WEATHERJEPA: JOINT EMBEDDING OF OBSERVATIONAL AND ATMOSPHERIC STATE REPRESENTATIONS FOR WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate weather forecasting depends not only on learning atmospheric dynamics, but also on determining how atmospheric information should be represented. Many leading data-driven weather models predict future gridded atmospheric fields from reanalysis-based training data, whereas observation-driven systems infer atmospheric representations directly from heterogeneous measurements. These approaches reveal a fundamental representation trade-off: atmospheric state representations provide coherent and physically structured information but inherit assumptions from assimilation systems, while observational representations preserve direct measurements but must discover atmospheric structure from incomplete evidence. We introduce WEATHERJEPA, a joint embedding framework that aligns complementary observational and atmospheric-state representations into a shared latent space. The atmospheric-state representation provides a structured view during training, while the observational representation defines the deployment pathway. Rather than reconstructing one representation from the other, WEATHERJEPA learns their relationship through joint embedding, grounding, and temporal prediction objectives. Frozen-state readouts and ten-day forecasts show that initial accessibility does not consistently predict forecast ranking. Under matched observation-driven comparisons, WEATHERJEPA achieves the lowest aggregate probe errors for both physical fields and in-situ observations. Across six representative station, radiosonde, and satellite targets, it reduces Day-10 forecast error by 4.4–12.7% relative to the strongest competing learned method for each target. On ICOADS marine zonal wind observations, its Day 1–10 mean MAE is 4.4% lower than IFS-HRES.

1 INTRODUCTION

Modern AI weather forecasters achieve strong performance when initialized from globally gridded atmospheric states (Bi et al., 2023; Lam et al., 2023; Bodnar et al., 2025). These states are products of data assimilation; the underlying observing system supplies sparse station measurements, unevenly distributed radiosonde profiles, and satellite radiances (Hersbach et al., 2020). Forecasting from observations therefore requires constructing the state that will be evolved. Viewed as a world-modeling problem, this means inferring an internal atmospheric state from partial measurements, evolving it without future observations, and decoding predictions of future measurements. *What should such a state preserve, and how should it be learned?*

A useful forecast state must make atmospheric information accessible and organize it for future prediction. Predictive representation learning, including DINO-WM and V-JEPA 2, motivates learning predictive structure in latent space (Zhou et al., 2025; Assran et al., 2025). For weather forecasting, we distinguish *semantic accessibility*, what physical fields and measurements can be decoded now, from *temporal retention*, what remains useful after rollout. These requirements must be tested separately: accurate current-state decoding does not establish whether that information remains useful for forecasting future measurements.

Existing observation-driven systems instantiate this state in different ways. Analysis-mediated systems such as Aardvark and FuXi Weather estimate gridded physical states from observations

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

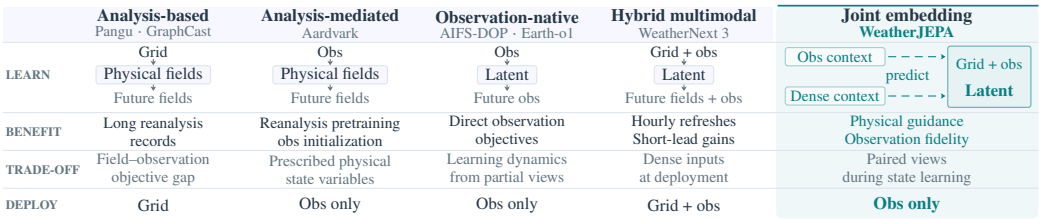

Figure 1: **Learning interfaces and deployment requirements.** Field fidelity need not imply observation fidelity. WeatherJEPA jointly grounds a shared latent target, predicts it from both contexts, and trains dynamics on longer dense-state records. Grid: atmospheric fields; obs: observations.

(Allen et al., 2025; Sun et al., 2025); GraphDOP and AIFS-DOP evolve latent states to predict measurements (Alexe et al., 2024; Pinnington et al., 2026); and Earth-o1 models evolution on a continuous, grid-free spatial domain (Gong et al., 2026). These designs differ in whether the evolving state is expressed as physical fields or learned latent features. Can a learned state benefit from dense physical guidance while retaining the freedom to represent what future observation prediction requires?

Reanalysis provides spatially complete context for learning from unevenly sampled observations. ERA5 combines observations with a physical model (Hersbach et al., 2020), although it remains an observation-constrained estimate rather than ground truth (Carrassi et al., 2018; Jin et al., 2024). We use it as a *dense training-time view and grounding target*. Joint embedding links this physical guidance to the observation-derived state, while observation reconstruction and temporal prediction jointly shape its learned features. This design provides dense atmospheric structure without restricting the state to selected physical variables, preserving the freedom to learn features for observation forecasting.

We introduce WEATHERJEPA, a joint-embedding framework for learning *predictive atmospheric states* for observation-driven world modeling. A paired observation–dense route learns reference coordinates, with grounding objectives that keep both physical fields and measurements decodable. An observation-only route learns to recover these coordinates through latent cross-view prediction. A restricted auxiliary dynamics predictor shapes the representation toward simple local evolution. Together, these objectives address readable state content, deployment recoverability, and latent forecastability. At inference, observations alone initialize the state; a forecast model rolls it forward, and observation readouts produce predictions. ERA5 supplies the dense view during training. Figure 1 contrasts the roles of dense fields in state learning and deployment.

We compare WEATHERJEPA with controlled adaptations of the three competing paradigms using matched frozen probes of 24 physical variables and nine observation families, ten-day forecasts, and objective ablations. Initial probe rankings do not consistently predict forecast rankings, and the strongest competitor changes with forecast horizon. Within this benchmark, WEATHERJEPA achieves the lowest mean physical-field probe error and the strongest *in situ* accessibility, while competing states retain advantages on several remote-sensing families. It also has the lowest forecast error among the compared learned methods on every evaluated IGRA variable–level target, both GridSat channels, and HadISD TAS at every lead from Day 1 to Day 10. On ICOADS marine zonal wind observations, its MAE averaged over Day 1–10 is 4.4% lower than that of IFS-HRES. Ablations connect readability to the two grounding terms and show that temporal shaping improves independent 24-hour latent predictability.

Contributions.

- We introduce WEATHERJEPA, which jointly embeds observation and reanalysis views during training to learn a predictive latent state that can be constructed from observations alone at deployment.
- We characterize atmospheric state constructions through frozen readouts and ten-day rollouts, finding rankings that vary with interface and horizon and consistent forecast gains over matched learned methods within this benchmark.
- We identify complementary roles for dense-field and observation grounding in preserving readable state content, and for temporal shaping in improving independent 24-hour latent predictability.

2 Related Work

Dense atmospheric state models. Most high-skill global AI forecasters evolve gridded atmospheric states initialized from analyses or reanalyses. Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), and AIFS (Lang et al., 2024) learn to forecast the evolution of these states. Beyond task-specific forecasting, foundation models such as Aurora (Bodnar et al., 2025) extend this paradigm toward broader Earth-system representations, while AtmoRep (Lessig et al., 2023) learns a task-independent stochastic model of atmospheric dynamics from historical dense-state trajectories. Together, these works learn representations or forecast dynamics from gridded physical fields, while largely relying on externally supplied fields rather than learning to construct them from heterogeneous raw observations.

Forecast states constructed from observations. Aardvark (Allen et al., 2025) and FuXi Weather (Sun et al., 2025) construct gridded atmospheric states from observations, with FuXi Weather also using its own cycled background forecasts; OMG-HD (Zhao et al., 2024) forecasts regional surface fields from station, radar, and satellite inputs. GraphDOP (Alexe et al., 2024) and AIFS-DOP (Pinnington et al., 2026) instead evolve observation-derived internal states and decode future observations without requiring a conventional analysis at deployment. DAWP (Gong et al., 2025) forecasts in satellite-observation space; OCELOT (Gholoubi et al., 2026) combines geometry-aware, instrument-specific encoders with a shared spherical latent; Earth-o1 (Gong et al., 2026) describes an observation-native atmospheric world model; and WeatherNext 3 (Rasp et al., 2026) combines analysis-derived atmospheric fields with recent satellite observations for hourly forecasting. Together, these methods span observation-only and analysis-augmented inference pipelines. Our focus is the state-learning problem within atmospheric world modeling: what do different state-construction paradigms make accessible, and what remains predictive after rollout?

Predictive latent world models. Latent world models learn predictive structure by making predictions in a learned representation space. DINO-WM (Zhou et al., 2025) predicts future pretrained visual features; V-JEPA 2 predicts masked video representations, and its action-conditioned extension V-JEPA 2-AC supports robotic planning (Assran et al., 2025). JEPA-WMs study how latent dynamics and planning design affect control performance (Terver et al., 2026), while LeWorldModel probes whether predictive latent states encode physical quantities (Maes et al., 2026). For observation-driven weather forecasting, the corresponding question is what atmospheric information a latent state retains through rollout and exposes to different observation readouts. WeatherJEPA addresses this question by using a dense physical view to organize the reference representation during training, while constructing its deployed state from observations alone.

3 Problem Formulation

Let $O_t = \{O_t^{(m)}\}_{m\in\mathcal{M}}$ denote heterogeneous observations, P_t a spatially dense atmospheric estimate, and C_t auxiliary information such as topography, climatology, and time features. Paired (O_t, P_t) samples are available on a limited set of times $\mathcal{T}_{\text{pair}}$, whereas the dense fields extend over $\mathcal{T}_{\text{dense}}$, with $\mathcal{T}_{\text{pair}} \subset \mathcal{T}_{\text{dense}}$ and $|\mathcal{T}_{\text{pair}}| \ll |\mathcal{T}_{\text{dense}}|$. A deployed forecast starts from (O_t, C_t) without P_t.

For an observation family m and forecast lead $k\Delta$, $k = 1, \ldots, k_{\max}$, where $k_{\max}$ is the maximum evaluated forecast step, let $Y_{t+k\Delta}^{(m)}$ denote the evaluated target variables from observation family m at time $t + k\Delta$, and $\widehat{Y}_{t+k\Delta}^{(m)}(O_t, C_t)$ its forecast made from the information available at deployment. The objective is to minimize the expected forecast error for each evaluated family and lead:

$$\mathcal{R}_{m,k}^{\text{deploy}} = \mathbb{E}_t\left[\mathcal{E}_m\left(\widehat{Y}_{t+k\Delta}^{(m)}(O_t, C_t), Y_{t+k\Delta}^{(m)}\right)\right], \tag{1}$$

where $\mathcal{E}_m$ measures error on the target observation interface.

4 METHOD

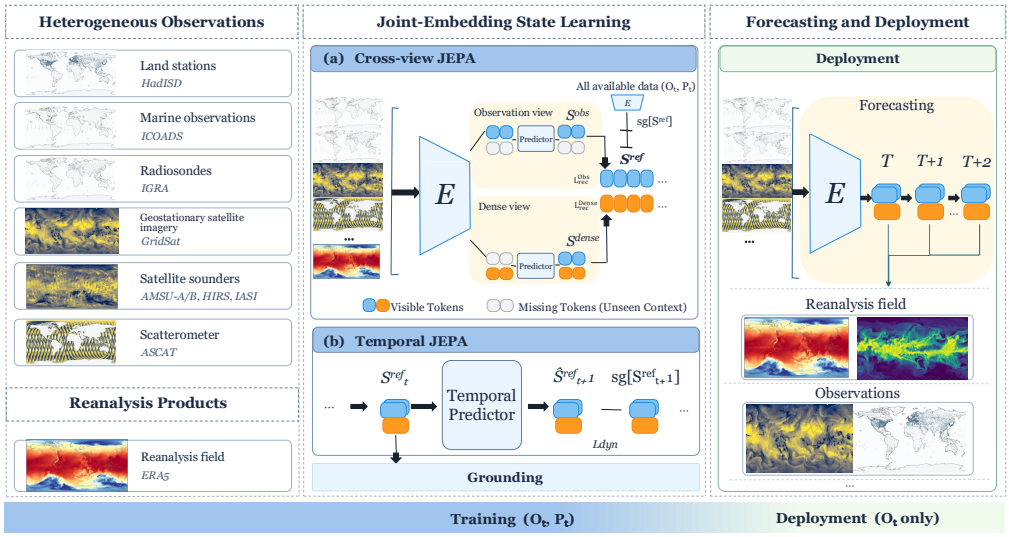

Figure 2: **WeatherJEPA framework.** During representation learning, a paired observation–dense route defines reference coordinates, while observation-only and dense-only routes learn to recover them. Grounding encourages decodable atmospheric information, and auxiliary temporal prediction favors simple local latent evolution. A separate forecasting processor is trained on dense-only state trajectories from a longer reanalysis archive. At deployment, observations initialize the state; target- and lead-specific forecast readouts decode its autoregressive rollout.

4.1 MULTI-VIEW STATE CONSTRUCTION WITH A DENSE REFERENCE

A single encoder E_φ includes the input branches and their shared fusion mechanism. It is evaluated with paired inputs (O_t, P_t, C_t), observation-only inputs (O_t, C_t), or dense-only inputs (P_t, C_t). With lightweight predictors Π_{obs} and Π_{dense} for the restricted views, the corresponding states are

$$S_t^{\text{ref}} = \mathcal{N}(E_\varphi(O_t, P_t, C_t)),$$
$$S_t^{\text{obs}} = \mathcal{N}(\Pi_{\text{obs}} E_\varphi(O_t, C_t)),$$
$$S_t^{\text{dense}} = \mathcal{N}(\Pi_{\text{dense}} E_\varphi(P_t, C_t)).$$

S_t^{ref} is the paired training-time reference representation, S_t^{obs} is the deployment representation, and S_t^{dense} is an auxiliary dense-only estimate. The three routes share the same encoder parameters; they differ in the available input views. The restricted-view predictors map observation-only and dense-only encodings into the coordinates of the paired reference state. Here "reference" denotes a learned coordinate system shared across sensing routes.

The normalization

$$\mathcal{N}(z) = \frac{z}{\sqrt{d^{-1}\|z\|_2^2 + \epsilon}}$$

fixes the scale ambiguity of a distance-based joint embedding at each spatial token, where d is the latent width and $\epsilon = 10^{-6}$ is a small constant for numerical stability. Normalized token RMS is approximately one by construction.

4.2 Cross-View Recovery under Asymmetric Availability

The paired reference representation is available only during representation learning. The observation-only and dense-only routes are trained to recover its detached coordinates,

$$\mathcal{L}_{\text{rec}} = \mathcal{D}_{\text{rec}}\big(S_t^{\text{obs}}, \text{sg}[S_t^{\text{ref}}]\big) + \mathcal{D}_{\text{rec}}\big(S_t^{\text{dense}}, \text{sg}[S_t^{\text{ref}}]\big)\,, \tag{2}$$

where $\mathcal{D}_{\text{rec}}$ is mean-squared error over normalized latent coordinates (Eq. (6)); the two routes have equal weight. The stop-gradient operation stabilizes cross-view recovery without freezing the reference representation, which continues to be optimized by temporal shaping and grounding. The observation route therefore learns an amortized map into a reference space formed from complementary evidence.

Because RMS normalization approximately fixes each token norm to $\sqrt{d}$, token-wise recovery MSE and cosine alignment carry nearly the same information:

$$d^{-1}\|s - r\|_2^2 \approx 2\big(1 - \cos(s, r)\big).$$

Cosine alignment provides an interpretable representation diagnostic.

4.3 Temporal Forecastability

Cross-view recovery makes the state observation-deployable, but does not determine how easily its coordinates evolve. We use a low-capacity auxiliary predictor K to predict the reference-state increment over a temporal separation Δ. For N spatial tokens of latent width d, the shaping loss is the mean-squared error

$$\mathcal{L}_{\text{dyn}} = \frac{1}{Nd}\left\|K(S_t^{\text{ref}}) - \text{sg}[S_{t+\Delta}^{\text{ref}} - S_t^{\text{ref}}]\right\|_F^2. \tag{3}$$

We use $\Delta = 24$ hours in the experiments. The auxiliary predictor shapes the representation during training; a separate autoregressive processor Φ_ψ produces the evaluated Day 1–10 forecasts.

4.4 Grounding of the Reference Representation

Cross-view recovery and temporal shaping alone do not ensure that the shared state retains meteorologically meaningful information. Let D_m be a readout for observation family m and D_{dense} a readout for the dense reference field. We define

$$\begin{aligned}
\mathcal{L}_{\text{obs}} &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \ell_m\Big(D_m(S_t^{\text{ref}}), O_t^{(m)}\Big)\,,\\
\mathcal{L}_{\text{dense}} &= \ell_{\text{dense}}\big(D_{\text{dense}}(S_t^{\text{ref}}), P_t\big)\,,\\
\mathcal{L}_{\text{gnd}} &= \mathcal{L}_{\text{obs}} + \mathcal{L}_{\text{dense}}.
\end{aligned} \tag{4}$$

Here ℓ_m and ℓ_{dense} denote reconstruction losses for observation family m and the dense reference field, respectively. Each ℓ_m is evaluated only where the corresponding target is available and respects the native geometry of that observation family. The observation term grounds the representation in direct measurements, while the dense-field term encourages spatially organized physical information to remain decodable. Grounding constrains *what information is recoverable from the state*.

Overall, representation learning combines grounding, cross-view recovery, and temporal shaping; the objective and continuation schedule are given in Appendix A.4.

4.5 Forecasting and Observation Readouts

After representation learning, we freeze the state-construction module and encode a longer reanalysis archive through the dense-only route to obtain S^{dense} trajectories. A separate processor Φ_ψ learns 24-hour transitions in this state space. Target- and lead-specific observation readouts are trained against the corresponding future measurements. At deployment, (O_t, C_t) produces S_t^{obs}, and the processor and readouts generate observation forecasts.

5 EXPERIMENTS

We compare four observation-driven state-construction paradigms using frozen-state readouts and Day 1–10 observation forecasts. On ICOADS marine observations, we also compare WEATHERJEPA forecasts with IFS-HRES as an operational reference. Controlled ablations examine the contributions of temporal shaping and observation and dense-field grounding. Evaluation details are given in Appendices A.2 and A.3.

5.1 WHAT STATE DOES EACH OBSERVATION-DRIVEN PARADIGM LEARN?

Before forecasting, we freeze each pre-forecast state and apply the same lightweight readout protocol. The probes decode 24 normalized ERA5 variables and all nine observation families; they diagnose what is accessible from the representation rather than serving as physical-unit forecast scores. For WEATHERJEPA, every probe starts from the observation-only deployment state S_t^{obs}.

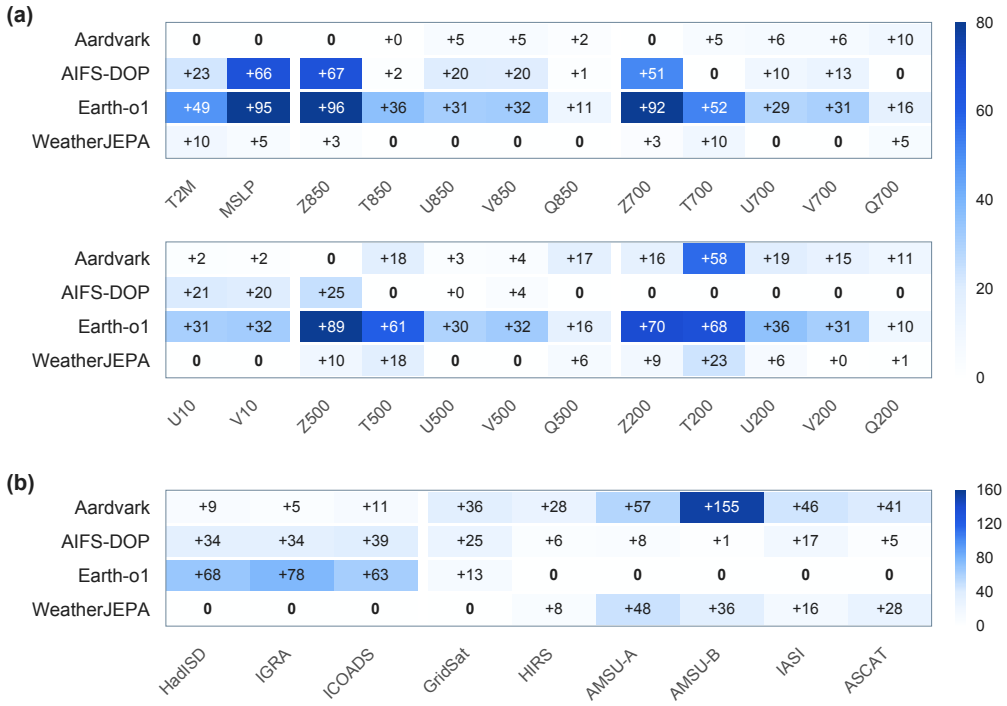

Figure 3: **State fingerprints under matched frozen readouts.** (a) 24 physical-field targets, arranged in two rows with surface variables on the left and complete pressure-level groups. (b) Nine observation families, with *in situ* families on the left and remote sensing on the right. Cells show integer-rounded percentage error above the best learned state for each target (bold 0 marks the best; darker blue is worse), avoiding comparisons of raw errors with different units.

Related measurements and fields can favor different states. Under matched frozen readouts, WEATHERJEPA leads HadISD, IGRA, ICOADS, and GridSat (Fig. 3) and the physical-field and *in situ* macro averages (Table 1(a)). Yet Earth-o1 decodes AMSU-A/B measurements better, while WEATHERJEPA decodes ERA5 temperature and specific humidity better at every evaluated level: sensor accessibility need not track related gridded fields. Likewise, WEATHERJEPA leads only seven of the 16 ERA5 height, temperature, and wind probes corresponding to IGRA forecast targets; Aardvark or AIFS-DOP leads the other nine (Fig. 3). These mismatches motivate testing whether initial-state rankings anticipate future observation errors in Section 5.2.

5.2 Which Information Remains Predictive under Rollout?

We decode future observations from Day 1–10 autoregressive state rollouts. Figure 4 shows six representative targets spanning three observation interfaces: localized surface temperature from HadISD, thermal, dynamical, and mass structure from IGRA radiosondes, and water-vapor-sensitive brightness temperature from GridSat. Table 1(b) summarizes the Day 1 and Day 10 relative gains. Full IGRA trajectories and additional GridSat results appear in Appendix B.

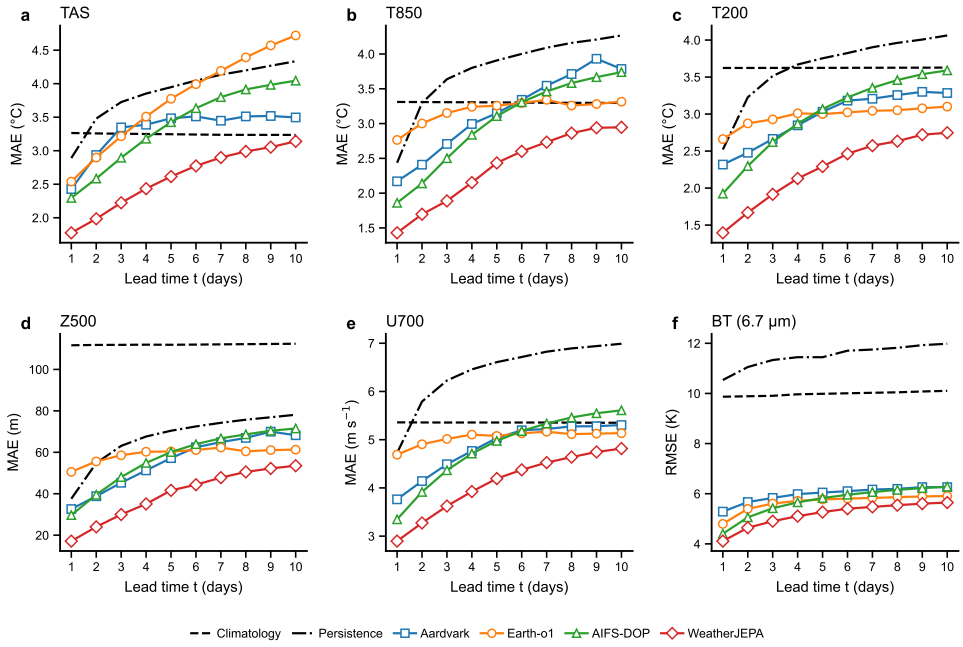

Figure 4: **Ten-day forecasts across complementary observation interfaces.** Representative targets include HadISD surface air temperature (TAS); IGRA lower- and upper-level temperature (T850, T200), geopotential height (Z500), and zonal wind (U700); and the GridSat 6.7-μm water-vapor-sensitive brightness-temperature channel. WeatherJEPA is compared with three learned state-construction paradigms and climatology/persistence references. Lower is better.

(a) Frozen-state error

Paradigm	Physical	*In situ*	Remote
Aardvark	0.2945	0.3826	0.8327
AIFS-DOP	0.2975	0.4761	0.5833
Earth-o1	0.3592	0.5969	**0.5370**
WeatherJEPA	**0.2780**	**0.3524**	0.6502

(b) Forecast gain

	TAS	T850	T200	Z500	U700	BT 6.7 μm
Day 1	22.7%	23.2%	27.5%	42.3%	13.5%	6.7%
Day 10	10.2%	11.1%	11.4%	12.7%	6.2%	4.4%

Table 1: **(a) Frozen-state accessibility; (b) forecast error reduction.** (a) Macro errors (lower is better): Physical averages 24 normalized LW-RMSEs, *in situ* averages HadISD/IGRA/ICOADS, and Remote averages six remote-sensing families. (b) WeatherJEPA's relative error reduction against the strongest competing learned method at each lead (higher is better); HadISD TAS and the IGRA targets use MAE, while GridSat BT uses RMSE.

Initial accessibility does not determine forecast rankings. At the observation-family level, WeatherJEPA leads the frozen IGRA and GridSat readouts (Fig. 3) and the forecasts on all 24 IGRA variable–level targets and both GridSat bands at every lead (Appendix Figs. 7 and 8). The gridded-field ranking differs: among 16 ERA5 height, temperature, and wind probes with corresponding IGRA targets, WeatherJEPA leads only seven static probes but all 16 forecasts at every lead.

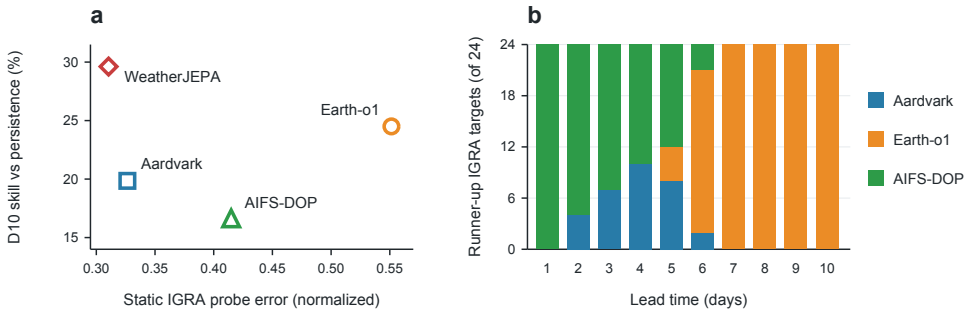

Figure 5: **IGRA accessibility and retention.** (a) Static probe error versus Day-10 skill over persistence. (b) Runner-up counts across 24 variable–level targets by lead.

Z500 and T200 illustrate this cross-interface reversal: Aardvark and AIFS-DOP, respectively, decode the frozen ERA5 fields better, whereas WeatherJEPA predicts the corresponding IGRA observations better. Because the probes and forecasts differ in target support, metric, and lead time, this comparison does not isolate the temporal model's effect. It does show that gridded-field decodability alone cannot determine observation-forecast ranking.

Static advantages can reverse with forecast horizon. AIFS-DOP has a lower frozen IGRA-family probe error than Earth-o1 (Fig. 3) and is runner-up on all 24 IGRA variable–level targets at Day 1. Earth-o1 becomes runner-up on 19 targets by Day 6 and all 24 from Day 7 onward (Fig. 5(b)). Its relative ranking improves even though its Day-10 MAE remains higher than its Day-1 MAE on every target (Appendix Fig. 7). WeatherJEPA remains first on all 24 targets, with Day-10 persistence-relative skill of 29.9/29.8/29.7/29.0% across 850/700/500/200 hPa (Appendix B.1).

The same horizon-dependent reordering extends beyond radiosondes. Earth-o1 has a better frozen GridSat-family readout than AIFS-DOP (Fig. 3), yet AIFS-DOP is the stronger forecast baseline on both evaluated bands through Day 4; Earth-o1 takes over from Day 5 (Appendix Fig. 8). WeatherJEPA remains best at every lead under RMSE. Thus even within one observation family, frozen-state accessibility does not establish the short- or long-range forecast ordering.

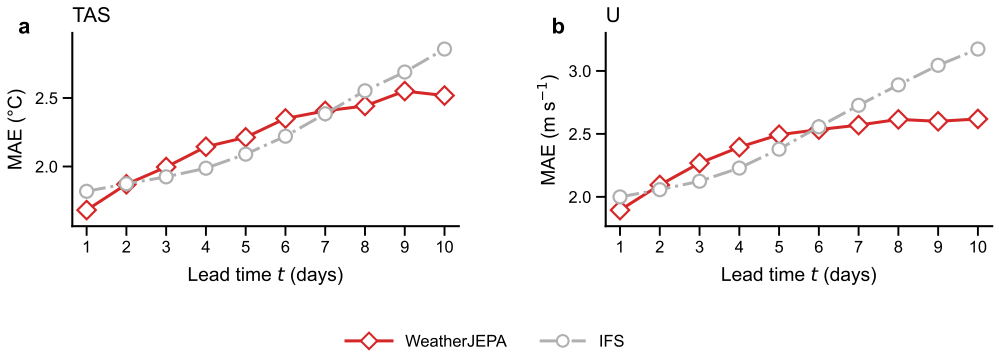

Figure 6: **ICOADS marine forecasts against IFS-HRES.** Day 1–10 MAE for marine air temperature (TAS, °C) and zonal wind (U, m s^{-1}). The IFS-HRES curves use bilinear extraction at the observation locations; lower is better.

The ICOADS readout in Section 5.1 favors WeatherJEPA on marine observations. Figure 6 extends the evaluation to Day 1–10 forecasts, using regridded IFS-HRES as an operational reference. Averaged equally over the ten leadwise MAEs, WeatherJEPA is lower by 1.0% and 4.4% on TAS and U, respectively. The leadwise rankings vary: IFS-HRES has lower MAE at several early leads, whereas WeatherJEPA leads on both variables from Day 8 onward, reaching 11.9% and 17.5% lower MAE at Day 10. The evaluation protocol appears in Appendix A.2.

5.3 What Shapes the Predictive State?

Sections 5.1 and 5.2 show that initial-state accessibility and forecast performance capture different properties. We now ablate temporal shaping and the two grounding terms to examine their respective contributions (Table 2).

(a) Temporal shaping

Metric	Full	w/o temporal
Skill (%)↑	**58.170**	38.610
Update cos.↑	**0.764**	0.623

(b) Grounding

Probe	w/o obs	w/o dense	w/o both
ERA5	-0.808%	$+67.959\%$	$+1149.057\%$
Observations	$+7.652\%$	-0.345%	$+248.327\%$

Table 2: **(a) Temporal objective; (b) grounding objectives.** (a) Fresh 24-hour probe metrics; skill is relative to latent persistence. (b) Probe-error change versus Full, rounded to three decimal places; positive is worse.

Temporal shaping improves independently probed predictability. With each representation frozen, a newly trained 24-hour latent predictor achieves 58.170% persistence-relative skill for the full model, versus 38.610% without the temporal objective; update cosine falls from 0.764 to 0.623. The effect therefore persists when the original jointly trained predictor is replaced by a fresh probe.

The grounding terms preserve complementary information. Removing observation grounding raises observation-probe error by 7.652% while leaving ERA5-probe error nearly unchanged. Removing dense-field grounding raises ERA5-probe error by 67.959% while leaving observation-probe error nearly unchanged; removing both sharply worsens both. The selective changes link each grounding term to the information accessible from the frozen state, while the temporal ablation tests how readily a new predictor can learn its 24-hour evolution.

6 Discussion and Conclusion

WeatherJEPA learns an observation-derived forecast state with dense physical guidance from reanalysis. Across matched observation-driven comparisons, it combines the lowest aggregate physical-field and *in situ* probe errors with the lowest Day 1–10 forecast errors on the evaluated HadISD TAS, IGRA, and GridSat targets. Its Day 1–10 mean marine TAS and U errors are also lower than the IFS-HRES reference on ICOADS.

These results suggest that dense physical guidance and an explicit physical-variable state need not be coupled. ERA5 agreement and observation fidelity favor different representations; the grounding terms likewise selectively preserve their corresponding readouts. Shared coordinates do not make these criteria interchangeable. Joint embedding connects the two views in learned coordinates, while separate reconstruction objectives constrain the content accessible through each interface. This lets reanalysis supply atmospheric structure while preserving scope for features supporting observation forecasting.

The temporal results extend this distinction from content to use. Positive fresh-probe skill without temporal shaping shows that present-state learning already yields predictive structure. Its improvement with temporal shaping persists for a new predictor, indicating greater predictability accessible from the frozen reference representation. Together with the horizon-dependent ranking changes, this suggests that a forecast state should be judged by both what it makes readable and what remains useful after evolution. Multi-day fresh probes could test how far the independently measured predictability extends through repeated state evolution.

AI Use Statement

Generative AI tools were used for literature-search assistance, language editing, document restructuring, and LaTeX drafting. The authors remain responsible for verifying technical claims, citations, equations, implementation descriptions, and experimental results against primary sources, code, and executed experiments.

REFERENCES

Mihai Alexe, Eulalie Boucher, Peter Lean, Ewan Pinnington, Patrick Laloyaux, Anthony McNally, Simon Lang, Matthew Chantry, Chris Burrows, Marcin Chrust, et al. Graphdop: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations. *arXiv preprint arXiv:2412.15687*, 2024.

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

R. J. H. Dunn, K. M. Willett, D. E. Parker, and L. Mitchell. Expanding hadisd: quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5 (2):473–491, 2016. doi: 10.5194/gi-5-473-2016. URL https://gi.copernicus.org/articles/5/473/2016/.

Robert JH Dunn, Kate M Willett, Peter W Thorne, Emma V Woolley, Imke Durre, Aiguo Dai, David E Parker, and RS Vose. Hadisd: A quality-controlled global synoptic report database for selected variables at long-term stations from 1973–2011. *Climate of the Past*, 8(5):1649–1679, 2012.

Imke Durre, Russell S. Vose, and David B. Wuertz. Overview of the integrated global radiosonde archive. *Journal of Climate*, 19(1):53 – 68, 2006. doi: 10.1175/JCLI3594.1. URL https://journals.ametsoc.org/view/journals/clim/19/1/jcli3594.1.xml.

EUMETSAT. Hirs level 1c fundamental data record release 1 - multimission - global, 2022a. URL https://user.eumetsat.int/catalogue/EO:EUM:DAT:0647.

EUMETSAT. Iasi principal components scores fundamental data record release 1 - metop-a and -b, 2022b. URL https://user.eumetsat.int/catalogue/EO:EUM:DAT:0758.

Eric Freeman, Scott D. Woodruff, Steven J. Worley, Sandra J. Lubker, Elizabeth C. Kent, William E. Angel, David I. Berry, Philip Brohan, Ryan Eastman, Lydia Gates, Wolfgang Gloeden, Zaihua Ji, Jay Lawrimore, Nick A. Rayner, Gudrun Rosenhagen, and Shawn R. Smith. Icoads release 3.0: a major update to the historical marine climate record. *International Journal of Climatology*, 37(5):2211–2232, 2017. doi: https://doi.org/10.1002/joc.4775. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4775.

RV Gelsthorpe, E Schied, and JJW Wilson. Ascat-metop's advanced scatterometer. *ESA bulletin*, 102:19–27, 2000.

Azadeh Gholoubi, Ronald McLaren, Mu-Chieh Ko, Xin Jin, Nicholas Esposito, Andrew Collard, Cory Martin, Russ Treadon, Daniel Holdaway, and Daryl Kleist. Ocelot: Direct atmospheric forecasting from heterogeneous earth observations using a graph-transformer hybrid model, 2026. URL https://arxiv.org/abs/2607.14196.

Junchao Gong, Jingyi Xu, Ben Fei, Fenghua Ling, Wenlong Zhang, Kun Chen, Wanghan Xu, Weidong Yang, Xiaokang Yang, and LEI BAI. Dawp: A framework for global observation forecasting via data assimilation and weather prediction in satellite observation space. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 148646–148682. Curran Associates, Inc., 2025. doi: 10.52202/085713-4972. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Junchao Gong, Kaiyi Xu, Wangxu Wei, Siwei Tu, Jingyi Xu, Zili Liu, Hang Fan, Zhiwang Zhou, Tao Han, Yi Xiao, et al. Earth-o1: A grid-free observation-native atmospheric world model. *arXiv preprint arXiv:2605.06337*, 2026.

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

Weixin Jin, Jonathan Weyn, Pengcheng Zhao, Siqi Xiang, Jiang Bian, Zuliang Fang, Haiyu Dong, Hongyu Sun, Kit Thambiratnam, and Qi Zhang. Weatherreal: a benchmark based on in-situ observations for evaluating weather models. *arXiv preprint arXiv:2409.09371*, 2024.

K. R. Knapp and S. L. Wilkins. Gridded satellite (gridsat) goes and conus data. *Earth System Science Data*, 10(3):1417–1425, 2018. doi: 10.5194/essd-10-1417-2018. URL https://essd.copernicus.org/articles/10/1417/2018/.

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.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. Aifs–ecmwf's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning. *arXiv preprint arXiv:2308.13280*, 2023.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Ewan Pinnington, Peter Lean, Mihai Alexe, Eulalie Boucher, Simon Lang, Patrick Laloyaux, Gert Mertes, Tomas Kral, Patricia de Rosnay, Matthew Chantry, et al. Aifs-dop: End-to-end medium-range weather prediction from observations alone with machine learning. *arXiv preprint arXiv:2606.19093*, 2026.

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

Stephan Rasp, Boris Babenko, Dominic Masters, Andrew El-Kadi, Samier Merchant, Guy Shalev, Ilan Price, Fred Zyda, Remi Lam, Sasha Shysheya, Matthew Willson, Stratis Markou, Shreya Agrawal, Suhani Vora, Mohammed Alewi Hassen, Sunny Mak, Tom R. Andersson, Megan Bela, Akib Uddin, Nofar Peled Levi, Ben Gaiarin, Ferran Alet, Aaron Bell, Peter Battaglia, and Alvaro Sanchez-Gonzalez. Weathernext 3: Increasing resolution and performance of global weather models with raw observations, 2026. URL https://arxiv.org/abs/2609.03582.

Philip W Rosenkranz. Retrieval of temperature and moisture profiles from amsu-a and amsu-b measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 39(11):2429–2435, 2001.

Xiuyu Sun, Xiaohui Zhong, Xiaoze Xu, Yuanqing Huang, Hao Li, J David Neelin, Deliang Chen, Jie Feng, Wei Han, Libo Wu, et al. A data-to-forecast machine learning system for global weather. *Nature Communications*, 16(1):6658, 2025.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=cHZn5Gdh8e. Reproducibility Certification.

Pengcheng Zhao, Jiang Bian, Zekun Ni, Weixin Jin, Jonathan Weyn, Zuliang Fang, Siqi Xiang, Haiyu Dong, Bin Zhang, Hongyu Sun, Kit Thambiratnam, and Qi Zhang. Omg-hd: A high-resolution ai weather model for end-to-end forecasts from observations, 2024. URL https://arxiv.org/abs/2412.18239.

Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

A Implementation and Evaluation Details

A.1 Data and Splits

We use the preprocessed 2007–2019 observation archive released with Aardvark Weather (Allen et al., 2025). It includes nine observation sources: HadISD (Dunn et al., 2012; 2016), IGRA (Durre et al., 2006), ICOADS (Freeman et al., 2017), GridSat (Knapp & Wilkins, 2018), HIRS (EUMETSAT, 2022a), IASI-PC (EUMETSAT, 2022b), AMSU-A, AMSU-B/MHS, and ASCAT. AMSU-A/B provide temperature- and moisture-sensitive microwave measurements (Rosenkranz, 2001), while ASCAT provides ocean-surface radar backscatter measurements (Gelsthorpe et al., 2000). The public archive provides daily observation snapshots, and we use one snapshot per day at 00 UTC. This sampling is temporally coarser than the six-hourly indexing used in Aardvark's original training pipeline. We use the $1.5°$, 6-hourly ERA5 product provided by WeatherBench 2 (Rasp et al., 2024).

Following Aardvark, we use observations from 2007–2017 for training, 2019 for validation, and 2018 for testing. The longer ERA5 archive used to train the forecasting processor also ends in 2017. Normalization and climatology are computed from training data only; missing targets are excluded using validity masks.

A.2 Baselines

We implement Aardvark (Allen et al., 2025), AIFS-DOP (Pinnington et al., 2026), and Earth-o1 (Gong et al., 2026) following the key architectural and training designs described in their papers, and evaluate them alongside WeatherJEPA using the same observation data and targets. These comparisons assess the methods under our experimental settings; differences in data and model scale may affect performance relative to the published results.

Persistence and climatology. Persistence and climatology provide simple references for forecast skill. Climatology uses training-split statistics. Both are evaluated on the same lead-dependent valid support as the learned forecasts.

Operational IFS-HRES reference. For the ICOADS (Freeman et al., 2017) comparison, we use the $1.5°$ IFS-HRES forecasts provided by WeatherBench 2 (Rasp et al., 2024). Forecasts are bilinearly interpolated to the observation locations and evaluated against the same observations at matched forecast times. We report MAE for each lead and its equally weighted mean over Day 1–10; relative reductions use the corresponding IFS-HRES error as the reference.

A.3 Evaluation Metrics

Let $e_i = \hat{y}_i - y_i$ denote the prediction error and $m_i \in \{0, 1\}$ the validity mask. We compute

$$\mathrm{MAE} = \frac{\sum_i m_i |e_i|}{\sum_i m_i}, \qquad \mathrm{RMSE} = \sqrt{\frac{\sum_i m_i e_i^2}{\sum_i m_i}},$$
$$\mathrm{LW\text{-}RMSE} = \sqrt{\frac{\sum_i m_i \cos\phi_i \, e_i^2}{\sum_i m_i \cos\phi_i}}, \tag{5}$$

where ϕ_i is latitude. ERA5 physical-field probes use LW-RMSE. HadISD, IGRA, and ICOADS forecasts use MAE, while GridSat forecasts use RMSE; forecast errors are reported in the corresponding variables' physical units. Comparisons use matched target support and validity masks.

Table 1(a) reports macro averages: Physical averages the 24 normalized ERA5-field LW-RMSEs; *In situ* averages the HadISD, IGRA, and ICOADS family scores; Remote averages the six remote-sensing family scores. Each target or family receives equal weight within its group.

A.4 WeatherJEPA Implementation and Training

Our implementation reuses the preprocessing and neural-network components of Aardvark Weather (Allen et al., 2025). Training first learns the representation, then trains the forecasting processor and observation decoders.

Representation learning. Paired observations and ERA5 jointly optimize grounding, cross-view recovery, and 24-hour latent prediction through $\mathcal{L}(p) = \mathcal{L}_{\mathrm{gnd}} + \mathcal{L}_{\mathrm{rec}} + \lambda(p)\mathcal{L}_{\mathrm{dyn}}$, where $p \in [0, 1]$ denotes normalized training progress. The temporal weight is zero for 15 epochs, rises linearly over the next 20, and then remains at 0.6. In Eq. (4), ℓ_m uses masked MAE for point observations and masked latitude-weighted MSE for gridded observations; ℓ_{dense} uses masked latitude-weighted MSE. All reconstruction losses exclude invalid targets. For RMS-normalized states $A, B \in \mathbb{R}^{N\times d}$, the recovery distance in Eq. (2) is mean-squared error over the N spatial tokens and d latent coordinates:

$$\mathcal{D}_{\mathrm{rec}}(A, B) = \frac{1}{Nd}\sum_{j=1}^{N}\|A_j - B_j\|_2^2. \tag{6}$$

Forecasting: processor and decoder. After freezing the representation, we train a forecasting processor on standardized 24-hour increments from longer S^{dense} reanalysis trajectories, with sequential adaptation across forecast leads. We then freeze the representation and processor and train observation decoders for each target and forecast lead against the corresponding future observations. Decoder checkpoints are selected by validation error in physical units.

A.5 Representation Diagnostics

Fresh temporal probe. For each ablation branch, we freeze the representation and train a newly initialized 3×3 convolutional predictor K_{probe} on 24-hour latent transitions. Here S_t denotes the frozen reference state S_t^{ref}. Let $\Delta S_t = S_{t+24\mathrm{h}} - S_t$ be the true increment and $\mathcal{Z}$ the increment standardization transform fitted on training data. The standardized target and prediction are $U_t = \mathcal{Z}(\Delta S_t)$ and $\widehat{U}_t = K_{\mathrm{probe}}(S_t)$. Latent persistence predicts no change; its standardized increment is $U_{\mathrm{pers}} = \mathcal{Z}(0)$. Temporal skill is

$$\mathrm{Skill} = 1 - \frac{\mathbb{E}_t\left[\|\widehat{U}_t - U_t\|_F^2\right]}{\mathbb{E}_t[\|U_{\mathrm{pers}} - U_t\|_F^2]}. \tag{7}$$

Positive skill indicates lower standardized-increment MSE than persistence; zero indicates equal error, and negative skill indicates higher error. Skill is reported as a percentage. Update cosine measures alignment in the same standardized space:

$$\mathrm{UpdateCos} = \mathbb{E}_t\left[\frac{\langle\widehat{U}_t, U_t\rangle_F}{\|\widehat{U}_t\|_F\,\|U_t\|_F}\right]. \tag{8}$$

Here $\mathbb{E}_t$ averages over evaluated transitions. Higher update cosine indicates better directional alignment in standardized coordinates, rather than agreement in increment magnitude. Probe results are evaluated on the 2019 validation split.

A.6 Model Size and Computational Cost

Representation learning uses bf16 on two NVIDIA A800 GPUs. The representation module and forecasting processor contain 53.2M and 57.5M parameters, respectively. Decoder size varies by task; an IGRA decoder contains 1.91M parameters. On NVIDIA A800 GPUs, representation learning, processor training, and lead-wise adaptation cost 21.5, 10.8, and 0.2 GPU-hours, respectively. For an IGRA target, decoder training across all ten forecast leads costs approximately 1.17 GPU-hours.

B SUPPLEMENTARY FORECAST FIGURES

B.1 COMPLETE IGRA FORECASTS

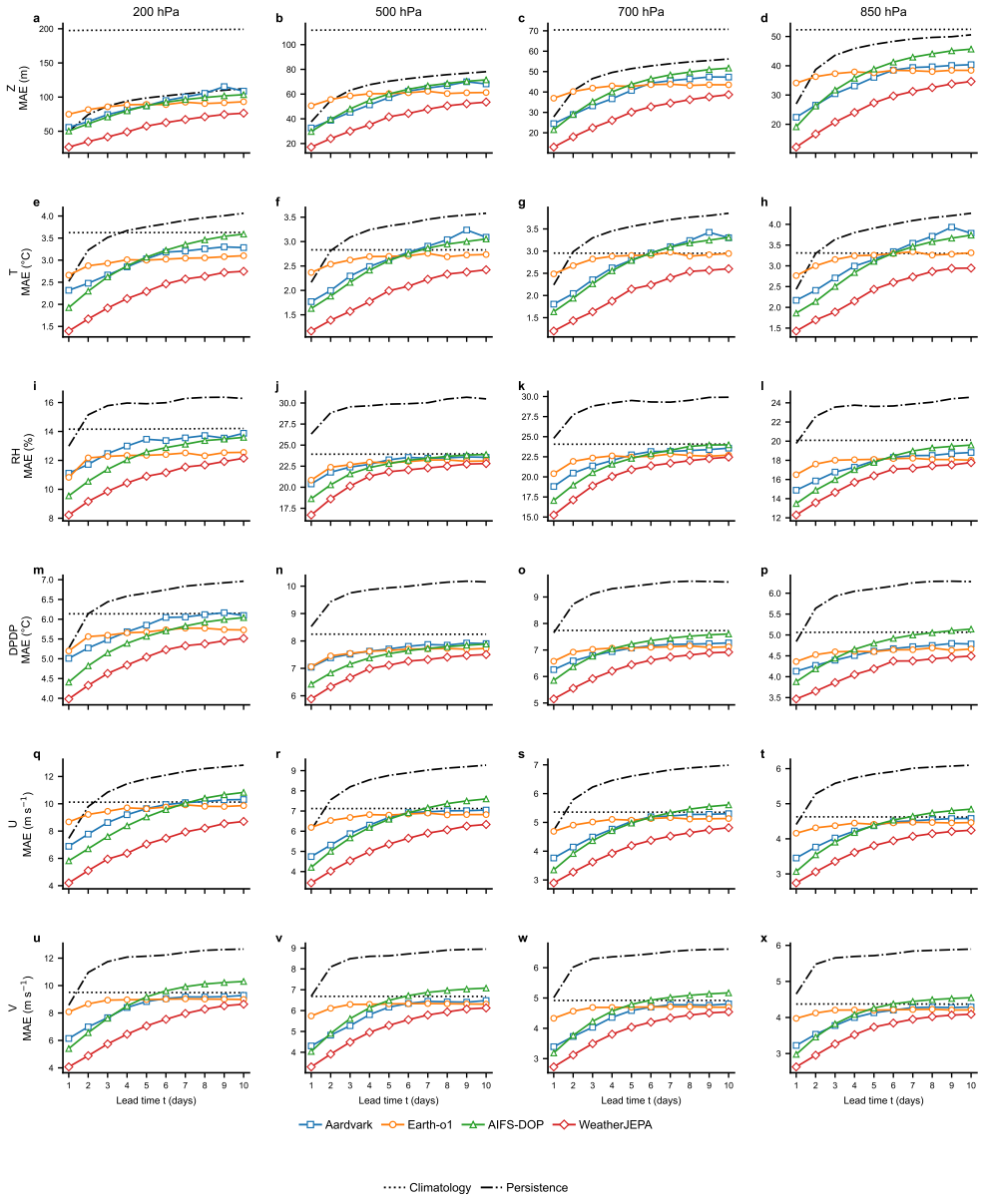

Figure 7: **IGRA forecasts across all 24 variable–level targets and ten lead days.** Each panel reports MAE in the physical unit shown on its row axis. Columns are pressure levels; curves compare WEATHERJEPA with Aardvark, AIFS-DOP, Earth-o1, climatology, and persistence. Lower is better.

At Day 10, persistence-relative skill at each pressure level is 100% minus the mean of the six variable-wise MAE ratios to persistence. WEATHERJEPA scores 29.9/29.8/29.7/29.0% at 850/700/500/200 hPa, respectively.

B.2 GridSat Forecasts on a Common Grid

GridSat is evaluated on common spatial support across methods. Figure 8 shows the Day 1–10 RMSE trajectories for both evaluated channels.

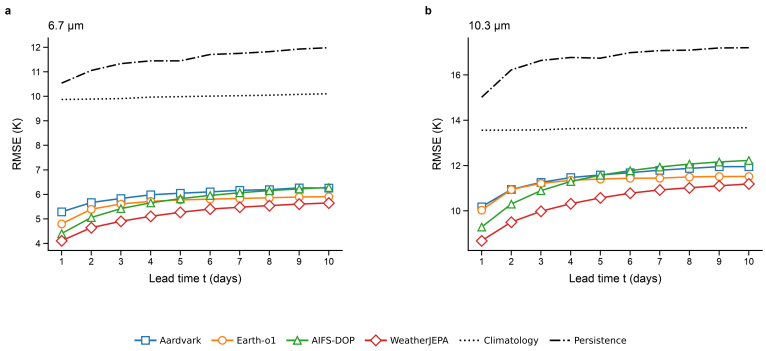

Figure 8: **GridSat brightness-temperature forecasts.** Day 1–10 RMSE on common spatial support for the 6.7-μm channel (left) and 10.3-μm channel (right). Curves compare WeatherJEPA with Aardvark, AIFS-DOP, Earth-o1, climatology, and persistence. Errors are in K; lower is better.