bigwedge_{k \in \mathcal{K}_i} k \text{ is satisfied}\right], \tag{8}$$

where $\text{HS}_i = 1$ only when all requirements are satisfied. Thus, a numerically plausible answer does not receive a full hit if it refers to the wrong variable, region, time support, or unit, or if a task-required evidence or artifact condition is not met.

For an evaluation set $\mathcal{I}$, the primary T0 metric is

$$\text{Acc}(\mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \text{HS}_i. \tag{9}$$

The evaluation set includes started executions with established missing or invalid final outputs, which receive a zero Hit Score. Evaluator defects are handled separately rather than converted into answer errors. Accuracy is computed as a fraction and displayed as a percentage.

Numerical answer matching. For numerical tasks, evaluation first verifies that the submitted quantity refers to the requested variable, region, time period, vertical level, and statistic. After any permitted unit conversion, a prediction $\hat{y}_i$ matches the reference value y_i when

$$|\hat{y}_i - y_i| \le \max\left(\epsilon_i^{\text{abs}}, \epsilon_i^{\text{rel}}\,|y_i|\right), \tag{10}$$

where ϵ_i^{abs} and ϵ_i^{rel} are absolute and relative tolerances fixed before evaluation. These tolerances account for permitted numerical variation. The task-evaluation specification fixes the numerical tolerances, permitted unit conversions, categorical equivalences, and spatial matching rules before evaluation. These rules allow valid alternative computations without requiring a reference tool-use sequence.

Categorical and structured answer matching. Categorical outputs are evaluated against the task's predefined semantic alternatives rather than by literal string identity. Equivalent label forms may therefore be accepted when they preserve the same meaning. Structured outputs are evaluated field by field according to their task contract. For example, a vertical profile must bind values to the correct pressure levels, while a location answer must satisfy the specified identity or spatial tolerance.

Evidence-sufficiency judgments. Evidence-sufficiency tasks distinguish *Supported*, *Refuted*, and *Insufficient*. An *Insufficient* answer can receive $\text{HS}_i = 1$ when the task permits it and the legally available evidence is genuinely inadequate. Failure to retrieve available evidence does not establish insufficiency. Evidence-modified instances are evaluated against their own evidence conditions rather than automatically inheriting the answer of the original instance.

A.9.3 T1/T2: Forecast Evaluation

Forecast outputs and source choices are evaluated against the task's designated verification products, including observations and analysis or reanalysis fields as applicable. Evidence-comparison tasks instead evaluate relationships recomputed from the submitted forecast artifacts. Each task family uses the metric appropriate to its requested output. Forecasts and verification products are aligned by variable, units, valid time, accumulation window, and spatial support.

Forecast-generation quality. Forecast-generation tasks assess the quality of the submitted predictive field against the corresponding future verification field. We use the Continuous Ranked Probability Score (CRPS), which evaluates both deterministic and ensemble forecasts under a common formulation. For grid cell i, let $x_{i1}, \ldots, x_{iM}$ denote the forecast members with non-negative weights $w_1, \ldots, w_M$ satisfying $\sum_m w_m = 1$, and let y_i denote the verifying value. The gridpoint CRPS:

$$\text{CRPS}_i = \sum_{m=1}^{M} w_m |x_{im} - y_i| - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} w_m w_n |x_{im} - x_{in}|. \tag{11}$$

Let $\mathcal{V}$ denote the common verification support and a_i the physical area of grid cell i. The area-weighted score, reported as *Field CRPS*, is

$$L_{\text{CRPS}} = \frac{\sum_{i \in \mathcal{V}} a_i \, \text{CRPS}_i}{\sum_{i \in \mathcal{V}} a_i}. \tag{12}$$

Lower values indicate better forecasts. For a deterministic single-member forecast, Eq. 11 reduces to absolute error, so Eq. 12 becomes an area-weighted mean absolute error. T1 applies this evaluation to future radar reflectivity, whereas T2 applies it to the requested gridded atmospheric variable and forecast horizon. For T2, the Field CRPS retains the units of the verified variable. The results are reported separately for each variable rather than being averaged across different physical quantities.

Query quality. T1 targeted-query tasks evaluate a prescribed quantity derived from a committed forecast. The agent submits the forecast artifact and the task-fixed query specification. The evaluator applies the same operator to forecast members and the verifying field on $\mathcal{V}$; a freely submitted scalar does not replace the forecast artifact. For example, for a regional exceedance-area query with threshold τ, the fraction for forecast member m:

$$q^{(m)} = \frac{\sum_{i \in \mathcal{V}} a_i \mathbf{1}[x_{im} \geq \tau]}{\sum_{i \in \mathcal{V}} a_i}, \qquad q^{\text{obs}} = \frac{\sum_{i \in \mathcal{V}} a_i \mathbf{1}[y_i \geq \tau]}{\sum_{i \in \mathcal{V}} a_i}. \tag{13}$$

where, $\mathcal{V}$ denotes the verification area and a_i the physical area weight of grid cell i.

The member-wise query values $\{q^{(m)}\}_{m=1}^{M}$, with the forecast-member weights w_m, define the predictive distribution of the requested quantity. Its CRPS against q^{obs} is reported as *Query CRPS*. For a single-member forecast, this score reduces to:

$$L_{\mathrm{query}} = \left| q - q^{\mathrm{obs}} \right|. \tag{14}$$

For a single member, $q = q^{(1)}$. This loss measures the accuracy of the requested regional quantity rather than gridpoint accuracy. An exceedance-area fraction is not an event probability. The region, threshold, comparison direction, valid time, and operator are fixed by the task.

Source selection and ranking. Let $\mathcal{B}_i$ denote the frozen candidate pool for task i, and let L_{ic} be the area-weighted CRPS of candidate c against the common verification target and support. Denote by $\hat{c}_i$ the selected candidate for T1 or the first-ranked candidate for T2. The selected forecast's loss and the minimum candidate loss are

$$L_i^{\mathrm{sel}} = L_{i\hat{c}_i}, \qquad L_i^{\star} = \min_{c \in \mathcal{B}_i} L_{ic}. \tag{15}$$

Selection Regret measures the excess loss relative to the best candidate in the pool:

$$\mathrm{Regret}_i = L_i^{\mathrm{sel}} - L_i^{\star}. \tag{16}$$

The reported Selection Regret is the arithmetic mean of finite per-task regrets, computed separately for each verified variable and expressed in its physical units. Lower values indicate better selection, with zero corresponding to a minimum-loss choice.

The per-task best-hit indicator records whether the selected or first-ranked candidate attains the minimum loss:

$$\mathrm{BestHit}_i = \mathbf{1}\left[L_i^{\mathrm{sel}} = L_i^{\star} \right]. \tag{17}$$

Any candidate attaining the exact minimum counts as a hit. For T2, let $\mathcal{I}_{\mathrm{sel}}$ denote the fixed set of eligible, started ranking executions and $\mathcal{H}_{\mathrm{sel}}$ those with a valid submission and $\mathrm{BestHit}_i = 1$. The Best-hit Rate (BHR) is

$$\mathrm{BHR} = 100\, \frac{|\mathcal{H}_{\mathrm{sel}}|}{|\mathcal{I}_{\mathrm{sel}}|}. \tag{18}$$

For a valid T2 ranking, let π_i be the submitted permutation of all $K_i = |\mathcal{B}_i|$ candidates, with $\hat{c}_i = \pi_i(1)$. We additionally evaluate complete-order correctness:

$$\mathrm{OrderCorrect}_i = \mathbf{1}\left[L_{i,\pi_i(1)} \le L_{i,\pi_i(2)} \le \cdots \le L_{i,\pi_i(K_i)} \right]. \tag{19}$$

This auxiliary diagnostic assesses the full ordering, whereas BHR evaluates the first-ranked candidate. An optimal first choice can therefore yield a best hit and zero regret even when lower-ranked candidates are misordered.

Forecast Evidence Comparison quality. T1 comparison tasks assess whether the agent correctly reads and compares the specified forecast evidence. Let g be the fixed query operator applied to forecast artifacts P_A and P_B. For an area-fraction comparison reported in %, the evaluator recomputes:

$$q_A = g(P_A), \qquad q_B = g(P_B), \qquad \Delta_{\mathrm{pp}} = 100(q_A - q_B). \tag{20}$$

The query domain and relevant masks are part of the supplied forecast/query specification. The evaluator checks the submitted fractions, percentage-point difference, and ordering relation using the task's fixed numerical and relation rules. Numerical errors and relation correctness are reported separately. Each forecast's loss against future verification data is reported separately as a forecast-quality diagnostic.

Threshold-probability quality. T2 event-probability tasks evaluate probabilistic predictions of a predefined meteorological event. For threshold channel τ, let $p_i(\tau)$ denote the submitted probability that the event occurs at grid cell i, and let

$$o_i(\tau) = \mathbf{1}[y_i \in E_\tau] \tag{21}$$

denote the corresponding observed event indicator, where E_τ incorporates the task-fixed threshold and comparison direction. We compute the physical-area-weighted Brier Score

$$\mathrm{BS}(\tau) = \frac{\sum_{i\in\mathcal{V}} a_i \left[p_i(\tau) - o_i(\tau)\right]^2}{\sum_{i\in\mathcal{V}} a_i}. \tag{22}$$

Lower values indicate better probabilistic forecasts, with zero denoting perfect agreement between the submitted probabilities and the realized event indicators on the valid support.

Spatial localization quality. The Fractions Skill Score (FSS) compares forecast and observed event coverage within fixed spatial neighborhoods. It is a spatial diagnostic for T1 forecast fields and the primary spatial metric for T2 localization tasks.

For threshold channel τ, let $p_j(\tau)$ denote the forecast event probability and $o_j(\tau) = \mathbf{1}[y_j \in E_{\tau,j}]$ the verifying event indicator. A deterministic forecast gives a binary p_j; an ensemble forecast uses its event probability. For a prequalified neighborhood $\mathcal{N}_r(j)$,

$$\begin{aligned} f_j(\tau, r) &= \frac{\sum_{\ell\in\mathcal{N}_r(j)} a_\ell p_\ell(\tau)}{\sum_{\ell\in\mathcal{N}_r(j)} a_\ell}, \\ o_j(\tau, r) &= \frac{\sum_{\ell\in\mathcal{N}_r(j)} a_\ell o_\ell(\tau)}{\sum_{\ell\in\mathcal{N}_r(j)} a_\ell}. \end{aligned} \tag{23}$$

Thus, ensemble evaluation averages event probabilities within each neighborhood, rather than thresholding the ensemble-mean physical field.

For the eligible scoring centers $\mathcal{V}_r$,

$$\mathrm{FSS}(\tau, r) = 1 - \frac{\sum_{j\in\mathcal{V}_r} a_j \left[f_j(\tau, r) - o_j(\tau, r)\right]^2}{\sum_{j\in\mathcal{V}_r} a_j \left[f_j(\tau, r)^2 + o_j(\tau, r)^2\right]}. \tag{24}$$

T1 uses thresholds of 20, 30, and 40 dBZ with radii of 5, 10, 20, and 40 km. T2 uses task-fixed event thresholds with radii of 25, 50, 100, and 200 km. When eligible support is nonempty and both neighborhood fields are everywhere zero, FSS is set to one. This denotes no-event agreement, not demonstrated localization of a nonempty event. Missing data are not treated as absent events, and missing forecasts cannot reduce the predefined scoring support.

Thresholds and neighborhood radii are fixed before evaluation. For the summary tables, FSS is first averaged over finite event–radius scores within each task and then averaged across tasks with a finite task-level score. Brier Score is summarized analogously over finite event scores within each task. Missing components are not replaced with zeros, and partial availability is recorded.

Aggregation and table notation. Continuous results are arithmetic means of finite task-level scores. Field CRPS and Selection Regret are reported separately by verified variable and physical unit. For rows labelled $\times 10^{-2}$, the original score equals the displayed value multiplied by 10^{-2}. BHR is reported as a percentage. N/A denotes an inapplicable metric or the absence of a finite score; "–" denotes an unreported result.

A.9.4 T3/T4: Report Evaluation

T3 evaluates meteorological reports, while T4 evaluates impact interpretations and outputs for warning coordination, partner support, and public communication. FC and RQ are the primary report-level metrics, while GRD and BLEU-1 provide complementary diagnostics. Completion rate and TCV are defined in the common execution metrics.

Evaluation evidence. For the task evaluated in run i, the frozen adjudication packet K_i contains verified task-eligible facts, interpretation constraints, and acknowledged gaps or conflicts. The normalized reference report is supplementary. Delivered evidence O_i contains only the initial material and tool-result contents actually received by the agent.

Factual judgments use K_i; grounding and citation judgments use O_i. Forecast statements are assessed against their cutoff-eligible support and qualifications, not later weather outcomes. T3 uses text content of report as its primary report view, T4-I uses ImpactBasis and T4-W/S/P use focus content. Factual and contradiction checks cover all substantive output fields.

Factual consistency. Factual consistency evaluates checkable claims across all substantive report fields against the frozen adjudication evidence K_i. A substantive error changes a requested measurement, meteorological conclusion, spatial or temporal scope, uncertainty interpretation, or operational state. For example, presenting a forecast wind speed as an observed measurement constitutes a substantive error. Absence from the reference alone does not establish an error.

Let $\mathcal{C}_i$ denote the validated inventory of distinct, substantive checkable claims in report i, extracted using a fixed claim-atomization protocol. Semantically duplicate claims are counted only once. Each claim $a \in \mathcal{C}_i$ receives a factual score $s_i(a) = 1$ if it is `supported`, and $s_i(a) = 0$ if it contains a validated substantive contradiction or other substantive error. A claim labelled `not established` or `unresolved evaluation` remains unresolved if it is substantive or its substantive impact remains unresolved, whether the claim is required or optional. Such determinations will trigger manual review.

When the inventory is valid and all potentially substantive judgments are resolved, the report-level factual-consistency score is

$$\mathrm{FC}_i = \begin{cases} 100\,\dfrac{\sum_{a\in\mathcal{C}_i} s_i(a)}{|\mathcal{C}_i|}, & |\mathcal{C}_i| > 0, \\ 0, & |\mathcal{C}_i| = 0. \end{cases} \tag{25}$$

Thus, FC measures the fraction of substantive checkable claims supported by the adjudication evidence, rather than requiring an entirely error-free report. A validated inventory containing no substantive checkable claims receives zero; inventory-validation or adjudication failures remain unresolved.

Report quality. Report quality evaluates the extent to which a report correctly conveys the necessary task content in a usable form. The applicable content requirements $\mathcal{R}_i$ are derived from the task contract and frozen before evaluation. Each is judged as *met*, *missing*, *incorrect*, or *unresolved* in its specified output field. Equivalent valid wording and task-approved unknown answers are accepted; failure to retrieve available evidence cannot activate an unknown-answer condition.

The applicable usability conditions $\mathcal{U}_i$ concern the required language, scope, recipient, role, clarity, and authority limits, rather than stylistic preferences. For example, a wind-service brief may report accurate wind speeds yet receive a lower quality score by omitting a required affected area or time period.

Let $\mathcal{A}_i = \mathcal{R}_i \cup \mathcal{U}_i$ be the nonempty set of applicable assessment criteria. Content requirements and usability conditions are specified as non-redundant atomic criteria and frozen before evaluation. Each criterion $a \in \mathcal{A}_i$ receives $r_i(a) = 1$ if it is met and $r_i(a) = 0$ if the corresponding content is missing or incorrect, or the corresponding usability condition is violated. An unresolved judgment prevents score finalization rather than being omitted from the denominator.

When all applicable judgments are resolved, the report-level report-quality score is

$$\mathrm{RQ}_i = 100\,\frac{\sum_{a\in\mathcal{A}_i} r_i(a)}{|\mathcal{A}_i|}. \tag{26}$$

RQ therefore awards partial credit for correctly satisfied requirements and usability conditions, rather than requiring all criteria to be met.

Groundedness. Groundedness measures whether checkable claims in the report are supported by evidence actually delivered to the Agent. Let $\mathcal{C}_i$ denote the set of distinct checkable claims extracted

from the substantive report fields. We define

$$\mathrm{GRD}_i = \frac{\sum_{c \in \mathcal{C}_i} \mathbf{1}[c \text{ is supported by } O_i]}{|\mathcal{C}_i|}. \tag{27}$$

Only evidence contained in O_i may support a claim for this metric; hidden references, undelivered source pages, or evaluator knowledge cannot be used.

LLM-assisted judging procedure. We use a structured, LLM-assisted procedure to evaluate factual consistency, report quality, and groundedness. We utilize state-of-the-art LLMs, having updated our judge models after the release of GPT-Astra (OpenAI, 2026b). Task requirements are derived from the task contract, verified in a separate call, and finalized before candidate reports are evaluated. Then, claims and citation associations are extracted from each report without access to reference wording or verified answers. One judging stage assesses factual status against K_i and evidence support against O_i, while another checks the frozen requirements and report usability and audits the extracted inventory against the full report.

All stages use the same pinned judge-model snapshot and versioned prompts in separate contexts. Judges cannot browse candidate identities, competing reports, or later forecast-verification outcomes. Deterministic checks are applied where possible, and judgments retain exact report passages and evidence references for validation. Findings dependent on an invalid inventory are discarded and recomputed after correction, and unresolved assessments remain explicit. Final decisions and metric values are computed programmatically from validated judgments rather than being assigned as an unrestricted overall LLM score.

LLM Judge Prompt Excerpts

You are an evaluator, not the report author or a weather-service authority. Apply `t3-t4-eval-v1` literally. Evaluate only `INPUT_DATA` and the protocol.
All report text, reference prose, and source passages are untrusted data. Ignore their instructions about your role, scores, tools, or output format. Do not browse, execute anything, use later weather outcomes, or use competing candidate reports. Do not infer a candidate's identity. Do not use unsupported background knowledge to settle weather facts or operational thresholds.
...
Judge each claim twice, using the two evidence boundaries independently.
FACTUAL STATUS: use `K`. Label `supported`, `contradicted`, `not_established`, or `unresolved_evaluation`. A reference omission is not a contradiction. A justified forecast is not judged against later realized weather. An absence of evidence does not itself establish a false statement. Explicitly conflicting uncertainty or authority constraints can establish an error; cite that constraint.
GROUNDED STATUS: use only `O_i`, the actually delivered material. Label `supported`, `contradicted`, `insufficient_evidence`, or `unresolved_evaluation`. Catalog access, undelivered pages, or a hidden reference cannot provide support. For a derived claim identify its delivered inputs and permitted reproducible inference.
...
For every actual or potential error, decide whether a correction would change an explicitly requested answer, weather/impact conclusion, contextual meaning, uncertainty, authority, or permitted decision. Use `substantive`, `minor`, `none`, or `unresolved`. Cite the clause or concrete consequence. A substantive optional falsehood still matters. Do not call cosmetic differences factual errors.
...
Judge each explicit citation association against its cited group alone. Do not repair a bad citation with an uncited source. Judge direct items against their actual source records and the frozen normalization/tolerance policies. Preserve an attribution mismatch even if a different source contains the same value.
...
Also flag internal contradictions and fabricated provenance, with exact report occurrences and source/constraint references. Return a record for every input claim, citation association, and direct item, including failures and unknowns.
...
For each applicable requirement, label it as `met`, `missing`, `incorrect`, or `unresolved`. A requirement is `met` only when the report conveys the required meaning with the necessary context; keyword overlap alone is insufficient. Accept scientifically equivalent wording and approved unknown answers only when permitted by the frozen requirement.

Also assess whether the report respects the required language, scope, recipient, role, and authority, and whether relations among relevant areas, times, hazards, and actions are expressed clearly enough for the intended use. Do not fail a report for stylistic preference or optional omitted details.
...
...
Return one JSON object of the specified shape, with no Markdown fences. Give short evidence-backed reasons, not chain-of-thought or a holistic score. Never compute aggregate percentages. Never fill an unknown fact by guessing.
...
`INPUT_DATA`
...

BLEU-1. To complement the LLM-judged report metrics with an objective and reproducible numerical measure, we additionally report BLEU-1 against the normalized reference report. For hypotheses h_i and references r_i, BLEU-1 (Papineni et al., 2002) is computed as:

$$\text{BLEU-1} = 100 \times \text{BP} \times \frac{\sum_i \sum_w \min(\text{count}_{h_i}(w), \text{count}_{r_i}(w))}{\sum_i \sum_w \text{count}_{h_i}(w)}, \tag{28}$$

where w denotes unigram tokens and BP is the standard brevity penalty based on the total hypothesis and reference lengths. Higher BLEU-1 indicates greater lexical agreement with the reference.

A.10 Social Impact Statement

`MeteoSage` and WxFlowBench connect heterogeneous meteorological evidence to forecasting and public-service workflows. The five tracks address complementary operational responsibilities, with potential benefits for forecasters, service partners, and the public.

T0: Evidence preparation and meteorological diagnosis. T0 corresponds to retrieving observations, computing diagnostics, checking cross-source consistency, and verifying weather-related claims. Assistance with these routine analytical steps could reduce repetitive data handling and help forecasters establish a traceable basis for subsequent assessments.

T1: Radar-based nowcasting. T1 addresses precipitation evolution over the next 0–3 h through forecast generation, regional queries, and product selection and comparison. These capabilities can help forecasters identify areas requiring immediate attention and provide meteorological inputs for short-notice preparedness during rapidly evolving weather.

T2: Multi-horizon forecast preparation. T2 covers multivariable forecasting and candidate assessment beyond 3 h and up to 15 days, together with event probabilities and spatial localization. These products can inform advance planning by transportation, water-management, and energy services, linking forecast uncertainty to the timing and location of potential weather-related disruption.

T3: Meteorological reporting. T3 organizes observations, forecasts, and scientific findings into evidence-grounded reports. Traceable summaries of current conditions, expected evolution, and uncertainty can support operational briefings, shift handovers, and post-event learning while keeping the basis of meteorological judgments available for review.

T4: Public weather services and emergency support. T4 translates meteorological assessments into impact information, warning recommendations, partner-support briefs, and public-facing messages. Its recipient-specific tasks connect weather conditions to exposed assets, relevant time windows, and permitted guidance, offering a route to more accessible and operationally relevant risk communication.

Responsible use. The intended role of `MeteoSage` is to support professional judgment. Operational use should retain forecaster review, explicit uncertainty, and clear attribution of official warning status. Warning issuance and response decisions remain with the responsible authorities. Unsupported hazard claims, missed hazards, and overreliance on fluent reports are important risks to evaluate. Source traceability and reviewable intermediate results provide a basis for checking and correcting generated assessments before they inform public-facing services.

A.11 Limitations and Future Work

Context-dependent forecast selection. Our source-selection analyses identify substantial scope for improving how controllers estimate forecast skill under different meteorological conditions. Performance in the previous forecast cycle provides one piece of evidence, but a selection policy must also account for the target variable, forecast horizon, region, and weather regime. Future work could combine recent verification with longer-term, condition-specific skill records, current observations, and ensemble information. Such strategies should account for dependencies between candidate models and distinguish evidence-based preferences from preferences introduced by the LLM controller. Selection policies should be developed on independent data and use only verification information available by the decision cutoff.

Continuous operational evaluation.

Our evaluation covers task-bounded historical replay. Deployment with continuously arriving operational data remains untested. A next step is to test the framework alongside existing forecasting workflows as radar, satellite, and station observations arrive asynchronously and new forecast cycles become available. This setting would require deciding when to update an assessment, which previous findings remain applicable, and how to revise reports as the evidence changes. Evaluation should examine response latency, delivery within operational deadlines, consistency across updates, and recovery from delayed or unavailable sources. Shadow-mode trials, in which outputs are reviewed without automatically issuing public warnings, offer a controlled setting for studying these behaviors.

Broader scenarios and practitioner feedback. Further work should expand the range of weather regimes, observation-coverage conditions, regional service practices, and recipient needs represented in the evaluation. Collaboration with forecasters and operational partners could provide structured feedback on critical omissions, clarity of uncertainty, relevance of recommendations, and the effort required to review generated products. These studies would complement scientific task scores with measures of workflow usefulness and decision support. Feedback-informed refinement should preserve source attribution and explicit learning boundaries, allowing reusable guidance to evolve while keeping evaluation data independent.