FROM DISCRETE FULL-FIELD LABELS TO ANY-TIME WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Full-field labels for data-driven weather forecasting are typically available only at discrete lead times, which ties the learned prediction mappings to predefined forecast intervals, while many weather processes and downstream applications require weather states at arbitrary times. To enable any-time prediction under this discrete supervision, we propose Physics-Supervised Any-Time Forecasting (PSAF). PSAF recomposes internal representation changes formed during fixed-endpoint forecasting into latent states for requested times, and maps them to weather fields through local decoding aligned with the forecast semantics of neighboring supervised lead times. To supervise intermediate times without labels, we use diagnostic-space physical surrogates as training targets. On WeatherBench 2, PSAF reduces the mean RMSE across 1–6 h lead times by 39.7% relative to endpoint field interpolation and by 10.1% relative to the strongest learned endpoint-reconstruction baseline. At intermediate lead times, PSAF outperforms the compared continuous-time methods on independent RTMA-RU analyses and downstream wind-power forecasting. Physics Surrogate Supervision further reduces the geostrophic and kinetic-energy target errors by 25.6% and 7.2%, while also improving physical diagnostics not included in the training objective.

1 INTRODUCTION

Data-driven weather forecasting has advanced substantially. Models such as FourCastNet, Pangu-Weather, GraphCast, FengWu, and FuXi have demonstrated competitive skill in global forecasting (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2025; 2023), while foundation models such as Aurora have extended these capabilities to a wider range of Earth-system tasks (Bodnar et al., 2025). Most mainstream forecasting models, however, learn from supervision available only at discrete lead times, so their predictions are typically organized on the same temporal grid. Such temporal discretization can be limiting because rapidly evolving weather processes and time-sensitive downstream applications such as wind power call for weather states at more flexible, finer-grained query times. We therefore ask: *How can a model predict global weather states at arbitrary query times given full-field labels only at discrete lead times?*

Existing approaches increase temporal flexibility through three main routes. Multi-lead training, represented by Stormer, broadens the available discrete forecast spans but does not directly support arbitrary-time queries (Nguyen et al., 2024). Continuous-time modeling, as in Continuous Ensemble Forecasting, enables forecasts at arbitrary lead times through a generative ensemble framework (Andrae et al., 2025), while endpoint-based temporal reconstruction, exemplified by ModAFNO, recovers intermediate fields between forecast endpoints (Leinonen et al., 2024). Our approach instead constructs query states from the internal representations already produced during fixed-endpoint forecasting.

Realizing this idea poses three challenges. First, these representations are trained for the forecast endpoint and do not explicitly encode intermediate query times. Second, intermediate query states must connect the distinct forecast semantics of their two surrounding supervised anchors without abrupt transitions. Third, labels are unavailable at intermediate times. We address these challenges with Physics-Supervised Any-Time Forecasting (PSAF), in which time–sample-conditioned Latent Recomposition constructs a query state from latent changes arising during its fixed 6-hour forecast, and a Local Decoder MoE decodes it using the Decoders corresponding to those anchors. For

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.

intermediate times without labels, Physics Surrogate Supervision constructs diagnostic-space targets based on geostrophic residuals and horizontal kinetic energy, while Continuity Regularization stabilizes how intermediate queries connect to the discrete anchors.

At supervised lead times on WeatherBench 2 (Rasp et al., 2024), PSAF achieves the lowest 1–6 h mean RMSE, with reductions of 39.7% and 10.1% relative to endpoint field interpolation and the best-performing learned endpoint-reconstruction baseline. At intermediate times, PSAF achieves leading results over the compared continuous-time methods in both independent RTMA-RU validation and downstream wind-power forecasting, while its weather fields show no additional smoothing or spurious high-frequency content. Physical-consistency evaluations show 25.6% and 7.2% reductions in geostrophic and kinetic-energy target errors, alongside improvements in diagnostics not included in the training objective. Mechanism analyses reveal distinct patterns of latent contribution and routing, while ablations confirm the contributions of the key components.

Our main contributions are as follows:

- We introduce PSAF for any-time weather forecasting under discrete full-field supervision, while leaving its fixed 6-hour forecasting path unchanged.
- We construct and decode query states from latent changes produced during fixed-endpoint forecasting, while physical diagnostics provide surrogate supervision at intermediate times without direct labels.
- We evaluate PSAF's forecast accuracy at supervised lead times and its effectiveness at intermediate times, with further analyses of physical consistency, query behavior, and component contributions.

2 Related Work

2.1 Fixed-Lead and Multi-Lead Weather Forecasting

Most data-driven weather models operate at discrete lead times. FourCastNet, GraphCast, FengWu, and FuXi advance the weather state at fixed intervals and extend forecasts through autoregressive rollout (Pathak et al., 2022; Lam et al., 2023; Chen et al., 2025; 2023). To increase lead-time flexibility, Pangu-Weather combines models with different prediction intervals, Stormer trains a single model using randomized dynamics intervals, and ARROW adaptively selects rollout step sizes according to the current state (Bi et al., 2023; Nguyen et al., 2024; Tian et al., 2026). These methods broaden the available forecast spans, but training on discrete lead times still ties their outputs to trained intervals and rollout combinations rather than arbitrary query times.

2.2 Continuous-Time Modeling and Temporal Refinement

One line of work models continuous or fine-grained temporal evolution from the initial state. Neural ODEs parameterize state derivatives to obtain continuous trajectories (Chen et al., 2018). ClimODE models weather transport as a continuous-time dynamical process, WeatherGFT combines fine-time-step physical evolution with data-driven correction, and CEF uses diffusion models to generate temporally consistent ensemble trajectories (Verma et al., 2024; Xu et al., 2024; Andrae et al., 2025). These approaches produce continuous or finer-grained weather outputs by explicitly modeling temporal evolution from the initial state.

Another line of work recovers intermediate weather states from forecast endpoints. ModAFNO uses a target-time-conditioned frequency-domain operator to reconstruct intermediate fields, while HourGlass performs probabilistic temporal downscaling between adjacent forecast times (Leinonen et al., 2024; Ingstad et al., 2026). PSAF instead constructs query states from latent representations generated while producing its fixed 6-hour forecast and directly outputs the target-time field within a single forecast window.

2.3 Physics-Informed Modeling in Weather Forecasting

Physics-informed learning incorporates physical knowledge through differential equations, conservation relations, or residual-based objectives (Raissi et al., 2019; Karniadakis et al., 2021). In

weather and climate forecasting, NeuralGCM couples a differentiable dynamical core with learned parameterizations, ClimODE uses value-conserving continuous-time dynamics, PASSAT combines spherical PDE solvers with graph learning, WeatherGFT embeds fine-time-step PDE evolution with learned correction, and WeatherODE introduces physics-driven ODE dynamics (Kochkov et al., 2024; Verma et al., 2024; Zheng et al., 2025; Xu et al., 2024; Liu et al., 2024). Across this literature, physics enters as dynamics, structural priors, or optimization constraints.

In PSAF, physical diagnostics instead provide surrogate supervision when labels are unavailable at intermediate times, rather than serving solely as consistency constraints.

3 METHOD

3.1 PROBLEM DEFINITION

Let $X_t \in \mathbb{R}^{69\times H\times W}$ denote the global weather field at time t. The model takes two weather states separated by 6 h as input, using historical states for the initial window and updating the pair through 6-hour autoregressive rollout for subsequent windows:

$$\mathcal{X}_t = (X_{t-6}, X_t). \tag{1}$$

Given any query lead time $\tau \in [0, 6]$, the model directly predicts the corresponding weather state within this 6-hour forecast window:

$$\widehat{X}_{t+\tau} = F(\mathcal{X}_t, \tau). \tag{2}$$

Full-field labels are available only at integer lead times $k \in \{0, 1, \ldots, 6\}$, denoted $Y_k = X_{t+k}$; integer queries therefore receive direct supervision, whereas fractional queries are unlabeled.

3.2 FRAMEWORK OVERVIEW AND FIXED 6-HOUR FORECASTING

PSAF performs any-time forecasting within a 6-hour window while preserving its fixed 6-hour forecasting path (Figure 1). Time–sample-conditioned Latent Recomposition forms each query latent state from internal representations generated along this path, after which Local Decoder MoE maps it to the target-time weather field. Training combines integer supervision, Physics Surrogate Supervision, and Continuity Regularization.

To produce its fixed 6-hour forecast, PSAF uses an architecture based on FengWu (Chen et al., 2025), with a multi-variable grouped Encoder, $L = 12$ Swin Transformer blocks, and a Decoder. Given $\mathcal{X}_t$, the Encoder produces the initial latent state z_0 and multi-scale skip features $\mathcal{S}$, after which the latent state passes through the blocks:

$$E(\mathcal{X}_t) = (z_0, \mathcal{S}), \qquad z_i = P_i(z_{i-1}), \quad i = 1, \ldots, L. \tag{3}$$

The original Decoder maps the terminal state z_L to the fixed 6-hour forecast:

$$\widehat{X}_{t+6} = D(z_L, \mathcal{S}). \tag{4}$$

3.3 TIME–SAMPLE-CONDITIONED LATENT RECOMPOSITION

During fixed 6-hour forecasting, successive blocks transform z_0 into z_L, with the latent change produced by block i defined as

$$C_i = z_i - z_{i-1}, \qquad z_L = z_0 + \sum_{i=1}^{L} C_i. \tag{5}$$

Because these transformations remain frozen during any-time training, each C_i retains the evolution semantics learned by block i in the original 6-hour forecasting process. We do not associate network depth with physical time; instead, we introduce $\alpha_i(\tau)$ as the cumulative adoption ratio of C_i in the query state.

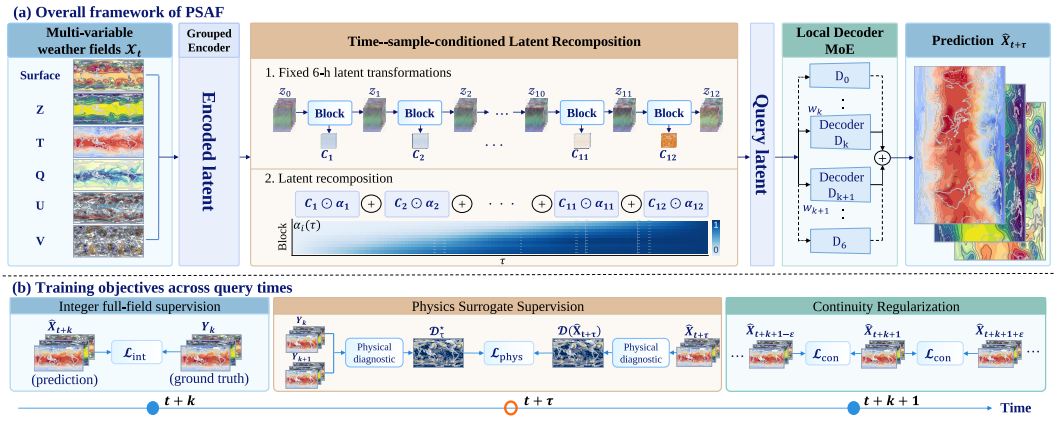


Figure 1: Overview of PSAF. (a) Any-time forecasting through Latent Recomposition and Local Decoder MoE. (b) Training objectives across query times.

To condition the adoption ratios on query time, sample-specific representations, and block position, we introduce a time–sample-conditioned controller. It first summarizes the block input state z_{i-1} and latent change C_i:

$$\begin{aligned} h_i^s &= W_s\ \mathrm{GAP}(\mathrm{LN}_s(z_{i-1}))\,, \\ h_i^c &= W_c\ \mathrm{GAP}(\mathrm{LN}_c(C_i))\,. \end{aligned} \tag{6}$$

The summaries are combined with the query-time encoding $e(\tau)$ and block embedding ℓ_i:

$$\begin{aligned} \beta_i(\tau) &= G([h_i^s, h_i^c, e(\tau), \ell_i])\,, \\ d_i(\tau) &= H_i(\beta_i(\tau))\,. \end{aligned} \tag{7}$$

Here G is shared across blocks and extracts a low-dimensional conditioning representation, while the block-specific channel expander H_i maps it to a modulation $d_i(\tau)$ matching the latent channels.

With normalized query time $s = \tau/6$, the cumulative adoption ratio is

$$\alpha_i(\tau) = s + \eta s(1-s)\tanh d_i(\tau), \qquad \eta = 2. \tag{8}$$

This parameterization fixes $\alpha_i(0) = 0$ and $\alpha_i(6) = 1$. Inside the interval, the block-wise responses remain query- and sample-dependent, without imposed monotonicity or temporal ordering. Using z_i for the latent state after block i and $z(\tau)$ for the recomposed state at query time τ, the final query state is

$$z(\tau) = z_0 + \sum_{i=1}^{L} \alpha_i(\tau) \odot C_i. \tag{9}$$

3.4 Local Decoder MoE

Local Decoder MoE maps the query state $z(\tau)$ to a full weather field using seven integer-time Decoders $D_0, \ldots, D_6$. D_0 reconstructs the current state, $D_1, \ldots, D_5$ correspond to the intermediate integer lead times, and D_6 is the original fixed 6-hour Decoder. All Decoders use the multi-scale skip features $\mathcal{S}$ from the Encoder.

For an integer query $k \in \{0, \ldots, 6\}$, hard routing directly selects the corresponding Decoder:

$$\widehat{X}_{t+k} = D_k(z(k), \mathcal{S}). \tag{10}$$

For a fractional query $\tau = k + \delta$, where $\delta \in (0, 1)$, the Router conditions on the query state and time to produce seven routing logits:

$$\begin{aligned} q(\tau) &= W_r\ \mathrm{GAP}(\mathrm{LN}_r(z(\tau)))\,, \\ r(\tau) &= R([e(\tau), q(\tau)])\,. \end{aligned} \tag{11}$$

Only the logits associated with the two endpoint Decoders D_k and D_{k+1} are jointly normalized:

$$[w_k, w_{k+1}] = \mathrm{softmax}([r_k(\tau), r_{k+1}(\tau)])\,. \tag{12}$$

The corresponding Decoder outputs are then combined in weather-field space:

$$\widehat{X}_{t+\tau} = w_k D_k(z(\tau), \mathcal{S}) + w_{k+1} D_{k+1}(z(\tau), \mathcal{S}). \tag{13}$$

This local mixture ties each fractional query to the forecast semantics of its two endpoint Decoders without globally mixing all seven.

3.5 Training Objectives under Discrete Supervision

The training objective combines integer full-field supervision, Physics Surrogate Supervision, and Continuity Regularization:

$$\mathcal{L} = \mathcal{L}_{\text{integer}} + \mathcal{L}_{\text{physics}} + \mathcal{L}_{\text{continuity}}. \tag{14}$$

Integer full-field supervision. Integer queries have direct full-field labels and define the discrete anchors. Any-time training supervises $k = 0, \ldots, 5$: $k = 0$ reconstructs the current state, while $k = 1, \ldots, 5$ use the corresponding ground-truth fields:

$$\mathcal{L}_{\text{integer}} = \frac{1}{6} \sum_{k=0}^{5} \ell_{\text{wMSE}}\Big(\widehat{X}_{t+k}, Y_k\Big)\,. \tag{15}$$

Here ℓ_{wMSE} is the variable-weighted mean squared error.

Physics Surrogate Supervision. For a fractional query $\tau = k + \delta$, where $\delta \in (0, 1)$, no corresponding field label is available. We select two physical diagnostics whose dominant variation is typically slow over adjacent hours and therefore use local linear interpolation between their values at Y_k and Y_{k+1} to construct fractional surrogate targets in diagnostic space.

Diagnostic-space interpolation is not equivalent to field-space interpolation. For example, because horizontal specific kinetic energy is quadratic in the wind field, evaluating it after interpolating the endpoint wind fields systematically underestimates the interpolation of their kinetic energies. We use the geostrophic residual $\mathcal{G}$, which measures the departure of horizontal wind from geostrophic balance, and the horizontal specific kinetic energy $\mathcal{K}$, which characterizes the nonlinear wind-energy state. For either $\mathcal{D} \in \{\mathcal{G}, \mathcal{K}\}$, the target and loss are

$$\begin{aligned} \mathcal{D}^\star_\tau &= (1-\delta)\mathcal{D}(Y_k) + \delta\mathcal{D}(Y_{k+1}), \\ \mathcal{L}_{\text{physics}} &= \sum_{\mathcal{D}\in\{\mathcal{G},\mathcal{K}\}} \lambda_{\mathcal{D}} \left\langle \mathrm{SmoothL1}\Big(\mathcal{D}(\widehat{X}_{t+\tau}), \mathcal{D}^\star_\tau\Big) \right\rangle_{\cos\phi}. \end{aligned} \tag{16}$$

Here $\langle\cdot\rangle_{\cos\phi}$ denotes the cosine-latitude weighted spatial average. Both diagnostics are computed from physical-unit weather fields, and their normalized forms enter the loss. For the geostrophic term, the target is constructed from the diagnostic residuals of the adjacent ground-truth states rather than set to zero.

Continuity Regularization. Continuity Regularization governs how fractional queries join the discrete integer anchors and suppresses isolated anomalous fluctuations through boundary constraints and temporal regularization:

$$\mathcal{L}_{\text{continuity}} = \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{br}}\mathcal{L}_{\text{boundary-router}} + \lambda_{\text{bv}}\mathcal{L}_{\text{boundary-value}} + \lambda_{\text{time}}\mathcal{L}_{\text{time}}. \tag{17}$$

The Router anchor preserves the lead-time identity of the integer Decoders. Near an anchor, the boundary-router term directs local routing toward the corresponding endpoint Decoder, while the boundary-value term connects the fractional output to the neighboring integer prediction through a stop-gradient without altering the anchor state. Within an interval, temporal regularization constrains the second-order temporal differences of the cumulative adoption ratios α and Router weights w.

3.6 TWO-STAGE TRAINING

Training proceeds in two stages. Stage 1 trains PSAF's fixed 6-hour forecasting component, producing its block-wise latent representations. In Stage 2, the Encoder, 12 Swin Transformer blocks, and D_6 remain frozen, while $D_0, \ldots, D_5$ are initialized from D_6. Time–sample-conditioned Latent Recomposition and Local Decoder MoE are then trained with the objectives defined above. This stage learns any-time queries while leaving the original 6-hour forecasting path unchanged.

4 EXPERIMENTS

4.1 SETUP

Data and task. We use hourly ERA5 data from WeatherBench 2 at 1.5° resolution (Rasp et al., 2024). Each state contains four surface variables $(\mathrm{U10}, \mathrm{V10}, \mathrm{T2m}, \mathrm{MSL})$ and five upper-air variables $(\mathrm{Z}, \mathrm{Q}, \mathrm{U}, \mathrm{V}, \mathrm{T})$ on 13 pressure levels, giving 69 channels. Stage 1 uses 1979–2015 and Stage 2 uses 2006–2015 for training; both use 2016–2017 for validation and held-out years for evaluation. All methods within each comparison share the same initialization times.

Training. Stage 1 trains for 70 epochs on four NVIDIA RTX PRO 6000 GPUs, followed by 2000 replay fine-tuning steps; Stage 2 trains for 50 epochs on three such GPUs. Both use FP32 and a per-device batch size of 2. Full data splits, evaluation protocols, and training configurations are given in Appendix A.1.

Overview. The experiments assess PSAF's accuracy at integer lead times and its effectiveness at unlabeled fractional times, then examine physical consistency, query behavior, and component contributions.

4.2 FORECAST ACCURACY

All models use the same ERA5 targets, variables, lead times, and latitude weights and are rolled out every 6 h to 240 h. We denote the base and replay-fine-tuned models as PSAF-Base and PSAF-FT. Against OneForecast (Gao et al., 2025), the best-performing external learned baseline, PSAF-FT reduces normalized RMSE by 4.0%, 8.4%, and 13.9% at 96, 168, and 240 h (Table 1). Figure 2 further shows growing wind-field advantages with lead time while retaining competitive error and correlation for Z500 and T850.

Table 1: Fixed-lead forecasting under a unified protocol. RMSE is normalized per channel by the training-set standard deviation, cosine-latitude weighted, and macro-averaged over 69 channels.

	Normalized RMSE ↓					ACC ↑				
Model	6 h	24 h	96 h	168 h	240 h	6 h	24 h	96 h	168 h	240 h
Pangu (retrained)	.0885	.1636	.3335	.4915	.6001	.9854	.9534	.8183	.5935	.3779
FuXi (retrained)	.1018	.1657	.3691	.5494	.6615	.9800	.9522	.7893	.5323	.3243
FourCastNet (retrained)	.0552	.1313	.3089	.4943	.6334	.9944	.9706	.8508	.6203	.3732
OneForecast (released)	.0546	.1223	.2720	.4458	.5910	.9944	.9742	.8846	.6940	.4553
PSAF-Base	**.0543**	**.1184**	.2638	.4390	.5869	**.9944**	**.9760**	**.8928**	**.7062**	.4695
PSAF-FT	.0552	.1190	**.2610**	**.4082**	**.5090**	.9942	.9757	.8898	.7054	**.4810**

Table 2 compares PSAF with field interpolation and three learned endpoint-reconstruction methods—ModAFNO, UPR-Net, and AMT—using the same current analysis and fixed 6-hour endpoint. We also train the same Stage 1 architecture at a fixed 1-hour interval, denoted PSAF (fixed-1h). PSAF reduces the mean 1–6-hour RMSE by 39.7% over field interpolation and by 10.1% over AMT. PSAF (fixed-1h) is more accurate from 1 to 3 h but requires fixed-step recursion and cannot directly answer fractional queries.

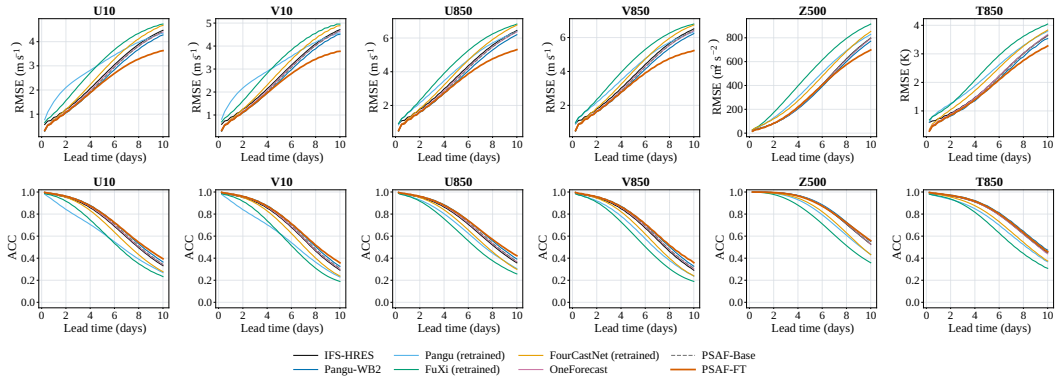


Figure 2: Variable-wise fixed-lead forecasting performance. Pangu-WB2 and IFS-HRES use forecast fields provided by WeatherBench 2; other baselines use released or retrained checkpoints.

Table 2: Integer queries within a 6-hour window. All methods except PSAF (fixed-1h) share the same fixed 6-hour endpoint prediction.

Method (RMSE ↓)	1 h	2 h	3 h	4 h	5 h	6 h	Mean 1–6 h
6h field interpolation	.05055	.08066	.09143	.08528	.06731	**.05517**	.07173
ModAFNO (Leinonen et al., 2024)	.03060	.04540	.05277	.05472	.05443	**.05517**	.04885
UPR-Net (Jin et al., 2025)	.03311	.05037	.05810	.05865	.05561	**.05517**	.05183
AMT (Li et al., 2023)	.02922	.04419	.05196	.05421	.05383	**.05517**	.04810
PSAF (fixed-1h)	**.01459**	**.02753**	**.03906**	.04941	.06042	.07306	.04401
PSAF	.02406	.03637	.04399	**.04860**	**.05136**	**.05517**	**.04326**

4.3 Validation at Intermediate Times

Following the integer-time comparison with endpoint-reconstruction methods, we evaluate fractional-time accuracy at $0.5, 1.5, \ldots, 5.5$ h against continuous-time forecasting methods using NCEP RTMA-RU near-surface analyses over the contiguous United States, available every 15 min at roughly 2.5 km resolution. We bilinearly sample RTMA-RU at forecast-grid centers and compute cosine-latitude-weighted MAE for T2m, U10, V10, and the wind vector. PSAF ranks first or second at both its native $1.5°$ resolution and after bilinear downsampling to $5.625°$ (Table 3), reducing the four MAEs against CEF by 5.9%–9.3% at matched resolution.

To assess downstream value, we bilinearly sample each weather forecast at the Kelmarsh and Penmanshiel wind-farm coordinates and use the same TFT configuration to predict 6 h of 10-minute power from the past 12 h of power and future weather at 10-minute resolution over the same 6-hour horizon (Lim et al., 2021). All future-weather inputs outperform History-only (Table 4). At matched $5.625°$ resolution, PSAF (bilinear) performs on par with CEF and improves over ClimODE; at native $1.5°$ resolution, it reduces NMAE and NRMSE against CEF by 14.84% and 13.78%, reaches 18.79% Persistence skill, and attains the best Ramp F1 and Ramp MAE.

Table 3: Fractional-time validation against RTMA-RU. Values are cosine-latitude-weighted MAE; T2m is reported in K and wind metrics in m s^{-1}. PSAF's $5.625°$ outputs are downsampled from its native $1.5°$ grid.

Method	Grid	T2m ↓	U10 ↓	V10 ↓	Wind vector ↓
ClimODE (Verma et al., 2024)	$5.625°$	3.0511	1.2667	1.3235	2.0333
CEF (Andrae et al., 2025)	$5.625°$	1.5462	1.1897	1.2286	1.8959
PSAF (conservative)	$5.625°$	1.6306	1.2729	1.3542	2.0636
PSAF (bilinear)	$5.625°$	1.4550	**1.0789**	1.1199	**1.7256**
PSAF (native)	$1.5°$	**1.3023**	1.1031	**1.1066**	1.7358

Table 4: 10-minute wind-power forecasting. Values are the mean and standard deviation over two wind farms and three random seeds.

Future weather input	Grid	NMAE ↓	NRMSE ↓	Persistence skill ↑
History-only	–	$.1176 \pm .0014$	$.1766 \pm .0021$	$-2.60 \pm 1.25\%$
ClimODE (Verma et al., 2024)	5.625°	$.1157 \pm .0045$	$.1703 \pm .0043$	$-.97 \pm 3.95\%$
CEF (Andrae et al., 2025)	5.625°	$.1093 \pm .0034$	$.1612 \pm .0036$	$+4.64 \pm 2.95\%$
PSAF (conservative)	5.625°	$.1096 \pm .0043$	$.1610 \pm .0033$	$+4.32 \pm 3.76\%$
PSAF (bilinear)	5.625°	$.1091 \pm .0045$	$.1605 \pm .0038$	$+4.82 \pm 3.92\%$
PSAF (native)	1.5°	$\mathbf{.0931 \pm .0037}$	$\mathbf{.1390 \pm .0060}$	$\mathbf{+18.79 \pm 3.25\%}$

Finally, we evaluate spatial structure using total variation, Laplacian RMS, high-frequency energy, and near-Nyquist energy. PSAF is closest to ERA5 on all four statistics at integer lead times, while fractional-midpoint ratios to linear interpolation between integer prediction anchors are .9970–1.0014. Figure 3 shows coherent evolution of large-scale systems, major isolines, and wind structures without obvious visual artifacts. Relative to the anchor reference, fractional fields show neither additional smoothing nor anomalous high-frequency growth.

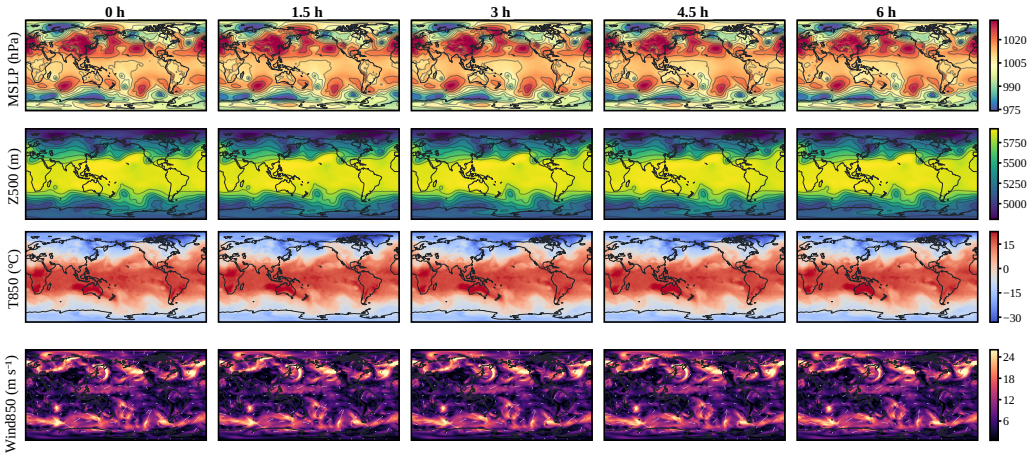


Figure 3: Direct queries across one 0–6-hour forecast window. Native-grid outputs use variable-specific color scales; wind shading and arrows denote speed and direction.

4.4 Physical Consistency

Table 5 shows that Physics Surrogate Supervision reduces the geostrophic and kinetic-energy target errors by 25.6% and 7.2%, respectively. Four held-out diagnostics improve by 3.8%–6.4%, demonstrating broader improvements in the physical consistency of the predicted fields. Similar gains appear for fractional queries and integer anchors separately.

Table 5: Physics Surrogate Supervision ablation with four held-out diagnostics.

Type	Metric	w/o Physics Surrogate	PSAF	Reduction
Physics target	Geostrophic target error	.01848	**.01376**	**25.6%**
	KE target error	.03115	**.02891**	**7.2%**
Held-out diagnostic	Thermal-wind target MAE ($\mathrm{m\,s^{-1}}$)	2.805	**2.699**	**3.8%**
	Vorticity target MAE ($10^{-5}\ \mathrm{s^{-1}}$)	.2862	**.2679**	**6.4%**
	Divergence target MAE ($10^{-5}\ \mathrm{s^{-1}}$)	.2760	**.2641**	**4.3%**
	Column-water target MAE ($\mathrm{kg\,m^{-2}}$)	.09123	**.08747**	**4.1%**

4.5 Query Behavior and Ablations

Figure 4 shows block-dependent cumulative adoption ratios that deviate from the uniform schedule (mean $|\alpha_i - s| = .179$): shallow-layer contributions strengthen earlier and deeper-layer contributions later. Router weights shift from the left to the right expert within each interval and approach the corresponding expert near integer anchors. Relative to w/o Continuity Regularization, PSAF reduces the boundary jump at $\pm.01$ h from integer anchors by 95.8% and the mean and maximum second-order differences by 18.4% and 50.1%; the larger maximum reduction indicates that the regularization mainly suppresses isolated anomalous fluctuations.

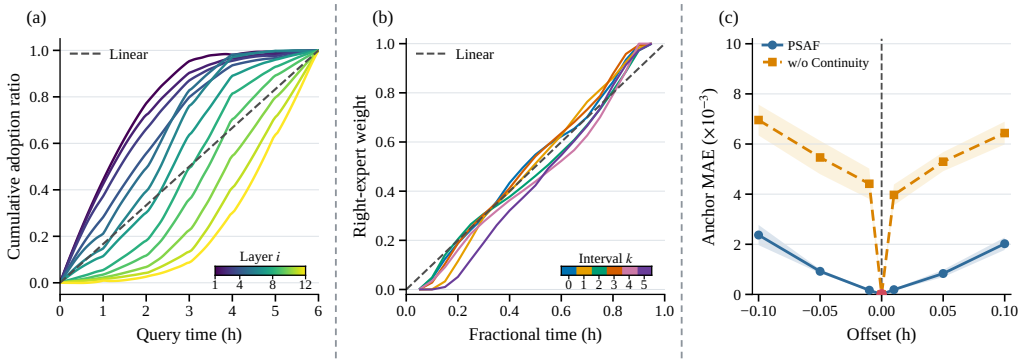


Figure 4: Query behavior of PSAF. (a) Block-wise cumulative adoption ratios. (b) Right-expert weights across local intervals. (c) Boundary MAE near integer anchors; shading denotes standard deviation.

Table 6 isolates the two architectural components. The w/o Latent Recomposition variant uses z_L for all queries while retaining Local Decoder MoE; its 1-hour RMSE rises by 12.9%, with the gap narrowing toward the 6-hour endpoint. The w/o Local Decoder MoE variant replaces the local experts and Router with a shared Decoder; its 1-hour RMSE rises by 10.8% and remains worse through 5 h.

Table 6: Ablations of Latent Recomposition and Local Decoder MoE, measured by normalized RMSE.

Model (RMSE ↓)	1 h	2 h	3 h	4 h	5 h	6 h
w/o Latent Recomposition	.02716	.03963	.04640	.04968	.05148	**.05517**
w/o Local Decoder MoE	.02665	.03828	.04518	.04929	.05241	**.05517**
PSAF	**.02406**	**.03637**	**.04399**	**.04860**	**.05136**	**.05517**

5 Conclusion

We introduced PSAF for direct weather prediction between supervised lead times. PSAF combines time–sample-conditioned Latent Recomposition and Local Decoder MoE to recompose internal representation changes produced during its fixed 6-hour forecast and decode the resulting states into weather fields, while Physics Surrogate Supervision and Continuity Regularization guide learning at intermediate times without labels. Experiments show improved accuracy at supervised lead times and strong performance at intermediate times on independent RTMA-RU analyses and downstream wind-power forecasting. Physical-consistency evaluations, query-behavior analyses, and ablations further support the proposed supervision and architectural designs. Overall, PSAF enables any-time weather forecasting from discrete full-field labels.

AI USE STATEMENT

The authors used large language models (LLMs) as general-purpose assistive tools during the preparation of this work. Specifically, LLMs were used to assist with code editing, organization of experiment scripts, and language editing of the manuscript. LLMs were not used to generate research ideas, design the method, make experimental decisions, or produce or interpret results. All research questions, methodological design, experimental decisions, verification of results, and final claims were determined and reviewed by the authors, who take full responsibility for the entire content of this paper.

REPRODUCIBILITY STATEMENT

We provide the data splits, preprocessing and evaluation protocols, training configurations, and additional implementation details in the appendix. The model components, training objectives, and evaluation settings used in the reported experiments are described in the main paper and appendix. Additional materials are available at `https://anonymous.4open.science/r/PSAF-CCBC/`.

REFERENCES

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations*, 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*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

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*, 641:1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Kang Chen, Tao Han, Fenghua Ling, Junchao Gong, Lei Bai, Xinyu Wang, Jing-Jia Luo, Ben Fei, Wenlong Zhang, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment*, 6:518, 2025. doi: 10.1038/s43247-025-02502-y.

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*, 6:190, 2023. doi: 10.1038/s41612-023-00512-1.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, Jiahao Wu, Qing Li, Hui Xiong, and Xiaomeng Huang. OneForecast: A universal framework for global and regional weather forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18658–18697. PMLR, 2025. URL `https://proceedings.mlr.press/v267/gao25r.html`.

Magnus Sikora Ingstad, Mariana C. A. Clare, Olav Ersland, Vera Gahlen, Håvard Homleid Haugen, Ophélia Miralles, Even M. Nordhagen, Thomas N. Nipen, Ivar A. Seierstad, John Bjørnar Bremnes, Michael Maier-Gerber, Zied Ben Bouallègue, Harrison Cook, Christian Lessig, Gert Mertes, Cathal O'Brien, Florian Pinault, Ana Prieto Nemesio, and Matthew Chantry. HourGlass: A probabilistic data-driven temporal downscaler for global and regional weather forecasting, 2026. URL `https://arxiv.org/abs/2607.11457`.

Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, Cheul-Hee Hahm, and Zhao-Min Chen. UPR-Net: A unified pyramid recurrent network for video frame interpolation. *International Journal of Computer Vision*, 133(1):16–30, 2025. doi: 10.1007/s11263-024-02164-x.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Jussi Leinonen, Boris Bonev, Thorsten Kurth, and Yair Cohen. Modulated adaptive Fourier neural operators for temporal interpolation of weather forecasts, 2024. URL https://arxiv.org/abs/2410.18904.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Peiyuan Liu, Tian Zhou, Liang Sun, and Rong Jin. Mitigating time discretization challenges with WeatherODE: A sandwich physics-driven neural ODE for weather forecasting, 2024. URL https://arxiv.org/abs/2410.06560.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2196.

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

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

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

Jindong Tian, Yifei Ding, Ronghui Xu, Hao Miao, Chenjuan Guo, and Bin Yang. ARROW: An adaptive rollout and routing method for global weather forecasting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59fb5cf4683eaec4c56918a69637c9d2-Abstract-Conference.html.

Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In *International Conference on Learning Representations*, 2024.

Wanghan Xu, Fenghua Ling, Wenlong Zhang, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Generalizing weather forecast to fine-grained temporal scales via physics-AI hybrid modeling. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0734.

Jiaqi Zheng, Qing Ling, and Yerong Feng. Physics-assisted and topology-informed deep learning for weather prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7958–7966, 2025. doi: 10.24963/ijcai.2025/885.

A Appendix

A.1 Data, Training, and Evaluation Details

Data representation and splits. Each WeatherBench 2 state contains four surface variables $(\mathrm{U10}, \mathrm{V10}, \mathrm{T2m}, \mathrm{MSL})$ and five upper-air variables $(\mathrm{Z}, \mathrm{Q}, \mathrm{U}, \mathrm{V}, \mathrm{T})$ at 13 pressure levels: 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50 hPa, giving 69 channels in total. The global grid has size 121×240 at 1.5° resolution. Stage 1, including replay fine-tuning, uses 1979–2015 for training, while Stage 2 and PSAF (fixed-1h) use 2006–2015; all use 2016–2017 for validation. Forecast-accuracy and integer-query experiments are evaluated in 2020 with initialization every 12 h, while mechanism, boundary, spatial-structure, and physical-consistency analyses use 2021 with initialization every 60 h. Training, validation, and test years do not overlap.

Evaluation metrics and aggregation. All normalized error metrics are computed with cosine-latitude weighting and then macro-averaged equally over the 69 channels. Let $e_c(\phi, \lambda)$ denote the normalized error for channel c and $w(\phi) = \cos\phi$ the latitude weight. The channel-wise normalized RMSE is

$$\mathrm{RMSE}_c = \left(\frac{\sum_{\phi,\lambda} w(\phi) e_c(\phi,\lambda)^2}{\sum_{\phi,\lambda} w(\phi)} \right)^{1/2}, \tag{18}$$

while normalized MAE follows the same aggregation with absolute error in place of squared error. For ACC, let a_c^{pred} and a_c^{true} denote forecast and ground-truth anomalies relative to the same daily climatology:

$$\mathrm{ACC}_c = \frac{\sum_{\phi,\lambda} w(\phi) a_c^{\mathrm{pred}} a_c^{\mathrm{true}}}{\sqrt{\sum_{\phi,\lambda} w(\phi)(a_c^{\mathrm{pred}})^2} \sqrt{\sum_{\phi,\lambda} w(\phi)(a_c^{\mathrm{true}})^2}}. \tag{19}$$

Physical-unit RMSE in the forecast-accuracy experiments is computed after converting predictions back to physical units and is reported by variable. Within each comparison, all methods use identical initialization times, targets, variable ordering, and latitude weighting.

Computational setup. All training and evaluation for PSAF and the reference models are conducted on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs in FP32, with TF32 disabled. The main training configurations are summarized below.

Table 7: Training configurations.

Component	Training years	Budget	GPUs	Batch/GPU	Initial LR / weight decay
PSAF Stage 1	1979–2015	70 epochs	4	2	5×10^{-4} / 10^{-5}
Replay fine-tuning	1979–2015	2000 steps	4	1	5×10^{-6} / 10^{-5}
PSAF Stage 2	2006–2015	50 epochs	3	2	5×10^{-4} / 10^{-5}
PSAF (fixed-1h)	2006–2015	50 epochs	4	2	5×10^{-4} / 10^{-5}

Optimization details. PSAF Stage 1 (including replay fine-tuning) and PSAF (fixed-1h) use the same variable-weighted negative log-likelihood (NLL), with weight .1 for geopotential channels and 1 for all other channels; their two input states are separated by 6 h and 1 h, respectively. Stage 2 uses variable-weighted MSE for integer full-field supervision, with the same channel weights. All training stages use AdamW with weight decay 10^{-5} and cosine learning-rate annealing. Stage 1, replay fine-tuning, and PSAF (fixed-1h) use $\beta_1 = \beta_2 = .9$, with initial learning rates of 5×10^{-4} for Stage 1 and PSAF (fixed-1h) and 5×10^{-6} for replay fine-tuning. In Stage 2, the initial learning rates are 5×10^{-4} for the Latent Recomposition controller and D_0–D_5, and 10^{-3} for the Router. The random seed is 20260809. The coefficient of $\mathcal{L}_{\text{integer}}$ is fixed to 1, while the weights, Smooth-L_1 parameters, and warm-up schedules for Physics Surrogate Supervision and Continuity Regularization are given in Appendix A.3 and Appendix A.4, respectively.

Stage-2 query sampling. Each Stage-2 update evaluates the full integer branch together with one sampled fractional query. For the fractional branch, one of the six adjacent integer intervals is first sampled uniformly, after which the interval interior, left boundary band, or right boundary band is selected with probabilities $70\%/15\%/15\%$ and the query time is sampled uniformly within the selected region. Interior samples use $\delta \in [.10, .90]$. This stratified sampling increases the frequency of queries near integer anchors relative to uniform sampling over the full temporal range.

A.2 Implementation Details

Module dimensions and parameter counts. The latent channel dimension is $C_z = 768$. The state and increment summaries h_i^s and h_i^c are each 128-dimensional, the Fourier query-time encoding is 128-dimensional, and the block embedding is 32-dimensional, giving a 416-dimensional input to the shared mapping G. G produces the 64-dimensional conditioning representation β_i, which the block-specific channel expander H_i maps to the 768-dimensional modulation d_i. The Router projects the query latent-state summary to 128 dimensions and concatenates it with a 128-dimensional time encoding before producing seven routing logits. PSAF contains 176.4M parameters in total: 123.8M frozen parameters in the grouped Encoder, 12 Swin Transformer blocks, and D_6, and 52.6M trainable parameters, including 0.88M for the Latent Recomposition controller, 0.15M for the Router, and 51.55M for D_0–D_5.

Cumulative adoption ratio range. We set $\eta = 2$. Since $\tanh d_i(\tau) \in [-1, 1]$, the cumulative adoption ratio $\alpha_i(\tau)$ has a theoretical range of $[-1/8, 9/8]$ over the full query window; for example, its lower bound at 2 h is $-1/9$, while its upper bound at 4 h is $10/9$. This range contains the uniform linear schedule while allowing nonlinear, non-monotonic, and moderate out-of-range adoption within the interval.

A.3 Physics Surrogate Supervision: Implementation and Additional Results

Physical-unit conversion and channel indexing. Both diagnostics are computed from weather fields restored to physical units. Each normalized channel is first converted as $X_c^{\text{phys}} = \sigma_c X_c^{\text{norm}} + \mu_c$, where μ_c and σ_c are computed from the training data. The required variables and pressure levels are then selected using the channel ordering defined in Appendix A.1.

Geostrophic diagnostic. Let Φ denote geopotential. The geostrophic wind components are defined as

$$u_g = -\frac{1}{f}\frac{\partial \Phi}{\partial y}, \qquad v_g = \frac{1}{f}\frac{\partial \Phi}{\partial x}, \tag{20}$$

where $f(\phi) = 2\Omega \sin \phi$ is the Coriolis parameter. The normalized geostrophic residual is

$$\mathcal{G}(X) = \left[\frac{u - u_g}{S_g}, \frac{v - v_g}{S_g}\right], \qquad S_g = \sqrt{u^2 + v^2 + \varepsilon^2} + \sqrt{u_g^2 + v_g^2 + \varepsilon^2} + v_0, \tag{21}$$

where $\Omega = 7.2921 \times 10^{-5}\,\mathrm{s}^{-1}$, $R = 6{,}371{,}000\,\mathrm{m}$, $\varepsilon = 10^{-6}$, and $v_0 = 1\,\mathrm{m\,s}^{-1}$.

The longitude spacing on the global grid is $\Delta\lambda = 2\pi/W$. We use periodic centered differences in the longitudinal direction:

$$\left.\frac{\partial \Phi}{\partial x}\right|_{j,i} \approx \frac{\Phi_{j,i+1} - \Phi_{j,i-1}}{2R \cos \phi_j \, \Delta\lambda}. \tag{22}$$

In implementation, neighboring longitude points are accessed cyclically using `roll(-1)-roll(+1)`, which naturally enforces the periodic boundary. Latitude derivatives are non-periodic. The latitude grid decreases from $+90^\circ$ to -90° according to $\phi_j = \pi/2 - j\Delta\phi$, where $\Delta\phi = \pi/(H-1)$. Interior points use centered differences:

$$\left.\frac{\partial \Phi}{\partial y}\right|_{j,i} \approx \frac{\Phi_{j-1,i} - \Phi_{j+1,i}}{2R\Delta\phi}, \tag{23}$$

with the corresponding one-sided differences at the two poles.

The geostrophic term uses 700, 600, 500, 400, 300, 250, and 200 hPa and applies a latitude mask of $|\phi| \in [20^\circ, 75^\circ]$. Angular derivatives in both longitude and latitude are first computed on the full grid, after which the latitude mask is applied and the derivatives are converted to spherical physical gradients. This avoids altering the finite-difference stencil near the mask boundaries. The latitude range excludes the near-equatorial region, where the Coriolis parameter becomes small, as well as the polar region, where discretization effects are more pronounced. The selected pressure levels also reduce the influence of boundary-layer, topographic, near-surface, and upper-stratospheric scale differences. The geostrophic surrogate target is constructed from the diagnostic residuals of the adjacent integer ground-truth states:

$$\mathcal{G}^\star_\tau = (1-\delta)\mathcal{G}(Y_k) + \delta\mathcal{G}(Y_{k+1}), \qquad \mathcal{L}_{\text{geo}} = \left\langle \text{SmoothL1}(\mathcal{G}(\widehat{X}_{t+\tau}), \mathcal{G}^\star_\tau) \right\rangle_{\cos\phi}. \tag{24}$$

Kinetic-energy diagnostic. The horizontal specific kinetic energy is normalized for each fractional query as

$$\mathcal{K}(X) = \frac{u^2 + v^2}{2S_K}. \tag{25}$$

Here S_K is a stop-gradient local scale computed from the two ground-truth endpoints:

$$S_K = (1-\delta)\frac{u_k^2 + v_k^2}{2} + \delta\frac{u_{k+1}^2 + v_{k+1}^2}{2} + K_0, \qquad K_0 = 25\,\text{m}^2\,\text{s}^{-2}. \tag{26}$$

The endpoint winds (u_k, v_k) and (u_{k+1}, v_{k+1}) are taken from Y_k and Y_{k+1}, respectively. This local scale balances contributions from strong- and weak-wind regions: scaling with the endpoint kinetic energies reduces the dominance of large absolute errors in strong-wind regions, while K_0 prevents excessive amplification when the local kinetic energy is small. With S_K shared by the prediction and both endpoint states for the current query, the surrogate target and corresponding loss are

$$\mathcal{K}^\star_\tau = (1-\delta)\mathcal{K}(Y_k) + \delta\mathcal{K}(Y_{k+1}), \qquad \mathcal{L}_{\text{KE}} = \left\langle \text{SmoothL1}(\mathcal{K}(\widehat{X}_{t+\tau}), \mathcal{K}^\star_\tau) \right\rangle_{\cos\phi}. \tag{27}$$

The kinetic-energy term uses 700, 500, 300, and 200 hPa and is evaluated over the full global grid without the geostrophic latitude mask. Unlike the geostrophic diagnostic, it does not depend on the Coriolis parameter and is therefore unaffected by the near-equatorial singularity associated with division by f. Both the geostrophic and kinetic-energy terms use Smooth-L_1 with $\beta = .01$ and cosine-latitude-weighted spatial aggregation. The geostrophic term serves as the primary cross-variable dynamical constraint, while the kinetic-energy term provides a complementary nonlinear constraint on the wind-field state. Their loss weights are $\lambda_{\text{geo}} = .01$ and $\lambda_{\text{KE}} = .004$, with both linearly warmed up over the first 5% of the total optimization steps.

Evaluation protocol. Using the 2021 evaluation set described in Appendix A.1, we evaluate fractional queries at $\delta \in \{.25, .50, .75\}$ within each of the six adjacent hourly intervals. This gives 18 fractional queries per initialization and 2,628 forecasts in total. Integer-anchor evaluation uses the five interior integer lead times, giving 730 forecasts. The ablation without Physics Surrogate Supervision uses the same initialization times, random seed, training budget, and remaining objectives as PSAF, with only λ_{geo} and λ_{KE} set to zero.

The main-text results are micro-averaged over fractional queries and integer anchors. We therefore report the two query regimes separately below. Physics Surrogate Supervision improves all six reported diagnostic metrics at fractional times and also yields improvements at integer anchors, where no physical loss is directly applied.

Table 8: Physics Surrogate Supervision ablation on fractional queries.

Type	Metric	w/o Physics Surrogate	PSAF	Reduction
Training surrogate	Geostrophic target error	.01799	**.01320**	**26.6%**
	KE target error	.03082	**.02828**	**8.2%**
Held-out diagnostic	Thermal-wind target MAE ($\mathrm{m\,s^{-1}}$)	2.750	**2.631**	**4.3%**
	Vorticity target MAE ($10^{-5}\,\mathrm{s^{-1}}$)	.2825	**.2612**	**7.5%**
	Divergence target MAE ($10^{-5}\,\mathrm{s^{-1}}$)	.2699	**.2568**	**4.9%**
	Column-water target MAE ($\mathrm{kg\,m^{-2}}$)	.09026	**.08602**	**4.7%**

Table 9: Physics Surrogate Supervision ablation on integer anchors.

Type	Metric	w/o Physics Surrogate	PSAF	Reduction
Physics diagnostic	Geostrophic truth error	.02026	**.01576**	**22.2%**
	KE truth error	.03235	**.03117**	**3.6%**
Held-out diagnostic	Thermal-wind truth MAE ($\mathrm{m\,s^{-1}}$)	3.001	**2.944**	**1.9%**
	Vorticity truth MAE ($10^{-5}\,\mathrm{s^{-1}}$)	.2994	**.2919**	**2.5%**
	Divergence truth MAE ($10^{-5}\,\mathrm{s^{-1}}$)	.2981	**.2905**	**2.5%**
	Column-water truth MAE ($\mathrm{kg\,m^{-2}}$)	.09471	**.09272**	**2.1%**

A.4 Continuity Regularization: Implementation Details

Regularization terms. Continuity Regularization combines Router anchor, boundary-router, boundary-value, and temporal regularization:

$$\mathcal{L}_{\text{continuity}} = \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{br}}\mathcal{L}_{\text{boundary-router}} + \lambda_{\text{bv}}\mathcal{L}_{\text{boundary-value}} + \lambda_{\text{time}}\mathcal{L}_{\text{time}}. \tag{28}$$

Router anchor is applied to integer queries by assigning the seven Router logits to the corresponding 0–6-hour anchors:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{7}\sum_{k=0}^{6} \mathrm{CE}(r(k), k). \tag{29}$$

For fractional samples within a boundary band, let $k^\star$ denote the adjacent integer anchor and $j^\star \in \{0, 1\}$ its index within the active Local Decoder MoE pair. The two boundary losses are

$$\begin{aligned} \mathcal{L}_{\text{boundary-router}} &= \rho(\delta)\,\mathrm{CE}([r_k(\tau), r_{k+1}(\tau)], j^\star), \\ \mathcal{L}_{\text{boundary-value}} &= \rho(\delta)\,\mathrm{SmoothL1}\Big(\widehat{X}_{t+\tau}, \mathrm{sg}(\widehat{X}_{t+k^\star})\Big). \end{aligned} \tag{30}$$

Here, sg denotes stop-gradient, preventing the boundary-value term from propagating gradients into the integer-anchor prediction. The weighting function $\rho(\delta)$ increases linearly from 0 to 1 as the query approaches the corresponding integer anchor within the .10-hour boundary band, and is zero throughout the interior sampling region $\delta \in [.10, .90]$.

For each query, we additionally construct $\tau - \epsilon$, τ, and $\tau + \epsilon$. The temporal regularization within each adjacent-anchor interval is defined as

$$\mathcal{L}_{\text{time}} = \|\alpha(\tau - \epsilon) - 2\alpha(\tau) + \alpha(\tau + \epsilon)\|_1 + \gamma\|w(\tau - \epsilon) - 2w(\tau) + w(\tau + \epsilon)\|_2^2. \tag{31}$$

We set

$$\epsilon = \min\{.05, \delta/2, (1 - \delta)/2\}\,\mathrm{h}, \tag{32}$$

which keeps all three queries within the same adjacent-anchor interval. The coefficient for the second-difference term on the Router weights is $\gamma = 1$.

Loss weights and scheduling. We use $\lambda_{\text{anchor}} = .01$, $\lambda_{\text{br}} = .001$, $\lambda_{\text{bv}} = .01$, and $\lambda_{\text{time}} = .001$. The boundary-value term uses Smooth-L_1 with $\beta = .01$. Router anchor and the two boundary terms are applied at their full weights from the start of training, while temporal regularization is linearly warmed up over the first 5% of the total optimization steps.

Boundary and temporal behavior evaluation. The matched w/o Continuity Regularization ablation removes Router anchor, both boundary constraints, and temporal regularization. The cumulative adoption ratio and local Router-weight curves in Fig. 4(a)–(b) are generated at query intervals of .05 h, while the boundary curves in Fig. 4(c) are evaluated at offsets of $\pm.01$, $\pm.05$, and $\pm.10$ h around the integer anchors from 1 to 5 h. Aggregating both sides of these anchors at the $\pm.01$-hour offsets, Continuity Regularization reduces the boundary output jump from .004192 to .000178, corresponding to a 95.8% reduction.

To compare boundary changes with typical temporal variation along the query path, we compute the mean MAE between adjacent queries over the full 0–6-hour window using a .1-hour grid. The full-window averages are .003908 for w/o Continuity Regularization and .003790 for PSAF, while the corresponding boundary changes at $\pm.10$ h from the anchors are .006697 and .002196. The boundary change of the ablated model is therefore $1.71\times$ its full-window average, whereas PSAF does not exhibit comparable boundary amplification.

Second-order temporal differences are evaluated on the same .1-hour grid over the full 0–6-hour window. Continuity Regularization reduces the mean second-order difference from .00172 to .00140, a reduction of 18.4%, and the maximum second-order difference from .00822 to .00410, a reduction of 50.1%. The substantially larger reduction in the maximum indicates that the regularization mainly suppresses anomalous fluctuations at isolated query positions.

We further examine the transitions between the boundary sampling bands and the interior region using a finer query spacing. At $\delta = .1$, the ratio between the left and right .001-hour step changes is 1.010, and the curvature-to-step ratio is .024; at $\delta = .9$, the corresponding values are .997 and .079. No clear numerical jump or anomalously sharp local kink is observed at either transition, indicating that the gradual weighting of $\rho(\delta)$ provides a stable transition between the two sampling regions.

A.5 Baseline Implementation Details

Fixed-lead forecasting baselines. The retrained FourCastNet baseline follows the official NVlabs AFNO implementation, Pangu-Weather follows the public NVIDIA Modulus/PhysicsNeMo implementation, and FuXi follows the implementation distributed with OneForecast. FourCastNet's input layer is adapted from one to two consecutive weather states, FuXi retains its native two-state input, and Pangu-Weather retains its original single-state input. Their channel dimensions and grid resolution are matched to the common 69-channel, 1.5° WeatherBench 2 setting.

Endpoint-reconstruction baselines. The 6h field-interpolation baseline linearly interpolates in output space between the current analysis X_t and the 6-hour prediction $\widehat{X}_{t+6}$ produced by PSAF Stage 1. AMT, UPR-Net, and ModAFNO take the normalized current analysis X_t, the same 6-hour prediction $\widehat{X}_{t+6}$, and the normalized query time $s = \tau/6$ as input, and are trained using only the 1–5-hour ERA5 integer labels from 2006–2015. We adapt only their RGB-dependent input and output layers to 69 channels while retaining the original temporal interpolation architectures. For each training sample, one target lead time is sampled uniformly from $\{1, 2, 3, 4, 5\}$ h. Validation uses 2016–2017, and checkpoints are selected by the mean MAE over the five integer lead times.

AMT is trained with AdamW using a learning rate of 2×10^{-4} and no weight decay. UPR-Net uses AdamW with a learning rate of 2×10^{-4}, weight decay 10^{-4}, and a 2000-step warm-up. ModAFNO uses Adam with a learning rate of 5×10^{-4}, $\beta = (.9, .95)$, and no weight decay. All three baselines use a batch size of 2 per GPU and are trained for up to 70 epochs.

PSAF (fixed-1h). PSAF (fixed-1h) is trained independently and recursively generates forecasts from 1 to 6 h starting from the two most recent states, using the checkpoint selected on the validation set. Its architecture, training years, and main optimization settings are summarized in Table 7.

A.6 Additional Analyses of Any-Time Queries and Spatial Structure

Spatial-structure statistics. Total variation measures first-order local spatial variation, Laplacian RMS measures the magnitude of second-order spatial variation, high-frequency energy characterizes fine-scale spectral content, and near-Nyquist energy captures high-frequency components close to the grid-resolution limit. All spatial-structure statistics are computed in normalized space and macro-averaged equally over the 69 channels, without the latitude weighting used for RMSE and MAE.

Table 10 compares the spatial-structure statistics of different methods at the 1–5-hour integer queries, using ERA5 ground truth as the reference. PSAF is closest to ERA5 across all four metrics.

Table 10: Spatial-structure ratios to ERA5 for 1–5-hour integer queries. Values closer to 1 indicate better agreement in structural magnitude, not pointwise accuracy.

Method	Total variation	Laplacian RMS	High-frequency energy	Near-Nyquist energy
6h field interpolation	.9075	.8048	.8129	.7711
ModAFNO	.9667	.8975	.9349	.8222
UPR-Net	.9450	.8556	.9018	.7638
AMT	.9625	.8925	.9349	.7975
PSAF	**.9673**	**.9081**	**.9479**	**.8258**

Table 11 reports the spatial-structure ratios of fractional midpoints relative to linear interpolation between PSAF's adjacent 1-hour integer-anchor predictions.

Table 11: Spatial-structure ratios of fractional midpoints relative to linear interpolation of adjacent predicted anchors. Values near 1 indicate similar structural magnitude, not phase or positional agreement.

Spatial statistic	PSAF / linear reference
Total variation	.9993
Laplacian RMS	.9992
High-frequency energy	.9970
Near-Nyquist energy	1.0014

Recursive composition test. PSAF is normally used to query the target lead time directly from the same historical context. As an additional test, we evaluate a two-step recursive composition using a 0.5-hour step. We first query the $-5.5\,\mathrm{h}$ and $+0.5\,\mathrm{h}$ states from the contexts $[-12\,\mathrm{h}, -6\,\mathrm{h}]$ and $[-6\,\mathrm{h}, 0\,\mathrm{h}]$, respectively, forming the shifted two-frame context $[-5.5\,\mathrm{h}, +0.5\,\mathrm{h}]$. A second 0.5-hour query from this context then produces the $+1\,\mathrm{h}$ state. On the 2021 test set initialized every $60\,\mathrm{h}$, the normalized MAE is .01578 for the direct $+1\,\mathrm{h}$ query and .02043 for the recursive $2 \times 0.5\,\mathrm{h}$ composition. All other any-time experiments in this paper use direct queries to the target lead time.

A.7 RTMA-RU External Validation: Setup and Additional Results

RTMA-RU evaluation uses the common domain covering approximately 24–50°N and 125–66°W. Forecasts are initialized every $60\,\mathrm{h}$ in 2021 and queried at fractional lead times of $0.5, 1.5, \ldots, 5.5\,\mathrm{h}$. The grid-relative U10/V10 winds in RTMA-RU are converted to Earth-relative winds using meridian convergence. ClimODE and CEF are evaluated using their officially released pretrained weights, which correspond to WeatherBench v1 inputs at $5.625°$ resolution, whereas PSAF uses WeatherBench 2 inputs at $1.5°$ resolution. PSAF outputs at $5.625°$ are obtained only at evaluation time by bilinear or conservative area downsampling from its native $1.5°$ predictions. The main-text results use pointwise evaluation: RTMA-RU is bilinearly sampled at the forecast-grid centers, and errors are computed over common valid times within the CONUS mask using cosine-latitude weighting.

Table 12 further reports grid-cell-average evaluation, where each forecast grid cell is approximated by averaging over a 7×7 set of subpoints. Under this protocol, PSAF with conservative area

downsampling achieves the lowest error across all four metrics, while the native 1.5° output ranks second in each case.

Table 12: RTMA-RU grid-cell-average results using 7×7 subpoint averages. Values are cosine-latitude-weighted MAE; T2m is reported in K and wind metrics in $\mathrm{m\,s^{-1}}$.

Method	Grid	T2m ↓	U10 ↓	V10 ↓	Wind vector ↓
ClimODE	5.625°	2.9904	1.0533	1.1201	1.7086
CEF	5.625°	1.5109	.9307	1.0187	1.5309
PSAF (conservative)	5.625°	**.6823**	**.4028**	**.4040**	**.6335**
PSAF (bilinear)	5.625°	1.0670	.6614	.6897	1.0588
PSAF (native)	1.5°	.8944	.6510	.6500	1.0217

The relative performance of the different output forms varies between the two evaluation protocols. Grid-cell-average evaluation spatially aggregates both the forecast field and the RTMA-RU analysis, making it more consistent with the regional-mean information preserved by conservative area downsampling. In contrast, the pointwise protocol used in the main text directly compares instantaneous values at forecast-grid centers without additional spatial aggregation. The two protocols therefore reflect complementary aspects of forecast error, corresponding to pointwise prediction and spatially aggregated evaluation, respectively, and we report both.

A.8 Wind-Power Forecasting: Setup and Additional Results

For the wind-power experiments, 10-minute SCADA turbine power from Kelmarsh and Penmanshiel is aggregated to the wind-farm level and aligned to a regular 10-minute time series. Each TFT sample uses 72 power observations from the previous 12 h together with 36 future weather inputs over the next 6 h to predict the subsequent 36 power values. TFT uses the default architecture and QuantileLoss in PyTorch Forecasting 1.7.0, with a batch size of 64, an initial learning rate of 10^{-3}, and a maximum of 100 epochs. All runs use early stopping with patience 10 based on validation `val_loss` and restore the best checkpoint. The same TFT configuration is used for all weather sources, with random seeds 2026, 2027, and 2028. The years 2016, 2017, and 2018 are used for training, validation, and testing, respectively. Persistence extends the latest observed power value across all 36 future time steps. NMAE and NRMSE are aggregated over the full 6-hour forecast horizon. Skill is defined relative to Persistence NMAE as $1 - \mathrm{NMAE}/\mathrm{NMAE}_{\mathrm{Persistence}}$, with positive values indicating lower error than Persistence.

Table 13 further reports ramp metrics. Ramp MAE measures the error in 1-hour power changes, while Ramp F1 identifies significant ramp events using a capacity-factor change threshold of .10. The TFT training objective does not directly optimize either ramp metric.

Table 13: Additional wind-power forecasting metrics, averaged over two wind farms and three random seeds.

Weather input	Grid	NMAE ↓	NRMSE ↓	Ramp F1 ↑	Ramp MAE ↓	Skill ↑
History-only	–	.11757	.17664	.00276	.07793	-2.60%
ClimODE	5.625°	.11569	.17027	.02295	.07815	$-.97\%$
CEF	5.625°	.10927	.16115	.05015	.07800	$+4.64\%$
PSAF (conservative)	5.625°	.10963	.16100	.03553	.07744	$+4.32\%$
PSAF (bilinear)	5.625°	.10906	.16048	.04477	.07759	$+4.82\%$
PSAF (native)	1.5°	**.09305**	**.13895**	**.11512**	**.07557**	$\mathbf{+18.79\%}$

A.9 Full Any-Time Weather-Field Sequences

Figures 5–6 show the direct-query sequence for the sample in Fig. 3, grouped into integer and fractional lead times. Both figures use the same color scale for each variable.

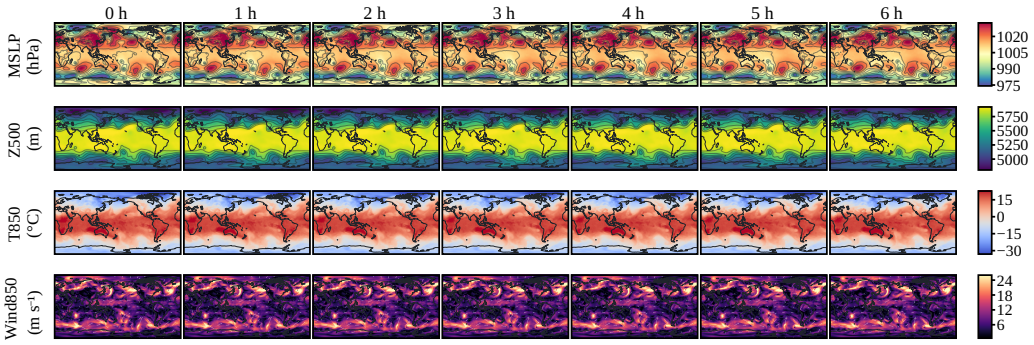


Figure 5: PSAF direct queries at integer lead times (0–6 h). Rows show MSLP, Z500, T850, and 850 hPa wind; wind shading and arrows indicate speed and direction.

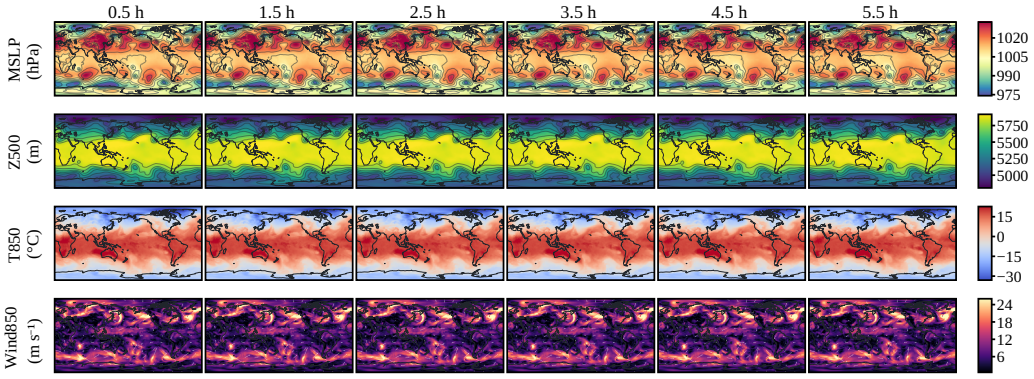


Figure 6: PSAF direct queries at fractional lead times (0.5, 1.5, . . . , 5.5 h). Rows show MSLP, Z500, T850, and 850 hPa wind; wind shading and arrows indicate speed and direction.