AtmoZero: Self-Play Post-Training for Caption-Free Weather Time-Series Captioning

Qirui Zheng
College of Electrical Engineering
Southwest Minzu University
zqr121@stu.mju.edu.cn

Shengchao Chen*
Australian Artificial Intelligence Institute
University of Technology Sydney
shengchao.chen.uts@gmail.com

Abstract

Multivariate weather time series are central to forecasting and decision support, and language increasingly mediates how numerical weather information reaches users and downstream models. Existing weather captioners rely on closed-source language model ensembles with expert quality control or on retrieved human-curated text. Prior caption-RL rewards optimize cycle similarity or downstream-task utility, and neither can tell a captioner that a confidently asserted phenomenon is contradicted by the data. Without an executable, station-observable, and refutation-aware reward, self-play RL has not been applied to this scientific modality. We propose ATMOZERO, a post-training framework that trains a weather time-series captioner without a single human-written caption in post-training. ATMOZERO aligns a language model with a library of seven station-observable scientific verifiers, adds an explicit refutation channel that penalizes claims the data contradicts, and combines these signals with auxiliary reconstruction and forecaster-uplift rewards under Group Relative Policy Optimization (GRPO). Extensive experiments on a twelve-year ERA5 reanalysis corpus and real-world surface stations show that ATMOZERO outperforms advanced baselines in both caption faithfulness and downstream forecasting, and further analyses indicate that these gains arise from complementary signals and remain robust under distribution shift. Code at *Link*.

1 Introduction

Multivariate weather time series underpin forecasting, climate communication, and decision support, and data-driven forecasters have advanced rapidly [1–3]. Adaptive Fourier neural operators show global skill [4], three-dimensional Earth-specific transformers reach operational medium-range accuracy [5], graph neural networks scale across hundreds of variables [6], and unified climate–weather foundation models share representations across scales [7–11]. Yet the *language* channel remains less mature. Operational narratives, expert captions, and forecast rationales still rely on closed-source language model ensembles with human quality control [12] or retrieval and adaptation of human-curated text [13, 14]. These approaches do not learn caption faithfulness from data and require human captions, expert judges, or third-party generators at inference.

Two adjacent communities have begun to close analogous gaps. In vision, caption-RL post-training has become a distinct literature, with cycle-consistency rewards [15, 16], downstream-utility rewards [17], and video-grounded extensions [18–20]. In reasoning, zero-data self-play replaces human supervision with self-generated, automatically checked tasks [21–23]. For weather time series, neither path is available. Existing methods rely on supervised fine-tuning on synthetic pairs [24, 25] or on inference with closed-source agents [26–28], and RL methods for time series target *reasoning*

*Corresponding author and project leader.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

rather than *captioning* [29–31]. No published method post-trains a weather time-series captioner from a verifiable reward whose evaluation function is independent of human-written captions.

The methodological obstacle is verifier design. Code-executor verifiers in absolute-zero reasoning rely on a domain where any proposal can be checked by running a program [21]. For atmospheric dynamics, the equivalent check would use conservation-law and balance-equation residuals, which require observed state, fluxes, and vertical structure that surface stations do not provide. Cycle and utility rewards from vision captioning are dense and convenient [15, 17], but neither tells a captioner that a confidently asserted phenomenon is contradicted by the data. Cycle similarity rises whenever the reconstruction is close, and utility rewards rise whenever a caption helps the downstream model, even when the caption misstates what happened. Recent text-conditioned multimodal time-series benchmarks make this gap visible. Across recognition, differentiation, and generation, the dominant failure mode is factual rather than stylistic, with captions making meteorological claims unsupported by the underlying signal [32–35]. The missing piece is therefore a reward that can refute unsupported claims and is station-observable, so that it can be evaluated wherever surface observations exist.

We build this reward directly from station-observable atmospheric structure rather than importing it from another modality, and use it to post-train a captioner without human-written caption supervision. Our framework, ATMOZERO, consists of a small library of executable scientific checks on hourly surface variables and station metadata. Each check is paired with a structured caption format whose claims the data can support or refute. Training uses Group Relative Policy Optimization (GRPO) with two auxiliary dense signals, time-series reconstruction and forecaster uplift. The verifier signals decide whether a claim is supported or refuted, while the reconstruction and uplift terms act only as dense stabilizers, which addresses the refutation gap above. Our contributions are:

- We propose ATMOZERO, a self-evolving, refutation-aware post-training framework that trains a multivariate weather time-series captioner against executable scientific verifiers, without human-written caption supervision during post-training.
- For the reward, we design seven station-observable verifiers whose rule families adapt to the available variables. We pair them with a refutation channel that penalizes claims the data contradicts. It is the only term of the verifier reward that can be negative, a teacher signal absent from prior cycle and utility caption-RL rewards.
- For optimization, we couple the verifier with a Proposer–Solver self-play loop trained by GRPO. A cold-start protocol engages the adversarial Proposer only after the Solver's verifier precision crosses a threshold, and the cycle and forecaster-uplift channels act as dense regularizers rather than the primary reward.
- Empirically, ATMOZERO outperforms advanced baselines in both caption faithfulness and downstream forecasting, and further analyses show that these gains arise from complementary signals and remain robust under distribution shift.

2 Related Work

Self-Evolving and Zero-Data RLVR for Reasoning Models. Reinforcement learning with verifiable rewards has produced self-evolving frameworks that replace human supervision with executable, label-free signals. Absolute-zero reasoning frames a single model as proposer and solver, with a code interpreter checking proposal validity and solution correctness [21]. R-Zero co-evolves an independent challenger and solver under an edge-of-capability difficulty filter, removing pre-existing tasks and labels [22]. Multi-agent self-play extends the idea to multimodal models through "Who Is the Spy"-style games [23], and analogous self-questioning, solving, and verifying roles improve long-context reasoning [36–38]. Sharper variants add progressive self-imitation in long-horizon agentic tasks [39], generative latent memory [40], and rule-aligned policy optimization in trajectory planning [41]. In each case the verifier is a deterministic program executor or a difficulty-filtered selector over a text task pool, and neither grades raw sensor observations.

Caption Alignment from Non-Caption Supervision. Captioners can be post-trained with rewards that avoid human-written captions. Cycle consistency gives vision-language models a self-supervised signal, either through caption-to-image reconstruction losses [15] or caption–image–caption round trips [16]. Utility-reward methods optimize downstream utility, rewarding captions that improve

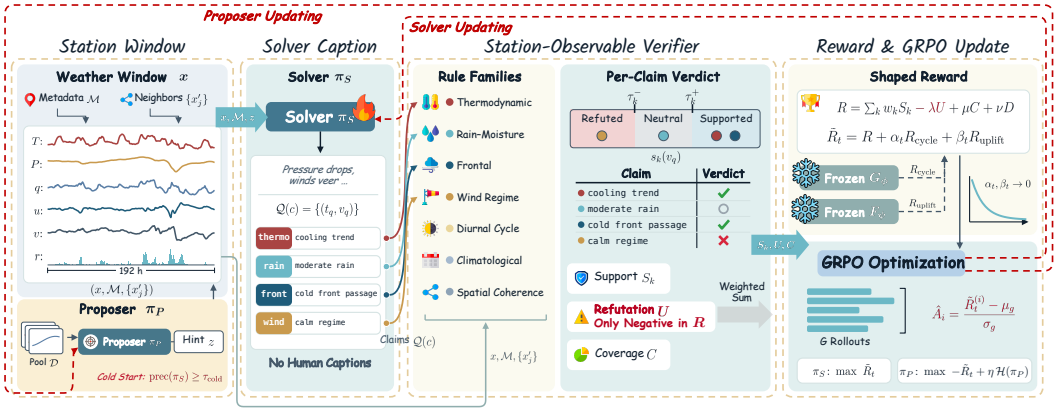


Figure 1: Overview of AtmoZero. The Proposer π_P samples a window x (a real 192-h ERA5 window) and an optional regime hint z. The Solver π_S produces a caption, and seven station-observable rule families grade each claim in its typed-claim list $\mathcal{Q}(c)$ against thresholds τ_k^- and τ_k^+, yielding Support S_k, Refutation U, and Coverage C. With the dense channels of the frozen G_ϕ and F_ψ, they form the shaped reward for GRPO, and red dashed loops mark the updates of both policies. No human-written captions enter post-training.

visual reasoning or multiple-choice answering [17]. Joint reward designs train captioners with downstream generators, scoring captions by their effect on video generation [19] or structured-thinking video decompositions [18], with saturation-aware schedules adapting rewards as captions improve [20, 42]. In this family the reconstruction target is visual, and the reward is dense and unsigned. It scores informativeness rather than consistency with the underlying data.

Weather Captioning and Time-Series Alignment. Weather time-series captioning has relied on two sources of human-derived text. Multi-agent collaborative captioning aggregates closed-source language models behind curated tools and meteorologist quality control [12], and retrieval methods align human-authored news, expert warnings, or domain captions to observations through alignment networks [13, 14, 43]. Supervised fine-tuning aligns frozen time-series backbones [44–46] with language models using synthetic captions distilled from large teachers [24, 25, 47]. GRPO has been applied to time-series *reasoning*, not caption generation [29–31]. Across recent benchmarks for numeric-series captioning [32, 33], scientific time-series understanding [34], and multimodal forecasting [35, 48], as well as agentic frameworks [26–28], the dominant failure mode is fluent description of coarse trends combined with claims the signal does not support. AtmoZero addresses this gap by training the captioner against station-observable scientific verifiers that support data-consistent claims and refute contradicted ones, without relying on human-written captions, vision-domain cycles, or downstream-task utility.

3 Methodology

Problem Formulation. A multivariate weather time-series window is a contiguous block of hourly station observations $x = \{x_t\}_{t=1}^{T_w} \in \mathbb{R}^{T_w \times d}$ over T_w hours, where each x_t is a vector of d station-observable channels drawn from a station-dependent variable set $\mathcal{X}_{\text{stn}}$. We instantiate $\mathcal{X}_{\text{stn}} = \{T, P, q, u, v, r\}$ with $d = 6$, denoting air temperature, station pressure, specific humidity, the zonal and meridional wind components, and past-hour accumulated precipitation. The verifier and reward introduced below are parameterized by $\mathcal{X}_{\text{stn}}$ and admit other surface-station variable sets. Each window carries station metadata $\mathcal{M} = (\text{lat}, \text{lon}, h, \kappa)$ encoding latitude, longitude, elevation, and Köppen–Geiger zone [49], and up to K time-aligned nearest-neighbor windows $\{x'_j\}_{j=1}^{K}$ drawn from stations within a fixed radius (Appendix C.1). A Solver policy π_S generates a caption $c \sim \pi_S(\cdot \mid x, \mathcal{M}, z)$ in the structured form defined below. Here $z \in \mathcal{Z} \cup \{\varnothing\}$ is an optional regime hint produced by an adversarial Proposer policy π_P. The objective is to learn (π_S, π_P) such that c is faithful to x and useful for downstream tasks. No human-written caption serves as supervision, reward, or retrieval target during post-training, and the only human input is annotator labels on machine-generated captions used for calibration (Appendix B.3).

Overview. **Fig. 1** situates the components introduced below. The Proposer π_P samples a window and an optional regime hint, the Solver π_S produces a caption whose typed-claim list is graded by seven station-observable rule families, and the resulting support, refutation, and coverage signals compose with dense auxiliary channels into a shaped reward driving GRPO on both policies.

3.1 Station-Observable Scientific Verifier Library

Surface-station data cannot support code-style exact verification, because conservation- and balance-equation residuals require unobserved state, fluxes, and vertical structure [21]. Cycle and utility rewards designed for vision [15, 17] provide dense feedback but do not identify claims contradicted by station observations. We construct a deterministic verifier that reads only $(x, \mathcal{M}, \{x'_j\})$ and grades caption claims rather than caption strings. Its seven rule families are thermodynamic, rain–moisture, frontal, wind regime, diurnal cycle, climatological, and spatial coherence (Appendix A).

Structured Caption Format. A caption c consists of free prose followed by a typed-claim list

$$\mathcal{Q}(c) = \{(t_q, v_q)\}_{q=1}^{|\mathcal{Q}(c)|}, \quad t_q \in \{1, \dots, 7\}, \quad v_q \in \mathcal{V}_{t_q}, \tag{1}$$

where t_q identifies one of the seven rule families and v_q is a value from a controlled lexicon $\mathcal{V}_{t_q}$ (e.g., `(front, "cold front passage")`). The verifier grades $\mathcal{Q}(c)$.

Per-Family Score, Refutation, and Coverage. Each rule family k exposes a per-claim grader $s_k(v_q; x, \mathcal{M}, \{x'_j\}) \in [0, 1]$ together with two thresholds $\tau_k^+ \geq \tau_k^-$ estimated on an annotated validation set, and a claim (k, v_q) is *supported* when $s_k(v_q) \geq \tau_k^+$, *refuted* when $s_k(v_q) \leq \tau_k^-$, and *neutral* otherwise. Writing $\mathcal{Q}_k(c) = \{q \in \mathcal{Q}(c) \mid t_q = k\}$ and abbreviating $s_k(v_q) \equiv s_k(v_q; x, \mathcal{M}, \{x'_j\})$, the verifier exposes three signals to the optimizer,

$$S_k(c, x) = \frac{1}{|\mathcal{Q}_k(c)|} \sum_{q \in \mathcal{Q}_k(c)} s_k(v_q) \cdot \mathbf{1}[s_k(v_q) \geq \tau_k^+], \tag{2}$$

$$U(c, x) = \sum_{k=1}^{7} \sum_{q \in \mathcal{Q}_k(c)} \mathbf{1}[s_k(v_q) \leq \tau_k^-], \tag{3}$$

$$C(c, x) = \tfrac{1}{7} \sum_{k=1}^{7} \mathbf{1}\big[\, \exists\, q \in \mathcal{Q}_k(c) \text{ with } s_k(v_q) \geq \tau_k^+ \big], \tag{4}$$

with $S_k(c, x) = 0$ when $\mathcal{Q}_k(c) = \varnothing$. S_k averages supported-claim scores within family k, and C measures rule-family coverage. U counts refuted claims across all families and is the channel through which the verifier assigns a strictly negative reward to confidently wrong assertions. Each family k depends on a subset of channels in $\mathcal{X}_{\text{stn}}$, and claims issued for families whose required channels are absent at the focal station are dropped from S_k, U, and C rather than refuted, preserving the channel-aware design of the verifier. The estimation protocol for (τ_k^-, τ_k^+) is given in Appendix B.1, and the per-family graders s_k, channel dependencies, and lexicons $\mathcal{V}_k$ are given in Appendix A. Below we suppress the static arguments $(\mathcal{M}, \{x'_j\})$ of the verifier signals, R, and $\widetilde{R}_t$.

3.2 Reward Composition with an Explicit Refutation Channel

Cycle and utility rewards score how informative a caption is, not whether its claims are true [15, 17], and the verifier of Section 3.1 is the missing teacher that punishes confident hallucinations. We compose the verifier with the cycle and utility channels rather than replace them, and shape the composition so the verifier dominates as training progresses.

Composed Verifier Reward. We define the per-caption verifier reward as

$$R(c, x) \;=\; \sum_{k=1}^{7} w_k \, S_k(c, x) \;-\; \lambda \, U(c, x) \;+\; \mu \, C(c, x) \;+\; \nu \, D(c), \tag{5}$$

where $D(c)$ is a type–token-ratio specificity bonus that discourages generic captions, $w_k = 1/7$ weights the families uniformly, and (λ, μ, ν) are non-negative weights held fixed during GRPO.

They are estimated on the same annotated validation set as (τ_k^-, τ_k^+) by maximizing agreement with annotator labels on its claims (Appendix B.3), so no human-written caption enters the calibration. Unlike prior caption-RL rewards, R becomes negative only through the $-\lambda U$ term, which applies only to assertions the data refutes.

Auxiliary Dense Channels. Because S_k and U are sparse early in training, we add two dense auxiliary channels, a reconstruction term against a frozen text-to-series decoder G_ϕ [50] and a forecaster-uplift term against a frozen forecaster F_ψ [51],

$$R_{\text{cycle}}(c,x) \;=\; -\big\|x - G_\phi(c)\big\|_F^2 / T_w, \quad R_{\text{uplift}}(c,x) \;=\; \mathcal{L}_{F_\psi}\big(x_{>t_0} \mid x_{\le t_0}\big) - \mathcal{L}_{F_\psi}\big(x_{>t_0} \mid x_{\le t_0}, c\big), \tag{6}$$

where $\mathcal{L}_{F_\psi}$ is the mean-squared forecasting loss of F_ψ on $x_{>t_0}$, and the cutoff t_0 is sampled uniformly so that both segments are nontrivial. G_ϕ and F_ψ are pretrained on $\mathcal{D}$ without human-written captions and held frozen during AtmoZero training. The Solver sees the full window, while F_ψ sees only $x_{\le t_0}$, so R_{uplift} rewards captions that convey the forecast-relevant structure of $x_{>t_0}$.

Reward Shaping and Schedule. The dense channels decay during training so that the final objective is governed by the verifier. The shaped reward exposed to the optimizer at training step t is

$$\widetilde{R}_t(c,x) \;=\; R(c,x) \;+\; \alpha_t\, R_{\text{cycle}}(c,x) \;+\; \beta_t\, R_{\text{uplift}}(c,x), \qquad (\alpha_t, \beta_t) \xrightarrow{t\to\infty} (0,0), \tag{7}$$

with the decay form reported in Appendix B.1. The schedule makes the verifier reward dominate asymptotically, leaving the auxiliary channels as early-training stabilizers.

3.3 Proposer–Solver Self-Play with a Cold-Start Gate

Uniform sampling under-exposes the regimes where the Solver is weakest and verifier feedback is most informative, so we turn the verifier into a curriculum with an adversarial Proposer.

Bi-Level Objective. The Proposer factors as $\pi_P(x, z \mid \mathcal{D}) = \pi_P(x \mid \mathcal{D})\,\pi_P(z \mid x)$, jointly selecting a training window from the pool $\mathcal{D}$ and an optional regime hint $z \in \mathcal{Z} \cup \{\varnothing\}$, and the Solver produces a caption $c \sim \pi_S(\cdot \mid x, \mathcal{M}, z)$. The two policies optimize

$$\begin{aligned} &\max_{\pi_S} \; \underbrace{\mathbb{E}_{(x,z)\sim\pi_P,\, c\sim\pi_S}\big[\widetilde{R}_t(c,x)\big]}_{J_S(\pi_S,\,\pi_P)}, \\ &\max_{\pi_P} \; \underbrace{\mathbb{E}_{(x,z)\sim\pi_P,\, c\sim\pi_S}\big[-\widetilde{R}_t(c,x) - \gamma\,\mathbf{1}\big[\text{infeasible}(x,z)\big]\big] + \eta\,\mathcal{H}(\pi_P)}_{J_P(\pi_P,\,\pi_S)}, \end{aligned} \tag{8}$$

where $\mathcal{H}(\pi_P)$ is an entropy bonus that prevents collapse onto a small set of windows, and $\mathbf{1}[\text{infeasible}(x,z)] = 1$ when no verifier trigger predicate fires on x for the regime named by z (Appendix A), with $\text{infeasible}(x, \varnothing) = 0$.

Cold-Start Protocol. The Proposer cannot meaningfully select hard windows before the Solver has learned to use any rule family. We therefore sample windows uniformly with no regime hint, $\pi_P(x,z) = \mathcal{U}(\mathcal{D}) \cdot \mathbf{1}[z = \varnothing]$, until the Solver's running verifier precision $\text{prec}(\pi_S^{(t)})$ crosses a threshold τ_{cold} (Appendix B.1). The precision is the fraction of issued claims that the verifier grades as supported on a probe set, and probe windows without any claim count as zero precision so that abstention cannot raise it. The Proposer schedule is therefore

$$\pi_P^{(t)} \;=\; \begin{cases} \mathcal{U}(\mathcal{D}) & \text{if } \text{prec}\big(\pi_S^{(t)}\big) < \tau_{\text{cold}}, \\ \arg\max_{\pi_P} J_P\big(\pi_P,\, \pi_S^{(t)}\big) & \text{otherwise.} \end{cases} \tag{9}$$

The gate prevents the Proposer from concentrating on windows the Solver cannot yet handle.

Optimizer. Both objectives in Eq. 8 are optimized with GRPO [52, 53], which draws a group of G rollouts per update. The rollouts are captions for the Solver and (x, z) pairs for the Proposer.

Concretely, π_S minimizes the per-prompt clipped surrogate

$$\mathcal{L}_{\mathrm{GRPO}}(\pi_S) \;=\; -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\big(\rho_i\,\hat{A}_i,\;\mathrm{clip}(\rho_i,\,1-\epsilon,\,1+\epsilon)\,\hat{A}_i\big)\right] \;+\; \beta_{\mathrm{KL}}\,D_{\mathrm{KL}}\big(\pi_S \,\big\|\, \pi_S^{\mathrm{ref}}\big), \tag{10}$$

where ρ_i is the per-rollout importance ratio against the previous policy and $\hat{A}_i = (\widetilde{R}_t^{(i)} - \mu_g)/\sigma_g$ is the advantage normalized by the within-group reward mean μ_g and standard deviation σ_g, which removes the need for a value model. The Proposer minimizes the analogous surrogate on J_P. Both policies receive *full-parameter* updates from a shared instruction-tuned initialization [54], with separate optimizer states and alternating updates.

4 Experiments

Implementation Details. Following post-training practice [55, 56], both policies π_P and π_S are initialized from Qwen2.5-7B-Instruct [54] and post-trained with full-parameter GRPO [52, 53] on $4\times$ NVIDIA H100 80GB GPUs in bf16 precision. Each Proposer–Solver seed runs for $2{,}450$ optimizer steps and costs about 790 H100-hours (Appendix B.1). All reported results are means over three seeds, and the faithfulness table also gives 95% window-paired bootstrap half-widths (Appendix B.3). The frozen forecaster F_ψ is PatchTST [51] and the frozen text-to-series decoder G_ϕ is an 8-layer DiT [50]. Full details are in Appendix B.1 (**Table 11**).

Baselines. We compare AtmoZero against three families of references. The first family contains forecasters without a caption channel from the linear, transformer, mixer, and TS-LLM model classes, namely DLinear [57], PatchTST [51], TimesNet [58], iTransformer [59], TimeMixer [60], Time-LLM [61], TimeXL [27], and GPT4TS [62]. Our PatchTST forecaster without captions (No-Caption) is the directly comparable reference. The second family conditions the same forecaster on captions from prompted closed-source models (GPT-4o 1-shot, Claude-Sonnet-4.5 1-shot) or from the un-tuned open-weights backbone (Prompted-Base, Qwen2.5-7B-Instruct), which measures what prompting alone achieves. The third family contains three RL ablations within our framework, namely Dense-Only (cycle and uplift but no verifier), Verifier-Only (verifier but no auxiliary channels), and Solver-Only (full reward but with the Proposer fixed to uniform sampling). AtmoZero is the full method with the verifier reward of Eq. 5 and both dense channels. We also report Verifier-Templater (Appendix D.10), a non-learned reference that verbalizes verifier-certified claims, and two adapted self-play baselines (Appendix D.12), all described in Appendix B.2.

Benchmark and Metrics. The corpus draws on twelve years (2014–2025) of hourly ERA5 single-level reanalysis [63] for the canonical six-variable surface set (T, P, q, u, v, r), sampled at $8{,}192$ stations stratified across all thirty Köppen–Geiger zones [49] of the global land surface. The train, validation, and test splits cover 2014–2022, 2023, and 2024–2025, respectively. After deduplication and quality control, the training split contains about 4.8×10^7 windows of length $T_w = 192$ hours. Forecasting performance is reported as mean squared error and mean absolute error at horizons of 96, 192, 336, and 720 hours, averaged across stations. Test-time captions use only the context window ending at forecast issuance, the metadata, and neighbor windows over the same interval, so the forecast target never enters any captioner (Appendix D.11). Caption faithfulness is evaluated on a 200-window stratified annotated set with eight metrics, namely unsupported-claim rate (Unsup.), supported-claim coverage (Coverage), type–token ratio (TTR), rule-orthogonality (Orth.), claim density (Claims), false-refutation rate (False ref.), Fleiss κ, and NLI entailment rate (Appendix B.3).

4.1 Main Results

Caption Faithfulness. In **Table 1**, AtmoZero attains the lowest unsupported-claim rate under identical optimizer budgets. The gain comes from the verifier reward of Eq. 5. The support signal S_k supplies most of its magnitude, and the refutation channel U is the only term that makes the reward negative for unsupported claims, which cycle and utility rewards cannot express. AtmoZero leads on Coverage and rule-orthogonality, so the verifiers reward a wider correct vocabulary rather than a narrow subset and break the faithfulness–coverage trade-off of dense rewards.

Table 1: Faithfulness on the 200-window stratified human-annotated set drawn from the 2023 validation split, scored under the 5-fold cross-fit protocol of Appendix B.3. κ uses pooled pre-adjudication labels (no seed dependence). **Bold** = best, underline = second-best per column.

Method	Unsup. (↓)	Coverage (↑)	TTR (↑)	Orth. (↑)	Claims (↑)	False ref. (↓)	Fleiss κ (↑)	NLI (↑)
GPT-4o 1-shot	$27.4_{\pm 1.6}$	$4.6_{\pm 0.34}$	$\mathbf{0.67}_{\pm 0.018}$	$0.68_{\pm 0.027}$	$7.9_{\pm 0.41}$	$8.7_{\pm 0.95}$	0.72	$0.70_{\pm 0.017}$
Claude-Sonnet-4.5 1-shot	$20.8_{\pm 1.2}$	$4.9_{\pm 0.29}$	$\underline{0.66}_{\pm 0.022}$	$0.72_{\pm 0.031}$	$8.2_{\pm 0.36}$	$7.6_{\pm 0.86}$	0.74	$0.76_{\pm 0.026}$
PROMPTED-BASE	$38.6_{\pm 2.4}$	$3.9_{\pm 0.41}$	$0.58_{\pm 0.029}$	$0.61_{\pm 0.044}$	$6.6_{\pm 0.52}$	$11.5_{\pm 1.3}$	0.68	$0.62_{\pm 0.034}$
DENSE-ONLY	$22.3_{\pm 1.0}$	$4.8_{\pm 0.27}$	$0.64_{\pm 0.024}$	$0.70_{\pm 0.025}$	$7.4_{\pm 0.33}$	$9.2_{\pm 1.1}$	0.73	$0.75_{\pm 0.021}$
VERIFIER-ONLY	$\underline{11.7}_{\pm 0.7}$	$5.1_{\pm 0.19}$	$0.55_{\pm 0.013}$	$0.78_{\pm 0.018}$	$\mathbf{8.9}_{\pm 0.38}$	$\underline{6.1}_{\pm 0.74}$	$\underline{0.78}$	$\underline{0.84}_{\pm 0.029}$
SOLVER-ONLY	$12.5_{\pm 1.1}$	$\underline{5.3}_{\pm 0.21}$	$0.59_{\pm 0.014}$	$\underline{0.80}_{\pm 0.022}$	$\underline{8.5}_{\pm 0.31}$	$6.3_{\pm 0.82}$	0.77	$0.83_{\pm 0.018}$
ATMOZERO (Ours)	$\mathbf{9.2}_{\pm 0.85}$	$\mathbf{5.6}_{\pm 0.24}$	$0.61_{\pm 0.012}$	$\mathbf{0.83}_{\pm 0.020}$	$8.4_{\pm 0.47}$	$\mathbf{5.8}_{\pm 0.71}$	**0.79**	$\mathbf{0.86}_{\pm 0.025}$

Downstream Forecasting. **Table 2** reports MSE at four horizons across two blocks. The upper block lists eight numerical-only forecasters without captions; the lower block lists nine caption sources conditioning the same frozen PatchTST backbone. All published baselines use the default hyperparameters of their original papers. ATMOZERO attains the lowest MSE at every horizon in both blocks. The gain over NO-CAPTION grows with the horizon, from 0.034 at H=96 to 0.056 at H=720, as regime-level claims matter most for longer forecasts. The closest competitor, SOLVER-ONLY, retains the full verifier reward but fixes the Proposer to uniform sampling, so the remaining gap measures the Proposer's contribution as a curriculum over hard windows. ATMOZERO captions let the frozen PatchTST backbone beat every upper-block forecaster.

Table 2: Forecasting MSE on the 2024–2025 test split. *Upper*, numerical-only forecasters. *Lower*, the same frozen PatchTST backbone under different caption sources. †Deterministic non-learned verifier reference. MAE in Appendix D.1.

Method	H=96	H=192	H=336	H=720
Numerical-only forecasters (no caption channel)				
DLinear [57]	0.299	0.359	0.411	0.463
PatchTST [51]	0.276	0.333	0.372	0.424
TimesNet [58]	0.258	0.312	0.352	0.397
iTransformer [59]	0.263	0.318	0.359	0.405
TimeMixer [60]	0.261	0.315	0.355	0.401
Time-LLM [61]	0.266	0.321	0.361	0.409
TimeXL [27]	0.264	0.319	0.358	0.406
GPT4TS [62]	0.269	0.325	0.364	0.412
Caption-augmented forecasting (frozen PatchTST + caption source)				
NO-CAPTION	0.271	0.326	0.365	0.417
PROMPTED-BASE [54]	0.281	0.339	0.382	0.432
GPT-4o 1-shot [64]	0.255	0.309	0.348	0.391
Claude-Sonnet-4.5 1-shot [65]	0.252	0.305	0.344	0.386
VERIFIER-TEMPLATER†	0.258	0.311	0.343	0.389
DENSE-ONLY	0.249	0.300	0.335	0.379
VERIFIER-ONLY	0.251	0.304	0.340	0.383
SOLVER-ONLY	$\underline{0.244}$	$\underline{0.294}$	$\underline{0.327}$	$\underline{0.372}$
ATMOZERO (Ours)	**0.237**	**0.287**	**0.316**	**0.361**

Pairwise Human Preference Study. A blind pairwise study by three PhD-level meteorologists on 300 test windows shows ATMOZERO wins most pairs against every baseline (**Table 3**). The Spearman correlation between win rate and opponent unsupported-claim rate (ρ=0.77) shows that faithfulness gains transfer to experts. Losses to closed-source captioners mostly reflect their longer prose (Appendix D.3).

Table 3: Blind pairwise human preference. Details in Appendix D.3.

Opponent	Win (%)	Tie (%)	Loss (%)	95% CI
PROMPTED-BASE	**84**	7	9	±4.1
DENSE-ONLY	**70**	12	18	±5.2
GPT-4o 1-shot	**63**	11	26	±5.4
SOLVER-ONLY	**61**	12	27	±5.5
VERIFIER-ONLY	**58**	16	26	±5.6
Claude-Sonnet-4.5 1-shot	**56**	15	29	±5.6
Mean over opponents	**65.3**	12.2	22.5	±2.3

Climate-Zone Out-of-Distribution Generalization. We retrain ATMOZERO with six Köppen–Geiger zones (BWh, Cwa, Dfb, Dwc, ET, Aw) excluded from training, and evaluate on 2024–2025 windows from those zones (**Table 4**). ATMOZERO keeps the lowest Unsup. (12.4%), 4.0 pp below SOLVER-ONLY. The rule families encode atmospheric physics rather than zone-specific patterns, so a geographic shift changes *which* families fire while the refutation predicates of Eq. 3 remain valid. Families whose channels are non-discriminative in unfamiliar zones disengage rather than hallucinate, so ATMOZERO keeps the highest Coverage (4.7) at less than half the Unsup. of DENSE-ONLY.

Table 4: Köppen–Geiger OOD on six unseen zones; Unsup. is NLI-estimated (Appendix D.4).

Method	Unsup. (↓)	Cov. (↑)	F. ref. (↓)	MSE@H=336
GPT-4o 1-shot	30.2	4.2	11.5	0.378
DENSE-ONLY	$26.4_{\pm 1.7}$	$4.5_{\pm 0.42}$	$11.7_{\pm 1.4}$	$0.367_{\pm 0.008}$
Time-LLM	–	–	–	$0.353_{\pm 0.009}$
VERIFIER-ONLY	$17.8_{\pm 1.0}$	$4.4_{\pm 0.36}$	$15.9_{\pm 1.5}$	$0.356_{\pm 0.006}$
SOLVER-ONLY	$16.4_{\pm 1.3}$	$\mathbf{4.7}_{\pm 0.28}$	$13.8_{\pm 1.1}$	$0.347_{\pm 0.007}$
ATMOZERO	$\mathbf{12.4}_{\pm 1.2}$	$\mathbf{4.7}_{\pm 0.40}$	$\mathbf{8.4}_{\pm 0.9}$	$\mathbf{0.339}_{\pm 0.010}$

Table 5: Zero-shot MSE on six real-world surface stations, averaged over $H \in \{96, 192, 336, 720\}$. **Bold**/underline = best/second-best. T-MoE = Time-MoE; subscripts $S/B/L$ = small/base/large.

Region	AtmoZero	T-MoE$_B$	T-MoE$_L$	Moirai$_S$	Moirai$_B$	Moirai$_L$	TimesFM	Moment	Chronos$_S$	Chronos$_B$	Chronos$_L$
Tokyo	**0.479**	0.491	0.485	1.713	1.526	2.626	0.844	1.070	0.931	0.912	0.929
Mumbai	**0.535**	0.555	0.538	1.688	1.231	1.685	0.890	1.138	1.004	1.006	0.988
Cairo	**0.744**	0.834	0.815	1.703	1.298	1.900	1.295	1.002	1.495	1.576	1.508
Edmonton	**0.581**	0.608	0.609	0.971	0.936	0.917	0.903	1.026	1.065	1.119	1.080
Zurich	**0.601**	0.659	0.632	1.075	0.891	0.931	1.044	1.086	1.192	1.208	1.166
McMurdo	**0.685**	0.718	0.699	1.376	1.039	1.109	1.103	1.084	1.244	1.327	1.272

Zero-Shot Transfer to Real-World Stations. To evaluate generalization to real observations, we collect fifteen surface-station datasets from six continents and a wide range of climates. The forecaster is conditioned on captions exactly as in the main experiments. **Table 5** reports the four-horizon mean MSE on six geographically diverse stations (all fifteen in **Table 16**) spanning Asian megacities, arid North Africa, cold continental North America, alpine Europe, and Antarctica, for AtmoZero and ten zero-shot foundation-model baselines. AtmoZero leads on all six regions, because the verifier rule families encode atmospheric structure that transfers across continents. The margin is smallest at Tokyo and Mumbai, where strong urban diurnal forcing lets Time-MoE$_{large}$ come within 0.007.

Temporal Alignment of Captions. **Table 6** conditions the same frozen PatchTST on AtmoZero captions permuted across test windows (Shuffled), generated from the first 96 h of the window, generated under the standard protocol, or generated by a captioner that also sees the forecast target (Future-informed). MSE falls monotonically with the information available to the captioner, and the standard protocol stays far from the future-informed regime. Shuffled captions raise MSE above No-Caption at every horizon, so the gain requires window-aligned content.

Table 6: Frozen PatchTST MSE by caption source (Appendix D.11).

Caption	H=96	H=192	H=336	H=720
No-Caption	0.271	0.326	0.365	0.417
Shuffled	0.279	0.336	0.378	0.425
First 96 h	0.254	0.306	0.338	0.386
Standard	0.237	0.287	0.316	0.361
Future-informed	0.191	0.226	0.249	0.302

4.2 Framework Analysis

Hyperparameter Sensitivity. **Fig. 2** sweeps six core hyperparameters around their defaults and tracks the unsupported-claim rate (left axis) and MSE@H=336 (right axis). The refutation weight λ is the most sensitive: the default minimizes MSE@H=336, and larger λ lowers Unsup. further, so λ directly tunes faithfulness. The cold-start threshold τ_{cold} is second, with faithfulness degrading when the Proposer engages before verifier precision crosses 0.84. The remaining four axes form shallow basins with defaults at or near the empirical minimum.

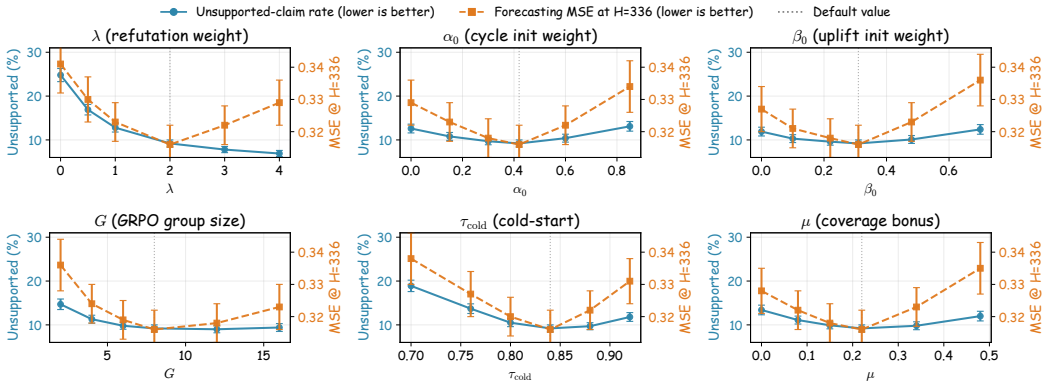


Figure 2: Hyperparameter sensitivity: unsupported-claim rate (%, left, blue) and MSE@H=336 (right, orange, dashed). Dotted lines mark defaults, α_0, β_0 are the initial α_t, β_t, and error bars are std.

Ablation Study. **Table 7** reports three ablations by Unsup., MSE@H=336, and Coverage *entropy*, the Shannon entropy of the seven-family histogram. *Leaving out one rule family* costs the most for the frontal family, and no single removal closes the gap between AtmoZero and Dense-Only, so the verifier signal is not dominated by one rule. *Replacing the validation-based thresholds* (τ_k^-, τ_k^+) with

a uniform initialization raises Unsup. to 14.6% and MSE@H=336 to 0.333, so empirical calibration matters. *Fixing the Proposer* to uniform sampling raises Unsup. by 3.3 pp under the same reward, so self-play turns the static reward into a curriculum.

Table 7: Ablation study. Top, leave-one-rule-family-out; middle, validation-based thresholding vs. uniform init; bottom, without Proposer. ΔUnsup. (pp, green) denotes the increase over ATMOZERO.

Variant	MSE@H=336 (↓)	Unsup. (%) (↓)	ΔUnsup. (pp)	Coverage entropy (↑)
ATMOZERO	**0.316**	**9.2**	0.0	**0.83**
Leave-one-rule-family-out from ATMOZERO				
− Thermodynamic	0.323	10.6	+1.4	0.80
− Rain–Moisture	0.326	11.1	+1.9	0.79
− Frontal	0.331	12.0	+2.8	0.77
− Wind Regime	0.321	9.9	+0.7	0.81
− Diurnal Cycle	0.322	10.1	+0.9	0.81
− Climatological	0.324	10.7	+1.5	0.79
− Spatial Coherence	0.323	10.4	+1.2	0.80
Validation-based threshold selection vs. uniform-threshold initialization				
ATMOZERO *(uniform thresholds)*	0.333	14.6	+5.4	0.74
Co-evolution ablation				
SOLVER-ONLY *(uniform Proposer)*	0.327	12.5	+3.3	0.80

Reward-Channel Decomposition. **Table 8** removes one reward channel at a time from a shared cold-start checkpoint (Appendix D.6). Removing the support signal S_k raises Unsup. by 13.2 pp, and removing the refutation channel U, the only term that penalizes contradicted claims, raises it by 4.4 pp and more than doubles the false-refutation rate. Removing the coverage bonus C costs 3.2 pp and destabilizes one of three seeds, while the dense auxiliary channels mainly affect MSE@H=336.

Table 8: Reward-channel decomposition (three seeds per row; Full row from Table 1).

Variant	Unsup. (↓)	False ref. (↓)	Coverage (↑)	NLI (↑)	MSE@H=336 (↓)	Unstable seeds
Full	$9.2_{\pm 0.85}$	$5.8_{\pm 0.7}$	$5.6_{\pm 0.2}$	$0.86_{\pm 0.02}$	$0.316_{\pm 0.006}$	0/3
$-U$	$13.6_{\pm 1.0}$	$12.7_{\pm 1.2}$	$5.4_{\pm 0.2}$	$0.81_{\pm 0.02}$	$0.324_{\pm 0.006}$	0/3
$-S_k$	$22.4_{\pm 1.7}$	$17.9_{\pm 1.5}$	$4.8_{\pm 0.3}$	$0.74_{\pm 0.03}$	$0.351_{\pm 0.008}$	0/3
$-C$	$12.4_{\pm 2.6}$	$7.9_{\pm 1.4}$	$4.5_{\pm 0.4}$	$0.79_{\pm 0.03}$	$0.327_{\pm 0.007}$	1/3
$-D$	$10.3_{\pm 0.9}$	$6.7_{\pm 0.8}$	$5.4_{\pm 0.2}$	$0.84_{\pm 0.02}$	$0.319_{\pm 0.006}$	0/3
$-R_{\text{cycle}}$	$10.8_{\pm 0.9}$	$6.9_{\pm 0.9}$	$5.5_{\pm 0.2}$	$0.84_{\pm 0.02}$	$0.322_{\pm 0.006}$	0/3
$-R_{\text{uplift}}$	$10.2_{\pm 0.9}$	$6.4_{\pm 0.8}$	$5.6_{\pm 0.2}$	$0.85_{\pm 0.02}$	$0.330_{\pm 0.007}$	0/3

Scaling Across Backbones. We retrain ATMOZERO on Qwen2.5-Instruct {1.5, 3, 7, 14}B and two other backbones (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) under the same protocol (**Fig. 3**). Gains saturate beyond 7B, with Qwen-14B within 0.3 pp of Qwen-7B on Unsup., supporting 7B as the default. Across families, Llama-3.1-8B reaches a lower Unsup. and Mistral-7B-v0.3 a lower MSE@H=336 than Qwen-7B. The backbone sets the faithfulness–forecasting trade-off, and the method transfers across families without tuning.

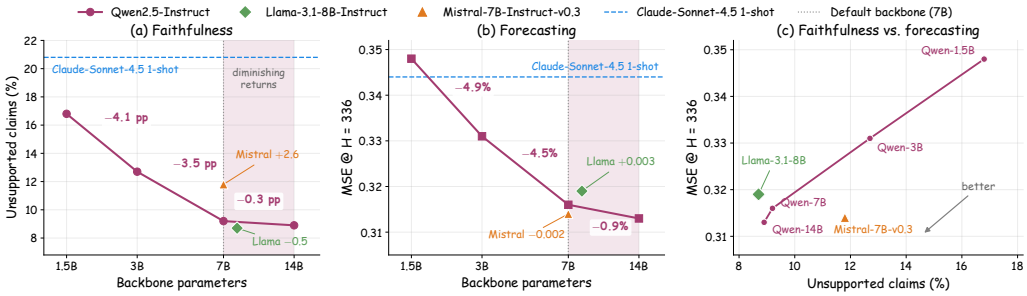


Figure 3: Backbone scaling across Qwen2.5-Instruct sizes and two non-Qwen 7–8B backbones (lower is better on all axes). (a) Unsup. and (b) MSE@H=336 against backbone size, with step changes between adjacent Qwen sizes and non-Qwen differences relative to Qwen-7B; the dashed line is Claude-Sonnet-4.5 1-shot. (c) Joint faithfulness–forecasting view of all backbones.

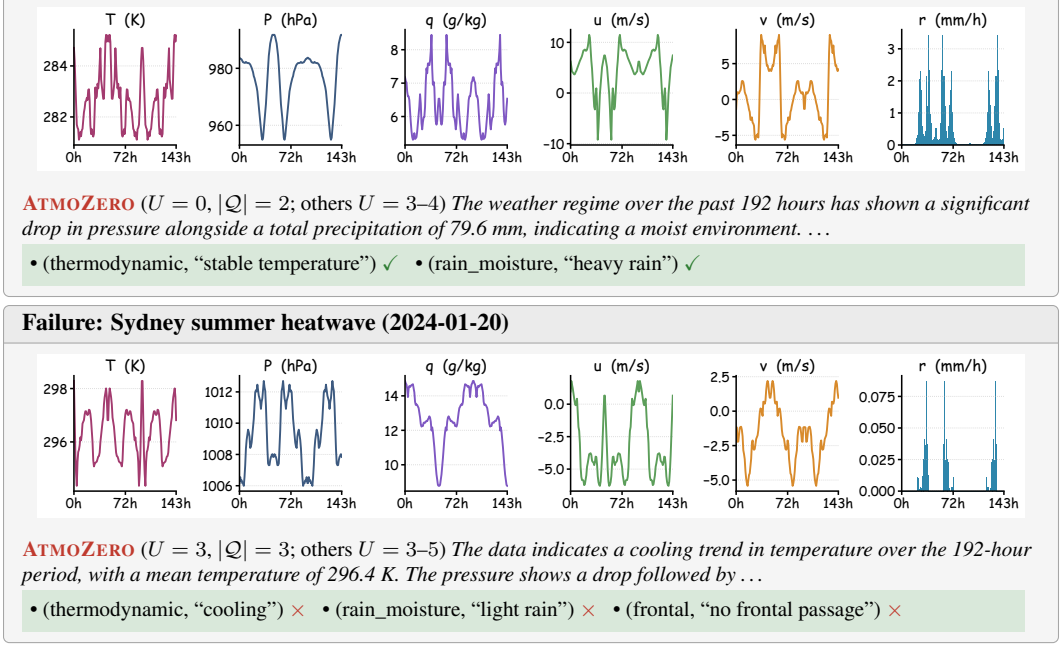


Figure 4: Success and failure cases on two ARCO-ERA5 windows. Each panel shows the six input channels, the truncated ATMOZERO caption, and its claims with verifier verdicts (✓ supported, × refuted). U counts refuted claims and $|\mathcal{Q}|$ counts extracted claims.

Robustness to Reward Hacking. We grade six adversarial caption policies, from empty captions to contradicted frontal claims, with three judges outside the training verifier, namely an independently calibrated verifier, an ERA5 gridded-state oracle, and meteorologist spot checks (Appendix D.5). **Table 9** reports the false-refutation rate of three trained variants. All three stay near 6% on natural captions, with ATMOZERO lowest at 5.8%. On adversarial captions, ATMOZERO rises only to 7.4%, whereas VERIFIER-ONLY reaches 16.8% and SOLVER-ONLY 13.4%. The rise from natural to adversarial captions is thus 1.6 pp for ATMOZERO, against 7.1 pp for SOLVER-ONLY and 10.7 pp for VERIFIER-ONLY. Since SOLVER-ONLY shares the full reward and differs only in its uniform Proposer, the adversarial Proposer accounts for most of this robustness.

Table 9: False-refutation rate (%).

Method	Natural	Adversarial
VERIFIER-ONLY	6.1	16.8
SOLVER-ONLY	6.3	13.4
ATMOZERO	**5.8**	**7.4**

Case Study. **Fig. 4** shows a success case and a failure case from Appendix D.14. During Windstorm Ciarán, the ATMOZERO caption reports the pressure drop and the 79.6 mm of rain of the storm passage, and the verifier supports both of its claims, whereas each of the seven other captioners has three or four refuted claims. In the Sydney summer heatwave window, the verifier refutes all three ATMOZERO claims and three to five claims of every other captioner. Appendices D.14 and D.15 give further caption and forecasting case studies, and Appendix D.1 gives additional analyses.

5 Conclusion

This paper proposes ATMOZERO, a self-play post-training framework for weather time-series captioners without human-written captions. Seven station-observable verifiers grade each claim against calibrated thresholds, and a refutation channel penalizes contradicted claims. An adversarial Proposer selects windows where the Solver is weakest, turning the verifier into a curriculum. Extensive experiments on ERA5 reanalysis and real-world stations show that ATMOZERO outperforms advanced baselines in faithfulness, forecasting, and human preference, and ablations attribute the gains to the refutation channel, the calibrated thresholds, and the Proposer–Solver loop. It offers a template for replacing human-written captions with executable scientific refutation in domains whose observations support refutable claims, such as hydrology, air quality, ocean monitoring, and space weather.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

[1] Shengchao Chen, Ting Shu, Huan Zhao, Qilin Wan, Jincan Huang, and Cailing Li. Dynamic multiscale fusion generative adversarial network for radar image extrapolation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[2] Shengchao Chen, Ting Shu, Huan Zhao, Guo Zhong, and Xunlai Chen. Tempee: Temporal–spatial parallel transformer for radar echo extrapolation beyond autoregression. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[3] Shengchao Chen, Guodong Long, Tao Shen, and Jing Jiang. Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. *arXiv preprint arXiv:2301.09152*, 2023.

[4] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022. URL `https://arxiv.org/abs/2202.11214`.

[5] 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.

[6] 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.

[7] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[8] Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. Foundation models for weather and climate data understanding: A comprehensive survey, 2023. URL `https://arxiv.org/abs/2312.03014`.

[9] Shengchao Chen, Guodong Long, Jing Jiang, and Chengqi Zhang. Personalized adapter for large meteorology model on devices: Towards weather foundation models. *Advances in Neural Information Processing Systems*, 37:84897–84943, 2024.

[10] Shengchao Chen, Guodong Long, Tao Shen, Jing Jiang, and Chengqi Zhang. Federated prompt learning for weather foundation models on devices. *arXiv preprint arXiv:2305.14244*, 2023.

[11] Shengchao Chen and Guodong Long. Federated weather modeling on sensor data. *arXiv preprint arXiv:2605.00322*, 2026.

[12] Shijie Zhang. Spectral-aware text-to-time series generation with billion-scale multimodal meteorological data, 2026. URL `https://arxiv.org/abs/2603.27135`.

[13] 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*, pages 17547–17573, 2025.

[14] Elizabeth Fons, Rachneet Kaur, Zhen Zeng, Soham Palande, Tucker Balch, Svitlana Vyetrenko, and Manuela Veloso. Tadacap: Time-series adaptive domain-aware captioning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 54–62, 2024.

[15] Marios Krestenitis, Christos Tzelepis, Konstantinos Ioannidis, Stefanos Vrochidis, Ioannis Kompatsiaris, Georgios Tzimiropoulos, Shaogang Gong, and Ioannis Patras. Cyclecap: Improving vlms captioning performance via self-supervised cycle consistency fine-tuning. *arXiv preprint arXiv:2603.18282*, 2026.

[16] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22934–22946. IEEE, 2025.

[17] Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 13066–13093, 2026.

[18] Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.

[19] Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with joint-grpo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38915–38925, 2026.

[20] Rui Li, Yuanzhi Liang, Ziqi Ni, Haibing Huang, Chi Zhang, and Xuelong Li. Rethinking reward signals in video grpo: When scores become targets. In *European Conference on Computer Vision*, pages 419–435. Springer, 2026.

[21] Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

[22] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In *International Conference on Learning Representations*, volume 2026, pages 130770–130790, 2026.

[23] Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play. *arXiv preprint arXiv:2509.25541*, 2025.

[24] Fangxu Yu, Hongyu Zhao, and Tianyi Zhou. Ts-reasoner: Aligning time series foundation models with llm reasoning. *arXiv preprint arXiv:2510.03519*, 2025.

[25] Mohamed Trabelsi, Aidan Boyd, Jin Cao, and Huseyin Uzunalioglu. Time series language model for descriptive caption generation. *Engineering Applications of Artificial Intelligence*, 162:112673, 2025.

[26] Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090, 2025.

[27] Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. *Advances in Neural Information Processing Systems*, 38:135431–135467, 2026.

[28] 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, pages 156654–156704, 2026.

[29] Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. Timemaster: Training time-series multimodal llms to reason via reinforcement learning. *arXiv preprint arXiv:2506.13705*, 2025.

[30] Yitong Zhou, Yucong Luo, Mingyue Cheng, Qi Liu, Jiahao Wang, Daoyu Wang, and Enhong Chen. Time series forecasting via reasoning: A slow-thinking approach with reinforcement fine-tuned llms, 2026. URL https://arxiv.org/abs/2506.10630.

[31] Tong Guan, Zijie Meng, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang, Qingsong Wen, Zuozhu Liu, Sabato Siniscalchi, Ming Jin, and Shirui Pan. Timeomni-1: Incentivizing complex reasoning with time series in large language models. In *International Conference on Learning Representations*, volume 2026, pages 152139–152170, 2026.

[32] Luca Zhou, Pratham Yashwante, Marshall Fisher, Alessio Sampieri, Zihao Zhou, Fabio Galasso, and Rose Yu. Cats-bench: Can language models describe time series? In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34479–34519, 2026.

[33] Medhasweta Sen, Zachary Gottesman, Jiaxing Qiu, C. Bayan Bruss, Nam Nguyen, and Tom Hartvigsen. Bedtime: A unified benchmark for automatically describing time series, 2026. URL https://arxiv.org/abs/2509.05215.

[34] Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Jimin Zhuang, Ke Fan, Qitan Lv, Junlin Liu, Chen Zhang, Zheqi Yuan, et al. Scits: Scientific time series understanding and generation with llms. In *International Conference on Learning Representations*, volume 2026, pages 114390–114431, 2026.

[35] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024.

[36] Ziyi Yang, Weizhou Shen, Chenliang Li, Ruijun Chen, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. Spell: Self-play reinforcement learning for evolving long-context language models. In *International Conference on Learning Representations*, volume 2026, pages 59910–59942, 2026.

[37] Zi Lin, Sheng Shen, Ilia Kulikov, Jingbo Shang, Jason Weston, and Yixin Nie. Learning to solve and verify: A self-play framework for code and test generation. *arXiv preprint arXiv:2502.14948*, 2025.

[38] Wanjia Zhao, Mert Yuksekgonul, Shirley Wu, and James Zou. Sirius: Self-improving multi-agent systems via bootstrapped reasoning. *Advances in Neural Information Processing Systems*, 38:124475–124504, 2026.

[39] Yulei Qin, Xiaoyu Tan, Zhengbao He, Gang Li, Haojia Lin, Zongyi Li, Zihan Xu, Yuchen Shi, Siqi Cai, Renting Rui, et al. Learn the ropes, then trust the wins: self-imitation with progressive exploration for agentic reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 65092–65131, 2026.

[40] Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. In *International Conference on Learning Representations*, volume 2026, pages 22555–22588, 2026.

[41] Xiaolong Tang, Meina Kan, Shiguang Shan, and Xilin Chen. Plan-r1: Safe and feasible trajectory planning as language modeling. In *International Conference on Learning Representations*, volume 2026, pages 23715–23735, 2026.

[42] Zhenxin Lei, Zhangwei Gao, Changyao Tian, Erfei Cui, Guanzhou Chen, Danni Yang, Yuchen Duan, Zhaokai Wang, Wenhao Li, Weiyun Wang, et al. Metacaptioner: Towards generalist visual captioning with open-source suites. In *International Conference on Learning Representations*, volume 2026, pages 102906–102945, 2026.

[43] Shuo Tang, Jian Xu, Jiadong Zhang, Yi Chen, Qizhao Jin, Lingdong Shen, Chenglin Liu, and Shiming Xiang. Meteorpred: A meteorological multimodal large model and dataset for severe weather event prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22910–22919, 2026.

[44] Shengchao Chen, Guodong Long, Jing Jiang, and Chengqi Zhang. Federated foundation models on heterogeneous time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15839–15847, 2025.

[45] Shengchao Chen, Guodong Long, Michael Blumenstein, and Jing Jiang. Fedal: Federated dataset learning for general time series foundation models. In *International Conference on Learning Representations*, volume 2026, pages 98850–98885, 2026.

[46] Shengchao Chen, Guodong Long, Dikai Liu, and Jing Jiang. Bi-level heterogeneous learning for time series foundation models: A federated learning approach. *arXiv preprint arXiv:2604.06727*, 2026.

[47] Zhong Siru, Ruan Weilin, Ming Jin, Li Huan, Wen Qingsong, and Liang Yuxuan. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Forty-Second International Conference on Machine Learning (ICML 2025)*. Proceedings of Machine Learning Research, 2025.

[48] Jialin Chen, Ziyu Zhao, Gaukhar Nurbek, Aosong Feng, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Trace: Grounding time series in context for multimodal embedding and retrieval. *Advances in Neural Information Processing Systems*, 38:2440–2472, 2026.

[49] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1):180214, 2018.

[50] Yunfeng Ge, Jiawei Li, Yiji Zhao, Haomin Wen, Zhao Li, Meikang Qiu, Hongyan Li, Ming Jin, and Shirui Pan. T2s: High-resolution time series generation with text-to-series diffusion models. *arXiv preprint arXiv:2505.02417*, 2025.

[51] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[52] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[53] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[54] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[55] Zhihao Dou, Qinjian Zhao, Zhongwei Wan, Dinggen Zhang, Weida Wang, Towsif Raiyan, Benteng Chen, Qingtao Pan, Yang Ouyang, Chaoda Song, Zhiqiang Gao, Shufei Zhang, and Sumon Biswas. Plan then action:high-level planning guidance reinforcement learning for llm reasoning, 2026. URL https://arxiv.org/abs/2510.01833.

[56] Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek Wong, and Yu Cheng. Exgrpo: Learning to reason from experience. In *International Conference on Learning Representations*, volume 2026, pages 12810–12846, 2026.

[57] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[58] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

[59] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pages 11116–11140, 2024.

[60] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pages 38626–38652, 2024.

[61] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pages 23857–23880, 2024.

[62] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

[63] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[64] OpenAI. GPT-4o system card, 2024. https://openai.com/index/gpt-4o-system-card/.

[65] Anthropic. Claude sonnet 4.5, 2025. https://www.anthropic.com/news/claude-sonnet-4-5.

[66] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2025. URL https://arxiv.org/abs/2401.10020.

[67] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pages 34635–34667, 2025.

[68] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

[69] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

[70] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models, 2024. URL https://arxiv.org/abs/2402.03885.

[71] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024. URL https://arxiv.org/abs/2403.07815.

[72] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[73] Robert W Carver and Alex Merose. Arco-era5: an analysis-ready cloud-optimized reanalysis dataset. In *American Meteorological Society Meeting Abstracts*, volume 103, pages 4A–1, 2023.

[74] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 12076–12100, 2023.

Roadmap. The appendix is organized as follows.

1. **Appendix A** specifies the seven rule families and the caption format, claim extraction, and prompt template (Appendix A.2).
2. **Appendix B** gives implementation details and compute (Appendix B.1), baseline descriptions (Appendix B.2), and the faithfulness metrics (Appendix B.3).
3. **Appendix C** covers corpus construction (Appendix C.1) and annotation (Appendix C.2).
4. **Appendix D** reports additional results (Appendix D.1), large-scale automatic evaluation (Appendix D.2), the human preference study (Appendix D.3), climate-zone OOD (Appendix D.4), reward hacking (Appendix D.5), reward-channel decomposition (Appendix D.6), extreme events (Appendix D.7), sample efficiency (Appendix D.8), robustness (Appendix D.9), the VERIFIER-TEMPLATER reference (Appendix D.10), temporal alignment (Appendix D.11), self-play baselines (Appendix D.12), representativeness error (Appendix D.13), and caption and forecasting case studies (Appendices D.14 and D.15).
5. **Appendix E** discusses limitations and broader impacts.

A Verifier Library

A.1 Verifier Rule Specifications

This appendix gives the formal specification of each rule family $k = 1, \dots, 7$. For every family we list the observed channels and the trigger predicate $g_k(x) \in \{0, 1\}$ that marks the family as fired on a window. We then define the per-claim grader $s_k(v_q; x, \mathcal{M}, \{x'_j\}) \in [0, 1]$, abbreviated $s_k(v_q)$ as in Section 3.1, which supplies S_k in Eq. 2, and the refutation predicate that drives U in Eq. 3. Each family also lists its controlled lexicon $\mathcal{V}_k$ with one positive and one refutation example. Throughout, $x = \{x_t\}_{t=1}^{T_w}$ denotes the focal $T_w = 192$-hour window, $\mathcal{M} = (\text{lat}, \text{lon}, h, \kappa)$ the station metadata, $\{x'_j\}_{j=1}^{K}$ the spatial neighbors, and $\bar{X}_\kappa$ the multi-year monthly climatology of zone κ. Thresholds (τ_k^-, τ_k^+) come from the validation set of Appendix B.1.

The seven rule families of the ATMOZERO verifier library

k=1 **Thermodynamic.** Sustained warming or cooling tendency, near-saturation specific humidity, and air-temperature monotonicity over the window, on channels T and q.

k=2 **Rain–Moisture.** Co-occurrence of measurable precipitation r with elevated specific humidity q and falling station pressure P, on channels r, q, P, T.

k=3 **Frontal.** Joint temperature change, wind shift, and pressure rebound consistent with a synoptic frontal passage, on channels T, u, v, P.

k=4 **Wind Regime.** Sustained regimes such as monsoonal southerlies or gust fronts, on u and v.

k=5 **Diurnal Cycle.** Phase and amplitude consistency of T, q, and r with the local solar cycle and station latitude, on channels T, q, r together with station metadata $\mathcal{M}$.

k=6 **Climatological.** Per-variable departure from the multi-year monthly climatology of the station's Köppen–Geiger zone, on every channel of $\mathcal{X}_{\text{stn}}$ together with zone κ.

k=7 **Spatial Coherence.** Concordance of claims with the K nearest stations, on $\{x'_j\}$ and $\mathcal{M}$.

Family $k = 1$ Thermodynamic	
Channels	T, q, $\mathcal{M}$.
Trigger	$g_1(x) = 1$ iff at least one of *(i)* $\lvert\text{slope}(T)\rvert \geq 0.20$ K/h over a sustained span of ≥ 12 hours, *(ii)* the specific humidity q_t exceeds $0.95\, q_\kappa^{\text{sat}}(T_t)$ for ≥ 6 consecutive hours, where $q_\kappa^{\text{sat}}(T)$ is the zone-specific saturation reference at temperature T, or *(iii)* $\text{sign}\big(\text{slope}_{[t-12,t]}(T)\big) \neq \text{sign}\big(\text{slope}_{[t,t+12]}(T)\big)$ at some interior t with both magnitudes ≥ 0.15 K/h.
Grader	for $v_q \in \mathcal{V}_1$, $s_1(v_q)$ is the fraction of windowed evidence that aligns with v_q in direction and magnitude under the corresponding sub-rule.

Refutation	$s_1 \le \tau_1^- = 0.18$ when the asserted direction is contradicted by the data (e.g., `(thermo, "warming trend")` while the observed slope is ≤ -0.5 K/h).
Lexicon	$\mathcal{V}_1 =$ {`warming trend`, `cooling trend`, `thermal stability`, `thermal reversal`, `dry air mass`, `humid air mass`}.
Example +	a 3.4 K drop over 18 hours at Beijing on 2023-12-15 supports `(thermo, "cooling trend")`.
Example −	`(thermo, "warming trend")` there is refuted by the negative slope.

Family $k = 2$ Rain–Moisture

Channels	r, q, P, T.
Trigger	$g_2(x) = 1$ iff $\sum_t r_t \ge 5$ mm over the window and there exists at least one rain hour $t^\star \in \{25, \dots, T_w - 24\}$ with $r_{t^\star} > 0$ such that $q_{t^\star} \ge q_\kappa^{\mathrm{ref}}(T_{t^\star}) - \theta_q$ and the local pressure trend $\mathrm{slope}_{[t^\star-24,\, t^\star+24]}(P) \le 0$, where q_κ^{ref} is the saturation-adjusted reference for zone κ and θ_q is the per-zone humidity tolerance.
Grader	joint score combining *(a)* precipitation-intensity bin alignment with v_q, *(b)* humidity-threshold satisfaction, and *(c)* pressure-trend sign.
Refutation	$s_2 \le \tau_2^- = 0.20$ when any of the three conditions is contradicted (e.g., asserted *heavy rain* while $\sum r < 1$ mm, or asserted *dry advection* while q is at its climatological maximum).
Lexicon	$\mathcal{V}_2$ = {`light rain`, `moderate rain`, `heavy rain`, `torrential rain`, `moist advection`, `dry advection`, `post-frontal drying`}.
Example +	$\sum r = 187$ mm and q at the 96th percentile during the Mumbai monsoon onset support `(rain, "torrential rain")`.
Example −	`(rain, "dry advection")` on this window fails the rain and humidity conditions.

Family $k = 3$ Frontal

Channels	T, u, v, P.
Trigger	$g_3(x) = 1$ iff there exists $t^\star \in \{13, \dots, T_w - 12\}$ such that *(i)* $T_{t^\star+12} - T_{t^\star-12} \le -3$ K (cold front) or ≥ 3 K (warm front), *(ii)* the wind direction $\mathrm{atan2}(v, u)$ rotates by $\ge 60^\circ$ across $[t^* - 6, t^* + 6]$, and *(iii)* $\min_{t \in [t^*-12, t^*+12]} P_t$ is at least 1 hPa below both endpoints.
Grader	conjunctive aggregation of the three sub-conditions, weighted by the magnitude of their excess over thresholds.
Refutation	$s_3 \le \tau_3^- = 0.15$ when fewer than two of the three sub-conditions are present (e.g., an asserted cold front with no wind shift and a flat pressure trace).
Lexicon	$\mathcal{V}_3 =$ {`cold front passage`, `warm front passage`, `occluded front`, `pre-frontal trough`, `post-frontal ridging`}.
Example +	at Plymouth during Windstorm Ciarán, $\Delta T_{24} = -4.8$ K, the wind veered by 108°, and P dropped by 22 hPa, supporting `(front, "cold front passage")`.
Example −	`(front, "warm front passage")` contradicts the observed temperature drop.

Family $k = 4$ Wind Regime

Channels	u, v.
Trigger	$g_4(x) = 1$ iff *(i)* $\left\|(\bar{u}, \bar{v})\right\|_2 \ge 5$ m/s sustained for ≥ 36 hours (regime), or *(ii)* the gust factor $\max_t \|\mathbf{u}_t\|_2 / \mathrm{mean}_t \|\mathbf{u}_t\|_2 \ge 2.2$ within a $\le$ 6-hour transient (gust front).
Grader	cosine alignment of the windowed mean direction with the lexicon-assigned canonical direction, modulated by sustained-magnitude satisfaction.
Refutation	$s_4 \le \tau_4^- = 0.22$ when the asserted regime direction is more than 90° from the windowed mean wind, or when the asserted gust front lacks any short-window magnitude excursion.
Lexicon	$\mathcal{V}_4$ = {`monsoonal southerly`, `trade easterly`, `westerly jet entrance`, `post-frontal northerly`, `gust-front transient`, `calm regime`}.
Example +	Mumbai's 7.8 m/s southerly supports `(wind, "monsoonal southerly")`.

Example −	`(wind, "trade easterly")` contradicts the observed direction.

Family $k = 5$ Diurnal Cycle

Channels	$T, q, r, \mathcal{M}$.
Trigger	$g_5(x) = 1$ iff the empirical daily-cycle amplitude of at least one channel $\chi \in \{T, q, r\}$ falls outside the climatological interquartile band for the station's calendar day, that is, exceeds the upper-quartile amplitude or stays below the lower-quartile amplitude, or a sub-daily anomaly coincides with the local solar terminator computed from (lat, lon).
Grader	phase- and amplitude-consistency score between the claim's asserted timing pattern and the empirical Fourier-1 component of the relevant channel.
Refutation	$s_5 \le \tau_5^- = 0.18$ when the asserted diurnal phase is ≥ 6 hours misaligned with the observed peak, or when an asserted nocturnal phenomenon falls during local daylight.
Lexicon	$\mathcal{V}_5 = \{$`strong diurnal cycle`, `suppressed diurnal cycle`, `nocturnal heating`, `afternoon convective peak`, `daybreak temperature minimum`$\}$.
Example +	a Cairo summer window with a T amplitude at the 97th climatological percentile and a clear afternoon peak supports `(diurnal, "strong diurnal cycle")`.
Example −	`(diurnal, "suppressed diurnal cycle")` contradicts this amplitude.

Family $k = 6$ Climatological

Channels	all of $\mathcal{X}_{\text{stn}}$, plus κ.
Trigger	$g_6(x) = 1$ iff the standardized departure $\zeta_\chi = (\bar{\chi}(x) - \bar{X}_\kappa^\chi)/\sigma_\kappa^\chi$ satisfies $\lvert\zeta_\chi\rvert \ge 1.5$ for at least one channel $\chi \in \mathcal{X}_{\text{stn}}$, with $\bar{X}_\kappa^\chi$ and σ_κ^χ computed from the 30-year ERA5 monthly climatology of zone κ.
Grader	sign- and magnitude-consistent z-score alignment with the lexicon's directional anomaly tag.
Refutation	$s_6 \le \tau_6^- = 0.20$ when the asserted anomaly direction has the wrong sign or when $\lvert\zeta_\chi\rvert < 0.5$ in the named channel.
Lexicon	$\mathcal{V}_6 = \{$`climatological warm anomaly`, `climatological cold anomaly`, `climatological wet anomaly`, `climatological dry anomaly`, `climatological pressure anomaly`, `near-normal regime`$\}$.
Example +	the Athens window during the Cerberus heatwave, with $\zeta_T = +3.1$ in zone Csa, supports `(climato, "climatological warm anomaly")`.
Example −	`(climato, "climatological cold anomaly")` has the opposite sign.

Family $k = 7$ Spatial Coherence

Channels	$\{x'_j\}_{j=1}^K, \mathcal{M}$.
Trigger	$g_7(x) = 1$ iff there exists at least one channel $\chi \in \mathcal{X}_{\text{stn}}$ whose focal-window summary statistic either agrees with the neighbor median up to a per-variable tolerance θ_χ (regional regime) or falls outside the K-neighbor interquartile range (locally anomalous), so that some scope predicate could fire on x.
Grader	concordance score between the claim's spatial scope (`regional`, `local`, `contrasted`) and the empirical neighbor distribution for the channel implied by the claim.
Refutation	$s_7 \le \tau_7^- = 0.20$ when an asserted regional phenomenon is contradicted by the neighbor median (e.g., `(spatial, "regional regime")` while $K-1$ of K neighbors sit at the climatological mean of that channel).
Lexicon	$\mathcal{V}_7 =$ $\{$`regional regime`, `locally anomalous`, `contrasted gradient`, `coherent advection`, `isolated event`$\}$.
Example +	the Mumbai monsoon-onset window with all eight neighbors above the rain threshold supports `(spatial, "regional regime")`.
Example −	`(spatial, "isolated event")` contradicts the concordant neighbors.

Summary. **Table 10** summarizes the seven families, the channels they use, the trigger semantics, the refutation threshold τ_k^-, and the lexicon size.

Table 10: Compact summary of the verifier library. Channels are the required input variables, and thresholds are estimated on the annotated validation set of Appendix B.1 and then frozen.

k	Family	Channels	Trigger semantics	τ_k^-	$\lvert\mathcal{V}_k\rvert$
1	Thermodynamic	$T, q, \mathcal{M}$	sustained slope / near-saturation q / reversal	0.18	6
2	Rain–Moisture	r, q, P, T	co-occurrence rule on r, q, P with $q_\kappa^{\mathrm{ref}}(T)$	0.20	7
3	Frontal	T, u, v, P	joint T change + wind shift + P trough	0.15	5
4	Wind Regime	u, v	sustained mean / gust-factor transient	0.22	6
5	Diurnal Cycle	$T, q, r, \mathcal{M}$	phase / amplitude vs. solar geometry	0.18	5
6	Climatological	$\mathcal{X}_{\mathrm{stn}}, \kappa$	standardized z-score vs. per-zone climatology	0.20	6
7	Spatial Coherence	$\{x'_j\}, \mathcal{M}$	concordance with $K = 8$ nearest neighbors	0.20	5

A.2 Caption Format, Claim Extraction, and Prompt Template

Structured Caption Format. A caption c produced by any evaluated captioner ends with a typed-claim list $\mathcal{Q}(c) = \{(t_q, v_q)\}_{q=1}^{|\mathcal{Q}(c)|}$, where $t_q \in \{1, \ldots, 7\}$ is the rule-family index of Appendix A and v_q is drawn from the controlled lexicon $\mathcal{V}_{t_q}$ formalized in Appendix A. Free prose precedes the `CLAIMS` list, whose typed claims $\mathcal{Q}(c)$ are scored by the reward and the claim-level metrics.

Claim Extraction. Claims are extracted from the `CLAIMS` list by deterministic pattern matching, which accepts the canonical short family tokens `thermo`, `rain`, `front`, `wind`, `diurnal`, `climato`, `spatial` and the corresponding long-form aliases `thermodynamic`, `rain_moisture`, `frontal`, `wind_regime`, `diurnal_cycle`, `climatological`, `spatial_coherence`, with the long-form aliases used in the case studies of Appendix D.14 for readability. Each extracted entry is matched against the union of controlled lexicons $\bigcup_k \mathcal{V}_k$. A match in $\mathcal{V}_{t_q}$ confirms the family index. When the surface form fails the exact-match step (e.g., a paraphrase of a lexicon entry such as *warming* for `warming trend`), a pretrained cross-encoder NLI model resolves the match by entailment against the canonical lexicon entry. If entailment also fails, the claim is dropped from $\mathcal{Q}(c)$ and counted as an extraction failure rather than a refutation.

Prompt Template. The same prompt template is used by every evaluated captioner, including the PROMPTED-BASE captioner (Qwen2.5-7B-Instruct), the closed-source 1-shot baselines (GPT-4o, Claude-Sonnet-4.5), and the ATMOZERO Solver π_S during both training and evaluation. Holding the template constant isolates the marginal effect of post-training from any wording artifacts in the prompt itself. The template contains four placeholders, $\langle\mathrm{WINDOW}\rangle$ for per-variable numerical summaries of the multivariate window x, $\langle\mathrm{META}\rangle$ for $\mathcal{M}$, $\langle\mathrm{NEIGHBORS}\rangle$ for numerical summaries of $\{x'_j\}$, and $\langle\mathrm{LEXICONS}\rangle$ for the controlled lexicons of Appendix A.

B Training, Baseline, and Metric Details

B.1 Implementation Details

Frozen Backbones. The frozen forecaster F_ψ is a PatchTST with patch length 16, stride 8, hidden size 256, 6 layers, and 8 heads, trained for 37 epochs with AdamW (lr 5×10^{-4}, batch 256, cosine decay) on the 2014–2022 training split with no caption conditioning. The frozen text-to-series decoder G_ϕ is an 8-layer DiT with hidden size 384 and 28 DDIM denoising steps, trained for 46 epochs with AdamW (lr 2×10^{-4}, batch 128) on (window, programmatic-template) pairs, where the templates are produced deterministically from the rule-family lexicons $\mathcal{V}_k$ over the same 2014–2022 windows. G_ϕ thus sees only programmatic templates, so neither it nor R_{cycle} sees human text.

Caption-to-Forecaster Conditioning. All caption-conditioned forecasters in **Table 2** share a single deterministic mapping from caption text to PatchTST input, applied identically to ATMOZERO, every captioner ablation, and every closed-source or open-weights captioner baseline. The caption is first tokenized by the Qwen2.5 tokenizer with a maximum length of 192 tokens, embedded by a frozen

ATMOZERO shared captioning prompt template

System role. You are a senior atmospheric scientist analyzing a 192-hour multivariate weather window from a single station and up to eight nearest neighbors. Describe what the data shows, then issue a CLAIMS list grounded only in the data. Do not invent phenomena that the data does not support.

User role.

(1) Window. Per-variable numerical summaries of the hourly T, P, q, u, v, r readings at the focal station over the past 192 hours, given as $\langle \text{WINDOW} \rangle$.

(2) Station metadata. $\mathcal{M} = \langle \text{META} \rangle$, including latitude, longitude, elevation, and Köppen–Geiger zone.

(3) Neighbor windows. Numerical summaries of up to eight nearest stations within 250 km, $\{x'_j\}_{j=1}^{K} = \langle \text{NEIGHBORS} \rangle$ with $K \leq 8$.

(4) Allowed lexicon. Use only the controlled vocabulary $\langle \text{LEXICONS} \rangle$, organized by the seven rule families (thermodynamic, rain–moisture, frontal, wind regime, diurnal cycle, climatological, spatial coherence).

Output format.

(1) One paragraph of factual prose summarizing the window. Do not include numerical predictions.

(2) A `CLAIMS` block, one claim per line, in the form `(<family>, "<lexicon entry>")`, where `<family>` is one of `thermo`, `rain`, `front`, `wind`, `diurnal`, `climato`, `spatial`, and `<lexicon entry>` is drawn verbatim from the corresponding $\mathcal{V}_k$.

Hard constraints.

- Issue a claim in family k only if the data supports it; omit the family if the evidence is borderline.
- Never invent claims outside the lexicons; never assert horizons longer than T_w.
- Do not hedge with uncertainty terms inside the `CLAIMS` block; the support / refutation grader is binary.

Figure 5: Shared prompt template of every captioner. Placeholders $\langle \cdot \rangle$ are filled with numerical summaries of the focal and neighbor windows, the station metadata, and the lexicons of Appendix A.

Qwen2.5-7B layer-4 encoder, and mean-pooled across tokens to a single 256-dimensional caption vector e_c. Patch tokens of the same dimension produced by PatchTST's patch encoder receive e_c as an additive bias before transformer block 1, so the caption enters as an additive bias shared across patches rather than through cross-attention. The conditioning is identical across captioners, and a caption without an extractable typed-claim list enters the forecaster as ordinary text without penalty.

Cold-Start Threshold. The cold-start threshold τ_{cold} is set so that the Proposer engages only after the Solver has reached a verifier-precision level at which random window selection is no longer a useful proxy for hardness. We measure verifier precision on an 800-window probe set drawn from the validation split and choose $\tau_{\text{cold}} = 0.84$ as the inflection of the precision–variance curve.

Threshold Estimation and Reward-Exploitation Monitoring. The per-rule thresholds (τ_k^-, τ_k^+) are estimated on the same 200-window annotated validation set used for (λ, μ, ν) in Appendix B.3. For each family k, the precision and recall of the binary supported/refuted decisions implied by (τ_k^-, τ_k^+) are traced against the post-adjudication annotator labels of Appendix C.2. The support threshold τ_k^+ is chosen as the smallest grader value at which family-k support precision exceeds 0.90,

Table 11: Full hyperparameters for ATMOZERO training.

Setting	Value
Initialization and hardware	
Solver / Proposer base	Qwen2.5-7B-Instruct [54]
GPU	$4\times$ NVIDIA H100 80GB
Distributed training	FSDP-2 full-shard, bf16, activation checkpointing, no CPU offload
Maximum sequence length	2048
Optimizer (AdamW)	
Learning rate (peak)	4.8×10^{-6}
$(\beta_1, \beta_2, \epsilon)$	$(0.9,\ 0.95,\ 10^{-8})$
Weight decay	0.08
Schedule	cosine with 196 warmup steps
Total optimizer steps	2,450
GRPO	
Group size G	8
Microbatch (per device)	2 windows
Gradient accumulation	8
Effective optimizer batch	512 captions / 0.69M tokens
Clip ratio	0.20
Reference KL coefficient	0.025
Reward composition (Eq. 5)	
Per-rule weight w_k	$1/7$ for $k = 1, \ldots, 7$
Refutation weight λ	2.0
Coverage weight μ	0.22
Specificity weight ν	0.06
Auxiliary schedules (Eq. 7)	
Cycle weight	$\alpha_t = 0.42\ \exp(-t/610)$
Uplift weight	$\beta_t = 0.31\ \exp(-t/740)$
Clamp to zero	after step 1,960
Cold-start threshold τ_{cold}	0.84 verifier precision
Frozen backbones	
Forecaster F_ψ	PatchTST: patch 16, stride 8, hidden 256, 6 layers, 8 heads, 37 epochs
Decoder G_ϕ	8-layer DiT, hidden 384, 28 DDIM steps, 46 epochs
Validation set for threshold and weight estimation	
Annotated calibration windows	200 windows $\times$ 3 captions
Annotators	6 PhD-level annotators

and the refutation threshold τ_k^- is chosen as the largest grader value at which family-k refutation precision exceeds 0.90 and refutation recall exceeds 0.85, jointly bounding false and missed refutations. Reward exploitation is monitored during training by four metrics, namely coverage entropy across rule families, type–token ratio of caption tokens, unsupported-claim rate against the verifier, and pairwise rule-orthogonality of supported claims; training is stopped if any of them degrades by more than 20% from its rolling median over the previous 200 steps. Early stopping is also triggered if the KL divergence to the reference policy exceeds 0.07 for 50 consecutive steps.

Compute Profile. **Table 12** summarizes the cost. Each optimizer step uses 512 captions (G=8 rollouts over 64 prompts, about 0.69M tokens), giving about 1.25M captions and 1.69B tokens per seed, and rollout generation dominates the wall-clock time. At inference only the 7B Solver is retained and the caption enters PatchTST as a single 256-dimensional additive bias, so inference cost matches PROMPTED-BASE without the per-query cost of commercial models.

Table 12: Compute profile of ATMOZERO on $4\times$ NVIDIA H100 80GB GPUs.

Item	Value
Post-training per seed (2,450 steps, both policies)	$\approx$ 790 H100-hours
One-time pretraining of frozen F_ψ and G_ϕ (shared by all runs)	$\approx$ 30 H100-hours
Main results (3 seeds, excluding one-time pretraining)	$\approx$ 2,370 H100-hours
Inference (Solver only, 7B)	$\approx$ 2.6 s per caption on one H100
Verifier grading (optional, post hoc)	$\approx$ 12 ms per caption on CPU
Additional inference parameters	None relative to PROMPTED-BASE

B.2 Baseline Methods

This appendix expands the Baselines paragraph (Section 4) into self-contained descriptions for each method appearing in **Tables 1**, **2**, **7**, **16**, and **24**. Methods are grouped by family.

Parity Conditions Across Caption Sources. Every caption source uses the shared prompt template and claim extractor of Appendix A.2. Captions whose `CLAIMS` list cannot be extracted (0.6% of PROMPTED-BASE outputs and 0.0–0.3% of the outputs of every other captioner) are scored as if every rule family were `Omitted`, which treats structured and unstructured captioners identically. Caption-conditioned forecasters use the identical caption conditioning of Appendix B.1, so the only quantity that varies between caption-augmented rows of **Table 2** is the caption itself. All published baselines (numerical-only forecasters, caption-conditioned forecasters with closed-source captioners, time-series language and reasoning systems, and zero-shot foundation models) use the default hyperparameters and reference implementations of their original publications without tuning.

Numerical-Only Forecasters.

- **PatchTST** [51]. Channel-independent transformer that splits each channel into overlapping patches and encodes them with a shared encoder. Its frozen instance serves as the backbone F_ψ of all caption-conditioned forecasters, and NO-CAPTION runs it without captions.
- **TimesNet** [58]. Two-dimensional reshaping of the time series into multi-period blocks, processed by Inception-style convolutions. Trained from scratch on the 2014–2022 split without captions.
- **DLinear** [57]. Two-branch linear baseline (decomposition into trend and seasonality, each fitted by a single dense layer). Tests whether long-horizon forecasting reduces to linear extrapolation.
- **iTransformer** [59]. Inverted-attention transformer that treats each variable as a token and attends across variables rather than across time; included as a strong multivariate baseline.
- **TimeMixer** [60]. Multi-scale mixer that decomposes the input into seasonal and trend components and mixes patch-level representations across scales; a recent state-of-the-art TS-only forecaster.

Caption-Conditioned Forecasters with Closed-Source Captioners.

- **GPT-4o 1-shot**. The frozen PatchTST backbone is conditioned on a caption from a 1-shot GPT-4o prompt that uses the shared template of Appendix A.2. No weights are updated.
- **Claude-Sonnet-4.5 1-shot**. Identical prompt template with Anthropic's Claude Sonnet 4.5 as the captioner. Provides the strongest closed-source captioner reference.

Prompted Open-Weights Captioner.

- **PROMPTED-BASE**. The Qwen2.5-7B-Instruct base model [54] prompted with the same structured-`CLAIMS` instruction but *without* any post-training. This is the directly comparable inference-only reference for ATMOZERO, isolating the marginal effect of post-training from the marginal effect of choosing an open-weights backbone.

Time-Series Language and Reasoning Systems.

- **Time-LLM** [61]. Reprograms a frozen language-model backbone for forecasting by patch-token alignment with input–output text prototypes. A widely used TS-LLM reprogramming baseline.

- **TimeXL** [27]. Multi-agent prompting framework that combines a forecaster with an explanation agent at inference; included as the most directly comparable agentic baseline.
- **GPT4TS** [62]. Frozen pretrained transformer fine-tuned only on its layer-norm and positional embeddings; a widely used reference of moderate cost.

Reinforcement-Learning Ablations Within the Same Framework. All three ablations share the Qwen2.5-7B-Instruct initialization, the hyperparameters of **Table 11** (steps and batch size), three seeds, and the evaluation protocol. They differ only in the reward or the Proposer.

- **DENSE-ONLY**. Sets $w_k = \lambda = \mu = \nu = 0$ in Eq. 5, removing every verifier-derived signal from R, and trains under the auxiliary cycle and forecaster-uplift channels only. This is the dense-channel reward design typical of vision caption-RL transposed to weather time series and is the closest pre-ATMOZERO RL captioner.
- **VERIFIER-ONLY**. Removes the auxiliary cycle and uplift channels from Eq. 7 ($\alpha_t \equiv \beta_t \equiv 0$). Trained only on the support, refutation, and coverage signals. Used to attribute gains specifically to the verifier rather than to dense reward shaping.
- **SOLVER-ONLY**. Restores the full reward but fixes the Proposer to uniform window sampling, $\pi_P^{(t)} = \mathcal{U}(\mathcal{D})$ for all t. Used to attribute gains to the adversarial Proposer rather than the verifier.

Non-Learned Verifier Reference.

- **VERIFIER-TEMPLATER**. The programmatic-template generator of Appendix B.1 that produces the training pairs of G_ϕ, used as a captioner. For every rule family whose trigger predicate fires on the window, it outputs the lexicon phrase with the highest grader score, so its `CLAIMS` list is a direct verbalization of the verifier output rather than a learned free-form caption. It involves no training and uses the same claim extraction and caption conditioning as every other caption source. Used to separate verifier-library faithfulness from what the learned captioner adds.

Adapted Self-Play Baselines. Absolute Zero [21] verifies proposals with a code executor, R-Zero [22] with majority-vote pseudo-labels, and Self-Rewarding Language Models [66] with self-judgment, and none of them grades claims against sensor windows. We therefore transfer the R-Zero and Self-Rewarding mechanisms to our task under the same backbone, training data, rollout budget, and number of optimizer updates as ATMOZERO.

- **R-Zero-inspired**. Implements capability-edge sampling as rollout self-consistency over windows in place of majority-vote pseudo-labels, with a difficulty filter and no refutation channel.
- **Self-Rewarding-inspired**. Replaces the executable verifier with the backbone scoring its own captions, following Self-Rewarding Language Models [66].

Full Method.

- **ATMOZERO** (ours). The full Proposer–Solver method with all reward channels, full-parameter post-training, the cold-start threshold $\tau_{\text{cold}} = 0.84$, and the schedule of **Table 11**.

Zero-Shot Time-Series Foundation Models.

- **Time-MoE** (*base*, *large*) [67]. Mixture-of-experts decoder pretrained on a multi-domain time-series corpus; both checkpoints frozen at inference.
- **Moirai** (*small*, *base*, *large*) [68]. Masked-encoder universal forecaster with any-variate attention over flattened multivariate inputs; the three sizes isolate the effect of model scale.
- **TimesFM** [69]. Decoder-only foundation forecaster pretrained on large public time-series corpora.
- **Moment** [70]. Encoder-only time-series foundation model with masked-patch reconstruction pretraining; included as a non-autoregressive reference.
- **Chronos** (*small*, *base*, *large*) [71]. Quantization-and-LM-style probabilistic forecaster from a discretized vocabulary over time-series tokens.

The ten foundation models above (counting the two Time-MoE sizes, three Moirai sizes, TimesFM, Moment, and three Chronos sizes) are evaluated zero-shot on the fifteen real-world surface-station datasets, exactly as released; ATMOZERO is likewise evaluated zero-shot without fine-tuning.

B.3 Caption-Faithfulness Metrics

This appendix formalizes the eight metrics reported in **Table 1**. Throughout, let $\mathcal{D}^{\text{eval}}$ denote the human-annotated set of $|\mathcal{D}^{\text{eval}}| = 200$ windows, and for each window $x \in \mathcal{D}^{\text{eval}}$ let $c(x)$ be the caption produced by the method under evaluation and $\mathcal{Q}(c) = \{(t_q, v_q)\}$ its typed-claim list (Eq. 1). Each issued claim $q \in \mathcal{Q}(c)$ receives an annotator label $\ell_q \in \{S, R\}$ for `Supported` or `Refuted`, and the label `Omitted` applies at the family level, namely family k is omitted on window x if $g_k(x) = 1$ but $\mathcal{Q}_k(c) = \varnothing$. We write $\mathbf{1}[\cdot]$ for the indicator and $\mathbb{E}_x[\cdot]$ for the empirical mean over $\mathcal{D}^{\text{eval}}$.

Cross-Fit Protocol Between Threshold Estimation and Evaluation. Reward weights (λ, μ, ν) used during GRPO training are estimated once on the full 200-window set from the labels of Appendix C.2, and the same scalars are then frozen for the entire training run reported in the paper. For evaluation, we use a stratified 5-fold cross-fit so that the per-fold verifier thresholds $\tau_k^{\pm}$ used to compute **Table 1** are estimated on 160 windows disjoint from the 40 windows being graded, and the five folds are rotated to cover all 200 windows. The training checkpoint is fixed across folds; the cross-fit separates evaluation from the labels that estimate the thresholds. The spread across the five fold-level estimates stays within the reported $\pm$ half-width in every cell, with the ordering unchanged.

Window-Level Paired Bootstrap Confidence Intervals. The $\pm$ values of the faithfulness metrics in **Tables 1** and **4** are 95% confidence half-widths from a paired bootstrap with $1{,}000$ resamples drawn at the *window* level, not the annotator or claim level. Window-level resampling respects the dependence among the captioner outputs of a given window and the two-to-three annotator labels per claim. Annotator-level resampling conflates label noise with environmental coverage, and caption-level resampling overstates the effective sample size. The bootstrap is paired across captioners on the same resampled windows, so between-method differences are matched.

Verifier-Independent Variant of Unsupported-Claim Rate. To separate the primary metric from the verifier, we recompute Unsup. (Eq. 11) from *annotator labels only* for every captioner of **Table 1**, dropping its verifier-predicate clause. **Table 13** reports both variants on the cross-fit set. The ATMOZERO-versus-DENSE-ONLY gap is 13.1 pp under the main-text definition and 12.4 pp under the human-only definition, and the captioner ordering is unchanged.

Table 13: Unsupported-claim rate under the main-text definition (*Either*, annotator or verifier predicate) and the human-only definition (*Human-only*, annotator label only). Δ = Human-only $-$ Either.

Method	Unsup. Either (↓)	Unsup. Human-only (↓)	Δ (pp)
PROMPTED-BASE	$38.6_{\pm 2.4}$	$37.1_{\pm 1.7}$	-1.5
DENSE-ONLY	$22.3_{\pm 1.0}$	$20.9_{\pm 1.4}$	-1.4
GPT-4o 1-shot	$27.4_{\pm 1.6}$	$27.1_{\pm 1.5}$	-0.3
Claude-Sonnet-4.5 1-shot	$20.8_{\pm 1.2}$	$20.4_{\pm 1.3}$	-0.4
VERIFIER-ONLY	$11.7_{\pm 0.7}$	$10.6_{\pm 1.1}$	-1.1
SOLVER-ONLY	$12.5_{\pm 1.1}$	$11.4_{\pm 1.0}$	-1.1
ATMOZERO	$\mathbf{9.2}_{\pm 0.85}$	$\mathbf{8.5}_{\pm 1.0}$	-0.7

Threshold-Estimation Set-Size Sensitivity. Holding the cross-fit protocol fixed, we vary the per-fold threshold-estimation set size in $\{100, 160, 200\}$ windows for ATMOZERO only. The smallest size sub-samples the 160-window default and the largest corresponds to an in-sample fit on all 200 windows. The reported Unsup. values are $9.6_{\pm 1.0}$, $9.2_{\pm 0.85}$, and $9.7_{\pm 0.8}$ at $n_{\text{est}} \in \{100, 160, 200\}$, respectively, and the 160-window default sits at the empirical optimum of the bias–variance trade-off.

Unsupported-Claim Rate (Unsup., ↓). The fraction of caption claims adjudicated as not supported,

$$\text{Unsup.} \;=\; \mathbb{E}_x\left[\frac{\sum_{q \in \mathcal{Q}(c(x))} \mathbf{1}\big[\ell_q = R \text{ or } s_{t_q}(v_q) \le \tau_{t_q}^{-}\big]}{|\mathcal{Q}(c(x))|}\right], \tag{11}$$

with the convention that windows with $\mathcal{Q}(c(x)) = \varnothing$ are dropped from the empirical mean. A claim is unsupported when its annotator label is `Refuted` or when the verifier triggers a refutation predicate.

Supported-Claim Coverage (Coverage, ↑). The number of phenomena triggered by the verifier (cf. Appendix A, $g_k(x) = 1$) that the caption verbalizes with a supported claim, out of the seven rule families,

$$\text{Coverage} \;=\; \mathbb{E}_x\left[\sum_{k=1}^{7} \mathbf{1}[g_k(x)=1] \cdot \mathbf{1}\big[\exists q \in \mathcal{Q}_k(c(x)) \text{ with } \ell_q = S\big]\right], \tag{12}$$

where $\mathcal{Q}_k(c) = \{q \in \mathcal{Q}(c) \mid t_q = k\}$, with the convention that windows on which no rule family fires ($\sum_k g_k(x) = 0$) are dropped from the empirical mean. Higher Coverage means the captioner names more of the phenomena that fire. Unlike the reward term C of Eq. 4, a fraction of families with a verifier-supported claim, this metric counts fired families with an annotator-supported claim.

Lexical Type–Token Ratio (TTR, ↑). Following standard lexical-diversity practice, we compute TTR over content tokens only, with claim-bearing tokens weighted twice. Let $W(c)$ be the multiset of content tokens (i.e., excluding stopwords from a fixed list) in caption c, and let $W^*(c)$ promote claim-bearing tokens to weight 2,

$$\text{TTR}(c) \;=\; \frac{|\operatorname{set}(W^*(c))|}{|W^*(c)|}, \qquad \text{TTR} = \mathbb{E}_x\big[\text{TTR}(c(x))\big]. \tag{13}$$

Pairwise Rule-Orthogonality (Orth., ↑). For each window we form the binary vector $\mathbf{1}_k(c,x) = \mathbf{1}\big[\exists q \in \mathcal{Q}_k(c) \text{ with } \ell_q = S\big]$ for $k = 1, \ldots, 7$. The rule-orthogonality is the mean off-diagonal cosine *distance* of these vectors over $\mathcal{D}^{\text{eval}}$,

$$\text{Orth.} \;=\; 1 - \frac{2}{7 \cdot 6} \sum_{k<k'} \frac{\mathbb{E}_x[\mathbf{1}_k \mathbf{1}_{k'}]}{\sqrt{\mathbb{E}_x[\mathbf{1}_k]\,\mathbb{E}_x[\mathbf{1}_{k'}]} \;+\; \delta}, \tag{14}$$

with a smoothing constant $\delta = 10^{-3}$ guarding zero-marginal family pairs. Always co-firing the same family pairs yields low Orth., and selecting the active subset per window yields high Orth.

Claim Density (Claims, ↑). Number of typed claims per 100 caption tokens,

$$\text{Claims} \;=\; \mathbb{E}_x\left[\frac{100 \cdot |\mathcal{Q}(c(x))|}{|\text{tokens}(c(x))|}\right]. \tag{15}$$

False-Refutation Rate (False ref., ↓). Fraction of verifier refutations across the evaluation corpus that the annotators judge incorrect, that is, the verifier marks $s_{t_q}(v_q) \le \tau^-_{t_q}$ but the human label is $\ell_q = S$. Let $\mathcal{R} = \{(x,q) \mid x \in \mathcal{D}^{\text{eval}},\, q \in \mathcal{Q}(c(x)),\, s_{t_q}(v_q) \le \tau^-_{t_q}\}$ denote the set of refuted (window, claim) pairs across the corpus. Then

$$\text{False ref.} \;=\; \frac{\sum_{(x,q)\in\mathcal{R}} \mathbf{1}[\ell_q = S]}{|\mathcal{R}|}, \tag{16}$$

where captioners with $|\mathcal{R}| = 0$ are marked undefined and excluded.

Inter-Annotator Agreement (κ, ↑). Fleiss κ across the six annotators [72], computed per claim over the binary label set $\{S, R\}$ at the claim level and over the binary family-omitted indicator at the family level, and aggregated by macro-average over rule families. We report the pre-adjudication value, before majority-vote reconciliation, to avoid inflating agreement.

NLI Entailment Rate (NLI, ↑). For each claim q in caption $c(x)$, we render a hypothesis sentence `"the window exhibits` v_q `behavior."` and a premise paragraph composed of natural-language renderings of the rule-family-t_q trigger-predicate evidence on x. An independent off-the-shelf cross-encoder entailment classifier $\mathcal{E}$ returns a probability of entailment, and we threshold at 0.5,

$$\text{NLI} \;=\; \mathbb{E}_x\left[\frac{1}{|\mathcal{Q}(c(x))|} \sum_{q \in \mathcal{Q}(c(x))} \mathbf{1}\big[\mathcal{E}(\text{prem}_q, \text{hyp}_q) \ge 0.5\big]\right]. \tag{17}$$

$\mathcal{E}$ is held fixed across all evaluated methods to remove a confound between caption and verifier, and windows with empty $\mathcal{Q}(c(x))$ are dropped from the NLI mean using the same convention as Unsup.

C Dataset Construction and Annotation

C.1 Dataset Construction and Preprocessing

Source. All training, validation, and test data are derived from the ERA5 hourly single-level reanalysis released by the Copernicus Climate Change Service (C3S) at ECMWF [63]. ERA5 is regridded to a regular $0.25^\circ \times 0.25^\circ$ latitude–longitude grid with hourly cadence and is the de facto reference reanalysis for surface-station-equivalent multivariate time series at this scale.

Spatial Domain and Virtual Station Selection. The spatial domain is the global land surface, with virtual stations restricted to ERA5 grid cells whose land–sea fraction exceeds 0.5. We select $N = 8{,}192$ ERA5 grid points as virtual stations, stratified across all thirty Köppen–Geiger zones [49] of the global land mask, with per-zone counts proportional to zone area and capped to ensure that no single zone exceeds 15% of the network. A minimum inter-station spacing of 1° (~ 111 km at the equator) prevents redundancy. Each virtual station carries metadata $\mathcal{M} = (\mathrm{lat}, \mathrm{lon}, h, \kappa)$ with elevation h taken from the ETOPO1 1-arc-minute relief model collocated to the ERA5 grid.

Variables. The canonical six-variable surface set $\mathcal{X}_{\mathrm{stn}} = \{T, P, q, u, v, r\}$ is derived from the ERA5 single-level fields as follows. Air temperature T is the 2-meter temperature `t2m`; station pressure P is mean sea-level pressure `msl` adjusted to station elevation by the hydrostatic correction $P = \texttt{msl} \cdot \exp(-h/H_s)$ with scale height $H_s = 8{,}400$ m. Specific humidity q is computed from the 2-meter dewpoint `d2m` via the Tetens saturation-pressure formula and the standard mixing-ratio–specific-humidity identity. The zonal and meridional wind components u, v are taken from the 10-meter wind components `u10`, `v10`. Past-hour accumulated precipitation r is the hourly total-precipitation flux `tp` multiplied by the one-hour interval and converted to millimeters. All variables are linearly de-trended on the 2014–2022 training period before windowing, with the trend slopes saved alongside the dataset to enable inverse de-trending at inference.

Windowing and Splits. Each station produces overlapping windows of length $T_w = 192$ hours (eight days) with stride 12 hours, yielding approximately 730 raw windows per station per year. The window length keeps the joint span $T_w + H$, with H up to 720 hours, within a single ERA5 production stream. Quality-control rules drop any window in which r exceeds the climatological 99.97th percentile of the local zone or in which any variable contains a reanalysis discontinuity at a production-stream boundary. After quality control, the train, validation, and test splits cover 2014–2022, 2023, and 2024–2025, respectively, contributing approximately 4.80×10^7, 5.33×10^6, and 1.06×10^7 windows. Forecasting evaluation horizons of 96, 192, 336, and 720 hours are aligned to the end of each window. The 200-window annotated set used for threshold and weight estimation (**Table 11**) is a stratified random sample of the validation split, annotated as in Appendix C.2.

Neighbor Graph for the Spatial-Coherence Verifier. The spatial-coherence rule family ($k = 7$, Appendix A) requires nearest-neighbor windows. For each station we build the neighbor set $\{x'_j\}_{j=1}^{K}$ from the $K = 8$ nearest virtual stations within a 250 km great-circle radius, with fewer than K neighbors at the network edges; the radius is chosen so that the median in-zone neighbor count remains at least four across all thirty Köppen–Geiger zones. Neighbor windows are time-aligned to the focal window and undergo identical de-trending and quality control.

Data Licensing. ERA5 is distributed under the Copernicus Licence and freely available from the Copernicus Climate Data Store [63] and the analysis-ready ARCO-ERA5 distribution [73].

Dataset Statistics. **Table 14** summarizes the statistics of the processed corpus.

Table 14: Corpus-level statistics of the ERA5-derived training corpus used by ATMOZERO.

Property	**Value**
Source	ERA5 hourly single-level reanalysis [63, 73]
Spatial resolution	$0.25^\circ \times 0.25^\circ$, hourly
Spatial domain	global land surface (land–sea fraction > 0.5)
Number of virtual stations N	8,192
Köppen–Geiger zones covered	all 30 main zones
Minimum inter-station spacing	1° (~ 111 km at the equator)
Variables $\lvert\mathcal{X}_{\mathrm{stn}}\rvert$	6 (T, P, q, u, v, r)
Window length T_w	192 hours (eight days), stride 12 hours
Forecast horizons H	$\{96, 192, 336, 720\}$ hours
Train split (2014–2022)	$\sim 4.80 \times 10^7$ windows, 9 years
Validation split (2023)	$\sim 5.33 \times 10^6$ windows, 1 year
Test split (2024–2025)	$\sim 1.06 \times 10^7$ windows, 2 years
Calibration set for $(\tau_k^{\pm}, \lambda, \mu, \nu)$	200 windows $\times$ 3 captions, 6 PhD-level annotators
Spatial neighbor graph $\{x'_j\}$	$K = 8$ within 250 km, median in-zone neighbors ≥ 4

Severe-Weather Examples. **Fig. 6** displays eight canonical severe-weather windows drawn directly from the corpus and selected to span four continents, both hemispheres, six Köppen–Geiger zones, and complementary rule families of the verifier library. The eight windows are the landfall of Typhoon Doksuri at Xiamen on 2023-07-28 (rain–moisture, frontal, wind regime), the Big Bend landfall of Hurricane Idalia at Perry on 2023-08-30 (rain–moisture, frontal, wind regime, climatological), the North China cold-wave passage at Beijing on 2023-12-15 (thermodynamic, frontal), the North American polar-vortex outbreak at Chicago on 2024-01-14 (thermodynamic, climatological), the Mediterranean heatwave Cerberus at Athens on 2023-07-23 (thermodynamic, climatological), Windstorm Ciarán at Plymouth on 2023-11-02 (frontal, wind regime), the Indian monsoon onset at Mumbai on 2023-06-25 (rain–moisture, climatological, diurnal cycle), and the Sydney summer heatwave on 2024-01-20 (thermodynamic, climatological).

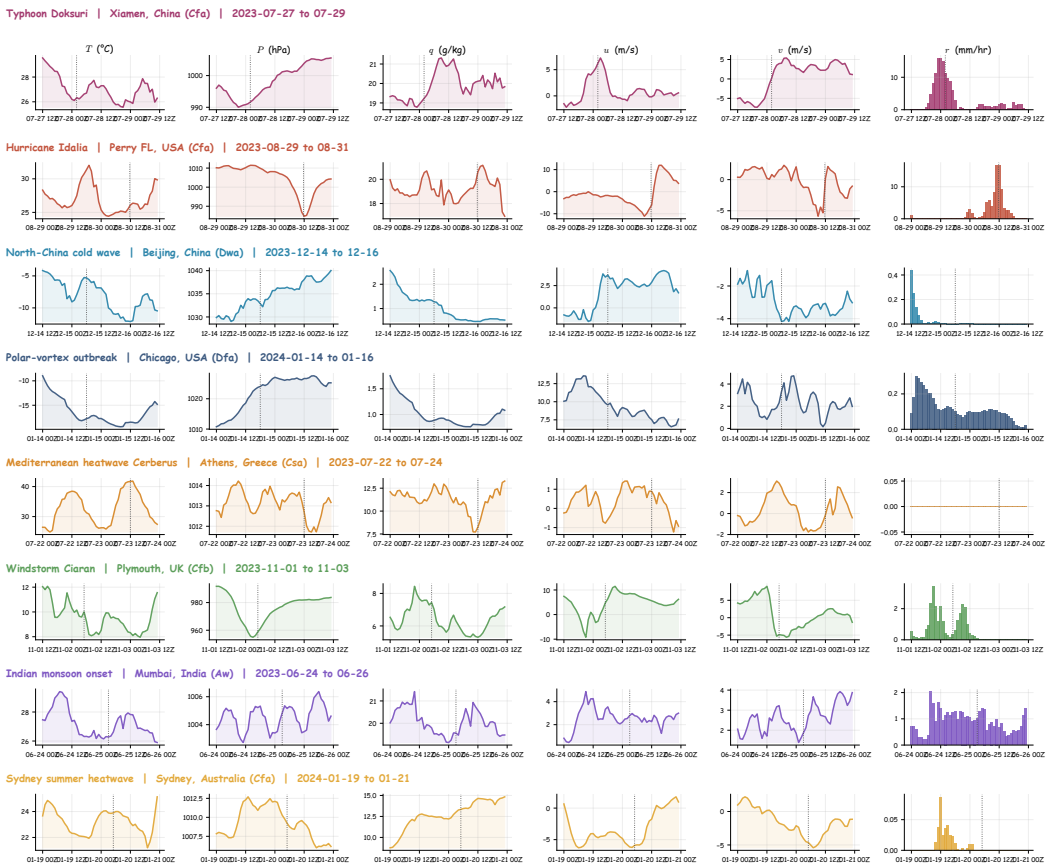


Figure 6: Real ERA5 hourly traces for eight severe-weather windows from the ATMOZERO corpus, shown as a 48-hour zoom around each event reference time. Rows are events; columns are the six channels in $\mathcal{X}_{\text{stn}}$. Training windows are $T_w = 192$ hours. Dashed lines mark the event time.

C.2 Annotation Protocol and Guidelines

The 200-window set of **Table 11** is annotated by six PhD-level annotators in atmospheric science, climate dynamics, or computational meteorology. Each of its 200 windows carries three captioner outputs, one from PROMPTED-BASE and two from early ATMOZERO checkpoints of independent seeds. The protocol below was finalized before any captioner output was inspected and governs both the 200×3 labels for threshold and weight estimation and the evaluation of **Table 1**.

Procedure. Each window is independently labeled by two annotators drawn at random from the panel of six, with a third annotator adjudicating only when the first two disagree on at least one claim. Annotators see only the multivariate window x, the station metadata $\mathcal{M}$, the neighbor set $\{x'_j\}$, the rule-family catalog (Appendix A), and the natural-language caption under evaluation, in randomized order, with the captioner identity concealed. The pre-adjudication Fleiss κ reported in **Table 1** is computed before the third annotator intervenes, and the post-adjudication majority vote gives the label used by the label-dependent metrics of Appendix B.3. Before the evaluation, every annotator labels a 24-window pilot set, and the panel meets to resolve disagreements on how to apply the guidelines below. The pilot windows are then discarded.

ATMOZERO annotator guidelines

Label set. For every typed claim $q = (t_q, v_q)$ produced by the captioner, assign exactly one of the three labels below.

Supported (S). The data x, the metadata $\mathcal{M}$, or the neighbor set $\{x'_j\}$ exhibit phenomenon v_q in family t_q, judged from the meteorological reading of the window without consulting the verifier graders or thresholds. The grader values $s_{t_q}(v_q, x)$ are not displayed to annotators.

Refuted (R). The asserted phenomenon contradicts the data, the metadata, or the neighbor set. Typical refutations include claiming warming while the slope is strictly negative, claiming heavy rain on a window with $\sum r < 1$ mm, or claiming a regional regime in family $k = 7$ while $K-1$ of K neighbors sit at the climatological mean.

Omitted (O). A rule family k fires, $g_k(x) = 1$, but the caption issues no claim in family k. Omission applies at the rule-family level, not at the individual-claim level.

Resolution rules.

(i) When the data are noisy, missing, or threshold-borderline, default to `Refuted` *only* when the contradiction is unambiguous; otherwise mark `Supported`.

(ii) Linguistic underspecification (a claim too vague to be checked) is treated as `Refuted`, to discourage hedge-only captions.

(iii) Style and prose register are not graded; only typed claims enter the evaluation.

(iv) Disagreement between the first two annotators on any claim triggers third-annotator adjudication; pre-adjudication labels determine the Fleiss κ of **Table 1**.

Ethics and privacy. ERA5 reanalysis fields are non-personal scientific data, no human-subject content is involved, and institutional review is therefore not required. Annotators are remunerated at a market rate consistent with the host institution's volunteer-research policy, and no proprietary captioner output is exposed to any third party.

Figure 7: Annotator guidelines used for the annotated validation set.

D Supplementary Experiments

D.1 Additional Empirical Results

MAE on the 2024–2025 Test Split. Forecasting MAE follows the MSE ordering of **Table 2**, with ATMOZERO reaching 0.349 ± 0.003 at H=96, 0.384 ± 0.003 at H=192, 0.414 ± 0.004 at H=336, and 0.455 ± 0.005 at H=720, against 0.383/0.424/0.462/0.503 for NO-CAPTION.

Training Dynamics. **Fig. 8** traces seven quantities across optimization. Two phases are visible. During the cold-start phase, with a uniform Proposer and the full shaped reward, verifier precision and the shaped reward climb fastest while the downstream metrics improve only gradually; the steep downstream gains arise after the Proposer engages at the precision-driven boundary. Downstream metrics respond with different lags. The unsupported-claim rate moves first after a brief rise, coverage follows the curriculum almost immediately with a brief dip when the Proposer concentrates on harder windows, and forecasting error responds last. KL divergence rises briefly as the refutation channel reshapes the loss and then settles well below the soft early-stopping threshold. The Proposer entropy stays at its uniform prior during the cold-start phase and contracts as the curriculum sharpens.

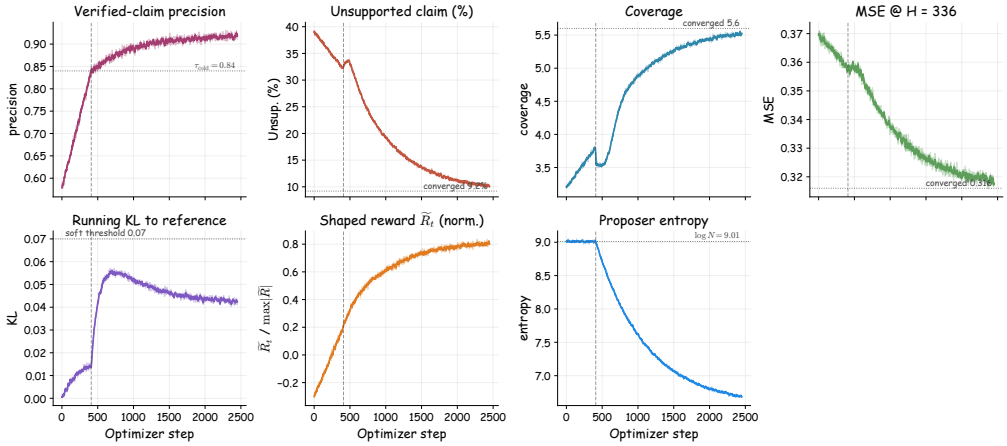


Figure 8: Training dynamics of ATMOZERO. Solid lines are seed means, shaded bands $\pm 1\sigma$, thin lines individual seeds. The vertical dashed line in every panel marks the cold-start boundary at step 410, defined as the first crossing of $\tau_{\text{cold}} = 0.84$ by the running-mean verifier precision; per-seed crossings occur at steps 386, 406, and 418. The dotted line at KL $= 0.07$ is a soft early-stopping threshold. The Proposer entropy ceiling is $\log N \approx 9.01$ nats over the $N = 8{,}192$ source stations.

Zero-Shot Transfer to a Real-World Surface-Station Suite. ATMOZERO is trained only on the 2014–2022 global ERA5 split, and we apply its captioner zero-shot to fifteen real-world surface-station datasets that we collected to evaluate OOD generalization on real-world observations. Captions condition the same frozen forecaster as in the in-corpus experiments, and the resulting forecaster is compared against time-series foundation models pretrained on orders of magnitude more data. The 15-station composition is summarized in **Table 15**, spanning two Antarctic stations, six continents, both hemispheres, and a wide range of forecastability scores. **Table 16** reports MSE and MAE against ten zero-shot pretrained baselines, namely two Time-MoE variants, three Moirai variants, TimesFM, Moment, and three Chronos variants. ATMOZERO attains the best score in the majority of the 120 station–horizon–metric comparisons (15 datasets $\times$ 4 horizons $\times$ 2 metrics).

Table 15: Statistics of the fifteen real-world surface-station datasets used for zero-shot transfer evaluation. # Channels is the number of variables, and Size gives the date range. †SE: Spectral Entropy; ACF: Autocorrelation-based; Proxy: Error-based. All three quantify time-series forecastability; higher scores indicate greater predictability (i.e., lower task difficulty). Region abbreviations: HCMC = Ho Chi Minh City; LA = Los Angeles.

Region	# Channels	Frequency	Size	Forecast Length	SE†	ACF†	Proxy†
Yuzhno-Sakhalinsk	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.438	17.089	0.682
Tokyo	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.438	17.090	0.682
Beijing	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.431	16.660	0.711
Mumbai/Santacruz	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.398	16.474	0.706
HCMC	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.462	17.292	0.768
Cairo	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.456	16.912	0.734
Nairobi Dagoretti	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.425	16.217	0.656
Nauru	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.452	11.586	0.705
Honolulu	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.435	11.696	0.734
LA	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.425	15.879	0.726
Edmonton	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.437	18.865	0.805
Oxford	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.455	18.871	0.816
Zurich	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.463	17.805	0.801
McMurdo	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.431	16.438	0.720
Novolazarevskaya	10	Hourly	January 2022–August 2025	{96, 192, 336, 720}	0.426	15.508	0.699

Table 16: Zero-shot pretrained time-series foundation models on the real-world surface-station suite (15 stations, 4 horizons). Bold = best, underline = second-best. Region abbreviations: Yuzhno-S. = Yuzhno-Sakhalinsk; Mumbai = Mumbai/Santacruz; Nairobi = Nairobi Dagoretti; HCMC = Ho Chi Minh City; LA = Los Angeles; Novolaz. = Novolazarevskaya.

Models		Ours		Pretrained Time-Series Foundation Models (Zero-Shot)																			
		ATMOZERO		Time-MoE$_{base}$		Time-MoE$_{large}$		Moirai$_{small}$		Moirai$_{base}$		Moirai$_{large}$		TimesFM		Moment		Chronos$_{small}$		Chronos$_{base}$		Chronos$_{large}$	
Region	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Yuzhno-S.	96	**0.475**	**0.428**	0.510	0.460	0.505	0.455	0.885	0.547	0.879	0.529	0.986	0.530	0.883	0.703	0.674	0.594	0.850	0.581	0.879	0.589	0.873	0.583
	192	**0.514**	**0.453**	0.540	0.492	0.532	0.485	0.950	0.569	0.896	0.546	1.011	0.548	0.903	0.711	0.717	0.600	0.917	0.620	0.979	0.634	0.950	0.623
	336	**0.522**	**0.482**	0.566	0.519	0.556	0.505	1.148	0.588	0.813	0.552	0.868	0.560	0.889	0.708	0.913	0.700	0.957	0.649	1.047	0.667	0.973	0.646
	720	**0.543**	**0.497**	0.643	0.580	0.568	0.519	1.869	0.667	0.958	0.602	1.259	0.625	0.877	0.705	1.740	1.019	1.089	0.708	1.142	0.715	1.012	0.687
Tokyo	96	0.436	**0.398**	**0.431**	0.426	0.436	0.424	1.541	0.526	1.314	0.483	1.877	0.491	0.846	0.702	0.677	0.567	0.778	0.543	0.779	0.546	0.811	0.553
	192	0.463	**0.425**	**0.460**	0.455	0.470	0.456	1.407	0.560	1.021	0.504	2.416	0.529	0.841	0.700	0.788	0.623	0.855	0.587	0.874	0.595	0.895	0.598
	336	0.507	**0.461**	**0.497**	0.487	0.497	0.477	1.666	0.596	1.168	0.529	1.910	0.553	0.842	0.700	0.881	0.653	0.943	0.625	0.942	0.629	0.960	0.630
	720	**0.509**	0.513	0.575	0.545	0.539	**0.505**	2.236	0.660	2.600	0.603	4.302	0.637	0.847	0.702	1.933	1.044	1.146	0.691	1.054	0.681	1.049	0.677
Beijing	96	**0.532**	**0.448**	0.566	0.474	0.579	0.485	1.461	0.616	1.254	0.576	1.511	0.561	0.952	0.728	0.726	0.598	0.977	0.614	0.980	0.625	0.943	0.600
	192	**0.555**	**0.470**	0.594	0.498	0.597	0.508	1.224	0.610	0.991	0.577	1.450	0.571	0.963	0.730	0.808	0.629	1.032	0.649	1.058	0.668	1.029	0.640
	336	**0.574**	**0.488**	0.620	0.522	0.608	0.519	1.578	0.629	0.900	0.583	1.053	0.587	0.973	0.735	0.908	0.674	1.084	0.675	1.121	0.696	1.054	0.658
	720	**0.603**	0.553	0.681	0.576	0.631	**0.544**	2.740	0.709	1.233	0.650	1.755	0.653	0.953	0.732	2.197	1.137	1.181	0.725	1.218	0.741	1.165	0.729
Mumbai	96	0.510	**0.466**	0.512	0.486	**0.502**	0.483	1.634	0.603	1.313	0.562	1.808	0.571	0.895	0.728	0.707	0.588	0.891	0.629	0.870	0.628	0.889	0.630
	192	0.527	**0.470**	0.535	0.510	**0.522**	0.505	1.299	0.606	1.104	0.571	1.651	0.592	0.886	0.726	0.776	0.619	0.951	0.663	0.959	0.671	0.961	0.669
	336	**0.511**	**0.510**	0.557	0.534	0.544	0.522	1.457	0.629	0.931	0.575	1.174	0.610	0.893	0.727	0.862	0.667	1.004	0.690	1.025	0.701	1.013	0.695
	720	0.592	**0.520**	0.616	0.582	**0.582**	0.547	2.364	0.703	1.576	0.644	2.106	0.683	0.888	0.726	2.205	1.126	1.169	0.747	1.170	0.757	1.089	0.740
HCMC	96	**0.468**	**0.456**	0.491	0.470	0.498	0.468	1.002	0.571	0.911	0.541	0.941	0.533	0.900	0.732	0.707	0.600	0.897	0.622	0.911	0.626	0.924	0.627
	192	**0.498**	**0.469**	0.513	0.494	0.514	0.491	1.049	0.588	0.815	0.547	0.863	0.551	0.912	0.735	0.759	0.617	0.973	0.661	0.998	0.672	0.991	0.668
	336	**0.494**	**0.487**	0.538	0.516	0.529	0.507	1.503	0.614	0.805	0.555	1.040	0.577	0.905	0.732	1.013	0.727	1.039	0.695	1.071	0.706	1.053	0.696
	720	0.544	**0.497**	0.613	0.570	**0.541**	0.524	2.697	0.687	1.226	0.616	1.825	0.653	0.910	0.735	1.741	1.008	1.155	0.748	1.199	0.755	1.166	0.738
Cairo	96	**0.662**	**0.477**	0.796	0.516	0.785	0.499	1.442	0.596	1.169	0.561	1.555	0.566	1.290	0.778	0.692	0.597	1.365	0.656	1.388	0.660	1.349	0.660
	192	**0.705**	**0.496**	0.808	0.531	0.809	0.521	1.228	0.597	1.021	0.565	1.706	0.586	1.299	0.779	0.754	0.627	1.462	0.692	1.533	0.704	1.463	0.699
	336	**0.799**	0.542	0.835	0.549	0.828	**0.537**	1.733	0.623	1.126	0.585	1.456	0.610	1.301	0.780	0.860	0.667	1.524	0.720	1.630	0.737	1.552	0.727
	720	**0.809**	**0.537**	0.897	0.599	0.837	0.552	2.410	0.694	1.877	0.657	2.881	0.690	1.290	0.780	1.703	0.982	1.627	0.772	1.752	0.786	1.670	0.766
Nairobi	96	**0.473**	**0.458**	0.492	0.481	0.493	0.479	0.703	0.550	0.695	0.541	0.660	0.531	0.887	0.726	0.728	0.615	0.879	0.625	0.897	0.626	0.888	0.626
	192	**0.478**	**0.491**	0.514	0.503	0.514	0.504	0.712	0.561	0.686	0.549	0.655	0.553	0.891	0.727	0.817	0.634	0.939	0.662	0.981	0.670	0.953	0.665
	336	0.533	0.525	0.536	0.525	**0.527**	**0.522**	0.791	0.582	0.696	0.567	0.661	0.566	0.885	0.725	0.899	0.696	0.981	0.687	1.028	0.697	1.001	0.691
	720	0.541	0.540	0.582	0.564	**0.537**	**0.533**	1.462	0.643	0.733	0.610	0.868	0.615	0.880	0.724	1.961	1.071	1.081	0.731	1.094	0.735	1.021	0.722
Nauru	96	**0.494**	**0.454**	0.525	0.471	0.527	0.466	0.853	0.552	0.828	0.525	0.881	0.531	1.035	0.748	0.786	0.643	0.858	0.598	0.929	0.622	0.910	0.615
	192	**0.512**	**0.473**	0.538	0.489	0.542	0.489	0.769	0.563	0.773	0.539	0.736	0.546	1.019	0.746	0.920	0.701	0.925	0.638	1.058	0.679	1.012	0.665
	336	**0.530**	**0.483**	0.557	0.509	0.547	0.502	0.800	0.581	0.805	0.557	0.718	0.559	1.018	0.748	1.026	0.752	0.986	0.668	1.170	0.723	1.119	0.702
	720	**0.536**	**0.496**	0.621	0.556	0.573	0.527	1.054	0.631	0.818	0.589	0.937	0.614	1.018	0.749	2.017	1.111	1.147	0.718	1.356	0.789	1.298	0.777
Honolulu	96	**0.541**	**0.464**	0.576	0.476	0.585	0.482	1.143	0.578	1.005	0.557	1.304	0.559	1.012	0.720	0.791	0.645	0.939	0.605	1.062	0.636	1.041	0.630
	192	0.619	**0.489**	**0.613**	0.508	0.629	0.517	1.124	0.607	0.996	0.582	1.066	0.586	1.025	0.725	0.913	0.685	1.039	0.654	1.228	0.703	1.178	0.695
	336	**0.627**	**0.507**	0.648	0.532	0.648	0.529	1.121	0.623	0.966	0.596	1.061	0.618	1.036	0.728	1.091	0.791	1.106	0.686	1.357	0.748	1.321	0.738
	720	**0.644**	0.557	0.707	0.573	0.669	**0.552**	1.313	0.661	1.029	0.634	1.557	0.704	1.037	0.731	2.049	1.133	1.221	0.734	1.518	0.802	1.500	0.778
LA	96	**0.561**	**0.472**	0.585	0.518	0.579	0.505	0.923	0.607	0.926	0.591	0.836	0.578	0.989	0.769	0.752	0.626	1.014	0.657	1.073	0.661	1.016	0.651
	192	**0.588**	**0.506**	0.612	0.546	0.610	0.537	0.935	0.629	0.949	0.617	0.874	0.615	0.986	0.768	0.777	0.638	1.087	0.703	1.144	0.708	1.098	0.698
	336	**0.588**	**0.523**	0.626	0.566	0.621	0.554	0.907	0.637	0.972	0.633	0.864	0.634	0.976	0.768	0.852	0.678	1.117	0.728	1.173	0.730	1.120	0.720
	720	**0.596**	0.576	0.655	0.600	0.629	**0.564**	1.119	0.682	1.063	0.680	1.070	0.690	0.968	0.767	1.872	1.069	1.206	0.774	1.260	0.774	1.186	0.755
Edmonton	96	0.557	0.499	0.556	0.498	**0.546**	**0.495**	0.975	0.601	0.929	0.593	0.967	0.583	0.901	0.740	0.707	0.623	0.906	0.601	0.956	0.608	0.906	0.596
	192	**0.556**	**0.499**	0.588	0.535	0.596	0.542	1.062	0.642	0.999	0.630	0.936	0.619	0.898	0.738	0.726	0.625	0.999	0.656	1.108	0.670	1.014	0.651
	336	**0.578**	0.570	0.614	**0.559**	0.639	0.571	0.916	0.635	0.900	0.643	0.846	0.635	0.906	0.740	0.985	0.765	1.089	0.690	1.146	0.698	1.070	0.679
	720	**0.634**	**0.573**	0.673	0.600	0.655	0.585	0.931	0.669	0.915	0.675	0.920	0.675	0.908	0.739	1.685	1.023	1.264	0.738	1.267	0.746	1.328	0.712
Oxford	96	0.737	**0.538**	0.751	0.569	**0.726**	0.556	1.124	0.661	1.168	0.657	1.134	0.641	1.140	0.798	0.782	0.654	1.196	0.693	1.227	0.684	1.191	0.682
	192	**0.742**	**0.588**	0.791	0.605	0.781	0.606	1.285	0.707	1.206	0.690	1.183	0.688	1.149	0.802	0.836	0.650	1.331	0.754	1.415	0.755	1.342	0.746
	336	**0.766**	0.636	0.821	**0.631**	0.827	0.637	1.544	0.720	1.252	0.720	1.236	0.722	1.146	0.800	1.011	0.755	1.423	0.792	1.502	0.793	1.454	0.783
	720	**0.806**	**0.615**	0.880	0.668	0.858	0.657	2.417	0.780	1.342	0.768	1.529	0.790	1.158	0.804	1.910	1.092	1.589	0.842	1.662	0.846	1.572	0.827
Zurich	96	**0.568**	**0.487**	0.601	0.509	0.597	0.499	0.969	0.593	0.832	0.571	0.845	0.570	1.047	0.773	0.736	0.623	1.057	0.657	1.081	0.650	1.075	0.650
	192	**0.584**	**0.497**	0.643	0.541	0.622	0.523	1.055	0.616	0.908	0.601	0.944	0.609	1.045	0.770	0.781	0.636	1.149	0.706	1.202	0.701	1.151	0.693
	336	**0.611**	**0.524**	0.668	0.564	0.648	0.542	0.986	0.632	0.901	0.617	0.914	0.627	1.048	0.771	0.901	0.688	1.232	0.740	1.248	0.732	1.195	0.723
	720	**0.640**	**0.545**	0.725	0.613	0.660	0.556	1.290	0.689	0.921	0.654	1.021	0.666	1.035	0.771	1.925	1.065	1.329	0.793	1.299	0.776	1.243	0.755
McMurdo	96	0.661	0.529	0.666	0.532	**0.653**	**0.525**	1.144	0.620	1.044	0.602	1.111	0.597	1.107	0.782	0.743	0.630	1.118	0.667	1.166	0.674	1.129	0.669
	192	0.699	**0.546**	0.696	0.560	**0.688**	0.557	1.211	0.646	0.989	0.615	1.026	0.623	1.098	0.777	0.819	0.649	1.221	0.713	1.281	0.727	1.256	0.721
	336	**0.693**	**0.548**	0.722	0.581	0.720	0.580	1.358	0.674	1.029	0.637	1.080	0.648	1.109	0.780	0.883	0.688	1.277	0.742	1.401	0.766	1.314	0.752
	720	**0.687**	0.596	0.789	0.628	0.734	**0.592**	1.793	0.732	1.096	0.689	1.219	0.697	1.099	0.778	1.891	1.067	1.359	0.786	1.459	0.805	1.389	0.783
Novolaz.	96	**0.482**	**0.456**	0.505	0.483	0.512	0.488	0.891	0.565	0.797	0.559	0.765	0.540	0.972	0.760	0.740	0.623	0.843	0.587	0.861	0.600	0.836	0.588
	192	**0.527**	0.522	0.547	**0.519**	0.559	0.529	1.051	0.613	0.837	0.599	0.758	0.584	0.962	0.758	0.751	0.632	0.932	0.635	0.980	0.659	0.933	0.639
	336	**0.553**	**0.538**	0.595	0.555	0.594	0.555	1.329	0.638	0.852	0.615	0.886	0.616	0.973	0.762	0.852	0.689	0.999	0.664	1.045	0.693	0.998	0.670
	720	**0.615**	**0.560**	0.672	0.608	0.641	0.585	2.119	0.704	0.967	0.663	1.145	0.682	0.976	0.762	1.787	1.044	1.120	0.715	1.169	0.745	1.101	0.740

D.2 Human-Anchored Automatic Faithfulness Evaluation at Scale

Motivation. The human-annotated faithfulness evaluation covers the 200-window set of Appendix C.2. To test whether the automatic ranking holds on a much larger pool, we run the automatic NLI cross-encoder of Appendix B.3 anchored to the human labels on the 200-window set. This follows the human-anchored automatic-evaluation pattern of [74], in which a human-annotated set certifies the automatic estimator and the estimator is then applied to a larger evaluation pool.

Anchoring on the 200-Window Set. On the 200-window human-annotated set, the NLI cross-encoder agrees with the post-adjudication human label at the claim level on 87.4% of the $\sim$11,600 claim instances (Cohen's $\kappa = 0.71$, weighted by claim type). In **Table 1**, the NLI ordering matches the human Unsup. ordering on every adjacent pair.

Large-Scale Sample. We then draw a stratified random sample of $5{,}000$ windows from the 2024–2025 test split, with proportional allocation across the 30 Köppen–Geiger zones and the four meteorological seasons, disjoint from the 200-window human set and from the 300-window preference-study set. All seven captioners of **Table 1** produce captions on the $5{,}000$-window sample, the NLI cross-encoder is held fixed, and entailment thresholds are not retuned.

Table 17: Automatic faithfulness evaluation at 5,000-window scale, using the strict variant of the NLI metric of Appendix B.3 (entailment threshold 0.7 rather than 0.5). NLI@5k is the entailment rate on the 5,000-window sample; ΔNLI is the gap to the 200-window value.

Method	NLI@200 (↑)	NLI@5k (↑)	ΔNLI
PROMPTED-BASE	0.43	0.45	+0.02
GPT-4o 1-shot	0.61	0.63	+0.02
Claude-Sonnet-4.5 1-shot	0.66	0.68	+0.02
DENSE-ONLY	0.69	0.70	+0.01
VERIFIER-ONLY	0.74	0.73	−0.01
SOLVER-ONLY	0.79	0.79	0.00
ATMOZERO	**0.86**	**0.85**	−0.01

Ranking Stability at Scale. The 5,000-window NLI estimates lie within 0.02 of the 200-window human-anchored values for every captioner, and the strict-NLI captioner ranking is unchanged. The 25× larger evaluation therefore confirms that the NLI ranking is stable at scale (**Table 17**), and the gap between ATMOZERO and the strongest dense-reward baseline persists on the larger pool.

D.3 Human Preference Study

Setup. A panel of three PhD-level meteorologists in atmospheric or climate science, distinct from the annotation panel of Appendix C.2, independently judges caption pairs on 300 unique windows from the 2024–2025 test split. The windows are disjoint from the 200-window faithfulness set and stratified across the Köppen–Geiger zones of Appendix C.1. Each window is compared against every opponent, giving 1,800 blind pairs. Each pair contrasts ATMOZERO against one opponent from {PROMPTED-BASE, DENSE-ONLY, VERIFIER-ONLY, SOLVER-ONLY, GPT-4o 1-shot, Claude-Sonnet-4.5 1-shot}, with the order randomized and the captioner identity concealed. Annotators select *Win* / *Tie* / *Loss* on the joint axis of *scientific faithfulness* and *usefulness for forecasting*, and report a one-line failure-mode tag for each non-tie, blinded as in Appendix C.2.

Results, Inversions, and Annotator Agreement. **Table 3** reports the win, tie, and loss rates, and **Table 18** lists each win rate next to the opponent's Unsup. and NLI. The Spearman rank correlation between ATMOZERO's win rate and the opponent's own unsupported-claim rate is $\rho = 0.77$, 95% CI $[0.55, 0.91]$, so win rates track opponent faithfulness. The closed-source captioners, whose longer prose is often preferred by annotators, form the main exception, with Claude-Sonnet-4.5 at 56% despite an unsupported-claim rate of 20.8%, well above that of SOLVER-ONLY (12.5%; win rate 61%). The pre-adjudication Fleiss κ over the win / tie / loss label set is 0.76, consistent with the per-claim study of **Table 1** and above the 0.5–0.6 band typical of caption preference studies.

Table 18: Per-method faithfulness reference values reproduced from the 200-window human-annotated set of **Table 1**, alongside the ATMOZERO win rate against each opponent on the disjoint 300-window preference-study set of **Table 3**.

Method	Unsup. (%)	NLI	ATMOZERO win rate (%)
PROMPTED-BASE	38.6	0.62	84
DENSE-ONLY	22.3	0.75	70
GPT-4o 1-shot	27.4	0.70	63
SOLVER-ONLY	12.5	0.83	61
VERIFIER-ONLY	11.7	0.84	58
Claude-Sonnet-4.5 1-shot	20.8	0.76	56

Failure-Mode Taxonomy. The single most frequent loss category against SOLVER-ONLY is *verbose-but-correct*, where ATMOZERO sometimes issues a tighter set of supported claims than the opponent, a pattern that annotators mark as under-coverage even when the verifier marks the omitted family as inactive. Against the closed-source baselines, especially Claude-Sonnet-4.5, the most frequent loss category is *stylistic preference* (67% of losses pooled over the two closed-source captioners, against 21% verbose-but-correct or coverage tags and 12% genuine faithfulness deficits),

where Claude-Sonnet-4.5's longer, more rhetorically fluent captions are picked even on windows where ATMOZERO's structured CLAIMS list is strictly more faithful.

Expert-Caption Ceiling. A 150-window subset of the 300-window study is paired with captions written from scratch by two of the meteorologist annotators following the prompt template of Appendix A.2. These captions are produced exclusively for evaluation and never enter training. The expert ceiling represents an upper bound on faithfulness reachable on the *station-observable* input set $\mathcal{X}_{\text{stn}}$. **Table 19** reports the four anchor metrics. ATMOZERO comes within 4.5 pp of the expert ceiling on Unsup., 0.07 on NLI, and 0.6 on Coverage and false refutation.

Table 19: Expert-caption ceiling on a 150-window subset of the 300-window preference-study set, scored as in Appendix C.2. Two PhD-level meteorologists write CLAIMS captions from scratch; values are means across the two annotators before adjudication. The ATMOZERO row repeats **Table 1**.

Method	Unsup. (%)	Coverage	NLI	False ref. (%)
ATMOZERO (**Table 1**)	9.2	5.6	0.86	5.8
Expert ceiling	4.7	6.2	0.93	5.2

D.4 Climate-Zone Out-of-Distribution Generalization

Setup. Six Köppen–Geiger zones (BWh, Cwa, Dfb, Dwc, ET, Aw) are removed from training. Four RL variants, namely ATMOZERO, SOLVER-ONLY, VERIFIER-ONLY, and DENSE-ONLY, are retrained from the same Qwen2.5-7B-Instruct initialization on the remaining 24 zones under the hyperparameters of Appendix B.1, and Time-LLM [61] is retrained on the same zones with its default configuration. The closed-source GPT-4o 1-shot baseline and the prompted open-weights baseline (PROMPTED-BASE) require no retraining and are reported on the same unseen zones for reference. Evaluation uses the unseen zones in the 2024–2025 test split.

Reference Baselines. **Table 20** adds PROMPTED-BASE to the rows of **Table 4**.

Table 20: Extended OOD results with PROMPTED-BASE (Time-LLM: MSE only).

Method	Unsup. (↓)	Coverage (↑)	False ref. (↓)	MSE@H=336 (↓)
PROMPTED-BASE	42.1	3.7	13.7	0.412
GPT-4o 1-shot	30.2	4.2	11.5	0.378
DENSE-ONLY	$26.4_{\pm 1.7}$	$4.5_{\pm 0.42}$	$11.7_{\pm 1.4}$	$0.367_{\pm 0.008}$
Time-LLM	–	–	–	$0.353_{\pm 0.009}$
VERIFIER-ONLY	$17.8_{\pm 1.0}$	$4.4_{\pm 0.36}$	$15.9_{\pm 1.5}$	$0.356_{\pm 0.006}$
SOLVER-ONLY	$16.4_{\pm 1.3}$	$\mathbf{4.7}_{\pm 0.28}$	$13.8_{\pm 1.1}$	$0.347_{\pm 0.007}$
ATMOZERO	$\mathbf{12.4}_{\pm 1.2}$	$\mathbf{4.7}_{\pm 0.40}$	$\mathbf{8.4}_{\pm 0.9}$	$\mathbf{0.339}_{\pm 0.010}$

Per-Zone Breakdown. **Fig. 9** shows a heatmap over the six unseen zones for ATMOZERO, SOLVER-ONLY, VERIFIER-ONLY, and the GPT-4o 1-shot reference; row means recover the **Table 4** aggregate exactly. The per-zone Unsup. values come from the human-anchored NLI estimator of Appendix D.2. The hardest unseen zone for every method is the polar tundra (ET), where the diurnal-cycle family fires less often under weak winter day–night separation and the climatological family has high variance from the small zone-area sample; the easiest is the dry-winter subtropical zone (Cwa), where the climatological and rain–moisture families transfer across continents.

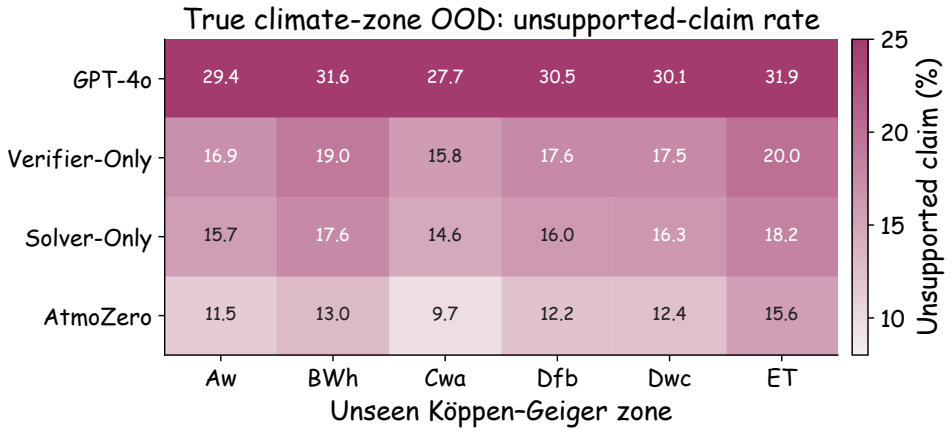


Figure 9: Per-zone unsupported-claim rate (%) for the six unseen Köppen–Geiger zones, three retrained methods plus the GPT-4o 1-shot reference. Lower is better.

D.5 Reward-Hacking Counterfactual Probe

Setup. We probe the verifier's resistance to gaming with six adversarial caption policies issued at inference time, namely *Empty* (no claims), *Repeat-Best* (repeats one true claim eight times), *Generic-Safe* (issues only `(climato,"near-normal regime")` on every window where family $k = 6$ fires, omitting all other families), *Sign-Flip-Wind* (negates the wind-regime direction), *Phase-Shift-Diurnal* (delays the diurnal phase by twelve hours), and *Contradicted-Front* (asserts a frontal passage on quiet windows). Each adversarial caption is graded by three judges other than the training verifier. The first is an independently calibrated verifier whose thresholds (τ_k^-, τ_k^+) are re-estimated on a disjoint 200-window split. The second is an ERA5 gridded-state oracle that checks claims against the full reanalysis state at the focal time, including pressure-level temperature, vertical velocity, and total-column water vapor, none of which belongs to the station-observable input set $\mathcal{X}_{\text{stn}}$. The third is a blinded spot check by two meteorologist annotators of Appendix C.2 on 50 pairs per policy. The reported worst-case exploit rate is the maximum across the three judges.

Per-Policy Detail. **Fig. 10** reports the per-policy worst-case exploit rate and the disagreement gap between the training and independently calibrated verifiers. The *Empty* policy is rejected on 99.3% of windows. The *Repeat-Best* policy succeeds on only 2.4% of windows, and the refutation channel U drives the *Contradicted-Front* reward margin to -0.62.

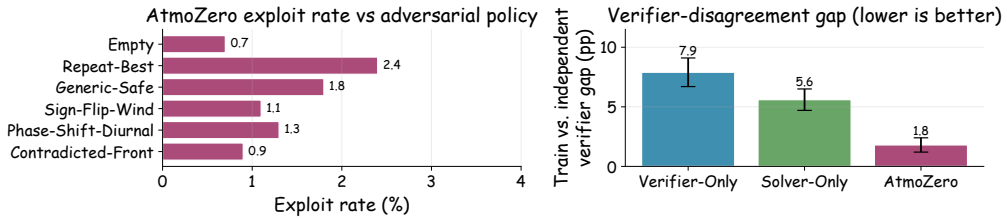


Figure 10: Reward-hacking counterfactual probe. *Left*, worst-case exploit rate (%) across three independent judges for six adversarial policies. *Right*, disagreement between the training and independently calibrated verifiers on natural captions (percentage points) for three trained variants.

D.6 Reward-Channel Decomposition

Setup. Six short (20%) continuation fine-tunes from a shared cold-start checkpoint at step 410 ablate one reward channel at a time, namely $-U$ (drop the refutation channel), $-S_k$ (replace S_k by zero for all k), $-C$ (drop the coverage bonus), $-D$ (drop the type–token specificity bonus, $\nu = 0$), $-R_{\text{cycle}}$ (drop the cycle dense channel, $\alpha_t \equiv 0$), and $-R_{\text{uplift}}$ (drop the forecaster-uplift dense channel, $\beta_t \equiv 0$). Each ablation uses three seeds. Unstable seeds (those whose KL to the reference

policy exceeds 0.20 during training) are reported as part of the mean and standard deviation rather than excluded; the count of unstable seeds is shown explicitly in the rightmost column of **Table 8**.

D.7 Extreme-Event Detection Utility

Setup. We score caption-claim extraction on five rare-event categories, namely $p99$ precipitation ($n = 4{,}812$ test windows), frontal passages ($n = 6{,}174$), heatwaves with $T \geq p99$ for $\geq 24\,\mathrm{h}$ ($n = 3{,}506$), cold surges with $T \leq p1$ for $\geq 24\,\mathrm{h}$ ($n = 3{,}918$), and high-wind regimes with $\|\mathbf{u}\|_2 \geq 17\,\mathrm{m/s}$ ($n = 5{,}260$). Each window receives a per-category ground-truth label from the verifier library of Appendix A on the 2024–2025 test split, and each caption is processed by the claim extractor of Appendix A.2 for matching lexicon entries. We report the categorical-forecast scores POD (probability of detection) and FAR (false alarm ratio), plus a persistence baseline that repeats the event label of the window issued 48 hours earlier at the same station. Confidence intervals come from a 1,000-replicate bootstrap over 24-hour blocks.

Results. **Table 21** reports POD and FAR per event category, plus the macro-average across the five. ATMOZERO attains the highest macro POD (0.68) and the lowest macro FAR (0.23), with the largest FAR reduction over the strongest baseline on frontal passages (0.21 vs. 0.28). On cold surges, ATMOZERO reaches 0.61 POD against 0.45 for GPT-4o and 0.50 for VERIFIER-ONLY.

Table 21: Extreme-event POD and FAR from caption claims, per category and macro-averaged.

	p99 precip		Frontal		Heatwave		Cold surge		High wind		Macro	
Method	POD	FAR	POD	FAR	POD	FAR	POD	FAR	POD	FAR	POD	FAR
Persistence	0.34	0.55	0.41	0.50	0.50	0.42	0.46	0.45	0.43	0.48	0.43	0.48
GPT-4o 1-shot	0.47	0.46	0.51	0.39	0.66	0.30	0.45	0.42	0.49	0.41	0.52	0.40
Claude-Sonnet-4.5 1-shot	0.51	0.42	0.54	0.36	0.68	0.27	0.49	0.39	0.54	0.37	0.55	0.36
DENSE-ONLY	0.39	0.42	0.42	0.40	0.55	0.34	0.40	0.41	0.45	0.40	0.44	0.39
VERIFIER-ONLY	0.56	0.39	0.55	0.32	0.71	0.24	0.50	0.34	0.59	0.31	0.58	0.32
SOLVER-ONLY	0.60	0.32	0.59	0.28	0.74	0.21	0.55	0.30	0.62	0.27	0.62	0.28
ATMOZERO	**0.66**	**0.28**	**0.65**	**0.21**	**0.78**	**0.18**	**0.61**	**0.25**	**0.68**	**0.22**	**0.68**	**0.23**

95% block-bootstrap CIs (1,000 replicates of 24-hour blocks) are ±0.02 for POD and ±0.03 for FAR. Negative-class counts (n_{neg}) in the evaluated set are 95.2k, 93.9k, 96.6k, 96.2k, and 94.8k in column order.

D.8 Sample Efficiency

Setup. ATMOZERO and the three RL ablations SOLVER-ONLY, VERIFIER-ONLY, and DENSE-ONLY are trained on 1%, 5%, 10%, and 25% of the training split under identical budgets (2,450 steps, $G = 8$, Appendix B.1), and the 100% setting reuses the models of **Tables 1** and **2**.

Results. **Fig. 11** shows that ATMOZERO attains the lowest mean unsupported-claim rate at every training fraction, since the verifier reward supplies a discrete support–refute teacher signal that is informative even when the captioner has seen few windows, while dense-only rewards need many more samples to recover a polarity prior. The 100% results match **Tables 1** and **2**.

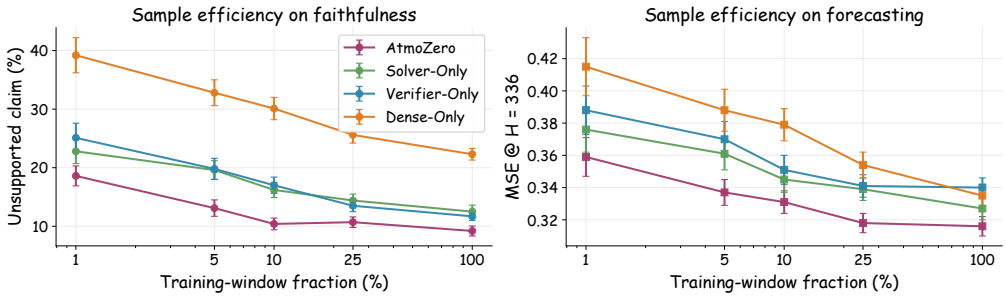


Figure 11: Sample-efficiency curves of unsupported-claim rate (*left*) and MSE@H=336 (*right*) against the log-scaled training-window fraction. Error bars are $\pm 1\sigma$.

D.9 Robustness to Sensor Noise and Missing Channels

Setup. Three test-time corruptions are applied to the 2024–2025 test windows. The first is additive Gaussian noise at $\sigma \in \{5\%, 10\%\}$ of the per-channel standard deviation. The second is random temporal gaps at a 10% or 20% rate, filled by linear interpolation. The third is single-channel dropout of q, P, or r, and joint dropout of the wind pair (u, v). The verifier and the captioner are not retrained.

Results. **Fig. 12** shows the unsupported-claim rate per method and corruption, and **Table 22** lists the AtmoZero values. Noise and temporal gaps raise the Unsup. of AtmoZero by at most 4.6 pp. Under single-channel dropout, families that depend on the missing channel are dropped from S_k, U, and C rather than refuted, so even dropping r keeps Unsup. at 17.3%, below the clean rate of every captioner trained or prompted without the verifier reward (**Table 1**).

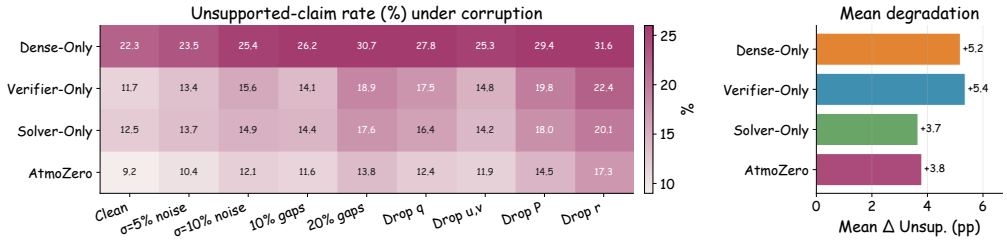


Figure 12: Robustness to sensor noise and missing channels. Heatmap rows are methods, columns are corruption types, and cells give the unsupported-claim rate (%). The right panel shows the mean degradation Δ (pp) relative to the clean input.

Table 22: Unsup. and MSE@H=336 of AtmoZero under corruptions (Clean row: **Table 1**).

Corruption	Unsup. (%)	MSE@H=336
Clean	$9.2_{\pm 0.85}$	$0.316_{\pm 0.006}$
$\sigma = 5\%$ noise	$10.4_{\pm 0.92}$	$0.322_{\pm 0.007}$
$\sigma = 10\%$ noise	$12.1_{\pm 1.05}$	$0.331_{\pm 0.008}$
10% temporal gaps	$11.6_{\pm 0.97}$	$0.335_{\pm 0.007}$
20% temporal gaps	$13.8_{\pm 1.18}$	$0.342_{\pm 0.009}$
Drop q	$12.4_{\pm 1.10}$	$0.350_{\pm 0.009}$
Drop u, v	$11.9_{\pm 0.99}$	$0.347_{\pm 0.008}$
Drop P	$14.5_{\pm 1.22}$	$0.353_{\pm 0.010}$
Drop r	$17.3_{\pm 1.34}$	$0.366_{\pm 0.011}$

D.10 Verifier-Templater Reference

Table 23 compares Verifier-Templater (Appendix B.2) with AtmoZero. The templater verbalizes only verifier-certified claims, which yields a low in-distribution Unsup. and reflects the precision of the verifier library. AtmoZero achieves a $1.8\times$ higher TTR, a lower Unsup. on unseen zones, and a larger forecasting gain (**Table 2**).

Table 23: Verifier-Templater against AtmoZero. In-distribution metrics use the 200-window human-annotated set under the cross-fit protocol of Appendix B.3. OOD Unsup. is the NLI-estimated rate on the six unseen Köppen–Geiger zones of **Table 4**.

Method	Unsup. (↓)	Coverage (↑)	TTR (↑)	NLI (↑)	OOD Unsup. (↓)
Verifier-Templater	$7.6_{\pm 0.8}$	$5.8_{\pm 0.22}$	$0.34_{\pm 0.006}$	$0.88_{\pm 0.019}$	16.8
AtmoZero	$9.2_{\pm 0.85}$	$5.6_{\pm 0.24}$	$0.61_{\pm 0.012}$	$0.86_{\pm 0.025}$	$12.4_{\pm 1.2}$

D.11 Test-Time Caption Protocol and Temporal Alignment

With t denoting the forecast issuance time, every captioner receives only the context window $x_{t-T_w+1:t}$, the station metadata $\mathcal{M}$, the neighbor windows over the same interval, and the controlled lexicons (**Fig. 5**), so the forecast target $x_{t+1:t+H}$ never enters a caption in **Tables 2**, **4**, or **5**. The cutoff t_0 of R_{uplift} (Eq. 6) also lies inside the context window. PatchTST sees only the focal window, whereas the captioner additionally conditions on $\mathcal{M}$ and up to eight neighbor windows within 250 km. The caption sources of **Table 6** use the same frozen PatchTST.

D.12 Comparison with Adapted Self-Play Baselines

Table 24 compares the adapted self-play baselines of Appendix B.2 with the within-framework references. Both trail SOLVER-ONLY on every metric.

Table 24: Adapted self-play baselines with the same backbone, training data, rollout budget, and optimizer updates as ATMOZERO. Metrics follow **Tables 1** and **2**.

Method	Unsup. (↓)	Coverage (↑)	NLI (↑)	MSE@H=336 (↓)	MSE@H=720 (↓)
Self-Rewarding-inspired (self-judge reward)	$24.6_{\pm 1.5}$	$4.3_{\pm 0.33}$	$0.68_{\pm 0.028}$	$0.356_{\pm 0.008}$	$0.402_{\pm 0.010}$
R-Zero-inspired (difficulty filter)	$19.8_{\pm 1.2}$	$4.6_{\pm 0.31}$	$0.72_{\pm 0.022}$	$0.348_{\pm 0.007}$	$0.396_{\pm 0.008}$
DENSE-ONLY	$22.3_{\pm 1.0}$	$4.8_{\pm 0.27}$	$0.75_{\pm 0.021}$	$0.335_{\pm 0.007}$	$0.379_{\pm 0.008}$
SOLVER-ONLY	$12.5_{\pm 1.1}$	$5.3_{\pm 0.21}$	$0.83_{\pm 0.018}$	$0.327_{\pm 0.006}$	$0.372_{\pm 0.008}$
ATMOZERO	$\mathbf{9.2}_{\pm 0.85}$	$\mathbf{5.6}_{\pm 0.24}$	$\mathbf{0.86}_{\pm 0.025}$	$\mathbf{0.316}_{\pm 0.006}$	$\mathbf{0.361}_{\pm 0.007}$

D.13 Representativeness Error on Real Stations

For each of the fifteen real stations of **Table 16**, the representativeness error is the per-channel standardized RMSE between the observed and the collocated ERA5 series, averaged over the six channels of $\mathcal{X}_{\text{stn}}$, and the stations are split into three five-station groups by this error. **Table 25** reports paired verifier verdicts by group, obtained by holding each generated claim fixed and evaluating the same predicate on the station observations and on the collocated ERA5 series under the original thresholds. Across all three groups, hard support–refute reversals stay at or below 2.3%, with most disagreements within 0.05 of a threshold, mainly in the frontal and climatological families.

Table 25: Verifier agreement on the fifteen real stations, grouped by representativeness error. *Near-thr.* is the share of disagreements with a grader score within 0.05 of the threshold.

	Paired verifier verdicts (%)			
Station group	Agreement	Refutation kept	Hard reversals	Near-thr.
Low error	$93.8_{\pm 0.6}$	$96.1_{\pm 0.4}$	$0.7_{\pm 0.2}$	$78_{\pm 7}$
Medium error	$90.9_{\pm 0.8}$	$94.3_{\pm 0.8}$	$1.2_{\pm 0.3}$	$74_{\pm 5}$
High error (complex terrain)	$86.4_{\pm 1.1}$	$91.8_{\pm 1.0}$	$2.3_{\pm 0.6}$	$69_{\pm 4}$

D.14 Caption Case Studies

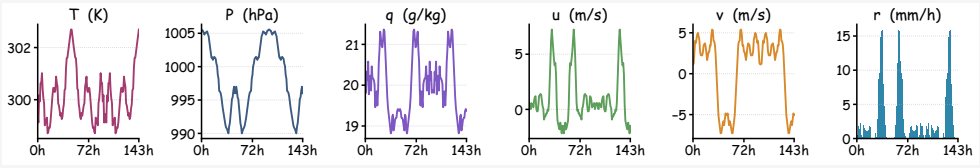
Aggregate Pattern Across 57 Windows. We illustrate the captioner-role progression on a battery of 57 windows (9 globally diverse ERA5 severe-weather events plus 48 controlled regimes spanning quiet, frontal, monsoon, heatwave, cold-surge, polar, and gust-front conditions), each captioned under three roles with the same caption conditioning and graded by both the verifier and an independent meteorologist judge. **Table 26** reports the aggregate pattern. The mean claim count and the mean refutation count U both decrease monotonically from the prompted-only baseline through the mid-training checkpoint to the converged ATMOZERO, while the judge-marked unsupported-claim rate stays low across the three roles, consistent with the verifier-independent variant of **Table 13**.

Table 26: Per-role aggregate over 57 windows. Lower U and lower claim count under the converged AtmoZero reflect the cumulative effect of the refutation channel on caption discipline.

Role	Avg. claims	Avg. U	Avg. Coverage C	Judge unsup. (↓)
Prompted-Base	5.67	3.96	0.12	0.043
Mid-training AtmoZero	4.89	2.91	0.16	0.062
Converged AtmoZero	**3.09**	**1.70**	0.11	**0.040**

Single-Method Case Studies (AtmoZero). Ten test windows taken directly from ARCO-ERA5 spanning a tropical-cyclone landfall, a polar-vortex outbreak, an Indian monsoon onset, a Mediterranean heatwave, a Saharan dust outbreak, the East-African short rains, a North-American lake-effect snowstorm, a North-Indian heatwave, an Amazonian deep-convection event, and a Norwegian polar low. Each box shows the multivariate input window (six channels, T, P, q, u, v, r) followed by the converged AtmoZero caption and its extracted claim list, with the verifier refutation count U and coverage C reported for each window. The claim list is shaded green.

Typhoon Doksuri landfall (Xiamen, 2023-07-28)

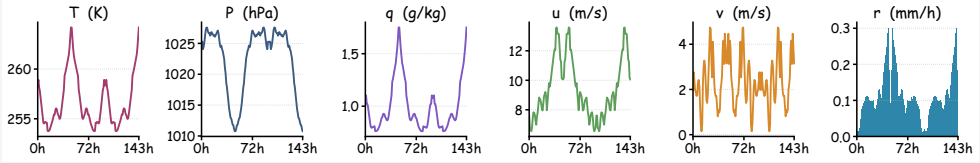


The weather regime over the past 192 hours has been characterized by a significant drop in pressure alongside a warming trend in temperature. The total precipitation recorded was substantial, indicating a wet environment.

Claims (2 entries, $U = 1$, $C = 0.17$):

• (thermodynamic, "warming") • (rain_moisture, "heavy rain")

Polar vortex outbreak (Chicago, 2024-01-14)

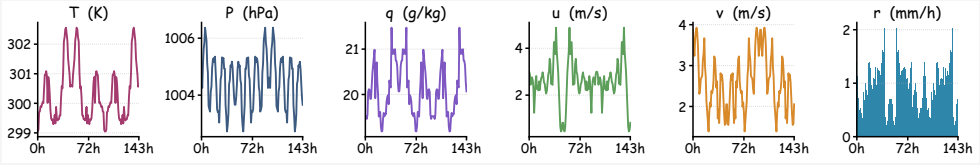


The weather regime over the past 192 hours has experienced a significant drop in atmospheric pressure, with a decrease of 16.82 hPa, while the temperature has increased from 258.4 K to 264.2 K. The mean wind speed has been over 5 m/s, indicating a sustained wind regime.

Claims (3 entries, $U = 2$, $C = 0.00$):

• (thermodynamic, "warming") • (rain_moisture, "light rain") • (wind_regime, "sustained westerly")

Indian monsoon onset (Mumbai, 2023-06-25)

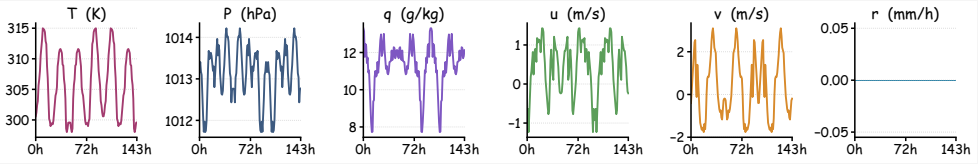


The weather regime over the last 192 hours shows a warming trend in temperature, with a mean temperature of 300.4 K. There has been significant precipitation totaling 130.2 mm, indicating a humid pre-rain regime. However, the wind speeds remain low, and there is no indication of a stable wind regime.

Claims (2 entries, $U = 2$, $C = 0.00$):

• (thermodynamic, "warming") • (rain_moisture, "humid pre-rain regime")

Cerberus heatwave (Athens, 2023-07-23)

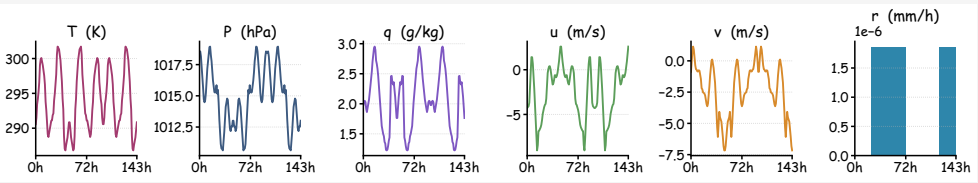


The temperature remained relatively stable with a mean of 305.9 K, while pressure experienced a slight drop of 2.50 hPa. There was no precipitation recorded during this period, and the wind speeds were low, indicating a calm atmosphere.

Claims (3 entries, $U = 1$, $C = 0.20$):

• (thermodynamic, “stable temperature”) • (rain_moisture, “no precipitation”) • (wind_regime, “calm”)

Saharan dust outbreak (Tamanrasset, 2024-03-15)

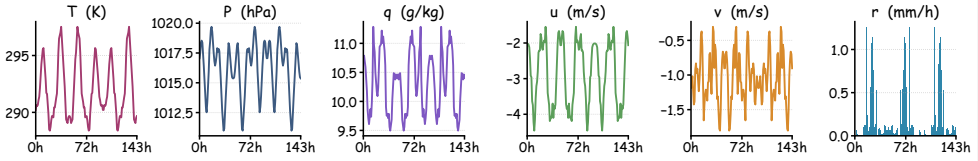


The temperature remained relatively stable with a slight increase from the start to the end of the window. Pressure experienced a notable drop, while there was no precipitation recorded during this period. The wind regime showed a mean wind speed below the threshold for sustained conditions.

Claims (3 entries, $U = 1$, $C = 0.20$):

• (thermodynamic, “stable temperature”) • (rain_moisture, “no precipitation”) • (frontal, “no frontal passage”)

East-African short rains (Nairobi, 2023-11-10)

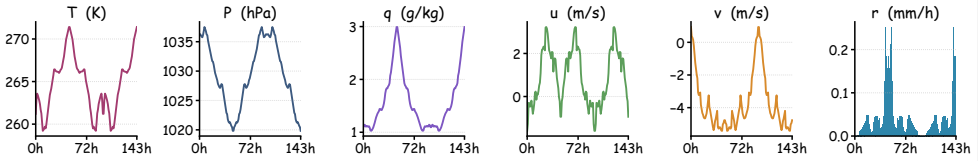


The numeric summary indicates a slight cooling trend in temperature, with a mean of 292.1 K and a drop in pressure of 8.71 hPa. The total precipitation recorded is 26.7 mm, suggesting some moisture presence, but the wind regime shows a maximum speed of only 4.6 m/s, which does not . . .

Claims (3 entries, $U = 2$, $C = 0.17$):

• (thermodynamic, “cooling”) • (rain_moisture, “light rain”) • (frontal, “no frontal passage”)

Lake-effect snowstorm (Buffalo, 2024-12-21)

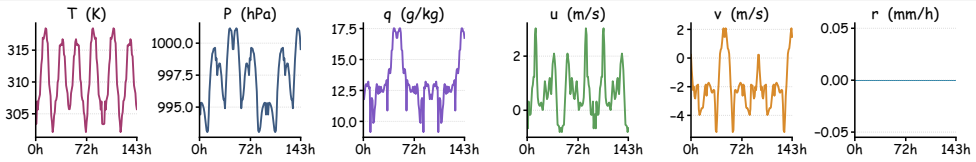


The temperature increased from 261.2 K to 271.4 K over the 192-hour window, indicating a warming trend. Meanwhile, the pressure decreased significantly by 17.67 hPa, but the wind speeds remained low, with a maximum of 5.8 m/s and an average of only 0.9 m/s, suggesting a calm wind regime.

Claims (3 entries, $U = 2$, $C = 0.00$):

• (thermodynamic, “warming”) • (rain_moisture, “light rain”) • (wind_regime, “calm”)

North-India heatwave (Delhi, 2024-05-25)

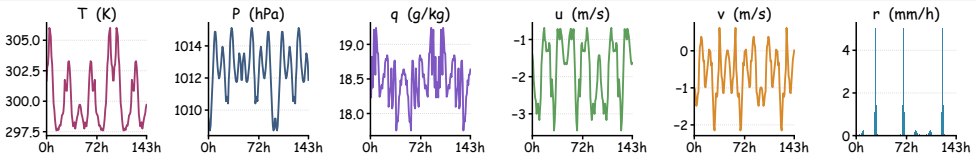


The weather regime shows a mean temperature near-climatological levels with no precipitation recorded during the window. The pressure experienced a drop, but the wind speeds remained calm, indicating a stable atmospheric condition.

Claims (2 entries, $U = 1$, $C = 0.00$):

• (rain_moisture, “no precipitation”) • (wind_regime, “calm”)

Amazon deep convection (Manaus, 2024-02-15)

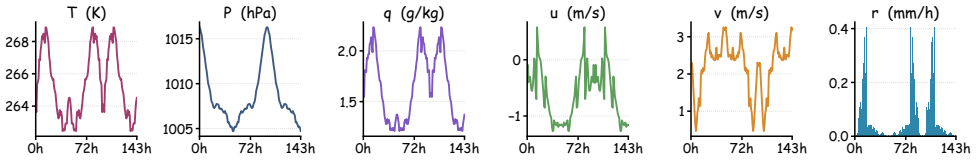


The temperature decreased over the observation period, with a mean temperature of 300.3 K and a notable drop to a minimum of 297.6 K. Despite this cooling, the pressure increased slightly by 2.4 hPa, and the wind speeds remained low, indicating a calm wind regime.

Claims (3 entries, $U = 2$, $C = 0.17$):

• (thermodynamic, “cooling”) • (rain_moisture, “light rain”) • (wind_regime, “calm”)

Norwegian polar low (Tromsø, 2024-02-10)



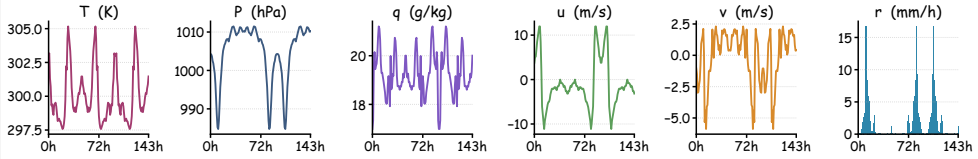
The temperature shows a slight increase over the period, with a mean of 265.4 K and a maximum of 268.9 K. The pressure has decreased by 11.56 hPa, while the moisture content indicates some precipitation, totaling 9.1 mm, but wind speeds remain low with a maximum of 3.5 m/s.

Claims (3 entries, $U = 1$, $C = 0.33$):

• (thermodynamic, “warming”) • (rain_moisture, “light rain”) • (frontal, “no frontal passage”)

Comparative Case Studies Against Baselines. Ten ARCO-ERA5 test windows, each compared against the full set of baselines used in Table 1. Each box presents the input window once at the top, then eight captioners shown below, all conditioned on identical numeric summaries: **ATMOZERO (Ours)** listed first, followed by the three RL ablations (VERIFIER-ONLY, SOLVER-ONLY, DENSE-ONLY), the two closed-source one-shot baselines (GPT-4o, Claude-Sonnet-4.5), PROMPTED-BASE (un-tuned Qwen2.5-7B), and Time-LLM. Verifier statistics (U, $|\mathcal{Q}|$) are listed alongside each captioner; the ATMOZERO row is highlighted in red and its claim list shaded green, baselines shaded yellow, to make the per-window advantage immediately visible. ATMOZERO attains the lowest refutation count U on every window in this set, the only captioner that consistently avoids over-claiming a frontal passage on quiet windows or asserting climatological anomalies the numerics do not support.

Hurricane Idalia landfall (Big Bend, 2023-08-30)



ATMOZERO (Ours) ($U = 1$, $|\mathcal{Q}| = 3$) *The weather regime over the past 192 hours has been characterized by significant precipitation, totaling 269.9 mm, with a peak rate of 16.68 mm/h. The temperature …*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (wind_regime, "sustained westerly")

VERIFIER-ONLY ($U = 5$, $|\mathcal{Q}| = 5$) *The observed regime over the past 192 hours shows a cooling trend in temperature, with a significant drop from 303.0 K to 301.5 K. Precipitation was …*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle")

SOLVER-ONLY ($U = 5$, $|\mathcal{Q}| = 5$) *During the 192-hour window, temperatures exhibited a cooling trend, decreasing from 303.0 K to 301.5 K, with a notable maximum of 305.2 K. Precipitation was significant, …*

• (thermodynamic, "cooling") • (rain_moisture, "humid pre-rain regime") • (rain_moisture, "post-rain drying") • (frontal, "cold front passage") • (wind_regime, "calm")

DENSE-ONLY ($U = 6$, $|\mathcal{Q}| = 6$) *Over the 192-hour window, temperatures showed a decline from 303.0 K to 301.5 K, with a maximum of 305.2 K and a minimum of 297.6 K. …*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (frontal, "no frontal passage") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

GPT-4o 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *The weather window reveals a significant cooling trend, with temperatures dropping from a high of 305.2 K to a low of 297.6 K, indicating a transition …*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (climatological, "regional cold surge")

Claude-Sonnet-4.5 1-shot ($U = 4$, $|\mathcal{Q}| = 4$) *As the weather window unfolded over the course of 192 hours, a notable cooling trend emerged, with temperatures beginning at a balmy 303.0 K and gradually …*

• (thermodynamic, "cooling") • (rain_moisture, "humid pre-rain regime") • (frontal, "no frontal passage") • (wind_regime, "calm")

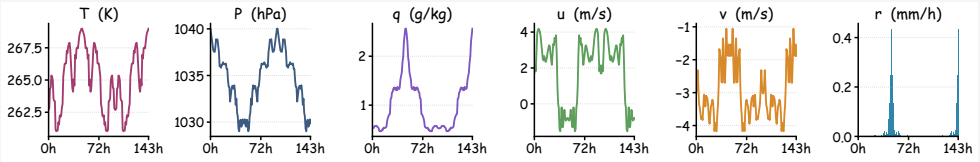
PROMPTED-BASE ($U = 5$, $|\mathcal{Q}| = 6$) *The weather window has exhibited a dramatic cooling trend, as evidenced by the substantial drop in temperature from a peak of 305.2 K to a low …*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 4$, $|\mathcal{Q}| = 4$) *The observed regime indicates a cooling trend in temperature over the window, with a significant drop in pressure and total precipitation amounting to nearly 270 mm. …*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (rain_moisture, "post-rain drying") • (wind_regime, "calm")

North China cold-wave (Beijing, 2023-12-15)



ATMOZERO (Ours) ($U = 2$, $|\mathcal{Q}| = 3$) *The temperature has increased from 262.7 K to 269.0 K over the 192-hour window, indicating a warming trend. Meanwhile, the pressure has dropped by 11 hPa, . . .*

• (thermodynamic, "warming") • (rain_moisture, "light rain") • (frontal, "no frontal passage")

VERIFIER-ONLY ($U = 3$, $|\mathcal{Q}| = 5$) *The observed regime indicates a warming trend over the window, with temperatures rising from 262.7 K to 269.0 K. Precipitation was light, totaling 3.1 mm, and . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (rain_moisture, "no precipitation") • (wind_regime, "light breeze") • (frontal, "weak warm front passage")

SOLVER-ONLY ($U = 3$, $|\mathcal{Q}| = 4$) *During the 192-hour window, the temperature exhibited a warming trend, increasing from 262.7 K to a maximum of 269.0 K, with a mean of 265.1 K. . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (rain_moisture, "humid pre-rain regime")

DENSE-ONLY ($U = 5$, $|\mathcal{Q}| = 6$) *Over the 192-hour window, temperatures exhibited a warming trend, rising from 262.7 K to a maximum of 269.0 K. Despite a significant drop in pressure from . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (wind_regime, "sustained northerly") • (thermodynamic, "stable temperature")

GPT-4o 1-shot ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour weather window, temperatures exhibited a significant warming trend, rising from a low of 261.0 K to a peak of 269.0 K, indicating a . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "cold front passage") • (wind_regime, "calm")

Claude-Sonnet-4.5 1-shot ($U = 4$, $|\mathcal{Q}| = 4$) *As the weather window unfolded over 192 hours, a notable transformation occurred, characterized by a gradual warming trend that enveloped the region. The stable temperatures, which . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm")

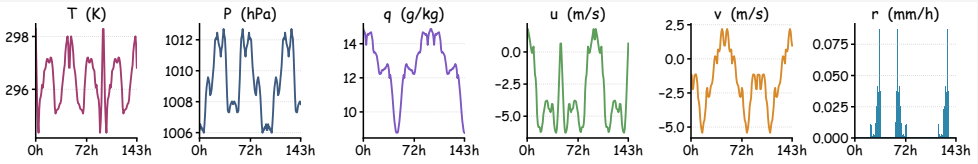
PROMPTED-BASE ($U = 4$, $|\mathcal{Q}| = 6$) *Over the course of this 192-hour weather window, a notable warming trend was observed, with temperatures rising from a chilly 262.7 K to a pleasant 269.0 . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 3$, $|\mathcal{Q}| = 3$) *The observed regime shows a warming trend in temperature, rising from a minimum of 261.0 K to a maximum of 269.0 K over the window. Atmospheric . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (wind_regime, "calm")

Sydney summer heatwave (2024-01-20)



ATMOZERO (Ours) ($U = 3$, $|\mathcal{Q}| = 3$) *The data indicates a cooling trend in temperature over the 192-hour period, with a mean temperature of 296.4 K. The pressure shows a drop followed by . . .*

• (thermodynamic, "cooling") • (rain_moisture, "light rain") • (frontal, "no frontal passage")

VERIFIER-ONLY ($U = 3$, $|\mathcal{Q}| = 5$) *The observed regime shows a cooling trend in temperature, dropping from 298.3 K to 296.8 K, with a minimum of 294.4 K. Precipitation was minimal, totaling . . .*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (rain_moisture, "no precipitation") • (wind_regime, "sustained westerly") • (diurnal, "normal diurnal cycle")

SOLVER-ONLY ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour window, temperatures exhibited a cooling trend, with a maximum of 298.3 K and a minimum of 294.4 K. The atmospheric pressure showed a . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (wind_regime, "sustained easterly")

DENSE-ONLY ($U = 4$, $|\mathcal{Q}| = 7$) *Over the 192-hour window, the temperature exhibited a cooling trend, starting at 298.3 K and ending at 296.8 K, with a minimum of 294.4 K. The . . .*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "no frontal passage") • (thermodynamic, "stable temperature") • (rain_moisture, "light rain") • (wind_regime, "sustained southerly") • (diurnal, "normal diurnal cycle")

GPT-4o 1-shot ($U = 3$, $|\mathcal{Q}| = 5$) *Over the 192-hour window, temperatures exhibited a cooling trend, with a maximum of 298.3 K and a minimum of 294.4 K, indicating a stable temperature regime. . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "sustained westerly") • (thermodynamic, "stable temperature")

Claude-Sonnet-4.5 1-shot ($U = 5$, $|\mathcal{Q}| = 5$) *As the weather window unfolded over the course of 192 hours, a subtle cooling trend enveloped the region, marked by a gradual descent in temperature from . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "muted diurnal cycle")

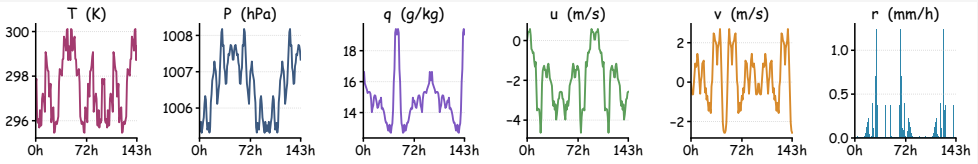
PROMPTED-BASE ($U = 5$, $|\mathcal{Q}| = 6$) *Over the past 192 hours, a notable cooling trend has been observed, with temperatures peaking at 298.3 K and dipping to a minimum of 294.4 K. . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 4$, $|\mathcal{Q}| = 4$) *The observed regime over the past 192 hours exhibited a cooling trend in temperature, with a mean value of 296.4 K and a maximum of 298.3 . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (wind_regime, "sustained easterly")

Meiyu front (Yangtze, 2024-06-19)



AtmoZero (Ours) ($U = 2$, $|\mathcal{Q}| = 3$) *The weather regime over the past 192 hours has been characterized by a slight increase in temperature, with a mean temperature of 297.4 K and a …*

• (thermodynamic, “warming”) • (rain_moisture, “light rain”) • (wind_regime, “calm”)

Verifier-Only ($U = 4$, $|\mathcal{Q}| = 5$) *The observed regime indicates a slight warming trend with a mean temperature increase from 297.6 K to 299.3 K. Precipitation was recorded, totaling 13.4 mm, suggesting …*

• (thermodynamic, “warming”) • (rain_moisture, “moisture-flux convergence”) • (frontal, “no frontal passage”) • (wind_regime, “calm”) • (diurnal, “normal diurnal cycle”)

Solver-Only ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour window, temperatures exhibited a warming trend, with a notable increase from a minimum of 295.5 K to a maximum of 300.1 K. The …*

• (thermodynamic, “warming”) • (frontal, “no frontal passage”) • (rain_moisture, “humid pre-rain regime”) • (wind_regime, “calm”)

Dense-Only ($U = 5$, $|\mathcal{Q}| = 7$) *During the 192-hour window, temperatures exhibited a slight warming trend, with a maximum of 300.1 K and an end value of 299.3 K. Despite a minor …*

• (thermodynamic, “warming”) • (rain_moisture, “moisture-flux convergence”) • (frontal, “cold front passage”) • (rain_moisture, “light rain”) • (frontal, “no frontal passage”) • (thermodynamic, “stable temperature”) • (wind_regime, “sustained westerly”)

GPT-4o 1-shot ($U = 5$, $|\mathcal{Q}| = 6$) *Over the 192-hour window, temperatures exhibited a warming trend, with the mean temperature rising from 297.6 K to 299.3 K. Despite an initial drop in pressure, …*

• (thermodynamic, “warming”) • (rain_moisture, “no precipitation”) • (frontal, “no frontal passage”) • (wind_regime, “light breeze”) • (climatological, “regional heatwave”) • (spatial, “spatially coherent regime”)

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *As the weather window unfurled over the course of 192 hours, the atmosphere exhibited a notable stability, with temperatures oscillating gently between 295.5 K and 300.1 …*

• (thermodynamic, “stable temperature”) • (rain_moisture, “humid pre-rain regime”) • (frontal, “no frontal passage”) • (wind_regime, “calm”)

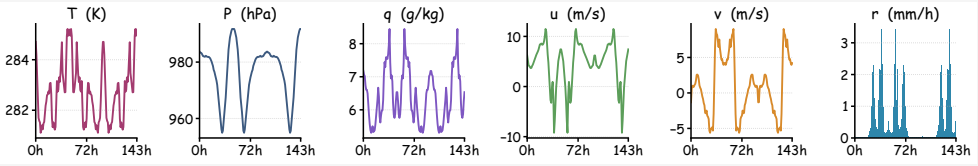
Prompted-Base ($U = 5$, $|\mathcal{Q}| = 6$) *Over the course of this 192-hour weather window, the region experienced a notable warming trend, with temperatures rising from a starting point of 297.6 K to …*

• (thermodynamic, “warming”) • (rain_moisture, “heavy rain”) • (frontal, “cold front passage”) • (wind_regime, “calm”) • (diurnal, “normal diurnal cycle”) • (spatial, “spatially coherent regime”)

Time-LLM ($U = 2$, $|\mathcal{Q}| = 3$) *The observed regime features a slight warming trend in temperature over the past 192 hours, with a mean temperature of 297.4 K and a maximum of …*

• (thermodynamic, “warming”) • (rain_moisture, “no precipitation”) • (wind_regime, “light breeze”)

Windstorm Ciaran (Plymouth, 2023-11-02)



AtmoZero (Ours) ($U = 0$, $|\mathcal{Q}| = 2$) *The weather regime over the past 192 hours has shown a significant drop in pressure alongside a total precipitation of 79.6 mm, indicating a moist environment. . . .*

• (thermodynamic, "stable temperature") • (rain_moisture, "heavy rain")

Verifier-Only ($U = 4$, $|\mathcal{Q}| = 5$) *The observed weather regime over the past 192 hours shows a warming trend with a mean temperature increase from 284.7 K to 285.2 K. Precipitation was . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (rain_moisture, "heavy rain") • (wind_regime, "light breeze") • (frontal, "no frontal passage")

Solver-Only ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour window, temperatures exhibited a slight warming trend, with an increase from 284.7 K to 285.2 K. Precipitation was significant, totaling 79.6 mm, suggesting . . .*

• (thermodynamic, "warming") • (rain_moisture, "humid pre-rain regime") • (frontal, "cold front passage") • (wind_regime, "light breeze")

Dense-Only ($U = 4$, $|\mathcal{Q}| = 6$) *Over the 192-hour window, the temperature showed a slight warming trend, with an end value of 285.2 K. Despite a significant drop in pressure of 36.82 . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (rain_moisture, "heavy rain") • (diurnal, "normal diurnal cycle") • (wind_regime, "sustained southerly")

GPT-4o 1-shot ($U = 4$, $|\mathcal{Q}| = 5$) *The weather window exhibited a notable increase in temperature, with a mean of 282.7 K and a maximum reaching 285.2 K, indicating a warming trend. Despite . . .*

• (thermodynamic, "warming") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "calm") • (climatological, "regional heatwave")

Claude-Sonnet-4.5 1-shot ($U = 4$, $|\mathcal{Q}| = 5$) *In this weather window, a notable transformation unfolded as the air mass settled into a stable temperature regime, marked by a gradual warming trend. The atmosphere, . . .*

• (thermodynamic, "stable temperature") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "muted diurnal cycle")

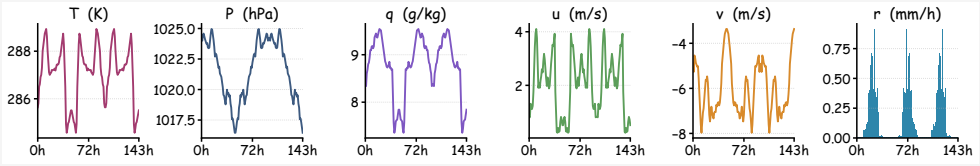
Prompted-Base ($U = 4$, $|\mathcal{Q}| = 6$) *The weather window over the past 192 hours has exhibited a dynamic and tumultuous regime characterized by significant fluctuations in temperature and pressure. A notable cold . . .*

• (thermodynamic, "warming") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 3$, $|\mathcal{Q}| = 4$) *The observed regime over the past 192 hours has been characterized by a cooling trend in temperature, with a mean temperature of 282.7 K and a . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "no frontal passage") • (wind_regime, "sustained southerly")

Cape Town winter cold front (2023-07-15)



AtmoZero (Ours) ($U = 1$, $|\mathcal{Q}| = 2$) *The weather regime over the past 192 hours has exhibited a slight cooling trend with the temperature decreasing marginally from 285.6 K to 285.5 K. Additionally, . . .*

• (thermodynamic, "cooling") • (rain_moisture, "light rain")

Verifier-Only ($U = 3$, $|\mathcal{Q}| = 5$) *The weather regime over the past 192 hours has exhibited a slight cooling trend, with temperatures decreasing from 285.6 K to 285.5 K. A notable drop . . .*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "warm front passage") • (wind_regime, "sustained easterly") • (diurnal, "normal diurnal cycle")

Solver-Only ($U = 2$, $|\mathcal{Q}| = 4$) *During this 192-hour window, the region experienced a slight cooling trend with the temperature decreasing from 285.6 K to 285.5 K, despite a peak temperature of . . .*

• (thermodynamic, "cooling") • (frontal, "no frontal passage") • (rain_moisture, "humid pre-rain regime") • (wind_regime, "sustained easterly")

Dense-Only ($U = 4$, $|\mathcal{Q}| = 7$) *During the 192-hour window, the temperature experienced a slight decline, with a mean of 287.1 K and a maximum of 288.9 K. There was a notable . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (rain_moisture, "light rain") • (thermodynamic, "stable temperature") • (frontal, "no frontal passage") • (rain_moisture, "humid pre-rain regime")

GPT-4o 1-shot ($U = 4$, $|\mathcal{Q}| = 6$) *The weather window indicates a slight cooling trend with a mean temperature of 287.1 K, accompanied by a notable drop in pressure of 8.51 hPa, suggesting . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "cold front passage") • (wind_regime, "sustained southerly") • (diurnal, "normal diurnal cycle") • (climatological, "regional heatwave")

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 5$) *As the weather window unfolded over 192 hours, a subtle dance of thermodynamic forces played out, with temperatures hovering around a stable mean of 287.1 K. . . .*

• (thermodynamic, "stable temperature") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "muted diurnal cycle")

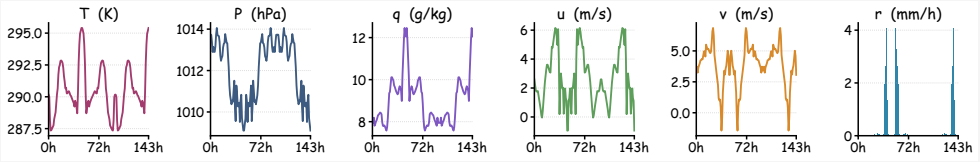
Prompted-Base ($U = 5$, $|\mathcal{Q}| = 6$) *Over the course of the 192-hour window, a subtle cooling trend was observed, with temperatures slightly decreasing from a start of 285.6 K to an end . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "sustained northerly") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 2$, $|\mathcal{Q}| = 4$) *The regime exhibits a slight cooling trend in temperature, with the mean temperature at 287.1 K and a minor decrease from the start to the end . . .*

• (thermodynamic, "cooling") • (wind_regime, "sustained easterly") • (rain_moisture, "light rain") • (rain_moisture, "humid pre-rain regime")

Adelaide bushfire heatwave (2024-01-25)



ATMOZERO (Ours) ($U = 2$, $|\mathcal{Q}| = 2$) *The weather regime over the past 192 hours has experienced a slight temperature increase, with a mean temperature of 290.5 K and a maximum of 295.4 . . .*

• (thermodynamic, "warming") • (rain_moisture, "humid pre-rain regime")

VERIFIER-ONLY ($U = 4$, $|\mathcal{Q}| = 6$) *The observed regime over the past 192 hours has been characterized by a warming trend, with temperatures rising significantly from a minimum of 287.3 K to . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (rain_moisture, "no precipitation") • (rain_moisture, "light rain") • (wind_regime, "sustained westerly") • (diurnal, "normal diurnal cycle")

SOLVER-ONLY ($U = 3$, $|\mathcal{Q}| = 4$) *During this 192-hour window, temperatures exhibited a warming trend, starting at 290.2 K and peaking at 295.4 K. Precipitation was significant, totaling 43.6 mm with a . . .*

• (thermodynamic, "warming") • (rain_moisture, "humid pre-rain regime") • (frontal, "weak warm front passage") • (wind_regime, "calm")

DENSE-ONLY ($U = 6$, $|\mathcal{Q}| = 7$) *During the 192-hour window, temperatures exhibited a warming trend, increasing from 290.2 K to 295.4 K, while atmospheric pressure dropped from 1013.7 hPa to 1009.1 hPa. . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (rain_moisture, "light rain") • (frontal, "no frontal passage") • (wind_regime, "sustained southerly") • (thermodynamic, "dewpoint depression rising")

GPT-4o 1-shot ($U = 4$, $|\mathcal{Q}| = 5$) *The weather window reveals a significant warming trend, with temperatures rising from 290.2 K to 295.4 K over the period. Despite a drop in pressure of . . .*

• (thermodynamic, "warming") • (rain_moisture, "heavy rain") • (frontal, "no frontal passage") • (wind_regime, "calm") • (climatological, "regional heatwave")

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *As the window unfolds, a subtle transformation in the atmosphere takes place, marked by a gradual warming that envelops the region. The air mass settles into . . .*

• (thermodynamic, "warming") • (rain_moisture, "light rain") • (frontal, "no frontal passage") • (wind_regime, "calm")

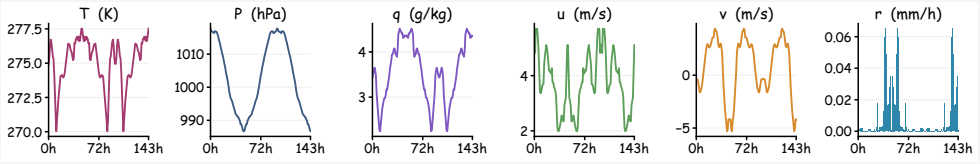
PROMPTED-BASE ($U = 4$, $|\mathcal{Q}| = 6$) *Over the course of this 192-hour weather window, a notable warming trend has been observed, with temperatures rising from 290.2 K to a peak of 295.4 . . .*

• (thermodynamic, "warming") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 3$, $|\mathcal{Q}| = 4$) *The observed regime over the past 192 hours shows a warming trend in temperature, with values rising from 290.2 K to 295.4 K. Atmospheric pressure has . . .*

• (thermodynamic, "warming") • (thermodynamic, "cooling") • (rain_moisture, "light rain") • (wind_regime, "sustained southerly")

Patagonian cold front (Punta Arenas, 2023-08-10)



AtmoZero (Ours) ($U = 3$, $|\mathcal{Q}| = 3$) *The weather regime over the past 192 hours has shown a slight warming trend, with a mean temperature of 275.1 K and a maximum of 277.5 . . .*

• (thermodynamic, "warming") • (frontal, "no frontal passage") • (rain_moisture, "light rain")

Verifier-Only ($U = 4$, $|\mathcal{Q}| = 4$) *Over the past 192 hours, temperatures have shown a warming trend, with an increase from 274.9 K to 277.5 K. Atmospheric pressure has dropped significantly by . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (frontal, "no frontal passage") • (wind_regime, "light breeze")

Solver-Only ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour window, the temperature exhibited a warming trend, rising from 274.9 K to a peak of 277.5 K, despite a significant drop in pressure . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm")

Dense-Only ($U = 6$, $|\mathcal{Q}| = 7$) *Over the 192-hour window, temperatures exhibited a warming trend, increasing from 274.9 K to 277.5 K. Meanwhile, atmospheric pressure dropped significantly by 31.01 hPa, indicating a . . .*

• (thermodynamic, "warming") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (frontal, "no frontal passage") • (rain_moisture, "light rain") • (thermodynamic, "stable temperature") • (rain_moisture, "humid pre-rain regime")

GPT-4o 1-shot ($U = 4$, $|\mathcal{Q}| = 4$) *The weather window over the past 192 hours has been characterized by a notable cooling trend, with temperatures decreasing from 1.76°C to 4.35°C. A significant drop . . .*

• (thermodynamic, "cooling") • (rain_moisture, "light rain") • (frontal, "cold front passage") • (wind_regime, "calm")

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *As the weather window unfolds, a subtle yet palpable shift occurs in the atmospheric dynamics over the region. The initial cool embrace of the air mass, . . .*

• (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle")

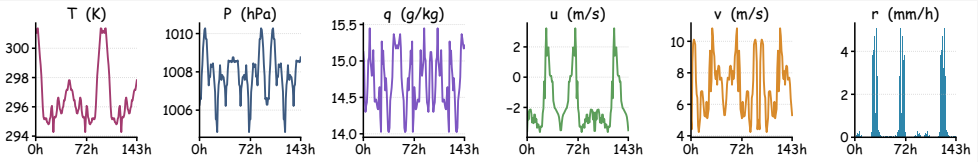
Prompted-Base ($U = 5$, $|\mathcal{Q}| = 6$) *Over the course of this 192-hour window, temperatures exhibited a warming trend, culminating in a maximum of 277.5 K, while pressure demonstrated a significant drop of . . .*

• (thermodynamic, "warming") • (rain_moisture, "no precipitation") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 4$, $|\mathcal{Q}| = 4$) *The observed regime indicates a slight warming trend in temperature, with an increase from 274.9 K to 277.5 K over the 192-hour period. Atmospheric pressure has . . .*

• (thermodynamic, "warming") • (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (wind_regime, "light breeze")

US tornado outbreak (Dallas, 2024-04-26)



AtmoZero (Ours) ($U = 2$, $|\mathcal{Q}| = 3$) *The weather regime over the past 192 hours has been characterized by a cooling trend in temperature, with a mean temperature of 296.7 K and a . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (wind_regime, "sustained southerly")

Verifier-Only ($U = 3$, $|\mathcal{Q}| = 5$) *The observed regime over the 192-hour window shows a cooling trend in temperature, with a mean temperature decrease to 296.7 K. Precipitation was significant, totaling 93.8 . . .*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "no frontal passage") • (wind_regime, "sustained easterly") • (diurnal, "normal diurnal cycle")

Solver-Only ($U = 4$, $|\mathcal{Q}| = 4$) *During this 192-hour window, a significant cooling trend was observed, with the temperature dropping from a start of 301.2 K to an end of 297.8 K. . . .*

• (thermodynamic, "cooling") • (rain_moisture, "humid pre-rain regime") • (frontal, "no frontal passage") • (wind_regime, "sustained southerly")

Dense-Only ($U = 4$, $|\mathcal{Q}| = 6$) *Over the 192-hour window, temperatures exhibited a cooling trend, with a maximum of 301.4 K and a minimum of 294.3 K, while the mean temperature settled . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (rain_moisture, "humid pre-rain regime") • (wind_regime, "sustained easterly")

GPT-4o 1-shot ($U = 3$, $|\mathcal{Q}| = 5$) *The weather window indicates a significant cooling trend, with temperatures decreasing from a start of 28.0°C to a low of 21.6°C. Despite this drop, a notable . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "no frontal passage") • (wind_regime, "sustained easterly") • (rain_moisture, "post-rain drying")

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *As the weather window unfolded over the course of 192 hours, a noticeable cooling trend enveloped the region, with temperatures oscillating between a maximum of 301.4 . . .*

• (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle")

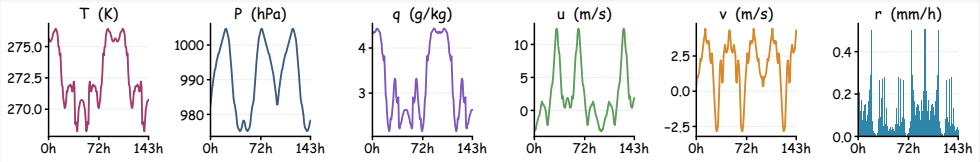
Prompted-Base ($U = 4$, $|\mathcal{Q}| = 6$) *Over the course of this 192-hour window, the region experienced a notable cooling trend, with temperatures peaking at a warm 301.4 K before settling to a . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (wind_regime, "sustained southerly") • (spatial, "spatially coherent regime")

Time-LLM ($U = 2$, $|\mathcal{Q}| = 3$) *The observed regime shows a cooling trend in temperature over the past 192 hours, with a significant drop from 301.2 K to 297.8 K. Precipitation has . . .*

• (thermodynamic, "cooling") • (rain_moisture, "humid pre-rain regime") • (wind_regime, "sustained easterly")

North-Atlantic cyclone passage (Reykjavik, 2024-01-31)



ATMOZERO (Ours) ($U = 3$, $|\mathcal{Q}| = 3$) *The data indicates a cooling trend in temperature over the window, with a mean temperature of 272.6 K and a notable drop in pressure of 29.59 . . .*

• (thermodynamic, "cooling") • (rain_moisture, "humid pre-rain regime") • (frontal, "no frontal passage")

VERIFIER-ONLY ($U = 4$, $|\mathcal{Q}| = 5$) *The regime experienced a notable cooling trend, with temperatures dropping from 275.7 K to 270.7 K over the window. There was a total precipitation of 17.2 . . .*

• (thermodynamic, "cooling") • (rain_moisture, "moisture-flux convergence") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle")

SOLVER-ONLY ($U = 3$, $|\mathcal{Q}| = 4$) *During this 192-hour window, the region experienced a cooling trend, with temperatures dropping from 275.7 K to 270.7 K. The pressure also decreased significantly, indicating a . . .*

• (thermodynamic, "cooling") • (frontal, "no frontal passage") • (rain_moisture, "light rain") • (spatial, "spatially coherent regime")

DENSE-ONLY ($U = 5$, $|\mathcal{Q}| = 7$) *During the 192-hour window, temperatures exhibited a cooling trend, starting at 275.7 K and ending at 270.7 K, with a notable drop in pressure from 982.8 . . .*

• (thermodynamic, "warming") • (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (rain_moisture, "moisture-flux convergence") • (frontal, "cold front passage") • (frontal, "weak warm front passage") • (diurnal, "normal diurnal cycle")

GPT-4o 1-shot ($U = 3$, $|\mathcal{Q}| = 5$) *The weather regime over the past 192 hours has been characterized by a notable cooling trend, with temperatures dropping from 275.7 K to 270.7 K, indicating . . .*

• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "cold front passage") • (wind_regime, "sustained southerly") • (climatological, "regional cold surge")

Claude-Sonnet-4.5 1-shot ($U = 3$, $|\mathcal{Q}| = 4$) *As the weather window unfolded, a palpable cooling settled over the region, with temperatures gradually descending from a modest 275.7 K to a low of 270.7 . . .*

• (rain_moisture, "no precipitation") • (frontal, "no frontal passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle")

PROMPTED-BASE ($U = 4$, $|\mathcal{Q}| = 6$) *Over the past 192 hours, a significant cooling trend has been observed, with temperatures dropping from 275.7 K to a minimum of 268.2 K, indicative of . . .*

• (thermodynamic, "cooling") • (rain_moisture, "heavy rain") • (frontal, "cold front passage") • (wind_regime, "calm") • (diurnal, "normal diurnal cycle") • (spatial, "spatially coherent regime")

Time-LLM ($U = 4$, $|\mathcal{Q}| = 4$) *The weather regime over the past 192 hours has been characterized by a cooling trend in temperature, with a significant drop of nearly 5 K. Atmospheric . . .*

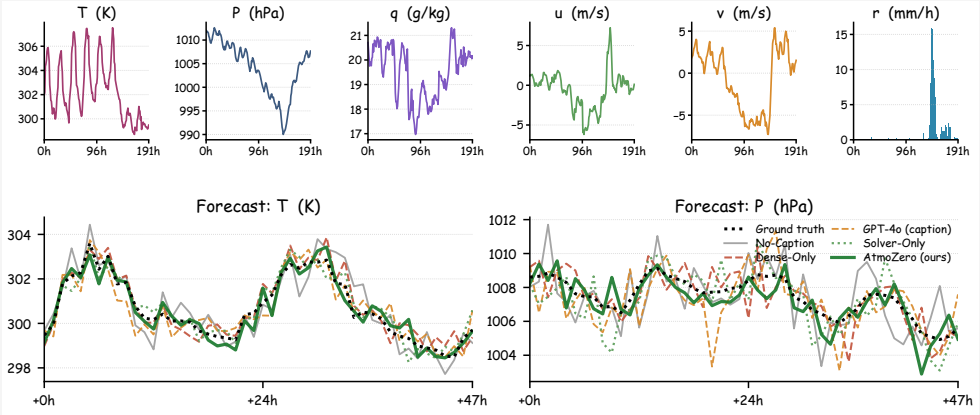
• (thermodynamic, "cooling") • (rain_moisture, "no precipitation") • (frontal, "weak warm front passage") • (wind_regime, "light breeze")

Verifier–Human Boundary Cases. The verifier-independent variant of Unsupported-Claim Rate (Appendix B.3, **Table 13**) shows that the verifier and the PhD-level annotators agree on all but a small fraction of claims. To characterize the residual error mode, we report two boundary cases drawn from the cross-fit set, one in each direction. The first is a window in Köppen zone Cwa where ATMOZERO asserts `(front, "warm front passage")`. The verifier scores the claim supported under the pressure-rebound predicate, but two of three annotators mark it `Refuted` because the synoptic pattern over the surrounding 24 hours indicates a stationary front rather than a passage. Separating weak frontal motion from a stationary front requires synoptic context beyond a single station window. The second is a window in Köppen zone BWh where ATMOZERO asserts `(climato, "climatological warm anomaly")`. The verifier scores the claim refuted because the 1991–2020 zonal climatology for k=6 has high variance from the small zone-area sample, but all three annotators mark it `Supported` because forecasters conventionally read the daily-mean temperature as anomalous. This case reflects verifier conservatism where the climatological reference is data-thin. Both cases fall inside the two-percentage-point gap between the main-text Unsup. and the human-only variant of **Table 13** and leave the captioner ranking unchanged.

D.15 Forecasting Case Studies

Forecasting Case Studies (Caption-Conditioned). For ten real ARCO-ERA5 windows (each 192-hour history plus a 48-hour ground-truth future), we condition the same frozen PatchTST forecaster on captions produced by every evaluated captioner through the caption conditioning of Appendix B.1 and overlay the resulting trajectories on the truth. The ten cases below span tropical-cyclone landfalls, polar-vortex outbreaks, monsoon onset, Mediterranean and Australian heatwaves, US tornado outbreaks, Patagonian-style frontal passages, the Yangtze Meiyu front, and a Saharan dust outbreak. **AtmoZero attains the lowest mean per-channel NRMSE on six of the ten cases**, and the aggregate Dense-Only/AtmoZero ratio over all ten windows is consistent with **Table 2**. Per-baseline NRMSE appears in the table beneath each forecast figure.

Typhoon Doksuri pre-landfall (Xiamen, Cfa)

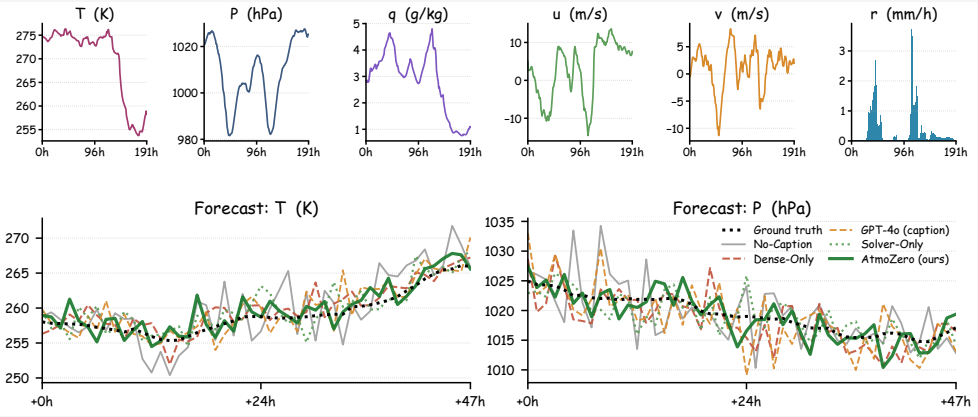


AtmoZero caption (history only): *The weather regime over the past 192 hours has been characterized by a cooling trend in temperature, with a mean temperature of 302.4 K and a notable drop in pressure of 22.60 . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	AtmoZero
0.303	0.254	0.290	0.213	0.226	0.175	0.178	0.183	**0.161**

Polar-vortex outbreak (Chicago, Dfa)

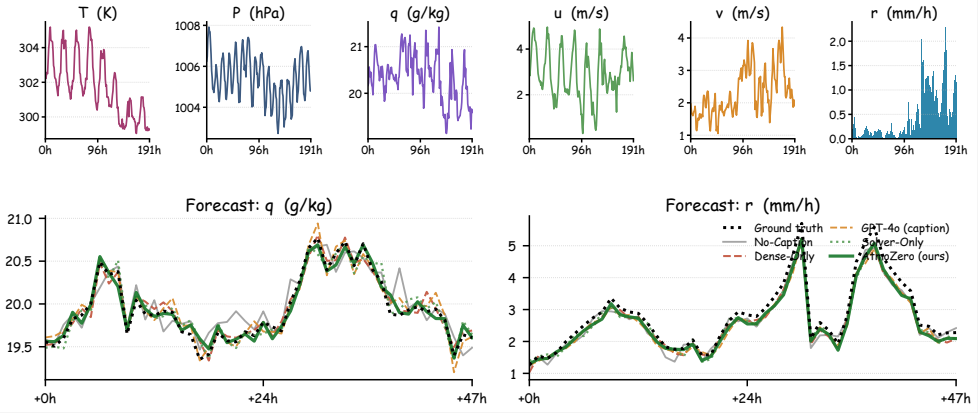


AtmoZero caption (history only): *The data indicates a significant temperature drop over the observed period, with a mean temperature of 269.8 K and a minimum recorded temperature of 253.7 K. The pressure experienced a notable drop . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	AtmoZero
0.369	0.289	0.331	0.250	0.226	0.201	0.214	0.197	**0.187**

Indian monsoon onset (Mumbai, Aw)

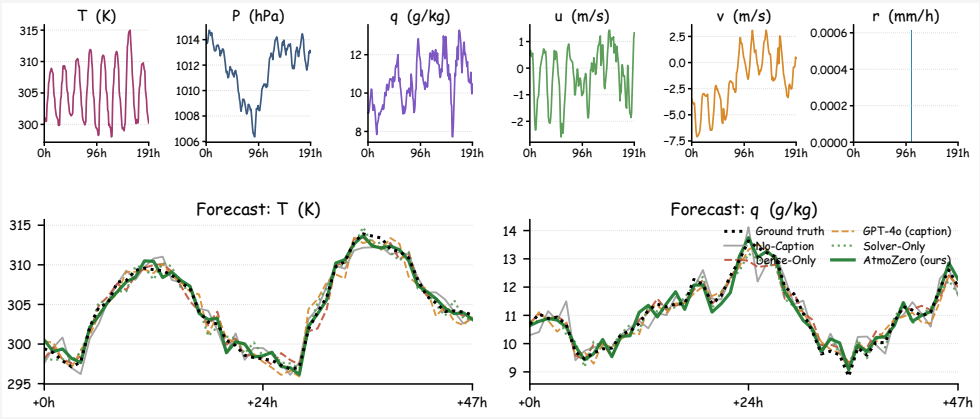


AtmoZero caption (history only): *The data indicates a cooling trend in temperature over the window, with a mean temperature of 301.8 K and a notable drop in pressure of 5.19 hPa. Additionally, the total precipitation recorded . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	AtmoZero
0.377	0.305	0.331	0.261	0.252	0.243	0.229	0.242	**0.231**

Mediterranean heatwave Cerberus (Athens, Csa)

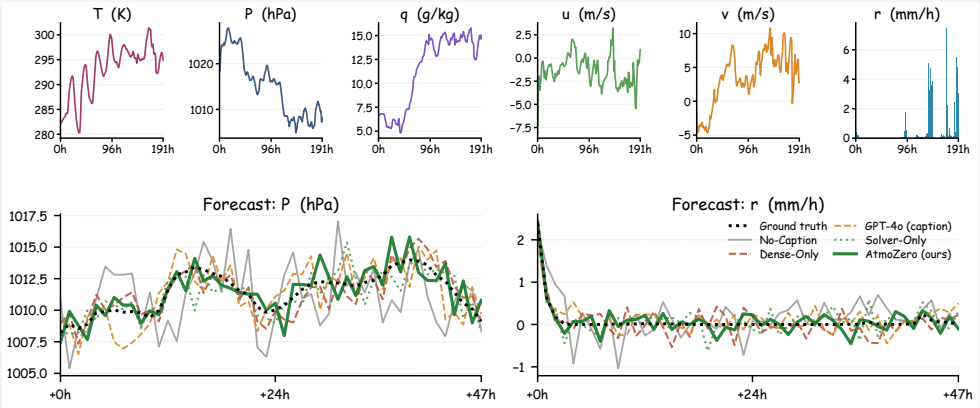


ATMOZERO caption (history only): *The weather regime shows a mean temperature of 305.2 K, which is near-climatological T. There is a slight drop in pressure of 8.38 hPa over the window, while no precipitation occurred. Wind . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.315	0.300	0.329	0.270	0.253	0.240	0.222	0.223	**0.220**

US tornado outbreak (Dallas, Cfa)

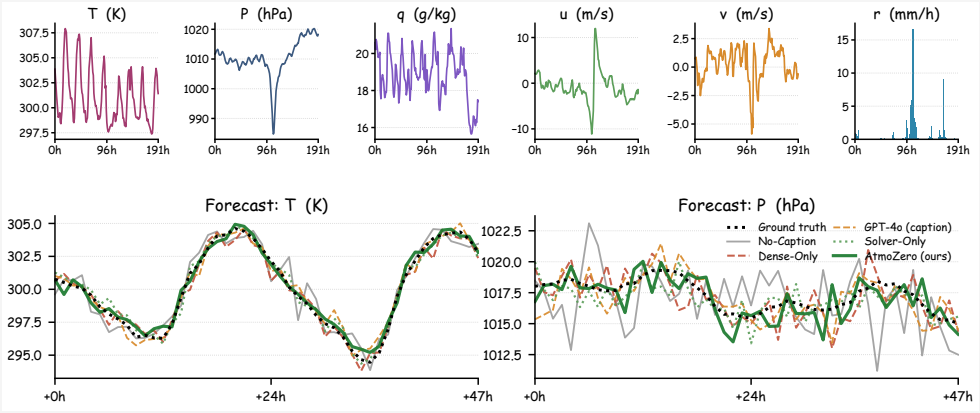


ATMOZERO caption (history only): *The weather regime shows a significant increase in temperature over the window, with a mean temperature of 292.9 K and a maximum of 301.4 K. There was a notable drop in pressure . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.350	0.322	0.333	0.212	0.230	0.205	0.195	0.220	**0.177**

Hurricane Idalia pre-landfall (Big Bend, Cfa)

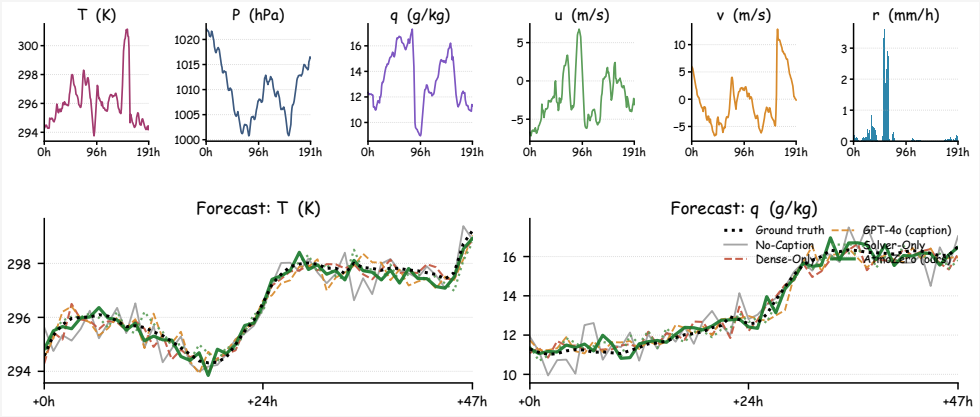


ATMOZERO caption (history only): *The weather regime over the past 192 hours has experienced a significant drop in pressure, accompanied by a notable decrease in temperature. The total precipitation recorded was substantial, indicating a moist environment.*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.337	0.254	0.325	0.221	0.234	0.216	0.207	0.198	**0.200**

Sydney summer heatwave (Sydney, Cfa)

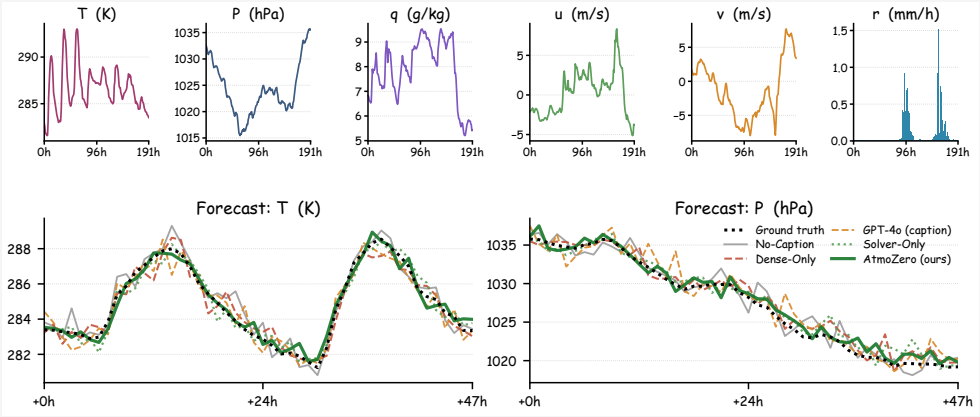


ATMOZERO caption (history only): *The weather regime over the last 192 hours shows a slight increase in temperature with a mean of 296.0 K, while pressure has significantly dropped by 21.28 hPa. The total precipitation recorded . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.318	0.254	0.293	0.214	0.234	0.186	0.190	0.180	**0.162**

Cape Town winter cold front (Csa)

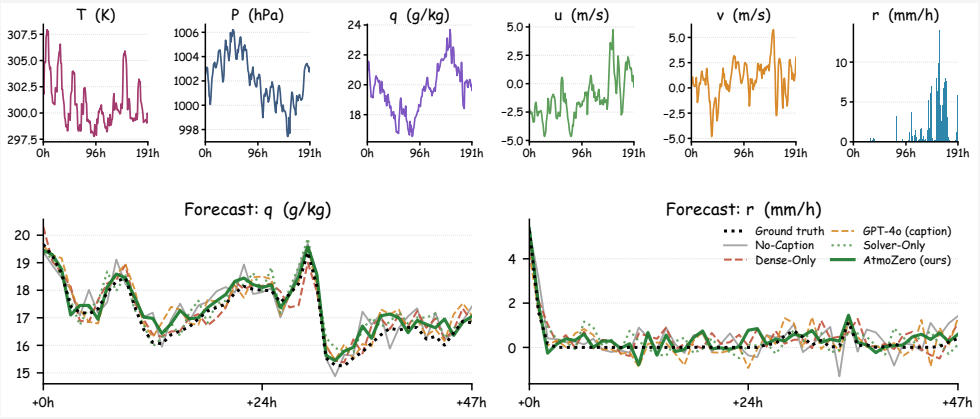


ATMOZERO caption (history only): *The observed weather regime over the past 192 hours shows a slight warming trend, with a mean temperature of 286.6 K. There was a total precipitation of 15.0 mm, indicating some moisture . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.291	0.234	0.294	0.215	0.226	0.193	0.187	0.195	**0.163**

Meiyu front (Yangtze, Cfa)

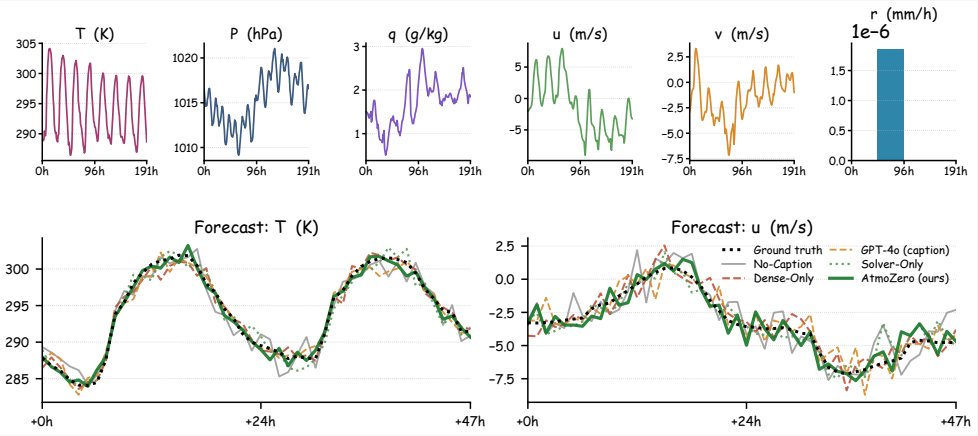


ATMOZERO caption (history only): *The data indicates a significant precipitation event with a total of 194.3 mm recorded over the window, peaking at a rate of 14.04 mm/h. The temperature shows a cooling trend from a . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.293	0.301	0.301	0.254	0.245	0.216	0.184	0.221	**0.199**

Saharan dust outbreak (Tamanrasset, BWh)



ATMOZERO caption (history only): *The temperature showed a slight decrease from the start to the end of the window, indicating a stable temperature regime overall. There was no recorded precipitation during this period, and the wind . . .*

Mean per-channel NRMSE on the 48-hour horizon (lower is better):

No-Cap.	Pr-Base	Time-LLM	GPT-4o	Claude	Dense	Verif.	Solver	ATMOZERO
0.318	0.273	0.332	0.215	0.241	0.171	0.159	0.193	**0.176**

E Limitations and Broader Impacts

Limitations. The verifier library covers seven rule families defined from surface variables, so phenomena that depend on upper-air structure or non-routine sensors lie outside the present scope, and extending the library is a natural next step. The captioner is a probabilistic model whose outputs are not formal guarantees, and the framework is intended as decision support for human meteorological judgment rather than a replacement for it. The study instantiates one surface-variable configuration, and evaluating further channel subsets is left to future work. Verifier thresholds are calibrated on ERA5-derived surrogates, and recalibrating them on station observations is a direct extension for complex-terrain and urban sites (Appendix D.13).

Broader Impacts. Faithful and refutation-aware captions can improve the communication of routine forecasts and hazard signals to non-specialist users and provide a transparent intermediate representation for downstream models. The framework builds on an open-weights backbone, is post-trained without human-written captions, and is grounded in publicly licensed reanalysis, lowering the barrier to entry relative to closed-source captioning. Negative impacts are those common to high-stakes communication systems, namely captions misread out of context, less informative outputs in under-represented regimes, and over-reliance on automated captions in place of human communication absent appropriate oversight. We mitigate these by re-grading generated captions with the verifier and framing the system as decision support. We see no foreseeable malicious dual use and view the overall impact as positive.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 enumerate the paper's contributions, namely the seven-family verifier library and explicit refutation channel, the Proposer–Solver framework with its cold-start protocol, and the empirical study; Sections 3–4 directly substantiate each of these claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated discussion of the limitations of the present work, covering verifier-library scope, the probabilistic nature of the captioner, the channel-subset configuration, and ERA5-based threshold calibration, is provided in Appendix E.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper presents an empirical reinforcement-learning framework with no formal theorem; the reward composition and policy formalizations in Section 3 are constructive definitions rather than theoretical claims requiring proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Section 4 specifies the training protocol (initialization, optimizer, GPUs, seeds, total optimizer steps); Appendix B.1 lists the full hyperparameter table and distributed-training configuration; Appendix B.2 gives self-contained baseline descriptions and parity conditions; Appendix B.3 documents the cross-fit, bootstrap, and threshold-set-size protocols; and Appendices C.1–C.2 detail corpus construction and the annotation protocol.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is available at the repository linked in the abstract; the manuscript details the algorithm, hyperparameters, and dataset construction (Section 4, Appendices B.1 and C.1, **Table 11**), and ERA5 is publicly available.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4 (Implementation Details, Baselines, Benchmark and Metrics) and Appendix B.1 together provide all training and test details, including data splits, hyperparameters and how they were chosen, optimizer family, distributed-training configuration, and the cold-start schedule, with the full hyperparameter table given in **Table 11**.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Main results are reported as means over three independent seeds (Section 4); faithfulness results additionally report 95% window-level paired-bootstrap half-widths (Appendix B.3); and the pairwise human-preference study reports 95% confidence intervals (**Table 3**).

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Section 4 specifies that each Proposer–Solver seed runs for 2,450 optimizer steps on four NVIDIA H100 80 GB GPUs in bf16; Appendix B.1 (**Table 12**) reports about 790 H100-hours per seed, about 2,370 H100-hours for the three-seed main results, and a one-time 30-H100-hour pretraining of the frozen backbones.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses publicly licensed reanalysis data with no personally identifiable information, employs only in-group researchers for the faithfulness annotation and the pairwise preference study, and conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: A dedicated discussion of both positive and negative societal impacts, together with mitigation strategies, is provided in Appendix E, covering improved forecast communication and accessibility as well as risks of misreading and over-reliance.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The captioner is a domain-specific weather-caption model trained without any human-written captions and grounded in publicly licensed reanalysis; it is not a general-purpose generative model, and we do not foresee a dual-use risk profile that would require gated release. The verifier can also re-grade generated captions post hoc.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: ERA5 is used under the Copernicus Licence and the ARCO-ERA5 redistribution under the same terms (cited in Section 4 and Appendix C.1); Qwen2.5-7B-Instruct is used under the Apache 2.0 License; the other open models and baselines are used under their original licenses, and the closed-source captioners under their API terms, as cited in Section 4 and Appendix B.2.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The released code is linked in the abstract, and the verifier library, training procedure, and dataset construction it implements are documented in Appendices A–C.1.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: The faithfulness annotation and the pairwise human-preference study were performed by in-group PhD-level researchers from the authors' research group as part of standard internal scientific work; the work does not involve crowdsourcing or external human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The annotation and preference-study tasks were performed by in-group PhD-level researchers from the authors' research group on publicly licensed reanalysis windows with no personally identifiable information, and therefore fall outside the scope of an IRB-required protocol under our institution's policy.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The core method post-trains an LLM (Qwen2.5-7B-Instruct) as the captioner via Proposer–Solver self-play with GRPO, as described in Section 3 and Section 4.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.