EARTH-JEPA: TOWARD EFFICIENT AND TRANSFERABLE EARTH-SYSTEM LATENT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning for Earth system forecasting has advanced rapidly, mostly through end-to-end supervision on reanalysis data for individual tasks. This training paradigm tends to be less efficient on spatiotemporally smooth fields, while making cross-task adaptation relatively expensive. To overcome these challenges, we introduce Earth-JEPA (E-JEPA), a reconstruction-free, multi-resolution, modality-wise Joint-Embedding Predictive Architecture for efficient and transferable Earth-system prediction. Modality-specific branches encode satellite observations, atmospheric reanalysis and geographic fields into a multi-resolution latent state. A latent-pyramid predictor with Scale-Specific Attention forecasts future representations, each supervised by its corresponding exponential moving average (EMA) target through a State-Transition Loss. On WeatherBench2, Earth-JEPA achieves 17.8% higher 24-hour U850 anomaly correlation coefficient (ACC) than a leading large-scale supervised weather forecaster using only about 12.8% of their estimated training compute. Experiments on SEVIR and our proposed EarthCast-Bench benchmark demonstrate transferability across time-intervals and sensors, robustness to missing reanalysis, and strong performance on dense tasks, e.g., precipitation nowcasting, and forecasting of surface and atmospheric variables. These results suggest a promising direction for efficient and transferable Earth-system latent representation learning. Code and data will be released upon publication.

1 INTRODUCTION

Deep learning models increasingly rival numerical weather prediction, produce skillful precipitation nowcasts, and support general-purpose weather and climate modeling (Price et al., 2025; Bodnar et al., 2025; Couairon et al., 2026; Pablos Sarabia et al., 2026). Much of this progress relies on end-to-end supervision: a large network is trained to reproduce task-specific physical fields from atmospheric reanalysis (Hersbach et al., 2020). This recipe leaves two challenges. Earth-system fields are smooth and redundant across space and time, so reproducing every grid point spends computation on largely predictable content. Masked pixel reconstruction inherits this burden (He et al., 2022; Tong et al., 2022). Further, coupling representation learning to a specific output makes transfer to new variables, sensors or forecast intervals costly.

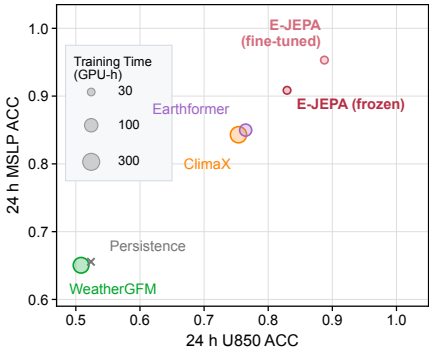

Figure 1: Training cost and 24-h forecast skill on WeatherBench2 (Rasp et al., 2024).

Reanalysis also limits the information available to the model. Although global and conveniently gridded, it inherits assimilation errors, has limited resolution and arrives with a delay. Geostationary satellites provide abundant direct observations at kilometer-scale resolution and minute-level cadence. Recent observation-driven forecasting and satellite-based nowcasting show the value of

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.

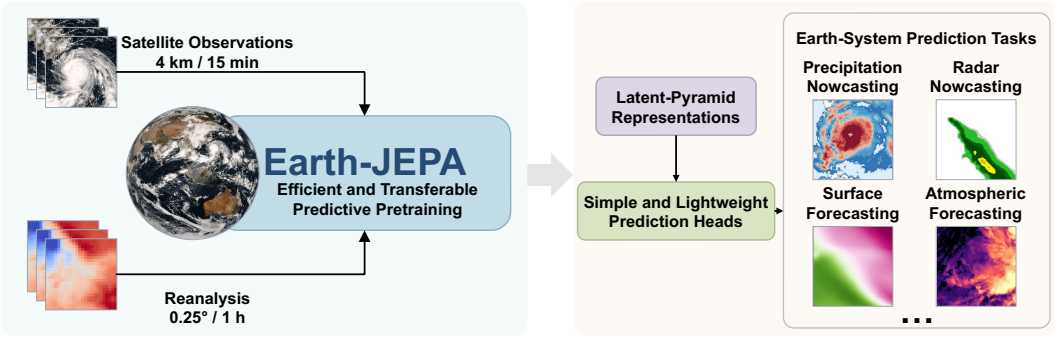

Figure 2: Earth-JEPA learns predictive latent pyramids from satellite observations, reanalysis and geographic conditioning, then adapts them through task-specific readouts.

these measurements (Allen et al., 2025; Agrawal et al., 2025). Learning directly from them can complement reanalysis-based modeling.

Earth-JEPA addresses these challenges through task-independent latent prediction (Fig. 2). A Modality-Wise Encoder organizes heterogeneous observations; a Latent-Pyramid Predictor forecasts their evolution across spatial scales; and a State-Transition Loss supervises each scale against EMA targets. Simple task readouts exploit the resulting representations, with reanalysis optional at inference. On WeatherBench2, Earth-JEPA achieves 17.8% higher 24-hour U850 anomaly correlation coefficient (ACC) than a leading large-scale supervised weather forecaster using only about 12.8% of their estimated training compute. Our contributions are:

- A reconstruction-free, multi-resolution JEPA framework connecting heterogeneous observation encoding, multi-scale future-state prediction and state-transition supervision.
- Empirical evidence that EMA targets prevent coarse-scale collapse: removing EMA reduces normalized effective rank at the coarsest scale by 92%, while state-transition supervision preserves more diverse representations than direct state matching.
- A transferable training recipe evaluated on missing-source inference, cross-sensor nowcasting and long-range forecasting at a different temporal interval, using task-specific readouts.
- EarthCastBench, a public-data benchmark aligning satellite brightness temperatures, reanalysis, geography and precipitation over 1.5 years, with 10,000 training sequences and dedicated validation/test splits.

2 Related Work

2.1 Deep Learning for Earth System Forecasting

One research line learns global atmospheric forecasting from reanalysis, progressing from Fourier operators to transformer and graph architectures (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023). Probabilistic and hybrid models extend forecast capabilities (Price et al., 2025; Kochkov et al., 2024; Couairon et al., 2026). A second line addresses dense precipitation nowcasting (Ravuri et al., 2021; Zhang et al., 2023; Gong et al., 2024; Agrawal et al., 2025; Pablos Sarabia et al., 2026), while weather foundation models seek reusable representations (Nguyen et al., 2023; Zhao et al., 2025; Bodnar et al., 2025; Zhu et al., 2026). Earth-observation models increasingly pretrain on satellite imagery (Cong et al., 2022; Fuller et al., 2023; Astruc et al., 2025; Jakubik et al., 2025), primarily for surface understanding.

These directions leave three related limitations: reconstructing redundant physical fields, task-dependent end-to-end training, and underuse of direct satellite observations for state prediction. Observation-driven forecasting demonstrates the value of the latter (Allen et al., 2025), but generally remains tied to a forecasting task. Earth-JEPA instead learns a satellite–atmosphere latent state that transfers through task-specific readouts.

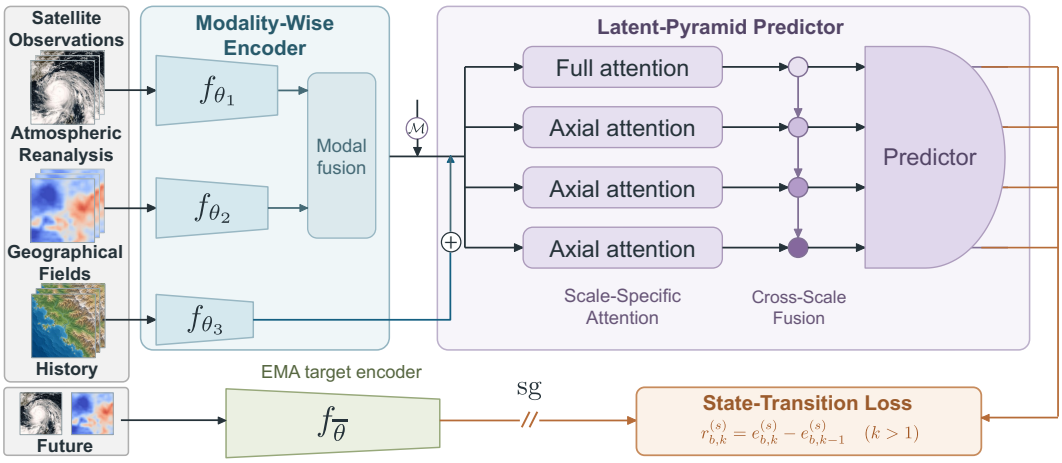

Figure 3: Earth-JEPA pretraining. Modality-specific encoders fuse satellite and reanalysis features; static geography conditions the masked latent-pyramid predictor. Scale-specific attention and cross-scale fusion produce future states supervised by State-Transition Loss against stop-gradient EMA targets.

2.2 JEPA

JEPAs predict unseen content in representation space rather than reconstructing pixels. Image and video variants (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025) motivate this approach for redundant Earth-system fields, while recent work studies explicit anti-collapse objectives (Balestriero & LeCun, 2025; Maes et al., 2026). AnySat learns across Earth-observation sensors and resolutions (Astruc et al., 2025); atmospheric JEPA explores future-state transfer on a single spatial grid (Anirudh et al., 2026); V-JEPA 2.1 supervises multiple encoder depths (Mur-Labadia et al., 2026). Earth-JEPA instead jointly predicts future states at distinct spatial resolutions, supervising every grid against its EMA target. Scale-Specific Attention limits dense-grid cost, while State-Transition Loss balances representation diversity and downstream accuracy; effective rank alone is not a measure of task performance (Garrido et al., 2023).

3 Earth-JEPA

3.1 Architecture

Earth-system fields are smooth over large spatiotemporal scales. To avoid repeatedly reconstructing their redundant content, Earth-JEPA first learns task-independent predictive representations and then adapts them through task-specific readouts. The pretraining stage (Fig. 3) combines a Modality-Wise Encoder, a masked Latent-Pyramid Predictor and an EMA target encoder. Every predicted scale receives state-transition supervision; downstream decoders consume the pyramid level by level. Notation and numerical configurations appear in Apps. A–C.

3.2 Modality-Wise Encoder

Heterogeneous observations require different encoders: a hierarchical Swin Transformer (Liu et al., 2021) processes fine-resolution satellite brightness temperatures and time encodings, a compact convolutional branch processes reanalysis and its availability indicator, and a smaller convolutional pyramid encodes static geography once per sample. Per-location two-token attention fuses the dynamic branches at each scale:

$$
\begin{aligned}
z_t^{(s)} &= \mathcal{F}_{\theta_F}^{(s)}\Big(f_{\theta_1}^{(s)}(x_t^{\mathrm{sat}}),\ f_{\theta_2}^{(s)}(x_t^{\mathrm{rea}})\Big) \in \mathbb{R}^{C_s \times H_s \times W_s}, \qquad c^{(s)} = f_{\theta_3}^{(s)}(x^{\mathrm{geo}}), \\
\bar{z}_k^{(s)} &= f_{\bar{\theta}}^{(s)}(x_{T+k}^{\mathrm{sat}},\ x_{T+k}^{\mathrm{rea}}), \qquad \bar{\theta} \leftarrow \mathrm{EMA}(\theta),\ \ \theta = \{\theta_1, \theta_2, \theta_F\},
\end{aligned}
\tag{1}
$$

Here s indexes scales from coarse to fine, t history frames and k future leads; $c^{(s)}$ conditions prediction. EMA updates only the dynamic branches and their fusion, excluding static conditioning. Reanalysis dropout with probability p_{drop} trains the fused state to remain predictive when analysis fields are unavailable, supporting the missing-source setting in Sec. 4.1.

3.3 Latent-Pyramid Predictor

Forecasting only the coarsest latent would discard detail needed by dense Earth-system tasks. Our predictor therefore jointly forecasts every spatial scale, with direct EMA supervision at each level. At scale s, a bidirectional masked transformer processes the dynamic history $Z^{(s)}$ and future queries $q_k^{(s)} = m^{(s)} + c^{(s)} + l_k^{(s)}$, where $m^{(s)}$ and $l_k^{(s)}$ are learned mask and lead tokens. Static features condition both history and future tokens.

Spatiotemporal Tube Masking. To discourage copying neighboring features, we partition the grid into $M \times M$ macro-cells and replace a fraction ρ by mask tokens. A spatial mask is shared across the last T_m history frames and all channels, forming tubes (Tong et al., 2022); masks are sampled independently by sequence and scale. This harder pretext task leaves probe accuracy and spectral diversity essentially unchanged in our ablation.

Scale-Specific Attention. Dense self-attention grows quadratically with token count. We reserve full attention for the compact coarsest grid and use sequential temporal, height and width attention (Ho et al., 2019) at finer scales:

$$\mathcal{A}_s(X) = \begin{cases} \mathrm{FullAttn}(X), & s = 1, \\ \mathrm{Attn}_W \circ\ \mathrm{Attn}_H \circ\ \mathrm{Attn}_T(X), & s = 2, \dots, S, \end{cases} \tag{2}$$

This changes per-block complexity from $\mathcal{O}((T+K)^2 H_s^2 W_s^2)$ to $\mathcal{O}(H_s W_s (T+K)^2 + (T+K)(W_s H_s^2 + H_s W_s^2))$. Even replacing coarse-grid full attention with axial attention has negligible effect in our ablation.

Cross-Scale Fusion. An FPN-style top-down pathway (Lin et al., 2017) exchanges information across resolutions. Projection $P^{(s)}$ and upsampling align coarse features; two-token attention mixes them with finer features, and a zero-initialized head $\psi^{(s)}$ makes fusion initially an identity. Prediction heads $\pi^{(s)}$ emit joint future increments:

$$\begin{aligned} \tilde{U}^{(s)} &= U^{(s)} + \psi^{(s)}\big(\mathrm{Attn}_2\big[U^{(s)};\ \mathrm{Up}\big(P^{(s)}\tilde{U}^{(s-1)}\big)\big]\big), \\ \hat{z}_k^{(s)} &= z_T^{(s)} + \textstyle\sum_{j=1}^{k} d_j^{(s)}, \qquad d_{1:K}^{(s)} = \pi^{(s)}\big(\tilde{U}^{(s)}\big), \end{aligned} \tag{3}$$

Here $\tilde{U}^{(1)} = U^{(1)}$. Cumulative increments anchor forecasts to the last observed state. Every scale retains its own supervision; removing fusion barely changes probe error, showing negligible accuracy cost, rather than proving that fusion is necessary or eliminates shortcuts.

3.4 State-Transition Loss

EMA and online encoders represent the same state differently because their parameters lag. Temporal differencing cancels the approximately time-invariant component of this discrepancy over a short horizon. We therefore align the first future state and supervise adjacent changes thereafter:

$$\begin{aligned} \mathcal{L}_{\mathrm{ST}} &= \frac{1}{BK} \sum_{b,k,s} \frac{v_{b,k}}{D_s} \left\| r_{b,k}^{(s)} \right\|_2^2, \qquad e_{b,k}^{(s)} = \hat{z}_{b,k}^{(s)} - \mathrm{sg}\Big(\bar{z}_{b,k}^{(s)}\Big), \\ r_{b,1}^{(s)} &= e_{b,1}^{(s)}, \qquad r_{b,k}^{(s)} = e_{b,k}^{(s)} - e_{b,k-1}^{(s)} \ \ (k > 1), \end{aligned} \tag{4}$$

The batch size is B, horizon K and scale dimension $D_s = C_s H_s W_s$; sg stops gradients. Validity weights retain observed target pairs: $v_{b,1} = q_{b,1}$ and $v_{b,k} = q_{b,k} q_{b,k-1}$. The nominal BK denominator remains fixed. Together with increment prediction, this objective learns state evolution; its empirical effect on diversity and accuracy is examined in Sec. 4.4.

Table 1: EarthCastBench precipitation nowcasting, averaged over 15–60 min. RMSE/MAE measure rain-rate error; CSI/POD/FAR measure event detection (App. D). Sup. rows use a separate 24-frame protocol. Earth-JEPA and Drop-ERA5 share the same readout. Throughout the tables, bold and underline indicate the best and second-best values within comparable protocol groups.

Method	RMSE ↓	MAE ↓	CSI ↑	POD ↑	FAR ↓
Persistence	0.8202	0.2797	0.7374	0.8557	0.1579
Earthformer (Gao et al., 2022b)	0.8169	0.2729	0.7515	0.8754	0.1585
ClimaX (Nguyen et al., 2023)	0.8082	0.2651	0.7571	0.8788	0.1546
Earthformer Sup. (Gao et al., 2022b)	0.7696	0.2842	0.6435	**0.9627**	0.3401
CasCast Sup. (Gong et al., 2024)	0.7588	0.2737	0.7063	<u>0.9562</u>	0.2701
WeatherGFM (Zhao et al., 2025)	0.7534	0.2515	<u>0.7693</u>	0.8901	0.1499
E-JEPA (Drop-ERA5)	<u>0.7179</u>	<u>0.2427</u>	0.7676	0.8807	<u>0.1433</u>
E-JEPA	**0.6898**	**0.2300**	**0.7788**	0.8931	**0.1411**

3.5 Task-Specific Decoder

A readout progressively upsamples coarse features and concatenates predicted states at matching resolutions, optionally incorporating observed target-variable history:

$$\hat{y}_{1:K} \;=\; g_\phi\Big(\big\{\hat{z}^{(s)}_{1:K}\big\}_{s=1}^{S},\; y_{1:T}\Big), \tag{5}$$

EarthCastBench uses multi-scale convolutional decoders, with event-sensitive supervision for rainfall. WeatherBench2 evaluates frozen and jointly fine-tuned backbones. SEVIR instead transfers only the frozen encoder and fusion modules into a shared radar prediction architecture. These simple adaptation interfaces expose representation quality; their capacity and objectives remain task-dependent, and stronger readouts can be substituted (App. C).

4 Experiments

We evaluate Earth-JEPA on EarthCastBench, WeatherBench2 (Rasp et al., 2024), and SEVIR (Veillette et al., 2020). EarthCastBench aligns 1.5 years of publicly available FY-4B infrared brightness temperatures, ERA5, land cover, viewing geometry, elevation and coordinates, with 10,000/1,000/1,000 training/validation/test sequences. The same pretrained Earth-JEPA backbone is used across these transfer tasks. For each comparison, learned baselines use the same input and prediction variables, data split, training budget, and evaluation protocol; only the minimal input/output mappings required by each model interface differ. The E-JEPA backbone uses 30 pretraining epochs and the separately reported E-JEPA-50M WeatherBench2 model uses 40. Data processing, optimization and metrics are specified in Apps. B–D.

4.1 Main Results on EarthCastBench

Precipitation nowcasting. We compare frozen-backbone readouts on common samples and report supervised task models separately. Earth-JEPA improves RMSE, MAE, CSI and FAR over the strongest non-Earth-JEPA baseline by 8.4%, 8.5%, 1.2% and 5.9%, respectively (Table 1). Supervised models attain higher POD but 91.4–141.0% higher FAR. In the out-of-period Ragasa case (Fig. 4), Earth-JEPA captures weakening in the highlighted regions; those illustrative selections do not establish unique or global advantage. With ERA5 inputs removed, RMSE rises by only 4.1%, supporting transfer from satellite features when reanalysis is unavailable.

Surface pressure. Under the shared historical-input and equal-setting protocol, Earth-JEPA leads the two reported pressure variables in Table 2. On high-rainfall scenes (13.9% of the test set), MSL RMSE/MAE improve over the strongest adapted baseline by 13.3%/15.3%, and surface-pressure errors by 13.0%/15.3%. The supplementary case shows localized pressure evolution during changing rainfall, with verifying-time rain provided only as context (Fig. 9). Drop-ERA5 denotes removal of ERA5 inputs.

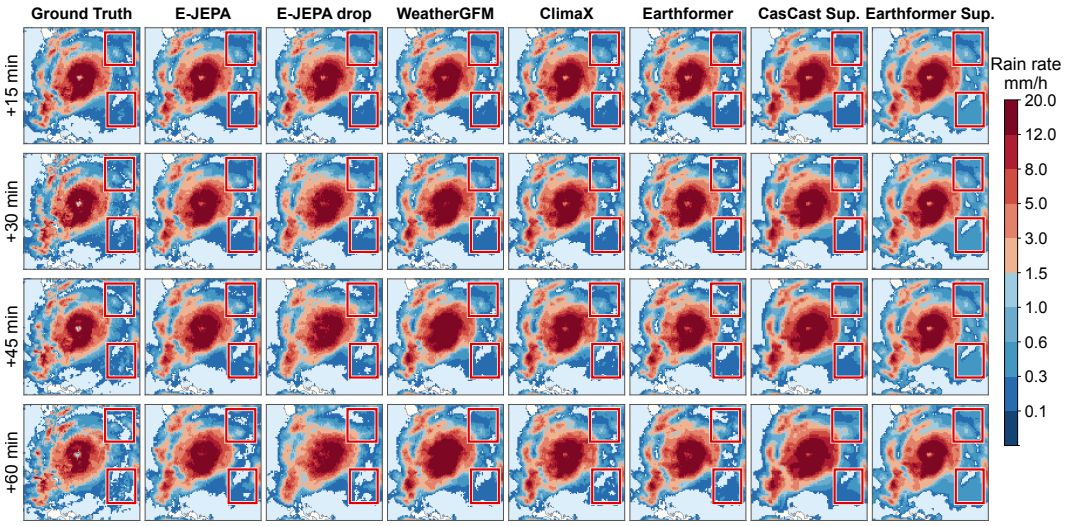

Figure 4: Typhoon Ragasa, initialized 2025-09-21 21:30 UTC, at 15–60 min leads. Models use native 0.04° tiles; rainfall is displayed on the 0.1° IMERG grid. Red regions were selected post hoc to illustrate local weakening, not independent or whole-scene superiority.

Table 2: Pressure errors (hPa), averaged over 15–60 min. High-rainfall samples have initialization-time IMERG rain rate averaged over all pixels, including dry pixels, above 1 mm h^{-1}.

Method	Mean sea-level pressure (msl), hPa				Surface pressure (sp), hPa			
	Full test samples		High-rainfall samples		Full test samples		High-rainfall samples	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
Persistence	0.7335	0.5370	0.7491	0.5568	0.6973	0.5099	0.7374	0.5464
ClimaX (Nguyen et al., 2023)	0.5785	0.4370	0.5840	0.4517	0.5571	0.4287	0.5841	0.4525
Earthformer (Gao et al., 2022b)	0.5672	0.4383	0.5798	0.4555	0.5578	0.4365	0.5800	0.4573
WeatherGFM (Zhao et al., 2025)	0.5687	0.4298	0.5761	0.4467	0.5449	0.4203	0.5740	0.4460
E-JEPA (Drop-ERA5)	0.5502	0.4157	0.5607	0.4301	0.5365	0.4121	0.5643	0.4338
E-JEPA	**0.5040**	**0.3796**	**0.4994**	**0.3784**	**0.5003**	**0.3800**	**0.4995**	**0.3778**

4.2 Time-Interval Transfer on WeatherBench2

We test transfer from 15-minute pretraining to eight 6-hourly histories and a 72-hour rollout. Each call predicts four future states, requiring three block calls. All adapted methods receive the same 121 historical fields and predict the same 121 target fields; each uses only the minimal input/output mapping needed by its architecture. The four standard variables U850, T850, Q700, and MSLP are the primary targets; the WeatherBench2 loss averages their mean with 0.1 times the mean loss over the remaining 117 fields, reflecting the finite readout capacity while retaining supervision over the complete state. All adapted baselines use equal downstream settings, including the same data, 40-epoch budget, and validation-based selection. ACC measures spatial agreement of forecast and observed anomalies from climatology.

ClimaX is the leading large-scale supervised weather forecaster used for the comparison in the abstract and Fig. 1. Against its 24-hour U850 ACC of 0.754, fine-tuned E-JEPA reaches 0.888, a 17.8% relative improvement (Table 3). Backbone pretraining uses 29.7 versus 232.3 A800 GPU-h (12.8%); Earth-JEPA fine-tuning adds 108.5 GPU-h, for 138.2 GPU-h in total. Figure 5 shows retained skill across leads. Fine-tuning outperforms freezing, and E-JEPA-50M and E-JEPA reach similar fine-tuned performance. Physical RMSE/MAE are reported in App. E. Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023), trained on multi-decade ERA5 at native 0.25° resolution, are reported separately as supervised references. Fine-tuned Earth-JEPA attains 89–96% of Pangu-Weather's 24-h ACC; the two supervised references require an estimated $147\times$ and $510\times$ more arithmetic per forecast, respectively, under their native configurations.

Table 3: WeatherBench2 ACC at 24/72 h on 732 global 2020 initializations; higher is better. All methods use one area-weighted evaluator. Native-resolution supervised references are shown separately. Parentheses give estimated FLOP ratios relative to E-JEPA for a 6–72 h forecast, not measured runtimes.

Method	24 h				72 h			
	U850	T850	Q700	MSLP	U850	T850	Q700	MSLP
Persistence	0.524	0.606	0.466	0.656	0.197	0.214	0.160	0.252
Supervised								
Pangu-Weather (≈147× FLOPs)	0.979	0.983	0.942	0.997	0.929	0.951	0.844	0.980
GraphCast (≈510× FLOPs)	**0.984**	**0.988**	**0.955**	**0.998**	**0.939**	**0.961**	**0.866**	**0.983**
Transfer								
WeatherGFM	0.508	0.625	0.456	0.650	0.188	0.226	0.156	0.250
ClimaX	0.754	0.807	0.738	0.843	0.308	0.372	0.308	0.338
Earthformer	0.765	0.811	0.742	0.850	0.329	0.415	0.281	0.350
E-JEPA-50M (frozen)	0.797	0.806	0.799	0.878	0.416	0.454	0.397	0.484
E-JEPA-50M (fine-tuned)	0.883	0.872	0.853	**0.953**	0.630	0.667	0.546	0.763
E-JEPA (frozen)	0.829	0.834	0.813	0.909	0.539	0.585	0.455	0.646
E-JEPA (fine-tuned)	**0.888**	**0.878**	**0.858**	0.953	**0.664**	**0.700**	**0.570**	**0.779**

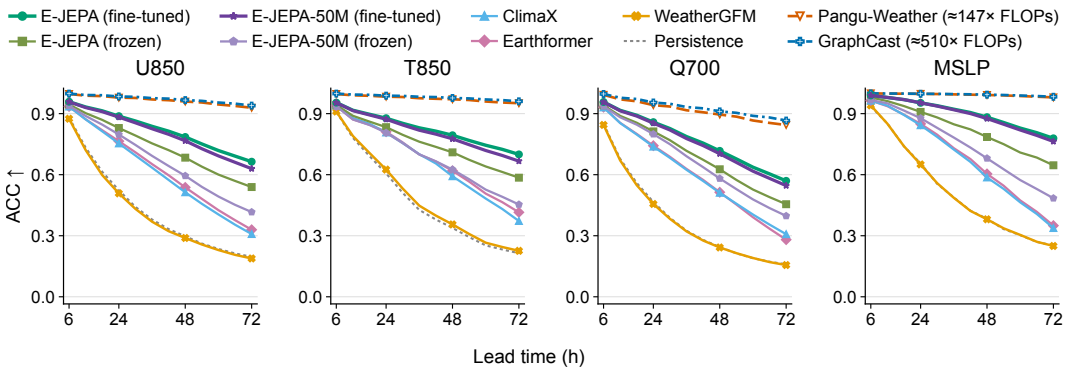

Figure 5: WeatherBench2 ACC versus lead time for wind, temperature, humidity and pressure. Earth-JEPA retains higher skill than the adapted baselines at longer leads; fine-tuning improves the frozen representation. RMSE/MAE curves appear in App. E.

4.3 Sensor Transfer on SEVIR

To test transfer to another sensor, we adapt the frozen infrared-trained encoder to radar vertically integrated liquid (VIL), using the same task prediction body as RainPro-2R (Pablos Sarabia et al., 2026). The results show that frozen Earth-JEPA encoder representations support cross-sensor prediction through this shared body. CSI-p4/p16 measure spatially pooled event detection, HSS adjusts for chance agreement, and LPIPS/SSIM measure perceptual similarity (App. D).

Earth-JEPA slightly improves CSI-p16 over the locally re-evaluated RainPro-2R but trails it on the other five metrics (Table 4). It remains competitive with published deterministic baselines, while generative models often favor pooled/perceptual scores. This comparison demonstrates cross-sensor usefulness of the frozen encoder representation under a shared prediction body; it does not evaluate transfer of the latent-pyramid predictor itself.

4.4 Ablation Study

We retrain learning-objective and architectural variants for 30 epochs, freeze them, and fit the same ridge readout on 10,000 training sequences, with 1,000 validation and 1,000 test sequences at 15–60 min leads. Effective ranks are entropy-based summaries of centered feature singular values,

Table 4: SEVIR nowcasting: five radar frames predict twenty (+100 min). *Published values from RainPro-8/DiffCast (Pablos Sarabia et al., 2026; Yu et al., 2024); †local RainPro-2R evaluation; ‡Earth-JEPA. The local RainPro-2R and Earth-JEPA rows form the comparable group, using the same 7,220 windows, 20-epoch training schedule, and validation-based checkpoint selection.

	Method	CSI ↑	CSI-p4 ↑	CSI-p16 ↑	HSS ↑	LPIPS ↓	SSIM ↑
Gen.	MCVD*	0.2148	0.3020	0.4706	0.2743	0.2170	0.5265
	PreDiff* (Gao et al., 2023)	0.2304	0.3041	0.4028	0.2986	0.2851	0.5185
	STRPM*	0.2512	0.3243	0.4959	0.3277	0.2577	0.6513
	DiffCast* (Yu et al., 2024)	0.3077	0.4122	0.5683	0.4033	0.1812	0.6354
Det.	ConvGRU*	0.2416	0.2554	0.3050	0.2834	0.3766	0.6532
	MAU*	0.2463	0.2566	0.2861	0.3004	0.3933	0.6361
	Earthformer* (Gao et al., 2022b)	0.2513	0.2617	0.2910	0.3073	0.4140	0.6773
	PhyDNet*	0.2560	0.2685	0.3005	0.3124	0.3785	0.6764
	SimVP*	0.2662	0.2844	0.3452	0.3369	0.3914	0.6304
	RainPro-2R† (Pablos Sarabia et al., 2026)	**0.3500**	**0.3762**	0.4012	**0.4474**	**0.2423**	**0.5912**
	RainPro-2R* (Pablos Sarabia et al., 2026)	0.3524	0.3834	0.4171	0.4501	0.2353	0.5966
	E-JEPA‡ (ours)	0.3343	0.3653	**0.4026**	0.4283	0.2485	0.5826

Table 5: Learning-objective and component ablations under a common 30-epoch pretraining and frozen-probe protocol. r_1–r_4: normalized centered effective ranks, coarse to fine; R_{tgt}: nominal target-dimension ratio. TBB RMSE and mean ERA5 nRMSE measure physical-probe accuracy. Higher rank denotes spectral diversity, not necessarily better task performance.

Variant	r_1	r_2	r_3	r_4	R_{tgt}	TBB RMSE (K) ↓	ERA5 nRMSE ↓
Learning objectives							
Online targets + SIGReg	0.326	0.522	0.585	0.643	18.13×	5.5509	0.09063
w/o EMA	0.036	0.144	0.377	0.435	18.13×	5.4323	0.09073
w/o state-transition loss	0.384	0.407	0.522	0.490	18.13×	5.3949	0.09042
Supervised	0.246	0.234	0.296	0.353	1.00×	**5.3219**	**0.08859**
Model components							
w/o fine-scale pathways	0.446	–	–	–	272.00×	5.5890	0.09125
w/o history masking	0.444	0.461	0.546	0.511	18.13×	5.3919	0.09042
w/o FPN	0.445	0.463	0.548	0.514	18.13×	5.3919	0.09041
Full-to-axial (coarse)	0.444	0.462	0.546	0.511	18.13×	5.3917	0.09042
E-JEPA	0.444	0.462	0.546	0.511	18.13×	5.3918	0.09041

normalized by channel count (Garrido et al., 2023); probes assess nine brightness-temperature and 195 ERA5 channels. This separates representation diversity, downstream accuracy and prediction-target compactness.

Learning objectives. Removing EMA reduces coarse-scale ranks r_1/r_2 by 92.0%/68.9% despite retaining stop-gradient. EMA is therefore central to avoiding collapse in this configuration. Replacing State-Transition Loss by direct state MSE lowers rank at every scale (13.5% at r_1) with almost unchanged probe error, supporting richer representations under temporal-difference supervision. Online targets with SIGReg (Balestriero & LeCun, 2025; Maes et al., 2026) shift diversity unevenly across scales and worsen TBB error by about 3%. Direct physical supervision improves TBB/ERA5 errors by 1.3%/2.0%, but has lower ranks than Earth-JEPA and a much larger target. Earth-JEPA thus approaches supervised probe accuracy with an $18.13\times$ smaller prediction target. Figure 6 shows close target/prediction structure; state-MSE yields a localized coarse-detail artifact, while no-EMA and SIGReg variants show discontinuities. Supervision retains richer fine-scale detail in this illustrative case.

Model components. Removing fine-scale paths raises the target ratio to $272\times$ but worsens TBB RMSE by 3.66%, confirming the cost of a coarse bottleneck for dense tasks. Removing tube masking or Cross-Scale Fusion, or replacing coarse full attention by axial attention, changes probe errors by less than 0.01%. Masking, top-down exchange and axial attention thus retain probe accuracy while jointly organizing information across modalities and scales.

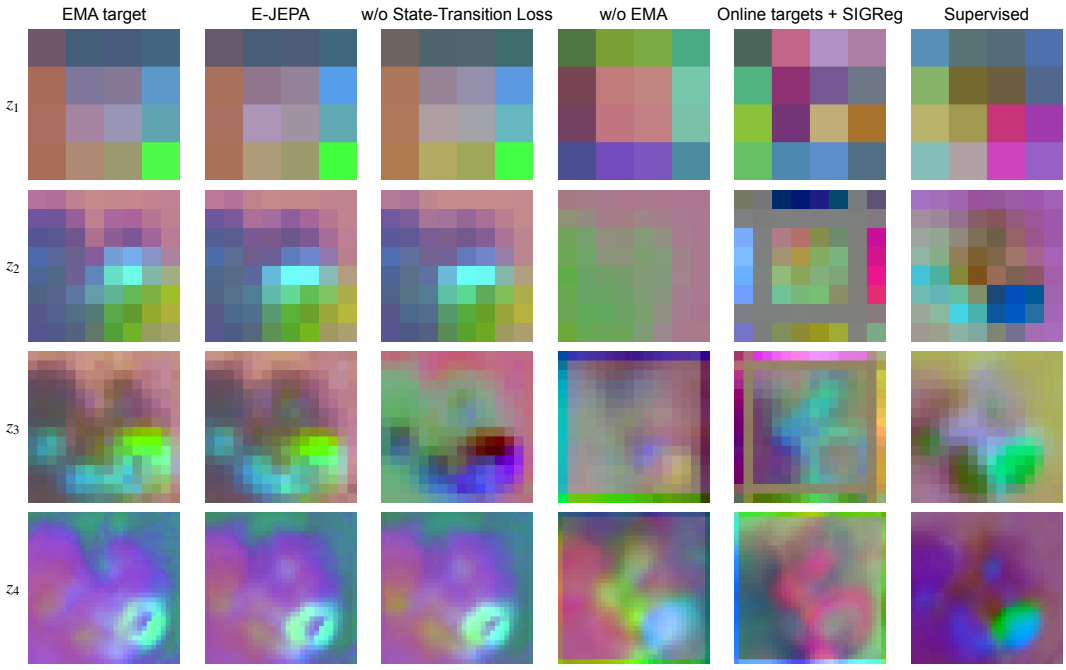

Figure 6: Latent pyramids at 60 min lead. Rows run from coarse to fine; RGB encodes three history-fitted PCA components. Only EMA targets and Earth-JEPA share PCA coordinates. Other variants use independent bases, so their colors cannot be compared as cross-model distances. Earth-JEPA preserves target spatial patterns across scales, whereas no-EMA predictions show marked discontinuities.

4.5 Computational Efficiency

Reconstruction-free supervision and Scale-Specific Attention give Earth-JEPA the lowest arithmetic cost in Table 6, 46.9% below WeatherGFM. Pretraining plus rainfall-head training uses 59.2% fewer GPU-hours than Earthformer and 86.9% fewer than ClimaX. Our readout has 43.7% fewer trainable parameters than the adapted baselines, so capacity is not matched. Unlike Fig. 1, these training costs include task adaptation.

Table 6: Rainfall-task computational cost. GPU-h include pretraining and readout training; arithmetic and parameter definitions follow App. D.

Method	GFLOPs↓	GPU-h↓	P_{task} (M)↓	R_{tgt}
ClimaX	1,002.1	235.7	21.86	0.95×
WeatherGFM	306.0	196.3	21.86	0.95×
CasCast Sup.†	326.1	108.1	126.59	–
Earthformer	6,332.1	75.8	21.86	0.95×
Earthformer Sup.†	19,252.1	68.8	119.25	–
E-JEPA	**162.4**	**30.9**	**12.31**	**18.13**×

5 Conclusion

Earth-JEPA combines modality-wise encoding, multi-scale latent prediction and state-transition supervision for efficient, transferable Earth-system representation learning. EarthCastBench results and transfer to WeatherBench2 and SEVIR demonstrate the promise of reconstruction-free pretraining, including missing-source inference and economical adaptation. EMA targets prevent coarse-scale collapse, while latent pyramids balance compactness with dense-task accuracy. Future work will evaluate more Earth-system tasks and scale the backbone toward operational model sizes. Our findings motivate further study of how Earth-system models are trained, beyond improving task-specific predictors alone.

AI Use Statement AI tools assisted English translation and language polishing. The authors reviewed the manuscript and take responsibility for its content and originality.

REFERENCES

Shreya Agrawal, Mohammed Alewi Hassen, Emmanuel Asiedu Brempong, et al. An operational deep learning system for satellite-based high-resolution global nowcasting. *arXiv preprint arXiv:2510.13050*, 2025.

Anna Allen et al. End-to-end data-driven weather prediction. *Nature*, 641:1172–1179, 2025. doi: 10.1038/s41586-025-08897-0.

K. M. Anirudh, S. Sandeep, and Hariprasad Kodamana. Tracing the unlabeled storm: Cross-variable transfer in a Lagrangian atmospheric JEPA framework. *arXiv preprint arXiv:2608.22358*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Mido Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Guillaume Astruc et al. AnySat: One Earth observation model for many resolutions, scales, and modalities. In *CVPR*, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Kaifeng Bi et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xinguang Xiang, and Wen Gao. MAU: A motion-aware unit for video prediction and beyond. In *NeurIPS*, volume 34, pp. 26950–26962, 2021.

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. STRPM: A spatiotemporal residual predictive model for high-resolution video prediction. In *CVPR*, pp. 13946–13955, 2022. doi: 10.1109/CVPR52688.2022.01356.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022.

Guillaume Couairon et al. ArchesWeatherGen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Sci. Adv.*, 2026. doi: 10.1126/sciadv.adx2372.

Mark Friedl and Damien Sulla-Menashe. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061. NASA EOSDIS Land Processes DAAC, 2022.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *NeurIPS*, 2023.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *CVPR*, pp. 3170–3180, 2022a. doi: 10.1109/CVPR52688.2022.00317.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for Earth system forecasting. In *NeurIPS*, 2022b.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Wang. PreDiff: Precipitation nowcasting with latent diffusion models. In *NeurIPS*, 2023.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*, volume 202, pp. 10929–10974. PMLR, 2023.

Junchao Gong, Lei Bai, et al. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *ICML*, volume 235 of *PMLR*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuolin Hsu, Robert Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In *Satellite Precipitation Measurement*, volume 67, pp. 343–353. Springer Nature, 2020. doi: 10.1007/978-3-030-24568-9_19.

Johannes Jakubik et al. TerraMind: Large-scale generative multimodality for Earth observation. In *ICCV*, 2025.

Dmitrii Kochkov et al. Neural general circulation models for weather and climate. *Nature*, 632: 1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Remi Lam et al. Learning skillful medium-range global weather forecasting. *Science*, 382:1416–1421, 2023.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, pp. 11474–11484, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017. doi: 10.1109/CVPR.2017.106.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, María Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, 13:4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

National Satellite Meteorological Center. FY-4B product catalogue. China Meteorological Administration, n.d. URL `https://satellite.nsmc.org.cn/DataPortal/en/code/FY4B.html`. Accessed September 2026.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *ICML*, volume 202 of *PMLR*, 2023.

NOAA National Centers for Environmental Information. ETOPO 2022 15 arc-second global relief model. NOAA NCEI, 2022.

Rafael Pablos Sarabia, Joachim Nyborg, Morten Birk, Jeppe Liborius Sjørup, Anders Lillevang Vesterholt, and Ira Assent. RainPro-8: An efficient deep learning model to estimate rainfall probabilities over 8 hours. In *ICLR*, 2026.

Jaideep Pathak et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price et al. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Stephan Rasp et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.*, 16:e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Suman Ravuri et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677, 2021.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS*, volume 30, 2017.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *ECCV*, pp. 459–479. Springer, 2022. doi: 10.1007/978-3-031-20053-3_27.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *NeurIPS*, volume 33, 2020.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, volume 35, 2022.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *CVPR*, pp. 27763–27773, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Yuchen Zhang et al. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619:526–532, 2023.

Xiangyu Zhao, Zhiwang Zhou, Wenlong Zhang, Yihao Liu, Xiangyu Chen, Junchao Gong, Hao Chen, Ben Fei, and Shiqi Chen. WeatherGFM: Learning a weather generalist foundation model via in-context learning. In *ICLR*, 2025.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, et al. On the foundations of Earth foundation models. *Commun. Earth Environ.*, 7:103, 2026. doi: 10.1038/s43247-025-03127-x.

A NOTATION

Table 7 defines the symbols used in the method description (Sec. 3) and Eqs. (1)–(5). Throughout, s indexes spatial scales from coarse to fine, t indexes history frames, and k indexes forecast leads; the batch index b is written only where the training objective requires it. All scale-s states lie in $\mathbb{R}^{C_s \times H_s \times W_s}$, and norms act on their vectorization. Numeric configuration values (grid sizes, depths, masking rate) are listed in App. C.

Table 7: Notation used in Sec. 3 and Eqs. (1)–(5).

Symbol	Meaning
Indices and dimensions	
s	spatial-scale index, coarse to fine, $s \in \{1, \dots, S\}$
t	history-frame index, $t \in \{1, \dots, T\}$
k	forecast-lead index, $k \in \{1, \dots, K\}$
b	sample index within a batch of size B
C_s, H_s, W_s	channels, height, and width of the scale-s latent grid
D_s	dimension of the vectorized scale-s state, $D_s = C_s H_s W_s$
Inputs and encoders	
x_t^{sat}, x_t^{rea}	satellite brightness-temperature and reanalysis inputs at history time t
x^{geo}	time-invariant geographic fields
f_{θ_1}, f_{θ_2}	satellite and reanalysis encoding branches
f_{θ_3}	geographic-field encoder (not EMA-updated)
θ_F	cross-modal fusion parameters
θ	online dynamic parameters, $\theta = \{\theta_1, \theta_2, \theta_F\}$
$f_{\bar{\theta}}$	EMA target encoder, $\bar{\theta} \leftarrow \mu\,\bar{\theta} + (1-\mu)\,\theta$ with momentum μ
$c^{(s)}$	static conditioning features at scale s, encoded once per sample by f_{θ_3}
p_{drop}	probability of dropping the reanalysis modality during pretraining
Latent states and prediction	
$z_t^{(s)}$	online fused dynamic state at history time t and scale s
$Z^{(s)}$	stacked history states, $Z^{(s)} \in \mathbb{R}^{T \times C_s \times H_s \times W_s}$
$\mathcal{M}(\cdot)$	spatiotemporal tube-masking operator on latent history tokens
M, ρ, T_m	macro-cell lattice side, masked macro-cell fraction, and masked trailing history frames
$m^{(s)}$, $l_k^{(s)}$	learnable mask token and lead code at scale s
$q_k^{(s)}$	future query token, $q_k^{(s)} = m^{(s)} + c^{(s)} + l_k^{(s)}$
$\mathcal{A}_s$	scale-specific attention at scale s (full self-attention at the coarsest scale; axial attention over time, height, and width otherwise)
$U^{(s)}$, $\tilde{U}^{(s)}$	raw and cross-scale-fused predictor outputs at scale s
$P^{(s)}$, $\psi^{(s)}$	1×1 projection and zero-initialized fusion head of Cross-Scale Fusion
$\pi^{(s)}$	per-scale prediction head emitting the latent increments
$d_k^{(s)}$	predicted latent increment at lead k
$\hat{z}_k^{(s)}$	predicted future state, $\hat{z}_k^{(s)} = z_T^{(s)} + \sum_{j=1}^{k} d_j^{(s)}$
$\bar{z}_k^{(s)}$	EMA target state at lead k
$\mathrm{sg}(\cdot)$	stop-gradient operator
Objective and adaptation	
$e_{b,k}^{(s)}$	state error, $e_{b,k}^{(s)} = \hat{z}_{b,k}^{(s)} - \mathrm{sg}\big(\bar{z}_{b,k}^{(s)}\big)$
$r_{b,k}^{(s)}$	transition residual, $r_{b,1}^{(s)} = e_{b,1}^{(s)}$ and $r_{b,k}^{(s)} = e_{b,k}^{(s)} - e_{b,k-1}^{(s)}$ for $k > 1$
$q_{b,k}$, $v_{b,k}$	valid-target indicator and pairwise validity weight, $v_{b,1} = q_{b,1}$, $v_{b,k} = q_{b,k} q_{b,k-1}$
$\mathcal{L}_{\mathrm{ST}}$	State-Transition Loss (Eq. (4))
$\|\cdot\|_2$	Euclidean norm over the vectorized state
g_ϕ	task-specific readout with parameters ϕ
$\hat{y}$, y	downstream prediction and reference target
R_{tgt}	prediction-target dimension ratio (App. D)

B Datasets and Preprocessing

B.1 EarthCastBench

Source modalities. EarthCastBench aligns four public data families over the FY-4B observing sector (60°S–60°N, 45°E–165°E). (i) *Geostationary radiances*: brightness temperatures—equivalent blackbody temperatures inferred from radiance, sensing cloud tops and atmospheric layers, and the surface in cloud-free pixels—from nine infrared bands (channels 7–15) of the AGRI imager aboard the geostationary FY-4B satellite (National Satellite Meteorological Center, n.d.), natively at 4 km and 15 min. (ii) *Reanalysis*: ERA5 (Hersbach et al., 2020) at 0.25° and hourly resolution, providing 15 upper-air variables (`t`, `u`, `v`, `q`, `w`, `r`, `cc`, `ciwc`, `clwc`, `crwc`, `cswc`, `d`, `pv`, `vo`, `o3`) on 13 pressure levels (1000–50 hPa), six single-level fields (`t2m`, `u10`, `v10`, `msl`, `d2m`, `sp`), and a high-resolution land 2 m temperature field (ERA5-Land (Muñoz-Sabater et al., 2021), reserved and unused in this study). (iii) *Precipitation*: half-hourly IMERG rain rates at 0.1° (Huffman et al., 2020); a derived binary convective-onset label is reserved but unused in this study. (iv) *Static geography*: ETOPO2022 elevation (NOAA National Centers for Environmental Information, 2022), the 17-class International Geosphere–Biosphere Programme (IGBP) land-cover classification from the 2024 MODIS MCD12Q1 layer (Friedl & Sulla-Menashe, 2022), latitude, longitude, and satellite zenith angle—the angle between the viewing direction and the local vertical, which controls parallax and atmospheric path length—complemented by sinusoidal encodings of local time and day of year.

Chronology and splits. Each sample is a 12-frame, 224-channel patch of 128×128 pixels at 15-min cadence: eight observed frames followed by four future frames (+15/+30/+45/+60 min). Anchors span 2024-03-04T16:00Z to 2025-08-31T15:00Z. Candidate anchors are first partitioned chronologically into training, validation, and test blocks; patches are then drawn with a fixed seed under two constraints: each (anchor, tile) pair is unique, and the anchor frame exhibits IMERG rain rates above 0.1 mm h^{-1} in more than 20% of patch pixels, yielding a precipitation-focused sampling. The final splits contain 10,000/1,000/1,000 samples, and every split passes checks on row counts, duplicate anchors, the rain-coverage condition, finite non-negative IMERG values, complete ERA5 records, and channel counts.

Geospatial and temporal alignment. The DEM uses the existing $0.04°$ global bedrock file and bilinear reprojection to $0.1°$. Class one-hot masks are average-reprojected to $0.1°$, not bilinear interpolation of class IDs. All fields are cropped around each patch with a two-pixel context margin and resampled to the $0.04°$ latitude–longitude grid, so that all channels are pixel-aligned. Frame times follow the 15-min AGRI cadence. All Earth-JEPA tasks use the same causal temporal-alignment rule: historical ERA5 and IMERG values are obtained using one-sided temporal interpolation/extrapolation from source records available no later than each frame. In particular, each later 15-min frame is extrapolated from preceding source records; no future source record is used to construct historical inputs. Future ERA5 and IMERG values are used only as prediction targets. Surface fields follow the same rule. The tasks of Sec. 4.1 read precipitation targets from ch. 214 and surface-variable targets from ch. 216–221.

Normalization. Brightness temperatures are standardized with a mean of 220 K and a standard deviation of 60 K. Elevation is divided by 5000 m and zenith angle by 70°; geographic coordinates are affinely rescaled from the sector extent, and the IGBP index is kept as a single categorical channel and divided by 17. ERA5 fields are standardized channel-wise with fixed statistics estimated on the training period. IMERG values are clipped to be non-negative, with missing entries set to zero. Pressure fields are converted from Pa to hPa for reporting.

B.2 WeatherBench2 Task Data

WeatherBench2 (Rasp et al., 2024) is a global, ERA5-based benchmark for data-driven weather models. We evaluate on its native 1.5° grid (121×240 cells, global coverage) under the 2020 protocol: 732 test initializations at 00 and 12 UTC. Inputs are eight six-hourly states spanning -42 h to the initialization time; forecasts are scored at six-hourly leads from +6 h to +72 h. Readout training and fine-tuning use a 10,000-row training subset with four training leads (6/12/18/24 h). Following the benchmark standard, we report ACC for U850, T850, Q700, and MSLP in the main text, with RMSE

Table 8: EarthCastBench sample contract. Each sample stacks 224 channels over eight observed frames and four future frames at 15-min cadence on the 4 km FY-4B grid. Channels 0–213 constitute the model input (TBB: brightness temperature); channels 214–223 carry downstream targets and reserved auxiliary fields.

Group	Ch.	n	Content and role
FY-4B AGRI TBB	0–8	9	Infrared brightness temperatures, bands 7–15; dynamic input.
Time context	9–12	4	Sine/cosine of local time and day of year; dynamic input.
Static geography	13–17	5	Elevation, IGBP class, latitude, longitude, zenith angle; static conditioning.
ERA5 upper air	18–212	195	15 variables $\times$ 13 pressure levels (1000–50 hPa); dynamic input.
ERA5 availability	213	1	1 if an ERA5 source exists for the frame, else 0.
IMERG precipitation	214	1	Rain rate (mm h^{-1}); target, and history input for precipitation readouts.
Convective onset	215	1	Reserved binary label (rain transitions $<1.0 \rightarrow >1.5$ mm h^{-1}); unused in this study.
ERA5 single level	216–221	6	`t2m`, `u10`, `v10`, `msl`, `d2m`, `sp`; surface targets.
Reserved	222	1	Unused loader slot.
ERA5-Land `t2m`	223	1	Reserved high-resolution land 2 m temperature; unused in this study.

Table 9: Dataset and task protocols. EarthCastBench is released with this work; WeatherBench2 and SEVIR are public benchmarks used here for transfer evaluation.

Dataset	Domain and grid	Cadence	History $\rightarrow$ horizon	Train/val/test
EarthCastBench	FY-4B sector; 128×128 patches at 4 km	15 min	8 frames $\rightarrow$ 4 frames (60 min)	10,000/1,000/1,000
WeatherBench2	Global; 121×240 at 1.5°	6 h	8 states $\rightarrow$ 12 steps (72 h)	10,000 training rows; 732 test initializations
SEVIR (VIL)	Contiguous US storm events; 128×128	5 min	5 frames $\rightarrow$ 20 frames (100 min)	standard splits; 7,220 test windows

and MAE in physical units deferred to App. E; all spatial averages use spherical cell-area weighting (definitions in App. D).

B.3 SEVIR Task Data

SEVIR (Veillette et al., 2020) is a dataset of storm events over the contiguous United States. We adopt the standard radar-only nowcasting protocol of prior work (Yu et al., 2024; Pablos Sarabia et al., 2026): the model receives five 5-min frames of vertically integrated liquid (VIL)—a radar-derived estimate of column liquid water used as a precipitation proxy—at 128×128 resolution and forecasts the subsequent 20 frames (+5 to +100 min). Each event contributes five windows starting at frames $\{0, 5, 10, 15, 20\}$, yielding 7,220 test windows on the release test split. VIL intensities are normalized to $[0, 1]$, and the dataset's standard train/validation/test partitions are used throughout.

C Architecture and Task-Adaptation Details

C.1 Backbone configuration

Input layout. The online encoder, Latent-Pyramid Predictor, and projectors contain 113.2M parameters at inference (the EMA target encoder is a momentum copy and is not counted).

Modality-Wise Encoder. The satellite branch processes the 13 high-rate channels (brightness temperatures + time encodings) with a Swin-V1 hierarchical transformer (Liu et al., 2021) (embedding dimension 96, depths $(2, 2, 6, 2)$, heads $(3, 6, 12, 24)$, window size 8, drop-path rate 0.1) and yields four feature maps that we reorder coarse-to-fine.

Table 10: Per-scale configuration of the latent pyramid, coarse to fine. Each scale uses a depth-6 predictor; the coarsest scale applies full self-attention and the three finer scales apply sequential temporal/height/width axial attention (Ho et al., 2019).

Scale	Grid	Channels	Attention	Heads	FFN dim	Projector hidden
z_1	4×4	768	full	8	2048	2048
z_2	8×8	384	axial (T, H, W)	6	1536	1536
z_3	16×16	192	axial (T, H, W)	3	768	768
z_4	32×32	96	axial (T, H, W)	3	384	384

E-JEPA-50M variant. The E-JEPA-50M backbone uses Swin embedding dimension 64, encoder heads $(2, 4, 8, 16)$, scale channels $(512, 256, 128, 64)$, predictor heads $(8, 4, 2, 2)$, and FFN/projector widths $(1344, 1024, 512, 256)$ on the same grids; it has 50.29M online parameters (68.67M including the stored EMA teacher) and was pretrained for 40 epochs (1,040 updates) rather than 30 (780). The E-JEPA-50M/E-JEPA comparison in Sec. 4.2 therefore differs in both capacity and pretraining budget and is not a controlled scaling study. Earth-JEPA without a size suffix refers to the default E-JEPA configuration.

Masking and dropout. Spatiotemporal Tube Masking partitions each scale's grid into a 4×4 macro grid ($M{=}4$) and masks two macro cells ($\rho{=}2/16$); the same spatial pattern is repeated across the final $T_m{=}4$ history frames and all feature channels, and is resampled independently per scale and per sample. Masked positions are replaced by a learned mask token, and each future query is the mask token plus the static features and a learned lead code. We additionally apply ERA5 modality dropout with probability $p_{\text{drop}}{=}0.3$ during pretraining, which supports the Drop-ERA5 inference setting; predictor attention dropout is 0.1.

EMA and loss weighting. The EMA target encoder tracks f_{θ_1}, f_{θ_2} and the cross-modal fusion parameters with a cosine momentum schedule from 0.996 to 1.0 over the 780-update run; the static encoder f_{θ_3} and the predictor are not EMA-updated. The State-Transition Loss (Eq. (4)) uses unit weights across the four scales and the four leads, the nominal BK denominator, and no target normalization. The shared backbone weights the first-state term by the target-frame validity $v_{b,1}{=}q_{b,1}$ and each transition term by $v_{b,k}{=}q_{b,k}q_{b,k-1}$.

C.2 SHARED PRETRAINING AND ABLATION PROTOCOL

The E-JEPA backbone and all ablation variants (Sec. 4.4) share the recipe in Table 11; ablations are retrained from fresh initialization and evaluated at the fixed final checkpoint (epoch 29 zero-based, i.e., after 30 completed epochs) without any validation- or test-based selection. For the frozen-probe comparison in Table 5, the four scales are bilinearly aligned to 32×32 and concatenated ($1{,}440$ channels); a ridge readout ($\lambda{=}0.1$) is fit on $10{,}000$ training sequences and scored once on $1{,}000$ validation and $1{,}000$ test sequences. TBB RMSE is reported in kelvin; the ERA5 score is the per-variable RMSE pooled over samples and the four leads, averaged over the 195 variables. The ranks r_1–r_4 are normalized centered effective ranks (Garrido et al., 2023) computed from the singular values of the centered feature matrices of 200 fixed sequences (App. D); they quantify spectral diversity, not overall quality.

Table 11: Shared pretraining recipe. Ablations re-run this exact recipe from fresh initialization with the same seed; the effective batch size, budget, and seed are identical across runs, while rank counts differ (e.g., 8 ranks $\times$ micro-batch 3 $\times$ accumulation 16 for the FPN ablation).

Item	Setting
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Learning rate	$10^{-4} \rightarrow 10^{-6}$ cosine, 52 warmup updates
Effective batch size	384 (micro-batch 4 $\times$ accumulation 16 $\times$ 6 GPUs)
Training rows / epoch	9,984 (26 updates/epoch)
Budget	30 epochs; 780 updates; 299,520 sample exposures
Precision / hardware	FP32; 6$\times$ NVIDIA A800-80GB (main runs)
Measured pretraining cost	29.7 A800 GPU-h
Seed	20260820

C.3 PER-TASK READOUTS

Table 12 summarizes the four downstream settings. The same pretrained Earth-JEPA backbone and checkpoint are used for precipitation nowcasting, surface pressure, WeatherBench2, and SEVIR. Unless stated otherwise, the backbone is frozen and only the task readout is trained.

Table 12: Per-task readout and optimization settings. "Selection" gives the reported checkpoint and its selection rule; EBS is the effective batch size. The two WeatherBench2 rows use the same readout; fine-tuning additionally updates the backbone.

Task	Readout and objective	Optimization and selection
Precipitation nowcasting (EarthCast-Bench)	multi-scale convolutional readout on the predicted latent pyramid; historical 12-term GroupA rainfall objective	All learned methods use the same 40-epoch budget and validation-based checkpoint selection
Surface pressure (EarthCast-Bench)	six single-level variables at $+15$–$+60$ min from the predicted pyramid plus eight observed surface fields; $50 \times (\mathrm{nMSE} + 0.05\,\mathrm{nMAE})$	AdamW, lr 10^{-3}, wd 10^{-4}; 40 epochs; EBS 512; seed 0 Selection: max. mean RMSE/MAE fractional gain over persistence on full validation (earliest epoch on ties)
WeatherBench2, frozen	121-field state readout; all learned methods use the same 121 historical fields and 121 prediction targets. The loss averages the four standard variables and adds 0.1 times the mean over the other 117 fields, reflecting finite readout capacity while retaining full-state prediction	Equal downstream data, 40-epoch budget, and validation-based selection across learned baselines
WeatherBench2, fine-tuned	same readout and 121-field input/output contract as the frozen row	Same data and validation-based selection protocol; cumulative E-JEPA cost includes 29.7 GPU-h pretraining and 108.5 GPU-h fine-tuning
SEVIR sensor transfer	radar adapter and input projections feed the frozen encoder representation into the shared task prediction body (Tu et al., 2022; Pablos Sarabia et al., 2026)	All compared methods use the same 20-epoch schedule and validation-based checkpoint selection; the locked test is evaluated after selection

Surface pressure. Here Drop-ERA5 removes ERA5 inputs when constructing the frozen backbone state; all rows with this label use the same removal rule.

WeatherBench2. All learned methods receive the same 121 historical fields and predict the same 121 WeatherBench2 fields, with only the minimal input/output mappings required by each model interface. The readout loss emphasizes the four standard reported variables because readout capacity is limited, while retaining a 0.1 weight on the mean loss over the other 117 fields. Learned baselines use equal downstream settings and validation-based selection. Readouts are trained on the 10,000-row training subset with four training leads (6/12/18/24 h) before evaluation on the 732 test initializations.

SEVIR. The frozen Earth-JEPA encoder representation is transferred to radar inputs through a task-specific input adapter and the same prediction body used by RainPro-2R (Veillette et al., 2020; Pablos Sarabia et al., 2026). This experiment evaluates the usefulness of the frozen encoder representation for cross-sensor prediction. Decoding thresholds and checkpoint selection use validation data; all compared methods follow the same 20-epoch schedule, and the locked test split is evaluated after selection.

Published SEVIR baselines include MCVD (Voleti et al., 2022), STRPM (Chang et al., 2022), ConvGRU (Shi et al., 2017), MAU (Chang et al., 2021), PhyDNet (Le Guen & Thome, 2020) and SimVP (Gao et al., 2022a).

D Evaluation Metrics

This appendix defines every quantitative metric reported in the main text. Unless stated otherwise, scores are computed per variable and per lead time and then averaged as indicated in each table; arrows mark whether higher (↑) or lower (↓) values are better.

D.1 Error norms

For a single physical variable with predictions $\hat{y}$ and targets y, evaluated over N test samples at one lead time on a grid of G cells with weights $w_g \geq 0$, $\sum_g w_g = 1$,

$$\mathrm{RMSE} = \Big(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} w_g\,(\hat{y}_{i,g} - y_{i,g})^2\Big)^{1/2}, \qquad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G} w_g\,|\hat{y}_{i,g} - y_{i,g}|. \tag{6}$$

The square root is taken *after* averaging squared errors over cases, so RMSE^2 is the mean per-case spatial MSE. The regional EarthCastBench grids use uniform per-pixel weights $w_g = 1/G$; the global WeatherBench2 grid uses the latitude-cell weights defined below. Brightness-temperature errors are reported in kelvin, averaged over the nine channels. For the 195-variable reanalysis readout (Table 5) we report the *mean normalized RMSE*: each channel is standardized, the squared errors of a channel are pooled over all samples and lead times before the root is taken, and the resulting per-channel RMSEs are averaged,

$$\overline{\mathrm{nRMSE}} = \frac{1}{C}\sum_{c=1}^{C}\Big(\frac{1}{NKG}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{g=1}^{G}\big(\hat{z}_{i,k,g,c} - z_{i,k,g,c}\big)^2\Big)^{1/2}, \tag{7}$$

so that variables with heterogeneous physical units contribute comparably.

D.2 Categorical precipitation scores

Thresholding a rain-rate field at τ (mm h^{-1}) yields binary events and contingency counts TP, FP, FN, TN (hits, false alarms, misses, correct negatives), accumulated over pixels and test cases. We report the critical success index (CSI, also known as threat score), probability of detection (POD) and false alarm ratio (FAR),

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \qquad \mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}, \tag{8}$$

which are standard verification measures in precipitation nowcasting (Ravuri et al., 2021; Zhang et al., 2023). The binarizing thresholds are listed with each table.

D.3 WeatherBench2 global scores

WeatherBench2 (Rasp et al., 2024) evaluates global fields on a 121×240 latitude–longitude grid. Latitude row i receives the spherical area weight

$$w_i \;=\; \frac{\sin\varphi_{i+\frac{1}{2}} - \sin\varphi_{i-\frac{1}{2}}}{\sum_{i'}\big(\sin\varphi_{i'+\frac{1}{2}} - \sin\varphi_{i'-\frac{1}{2}}\big)}, \tag{9}$$

where $\varphi_{i\pm\frac{1}{2}}$ are the cell boundaries (midpoints between adjacent rows, $\pm\pi/2$ at the poles), and the weighted spatial mean is $\langle x\rangle_w = \frac{1}{240}\sum_{i=1}^{121} w_i \sum_{j=1}^{240} x_{ij}$. For each initialization t, variable and lead time, the anomaly correlation coefficient (ACC) compares the spatial patterns of the forecast and observed anomalies $f'_t = \hat{y}_t - c_t$ and $a'_t = y_t - c_t$ relative to the climatological field c_t of the valid time,

$$\mathrm{ACC}_t = \frac{\langle f'_t\, a'_t\rangle_w}{\sqrt{\langle {f'_t}^2\rangle_w\,\langle {a'_t}^2\rangle_w}}. \tag{10}$$

ACC thus measures whether a forecast places above- and below-normal conditions in the right locations. We report the unweighted mean of ACC_t over the 732 initializations of the 2020 evaluation set; an exactly zero forecast anomaly field, for which the ratio is indeterminate ($0/0$, e.g., the climatology reference), is assigned $\mathrm{ACC} = 0$. WeatherBench2 RMSE and MAE apply the same latitude weights in the definitions above.

D.4 SEVIR nowcasting scores

SEVIR (Veillette et al., 2020) nowcasts vertically integrated liquid (VIL), a radar-derived measure of column liquid water. Following the protocol of Yu et al. (2024) and Pablos Sarabia et al. (2026),

we evaluate the six VIL thresholds $\{16, 74, 133, 160, 181, 219\}$ with decoding calibrated on the validation split; contingency counts are pooled over all 7,220 test windows. CSI and the Heidke skill score (HSS),

$$\mathrm{HSS} = \frac{2\,(\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN})}{(\mathrm{TP}+\mathrm{FP})(\mathrm{FP}+\mathrm{TN}) + (\mathrm{TP}+\mathrm{FN})(\mathrm{FN}+\mathrm{TN})}, \tag{11}$$

are computed per VIL threshold and lead time, then averaged over the 20 lead frames and the six thresholds. The pooled variants CSI-pk, $k \in \{4, 16\}$, aggregate the predicted and observed fields with a $k \times k$ spatial max-pooling window before binarization, thereby crediting event detection at coarser spatial scales and reducing the double penalty of small displacement errors (Yu et al., 2024). Two perceptual scores complement the categorical ones: LPIPS (Zhang et al., 2018) is a weighted distance between unit-normalized deep features of predicted and observed frames (lower is better), and SSIM (Wang et al., 2004),

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{12}$$

compares local luminance, contrast and structure (higher is better); both are averaged over frames and test cases.

D.5 Spectral diversity of the representation

To diagnose representational collapse (Table 5) we compute the normalized centered effective rank r_s of the forecast latents at each pyramid scale $s \in \{1, \ldots, 4\}$ from coarse to fine, following the effective-rank diagnostic of Garrido et al. (2023) as used for joint-embedding models by Bardes et al. (2024). At each scale we pool $N = 12{,}800$ tokens—200 fixed evaluation sequences $\times$ 4 leads $\times$ 16 fixed spatial positions—into a matrix $Z \in \mathbb{R}^{N \times C}$ centered by its column mean, with singular values $\sigma_1 \geq \cdots \geq \sigma_J$, $J = \min(N, C)$, and $p_j = \sigma_j / \sum_i \sigma_i$. Then

$$r_s = \frac{1}{C} \exp\Big(-\sum_{j=1}^{J} p_j \log p_j\Big) \in (0, 1]. \tag{13}$$

r_s approaches 1 for a flat spectrum (all feature directions equally expressed) and drops toward 0 when the representation concentrates in a few directions; it characterizes spectral diversity, not downstream accuracy by itself.

D.6 Prediction-target dimension ratio

R_{tgt} compares the dimensionality of a 204-channel physical-state target at 128×128 resolution with the dimensionality D_{pred} of the target a model actually predicts,

$$R_{\mathrm{tgt}} = \frac{204 \times 128 \times 128}{D_{\mathrm{pred}}}. \tag{14}$$

Earth-JEPA predicts the four-scale latent pyramid ($R_{\mathrm{tgt}} = 18.13\times$); the supervised variant predicts the physical state directly ($1.00\times$); adapted baselines predicting all 214 physical channels give $204/214 \approx 0.95\times$. R_{tgt} is a nominal ratio of target sizes: it scales the output dimensionality of the predictor, and hence part of its arithmetic cost, but it is neither a measured information-compression rate nor, by itself, a speedup factor.

D.7 Compute accounting

GPU-hours measure cumulative device time, (number of devices) $\times$ (wall-clock hours), summed over the recorded stages indicated in each table or figure caption (pretraining and downstream training; unrecorded feature extraction and evaluation are excluded). Because they scale with device count, GPU-hours are not elapsed training time. For external baselines trained on other hardware, Fig. 1 converts published training resources to A800-equivalent GPU-hours through the ratio of dense peak tensor throughputs, $H_{\text{A800-eq}} = N\, t_{\mathrm{h}}\, P_{\mathrm{source}} / P_{\mathrm{A800}}$; these entries (marked with an asterisk) are capacity proxies, not measured runtimes.

Counted FLOPs use the convention of 2 FLOPs per multiply–accumulate and cover registered dense arithmetic—convolutions (including transposed convolutions), matrix products and attention—evaluated at batch size 1 and the input resolution of the corresponding experiment. Normalization, nonlinearities, elementwise operations, reshaping, interpolation and probability decoding are not counted. Counted FLOPs are a deterministic function of architecture and input shape and are reported as an arithmetic-cost proxy, not as measured latency or energy.

E Supplementary Results

This appendix collects the quantitative WeatherBench2 results and an illustrative pressure case from EarthCastBench.

E.1 WeatherBench2 RMSE and MAE

Tables 13 and 14 report RMSE and MAE at 24 h and 72 h for the transfer comparison of Sec. 4.2, and Figs. 7–8 show the corresponding error curves over the full 6–72 h rollout. The evaluation protocol is identical to the ACC results in Table 3: 732 global initializations at 00/12 UTC throughout 2020 on the 1.5° (121×240) grid with spherical-cell-area weighting (Rasp et al., 2024). The adapted learned methods use the same input and target variables and equal downstream settings, with only minimal input/output mappings for their interfaces. The ranking within the transfer group is stable across all three metrics: fine-tuned readouts consistently improve on their frozen counterparts, and E-JEPA (fine-tuned) is the most accurate transfer method on all four variables at 72 h, reducing RMSE by 20–47% relative to the strongest adapted baseline depending on the variable (18–43% at 24 h).

Table 13: RMSE (lower is better) at 24 h and 72 h on the global 1.5° grid (121×240), using 732 test initializations at 00/12 UTC throughout 2020 with spherical-area weighting. Rows are grouped as in Table 3, with Earth-JEPA rows closing the table; best per column in **bold** among adapted methods.

	Method	24 h				72 h			
		U850 ($m\,s^{-1}$)	T850 (K)	Q700 ($g\,kg^{-1}$)	MSLP (hPa)	U850 ($m\,s^{-1}$)	T850 (K)	Q700 ($g\,kg^{-1}$)	MSLP (hPa)
	Persistence	5.540	3.051	1.650	5.980	7.206	4.328	2.074	8.856
	Climatology	5.684	3.444	1.600	7.241	5.687	3.442	1.600	7.247
Transfer	WeatherGFM	5.183	2.912	1.555	5.635	6.732	4.226	1.966	8.359
Transfer	ClimaX	3.747	2.047	1.082	3.892	6.743	3.812	1.756	8.682
Transfer	Earthformer	3.667	2.019	1.071	3.816	6.426	3.538	1.832	8.826
E-JEPA	E-JEPA-50M (frozen)	3.440	2.063	0.961	3.477	5.600	3.455	1.601	6.888
E-JEPA	E-JEPA-50M (fine-tuned)	2.667	1.687	0.833	**2.181**	4.607	2.716	1.433	4.807
E-JEPA	E-JEPA (frozen)	3.191	1.924	0.931	3.035	5.244	3.077	1.532	5.884
E-JEPA	E-JEPA (fine-tuned)	**2.613**	**1.651**	**0.820**	2.185	**4.390**	**2.595**	**1.398**	**4.639**

E.2 Representative surface-pressure variables on EarthCastBench

The main paper reports mean sea-level pressure and surface pressure as the two representative surface variables. Fig. 9 shows one illustrative pressure case.

Table 14: MAE (lower is better) at 24 h and 72 h on the global 1.5° grid (121 × 240), using 732 test initializations at 00/12 UTC throughout 2020 with spherical-area weighting. Rows are grouped as in Table 3, with Earth-JEPA rows closing the table; best per column in **bold** among adapted methods.

	Method	24 h				72 h			
		U850 ($m\,s^{-1}$)	T850 (K)	Q700 ($g\,kg^{-1}$)	MSLP (hPa)	U850 ($m\,s^{-1}$)	T850 (K)	Q700 ($g\,kg^{-1}$)	MSLP (hPa)
	Persistence	3.803	1.973	1.123	3.557	5.181	2.939	1.468	5.508
	Climatology	4.143	2.440	1.201	4.573	4.145	2.439	1.200	4.577
Transfer	WeatherGFM	3.616	1.926	1.086	3.453	4.861	2.888	1.404	5.241
Transfer	ClimaX	2.714	1.447	0.770	2.605	4.767	2.656	1.291	5.681
Transfer	Earthformer	2.648	1.423	0.763	2.555	4.498	2.466	1.321	5.581
E-JEPA	E-JEPA-50M (frozen)	2.515	1.465	0.684	2.350	4.131	2.493	1.156	4.623
E-JEPA	E-JEPA-50M (fine-tuned)	1.960	1.203	0.587	**1.457**	3.346	1.944	1.010	3.116
E-JEPA	E-JEPA (frozen)	2.359	1.364	0.659	2.042	3.932	2.208	1.091	3.929
E-JEPA	E-JEPA (fine-tuned)	**1.923**	**1.180**	**0.576**	1.459	**3.188**	**1.862**	**0.984**	**3.026**

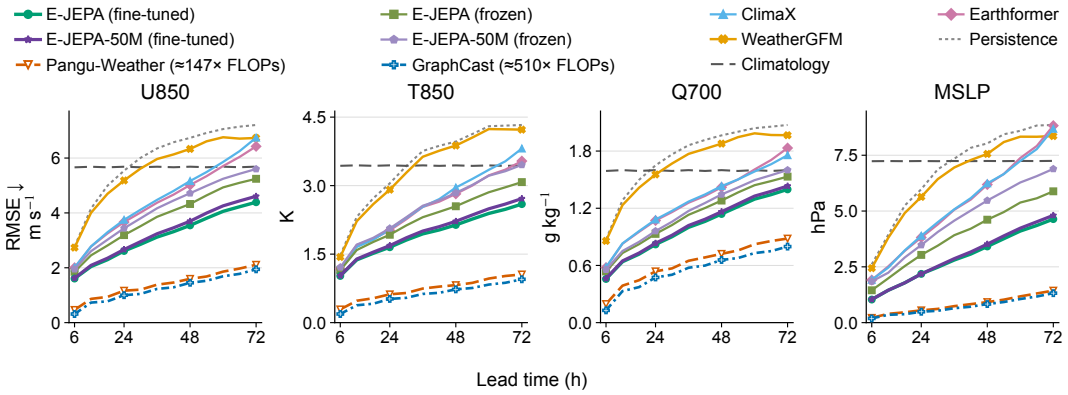

Figure 7: RMSE (↓) from 6 h to 72 h for U850, T850, Q700 and MSLP, the curve-level counterpart of the ACC curves in Fig. 5 under the protocol of Table 13.

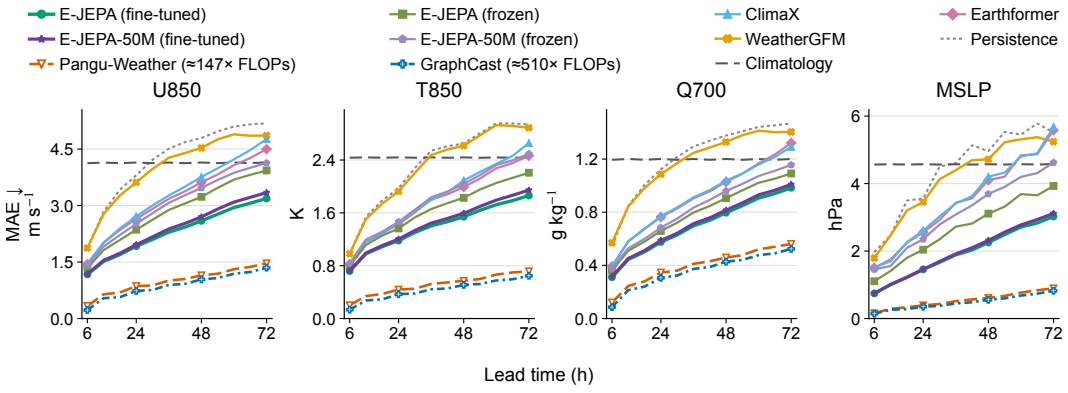

Figure 8: MAE (↓) from 6 h to 72 h for U850, T850, Q700 and MSLP under the protocol of Table 14.

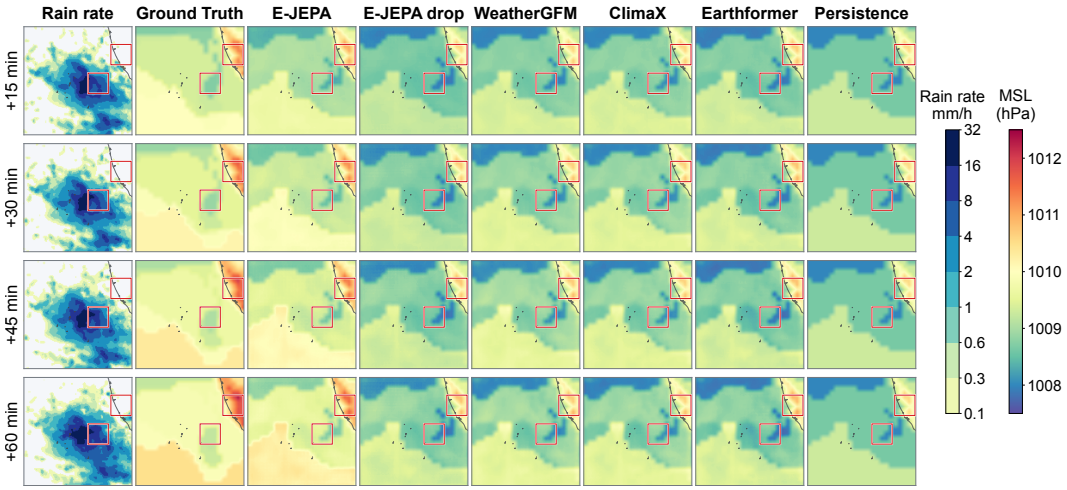

Figure 9: Mean sea-level pressure (MSL) nowcasts for a strong-precipitation test case initialized at 13:52:29 UTC on 2025-08-01 over a region west of India. Rows give the +15/+30/+45/+60 min leads; columns give, from left, the IMERG rain rate at each verification time (Huffman et al., 2020), the MSL ground truth, and the forecasts of Earth-JEPA, Earth-JEPA (Drop-ERA5), WeatherGFM, ClimaX, Earthformer, and Persistence. The rain panels are context at verification time—neither model inputs nor forecasts—and all MSL panels share the 1007.5–1012.5 hPa scale. The two red 24×24-pixel regions (upper right and middle right; identical across methods, leads, and rain panels) were selected to illustrate local forecast structure. This single case is illustrative only.