ONE ENCODING, CONTINUOUS LEAD TIMES: COHERENT ATMOSPHERIC REPRESENTATIONS WITH COMPOSITIONAL FINITE-TIME FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weather forecasts are often required at lead times that do not coincide with the temporal cadence of available data or trained forecasters. Although atmospheric evolution is continuous, most data-driven models learn transitions on a predefined temporal grid. Predictions between supervised horizons therefore rely on interpolation, repeated fixed-step rollout, or separately conditioned mappings that neither explicitly represent intermediate evolution nor ensure one atmospheric trajectory. We present PhiCast, which models forecasting as the continuous evolution of one shared latent atmosphere. It encodes the observed multivariate state once, represents forecast duration as a smooth state-aware coordinate, and reaches each queried horizon through a shared solver-free finite-time flow. Composition-aware learning further aligns direct and temporally decomposed paths in latent and forecast spaces. PhiCast therefore provides intermediate forecasts without interpolation, autoregressive feedback, numerical integration, or repeated encoding. Across lead times without direct forecast supervision, it achieves the best RMSE and ACC for all four evaluated variables, reducing variable-wise RMSE by 4.4–20.4% relative to the strongest competing result. By organizing different lead times as coherent states of one evolving atmosphere, PhiCast decouples forecast query cadence from the discrete supervision used for training.

1 INTRODUCTION

Weather is a continuous and multiscale dynamical process, whereas atmospheric observations, reanalysis products, and forecast supervision are available only at discrete times (Rasp et al., 2020; 2024). Forecasting between supervised horizons therefore requires more than producing outputs on a finer grid: a model must represent how the same atmosphere evolves between sparse temporal targets. This exposes three coupled requirements: (i) *continuous queryability under sparse supervision*; (ii) *direct target-time access* without interpolation, numerical integration, or autoregressive rollout; and (iii) *cross-horizon compositional coherence*, so independently queried forecasts remain compatible with one atmospheric trajectory (Bengio et al., 2015; Chen et al., 2018; Chen & Wu, 2023). Yet standard forecasting objectives still organize time as a transition step, rollout count, or external horizon condition. *The fundamental limitation is that existing paradigms fragment a single atmospheric evolution into disconnected lead-time predictions.*

Existing approaches satisfy these requirements only in part. Fixed-cadence systems learn discrete transitions or rollouts, while lead-conditioned models directly select a target horizon but do not couple separate queries through one evolving representation (Kurth et al., 2023; Nguyen et al., 2023). Continuous latent dynamics provide states beyond a fixed grid, but commonly rely on numerical integration, as in ClimODE, or emphasize queryability without enforcing cross-horizon agreement (Chen et al., 2018; Verma et al., 2024; Biloš et al., 2021). Semigroup learning constrains temporal composition, but has not been formulated to combine one encoded weather state with direct solver-free queries of future atmospheric fields (Chen & Wu, 2023). Thus, no existing formulation jointly combines a shared atmospheric origin, direct continuous-time querying, and compositional coherence.

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

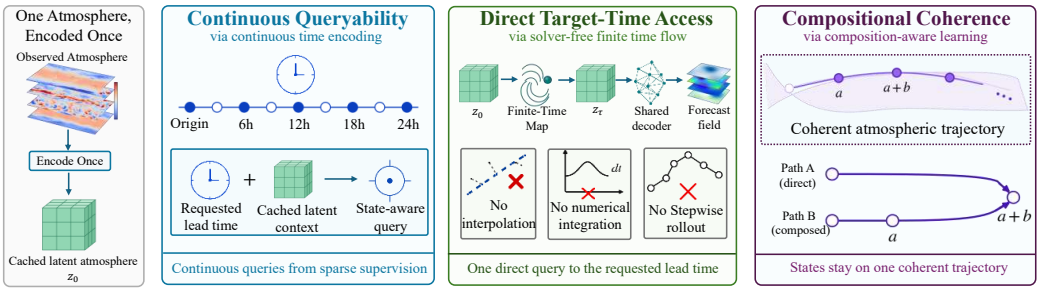

Figure 1: PhiCast addresses three temporal requirements from one atmosphere encoded once: continuous queryability under sparse supervision, direct target-time access through a solver-free finite-time flow, and cross-horizon compositional coherence. Here, a and b are successive normalized subintervals; the composed path is used only during training.

PhiCast models forecasting as the continuous evolution of one shared latent atmosphere. To support *continuous queryability under sparse supervision*, it combines a smooth duration encoding with the current latent-state context while reusing the atmosphere encoded at the forecast origin. To provide *direct target-time access*, a shared solver-free finite-time flow maps that cached representation to each requested state without constructing intermediate forecasts. To enforce *cross-horizon compositional coherence*, latent- and forecast-space objectives align direct queries with temporally decomposed paths reaching the same target. Together, these mechanisms organize independently queried forecasts as coherent states of one evolving atmosphere, as summarized in Figure 1.

Our main contributions are:

- We introduce PhiCast, which encodes the atmosphere once and directly queries continuous lead times through a shared, solver-free finite-time flow.
- We propose latent- and forecast-space composition objectives that align direct and decomposed queries into one coherent atmospheric trajectory.
- On WeatherBench, PhiCast achieves the best average RMSE and ACC for all four variables across 20 lead times without direct forecast supervision, while substantially reducing composition inconsistency.

2 Related Work

We briefly review three lines of related research, with an expanded discussion in Appendix A.

Data-driven weather forecasting. Modern weather models use spectral operators, Transformers, graph networks, or Earth-specific architectures to predict global atmospheric fields (Kurth et al., 2023; Nguyen et al., 2023; Bi et al., 2023; Lam et al., 2023). Many retain fixed-cadence transitions or rollout, while direct lead-time conditioning does not by itself couple queried horizons through one compositionally regularized evolution (Espeholt et al., 2022; Nguyen et al., 2024).

Neural operators and spatiotemporal dynamics. Neural operators learn physical field mappings through spectral, graph-based, and multiscale transformations (Li et al., 2020; Pfaff et al., 2020; Brandstetter et al., 2022; Zhang et al., 2024). Continuous representation and evolution variants have also emerged (Yin et al., 2022; Hou et al., 2026), but continuous lead-time access and cross-horizon consistency generally remain separate concerns.

Continuous-time and compositionally consistent evolution. Neural ODEs, Latent ODEs, and Neural Flows provide continuous or finite-time latent transformations (Chen et al., 2018; Rubanova et al., 2019; Biloš et al., 2021). ClimODE integrates learned weather dynamics, WeatherGFT combines fine-scale PDE evolution with neural correction, and continuous ensemble forecasting conditions diffusion on lead time (Verma et al., 2024; Xu et al., 2024; Andrae et al., 2025). Semigroup learning separately constrains temporal composition (Chen & Wu, 2023), leaving direct latent querying and compositional agreement only partially connected across existing formulations.

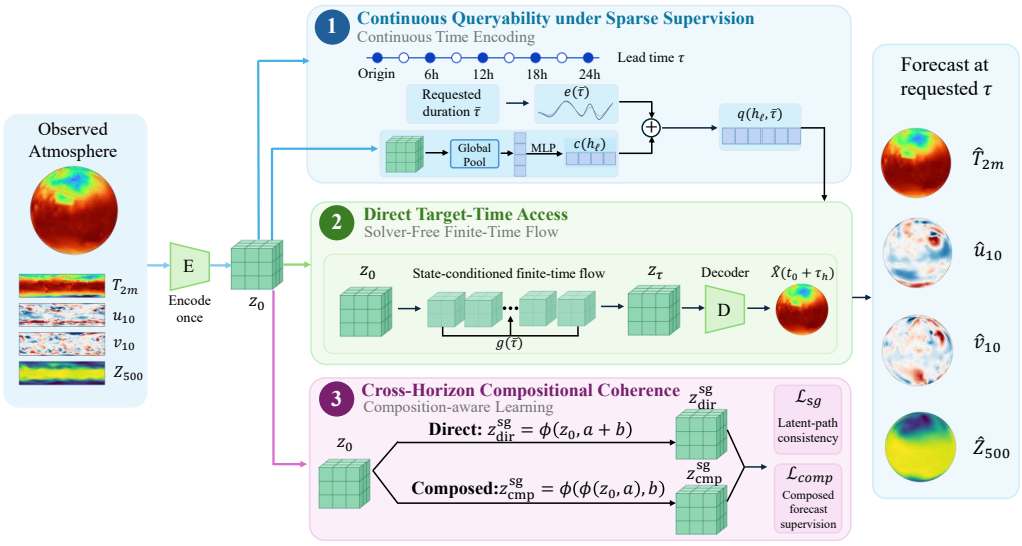

Figure 2: PhiCast framework. **(1) Continuous queryability:** continuous time encoding combines the requested duration with the current latent-state context. **(2) Direct target-time access:** a solver-free finite-time flow maps the atmosphere encoded once to the requested latent state and forecast. **(3) Cross-horizon compositional coherence:** training aligns direct and composed queries in latent and forecast spaces, while inference uses the direct path.

3 Method

3.1 Model Overview

Given an atmospheric state $X(t_0) \in \mathbb{R}^{B\times C\times H\times W}$ at forecast origin t_0, PhiCast predicts the multivariate field at a requested future time, where B, C, H, and W denote batch size, physical variables, latitude, and longitude. The observed field is standardized, augmented with spherical coordinates, and encoded once as the spatial latent atmosphere $z_0 = E(\widetilde{X}(t_0))$. Reusing this same z_0 for every lead time gives all forecasts a common atmospheric origin and prevents the representation itself from changing merely because a different horizon is requested. Continuous Time Encoding places supervised and intermediate durations in one smooth state-aware coordinate system, so an intermediate lead is handled by the same learned rule rather than by a separate head or post-hoc interpolation. Continuous Flow and Final Mapping converts that coordinate into the complete displacement from z_0 to the requested latent state in one solver-free query, which removes the need for fixed-step integration, autoregressive rollout, and repeated encoding. Composition-aware learning aligns the direct query with temporally decomposed queries to the same valid time, so direct access at many horizons does not fragment the forecasts into unrelated states. PhiCast therefore comprises three coupled mechanisms. Figure 2 details the two modules executed at inference, Continuous Time Encoding and Continuous Flow and Final Mapping. The third mechanism is composition-aware learning, which is applied only during training and detailed in Appendix B.2. We use four repeated multiscale blocks as the fixed depth of the finite-time map throughout all experiments. Together, they progressively refine the requested latent transformation. The initial-state tendency supplies directional information from z_0, and one shared decoder maps the queried latent state to weather variables. These mechanisms jointly turn a single encoding into a continuously queryable and compositionally coherent atmospheric representation.

Let $\tau_{\mathrm{h}} \geq 0$ denote lead time in hours and $\bar{\tau} = \tau_{\mathrm{h}}/24$ its normalized value. Valid time is $t_0 + \tau_{\mathrm{h}}$, while latent evolution is conditioned on $\bar{\tau}$. The next two subsections explain the inference modules in Figure 2. The surrounding spatial interfaces and training objectives are detailed in Appendix B, and the complete training and direct-query inference procedure is summarized in Appendix Table 3.

3.2 CONTINUOUS TIME ENCODING

A central difficulty in continuously queryable weather forecasting is that encoding the atmosphere once provides a shared initial state, but does not tell the model where a requested forecast lies along the subsequent evolution. This becomes particularly important under sparse lead-time supervision, where unseen intermediate horizons cannot simply be treated as predefined prediction targets. PhiCast addresses this difficulty by introducing a smooth duration coordinate that remains well defined between supervised horizons and combining it with the current latent atmosphere before conditioning the shared evolution map. The duration specifies where along the future evolution the forecast is requested, while the latent context determines how the current weather regime should respond to that duration. As a result, changing the lead time changes the evolution instruction applied to the same cached z_0, rather than selecting another prediction head or constructing a separate representation. This gives supervised and intermediate horizons a common temporal reference and encourages nearby forecast durations to correspond to nearby changes of the same evolving latent atmosphere.

The first step is to construct a smooth duration coordinate with sufficient temporal resolution to express both gradual trends and variation across multiple time scales. The upper panel of Figure 2 groups this coordinate into nonperiodic and periodic components; explicitly, the physical duration is normalized and expanded into the following 12-dimensional feature set:

$$\begin{aligned} \bar{\tau} &= \frac{\tau_{\mathrm{h}}}{24}, \\ \psi(\bar{\tau}) &= \Big[\bar{\tau},\ \log(1+\bar{\tau}),\ \frac{\bar{\tau}}{1+\bar{\tau}},\ \frac{\bar{\tau}^2}{1+\bar{\tau}^2}, \\ &\qquad \{\sin(\pi f_k \bar{\tau}), \cos(\pi f_k \bar{\tau})\}_{k=1}^{4} \Big], \qquad f_k \in \{0.25, 0.5, 1, 2\}. \end{aligned} \tag{1}$$

Here, τ_{h} is the requested duration in hours, $\bar{\tau}$ is its normalized value, $\psi(\bar{\tau}) \in \mathbb{R}^{12}$ is the continuous duration-feature vector, and π is the circle constant. The frequency set is fixed and deterministic: $f_1 = 0.25$, $f_2 = 0.5$, $f_3 = 1$, and $f_4 = 2$. These frequencies are neither learned nor randomly sampled; each f_k contributes exactly one sine–cosine pair, which together with the four nonperiodic terms gives 12 features. The nonperiodic components expose elapsed-time magnitude and saturation, while the fixed-frequency pairs provide complementary temporal scales. Linear layers of sizes $12 \to 128$ and $128 \to 128$, each followed by SiLU, learn how to combine these smooth features into the time embedding $e(\bar{\tau}) \in \mathbb{R}^{128}$. Consequently, an unsupervised intermediate lead changes the same continuous coordinate rather than activating a separate predictor. Together with the state conditioning below, this coordinate allows sparse lead-time supervision to constrain intermediate queries without changing the encoded atmospheric origin.

Time alone does not determine atmospheric change: the same duration can correspond to different tendencies under different atmospheric regimes. In a representative flow block, let $h = \mathrm{GroupNorm}(z)$ denote the normalized feature of the current latent state z. Global average pooling followed by a state multilayer perceptron extracts the context $c(h)$ from this same feature. The time–state code, affine head, and FiLM modulation are exactly

$$q(h, \bar{\tau}) = e(\bar{\tau}) + c(h), \qquad (\gamma, \beta) = A(q(h, \bar{\tau}))\,, \qquad h_{\mathrm{cond}} = h \odot (1 + 0.25 \tanh(\gamma)) + \beta. \tag{2}$$

Here, $c(h) \in \mathbb{R}^{B\times 128}$ is the pooled state context, $q(h, \bar{\tau})$ is the joint time–state code, A is the affine head, and $\gamma, \beta \in \mathbb{R}^{B\times 128}$ are channel-wise scale and shift vectors. The symbol $\odot$ denotes multiplication with spatial broadcasting. The conditioned feature h_{cond} forms branch C in Figure 2 so that the full- and coarse-scale spatial transformations in branches A and B are fused with an explicit instruction determined jointly by duration and the current atmosphere. This branch is therefore what makes a cached latent state respond differently to the same lead time under different weather regimes, allowing one encoded atmosphere to be queried repeatedly without constructing or re-encoding a new horizon-specific representation.

3.3 CONTINUOUS FLOW AND FINAL MAPPING

The state-aware duration representation developed in Section 3.2 provides the shared evolution rule with a continuous query coordinate, specifying both the requested horizon and how the current atmospheric regime should condition it. This conditioning makes intermediate horizons identifiable, but it does not yet define the transformation from the cached z_0 to the requested future state. Without an

appropriate evolution structure, the model would still need interpolation, numerical integration, or autoregressive transitions to reach that state. PhiCast closes this gap by learning a single finite-time map ϕ shared across all forecast horizons, which transforms the cached z_0 directly into the complete latent change associated with the requested duration. Its multiscale spatial pathways capture full-resolution, coarse-scale, and coordinate-aware local atmospheric interactions, while the time–state condition modulates how the same shared map transforms the current atmosphere for different durations. Importantly, the internal flow blocks are representation transformations within one finite-time query rather than successive numerical time steps. Consequently, one encoded atmosphere can be mapped directly to any requested state within the modeled interval through a single solver-free evaluation and a shared decoder, without constructing, decoding, feeding back, or re-encoding intermediate forecasts.

The lower panel of Figure 2 implements this finite-time map, transforming the initial weather latent z_0 into the latent atmosphere $z(\bar{\tau})$ associated with the physical valid time $t_0 + \tau_{\mathrm{h}}$. To preserve weather interactions across spatial scales while changing the displacement continuously with duration, PhiCast uses four repetitions of the representative finite-time flow block shown in the figure. These blocks provide network depth for one direct query and are not numerical time steps. In block ℓ, GroupNorm produces the current normalized feature from z_ℓ. Branch A sums a full-resolution spectral convolution with 16×16 modes and a parallel 3×3 cylindrical convolution, combining global spectral coupling with coordinate-aware local structure. Branch B applies 2×2 average pooling, sums an 8×8-mode coarse spectral convolution with a parallel coarse cylindrical convolution, and bilinearly upsamples the result, exposing broader low-frequency organization at lower spatial cost. Branch C supplies the time–state FiLM feature h_{cond} from Equation 2. Concatenating the three branches and applying the 1×1 fusion network produces one duration- and state-dependent displacement:

$$\Delta z_\ell = F(z_\ell, \bar{\tau}), \qquad z_{\ell+1} = z_\ell + g(\bar{\tau})\Delta z_\ell. \tag{3}$$

Here, $\ell \in \{0, 1, 2, 3\}$ indexes the four blocks, F denotes the concatenation-and-fusion mapping represented in Figure 2, Δz_ℓ is the proposed latent displacement, and $g(\bar{\tau})$ is the duration gate. All queried horizons share these four blocks and differ only through the continuous duration and the state it conditions; hence changing the requested time evolves the same latent object rather than invoking a separate transition network. Because the blocks are shared and evaluated within one finite-time query, target-time access does not accumulate the interpolation or stepwise-transition errors associated with autoregressive rollout.

The gate shown in Figure 2 is

$$g(\bar{\tau}) = \frac{\bar{\tau}}{\sqrt{1 + \bar{\tau}^2}}, \qquad g(0) = 0 \;\Rightarrow\; \phi(z_0, 0) = z_0. \tag{4}$$

Here, $\bar{\tau}$ is the normalized nonnegative duration and $g(\bar{\tau}) \in [0, 1)$ scales every block displacement. This gate connects direct queryability to a valid trajectory origin: the zero-duration equality makes the learned finite-time flow start exactly from the encoded atmosphere, while bounded growth prevents the explicit multiplier on each displacement from increasing linearly without limit. It therefore makes different lead-time queries departures from one shared initial state rather than disconnected horizon-specific predictions. After the fourth block, the result is the transient state $\phi_{\mathrm{transient}}(z_0, \bar{\tau})$.

The multiscale blocks describe nonlinear, lead-dependent adjustment, but a single initial atmosphere can also carry a directional tendency that should persist across the queried interval. PhiCast therefore complements the transient path with an initial-state-tendency branch derived only from z_0. The complete finite-time map corresponding to the two-path merge in Figure 2 is

$$\phi(z_0, \bar{\tau}) = \phi_{\mathrm{transient}}(z_0, \bar{\tau}) + \bar{\tau}\,\sigma(\alpha)V(z_0), \qquad z(\bar{\tau}) = \phi(z_0, \bar{\tau}). \tag{5}$$

Here, ϕ is the complete finite-time map, $\phi_{\mathrm{transient}}$ is the four-block transient path, $V(z_0) \in \mathbb{R}^{B \times d \times H \times W}$ is the learned initial-state velocity obtained from the initial latent state, α is a learned scalar gain logit, σ is the logistic sigmoid, and $z(\bar{\tau})$ is the queried latent state. The transient path models nonlinear adjustment, whereas $\bar{\tau}\sigma(\alpha)V(z_0)$ is the initial-state tendency that accumulates an initial-state-dependent direction with duration. At $\bar{\tau} = 0$, the transient displacement and the initial-state tendency both vanish, so the complete map exactly recovers the shared origin z_0 while retaining complementary short- and longer-duration behavior away from the origin.

Finally, the shared decoder uses the residual anchor shown in Figure 2.

$$\begin{aligned} \Delta\hat{\bar{X}}(\bar{\tau}) &= D(z(\bar{\tau})) - D(z_0), \\ \hat{\bar{X}}(t_0+\tau_{\mathrm{h}}) &= \bar{X}(t_0) + \Delta\hat{\bar{X}}(\bar{\tau}), \\ \hat{X}(t_0+\tau_{\mathrm{h}}) &= \mathcal{N}^{-1}\Big(\hat{\bar{X}}(t_0+\tau_{\mathrm{h}})\Big) \\ &= \Big[\hat{T}_{2m}, \hat{u}_{10}, \hat{v}_{10}, \hat{Z}_{500}\Big]_{t_0+\tau_{\mathrm{h}}}. \end{aligned} \tag{6}$$

Here, D is the decoder shared by all lead times, $\Delta\hat{\bar{X}}(\bar{\tau})$ is the standardized forecast increment, $\bar{X}(t_0)$ and $\hat{\bar{X}}(t_0+\tau_{\mathrm{h}})$ are the standardized observation and prediction, and $\mathcal{N}^{-1}$ denotes channel-wise inverse standardization. The variables $\hat{T}_{2m}$, $\hat{u}_{10}$, $\hat{v}_{10}$, and $\hat{Z}_{500}$ are the predicted 2-m temperature, 10-m zonal wind, 10-m meridional wind, and 500-hPa geopotential fields in physical units. The residual anchor references every decoded forecast to the same observed atmosphere, preventing the output mapping from introducing a separate origin for each horizon. The encoder and decoder are therefore time independent and shared, while all horizon dependence is confined to the continuous, state-conditioned map ϕ. This division of roles is what makes one-time encoding and direct multi-horizon querying compatible: z_0 is cached once, $\bar{\tau}$ selects a state along its learned evolution, and no intermediate forecast must be decoded, fed back, or re-encoded.

Direct finite-time access at each duration still does not guarantee that independently queried forecasts belong to the same evolution. If horizons are constrained only by their individual forecast targets, a direct query $\phi(z_0, a+b)$ can disagree with the decomposed query $\phi(\phi(z_0, a), b)$ even though both paths represent the same valid time, where $a, b \geq 0$ are normalized subintervals. PhiCast therefore directly penalizes latent disagreement between these paths and applies forecast supervision to the weather field decoded from the composed path. The two objectives encourage alternative decompositions to represent the same target without claiming an exact autonomous semigroup for the externally forced atmosphere. When a second subinterval is queried, the same map is applied to the intermediate latent, so its state-dependent velocity is recomputed from that intermediate state rather than cached from the original z_0. The complete objectives are given in Appendix B.2, and the end-to-end training and direct-query inference procedure is summarized in Appendix Table 3. At inference, the observed atmosphere is encoded once and each requested lead time is produced by a single direct finite-time query. The composed path is used only to regularize training and is not evaluated at inference.

4 Results

4.1 Forecast Accuracy at Continuously Queried Lead Times

Table 1 compares PhiCast with seven baselines and separates the four forecast-supervised leads (6, 12, 18, and 24 hours) from the remaining 20 integer leads without direct forecast supervision. Entries report the mean and sample standard deviation over seeds 42, 43, and 44. On the latter group, PhiCast obtains the lowest RMSE and highest ACC for every variable. On supervised leads it performs best for temperature and both wind components, while ClimaX is slightly better for Z_{500}. The unsupervised group contains many short leads, so its absolute error should not be compared with the supervised group as though they were equally difficult; comparisons are made between models within each group. The lower RMSE alongside higher ACC indicates that PhiCast reduces physical-amplitude error while preserving non-climatological spatial patterns. Appendix Tables 9 and 10 report individual-seed results.

Table 1 aggregates performance across lead times, while Figure 3 resolves a representative seed-42 run at every integer hour. Only 6, 12, 18, and 24 hours receive direct forecast supervision. PhiCast remains strongest across the four variable-wise curves over most or all of the interval, and its error changes smoothly between supervised and unsupervised times. Thus, the aggregate gains are supported by a well-behaved per-lead profile rather than isolated favorable horizons. Appendix Figure 7 provides the macro-average curve.

Table 1: Forecast accuracy over the complete 2018 test set, separated into forecast-supervised leads (6, 12, 18, and 24 hours) and leads without direct forecast supervision (the other integer hours in 1–24). Entries are mean ± sample standard deviation over three independent training seeds (42, 43, and 44).

Forecast-supervised leads

	RMSE ↓				ACC ↑			
Model	T_{2m}	u_{10}	v_{10}	Z_{500}	T_{2m}	u_{10}	v_{10}	Z_{500}
PhiCast	**1.246±0.008**	**1.532±0.009**	**1.563±0.009**	148.04±1.26	**0.88961±0.00125**	**0.92615±0.00086**	**0.92583±0.00085**	0.98391±0.00028
ClimODE	1.821±0.016	1.740±0.002	1.758±0.002	199.00±1.52	0.78383±0.00258	0.90365±0.00026	0.90506±0.00027	0.97081±0.00043
ClimaX	1.421±0.002	1.575±0.001	1.599±0.001	**139.67±0.12**	0.85300±0.00042	0.92172±0.00012	0.92215±0.00009	**0.98572±0.00003**
ResNet	1.491±0.002	1.579±0.003	1.602±0.002	173.29±0.54	0.84728±0.00031	0.92137±0.00033	0.92188±0.00025	0.97795±0.00012
U-Net	1.580±0.014	1.745±0.002	1.760±0.002	191.73±0.96	0.82679±0.00246	0.90303±0.00029	0.90490±0.00023	0.97287±0.00028
FNO	1.624±0.004	1.885±0.002	1.926±0.003	218.67±3.56	0.81703±0.00033	0.88583±0.00024	0.88487±0.00045	0.96457±0.00118
SFNO	2.634±0.031	2.921±0.015	3.168±0.020	333.94±4.64	0.62590±0.01032	0.72670±0.00391	0.68503±0.00508	0.91584±0.00256
WeatherGFT	1.775±0.028	2.042±0.015	1.957±0.004	246.75±5.92	0.79747±0.00512	0.86705±0.00208	0.88138±0.00056	0.95524±0.00207

Leads without direct forecast supervision

	RMSE ↓				ACC ↑			
Model	T_{2m}	u_{10}	v_{10}	Z_{500}	T_{2m}	u_{10}	v_{10}	Z_{500}
PhiCast	**1.222±0.006**	**1.414±0.007**	**1.446±0.007**	**132.93±0.77**	**0.89540±0.00083**	**0.93750±0.00061**	**0.93689±0.00060**	**0.98705±0.00015**
ClimODE	1.892±0.013	1.772±0.004	1.835±0.004	194.43±0.82	0.77611±0.00246	0.90000±0.00042	0.89635±0.00046	0.97211±0.00023
ClimaX	1.664±0.004	1.608±0.006	1.694±0.013	157.57±2.58	0.80482±0.00097	0.91839±0.00062	0.91248±0.00139	0.98181±0.00061
ResNet	1.535±0.010	1.482±0.004	1.512±0.004	160.62±0.62	0.84269±0.00178	0.93103±0.00038	0.93068±0.00036	0.98110±0.00013
U-Net	1.711±0.019	1.638±0.008	1.661±0.014	178.21±1.60	0.80665±0.00377	0.91514±0.00088	0.91580±0.00145	0.97663±0.00042
FNO	1.603±0.001	1.741±0.002	1.784±0.004	197.51±2.41	0.82518±0.00107	0.90353±0.00029	0.90216±0.00050	0.97119±0.00072
SFNO	2.553±0.032	2.731±0.020	3.011±0.032	314.62±6.81	0.64786±0.01021	0.76147±0.00429	0.71581±0.00700	0.92538±0.00341
WeatherGFT	1.791±0.001	1.886±0.012	1.820±0.003	222.05±4.82	0.80026±0.00038	0.88759±0.00151	0.89820±0.00040	0.96385±0.00150

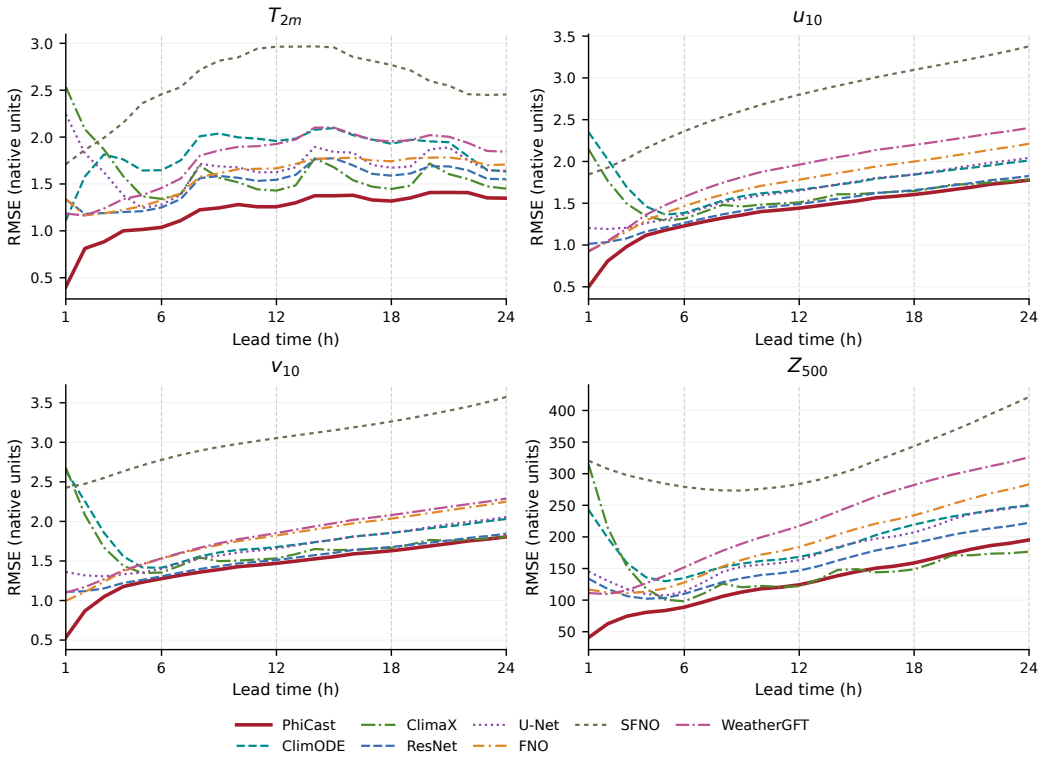

Figure 3: Variable-wise latitude-weighted RMSE at 1–24 hours for the seed-42 checkpoints. Dashed lines mark the supervised leads (6, 12, 18, and 24 hours); the other integer leads receive no direct forecast supervision. Every model starts from the same observed origin. ClimODE integrates its ODE, whereas other models query directly; no method uses interpolation or autoregressive rollout.

4.2 Continuous Latent Trajectory and Time Composition

We next test whether predictions at different lead times behave as states of one continuous trajectory rather than independent horizon-specific outputs. These diagnostics use ClimaX and ClimODE as deterministic temporal references with accessible hidden states, contrasting direct horizon conditioning with solver-based continuous evolution. Baselines without an intrinsic composable latent map

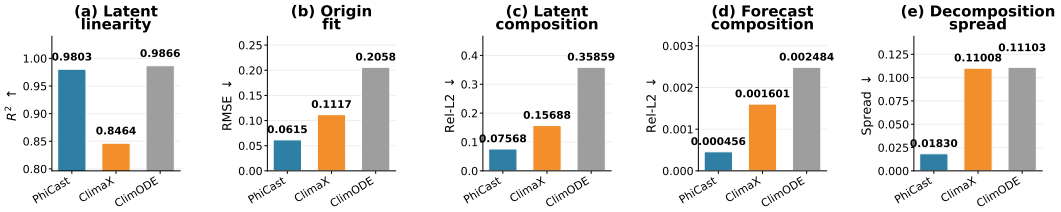

Figure 4: Continuous-time representation diagnostics for PhiCast, ClimaX, and ClimODE on 32 fixed 2018 forecast origins. Panels (a)–(b) measure how normalized latent displacement varies with temporal separation. Panels (c)–(d) compare direct and staged latent or decoded predictions, and panel (e) measures output variation across decompositions of the same target duration. For ClimODE, the staged path reconstructs its augmented state from the intermediate predicted weather field before integrating the second interval. Higher is better for R^2 in (a); lower is better for the other metrics. Values above the bars are the unnormalized measurements.

remain in the forecast-accuracy comparison but are omitted here to avoid architecture-dependent auxiliary states.

For latent smoothness, we compare displacement trends using affine and through-origin fits because latent coordinate systems and scales differ across architectures. Figure 4(a)–(b) shows smooth temporal trends for both PhiCast and ClimODE, while PhiCast aligns most closely with the encoded origin. This supports approximately duration-proportional evolution over the measured interval without assuming a globally linear trajectory. Panels (c)–(e) compare direct and staged queries: PhiCast composes latent maps, ClimaX feeds back an intermediate forecast, and ClimODE reconstructs its augmented state from the intermediate predicted weather field. PhiCast shows consistently lower latent and decoded disagreement and less variation across decompositions, supporting stronger cross-horizon agreement. ClimODE's native continuation from its complete augmented state remains nearly exactly compositional, so its staged result measures sensitivity to a weather-field-only restart rather than the composition of the ODE solver itself.

4.3 Ablation Analysis of Core Mechanisms

Figure 5 isolates how each mechanism contributes to continuously queryable forecasting while leaving the surrounding model structure unchanged. The full time encoding maintains a richer temporal spectrum than scalar conditioning, so different lead times can express distinct yet smoothly related evolution instructions instead of collapsing onto nearly the same temporal direction. Replacing the structured finite-time update with the Direct Mapper weakens alignment with the encoded origin and degrades forecasts at leads without direct supervision. Because the conditioned block depth and width remain unchanged, this trend attributes the benefit to the identity-preserving gate and initial-state velocity rather than additional capacity. Removing the composition objectives produces a marked increase in direct-versus-composed disagreement even though pointwise forecast supervision is retained. This shows that accurate individual targets do not by themselves organize independently queried horizons into a compatible trajectory. Read together, the panels connect representation geometry to forecast behavior: temporal encoding distinguishes requested horizons, the origin-aware flow structures their displacement, and composition-aware learning couples otherwise independent queries. The ablations therefore expose complementary failure modes rather than three measurements of the same effect. Together, the mechanisms locate a requested horizon, reach it directly from the shared atmosphere, and align it with the same evolving trajectory. Further ablations are provided in Appendix F.1.

4.4 Between-Supervision Lead-Time Accuracy

Table 2 isolates the three intervals strictly between supervised horizons, excluding the earliest forecasts for which proximity to the observed origin can favor persistence-like behavior. To compare variables with different physical scales without allowing one field to dominate the aggregate, RMSE is normalized per variable before averaging, whereas ACC is averaged directly. Forecasting becomes harder for every method as the requested horizon moves farther from the origin, yet PhiCast remains

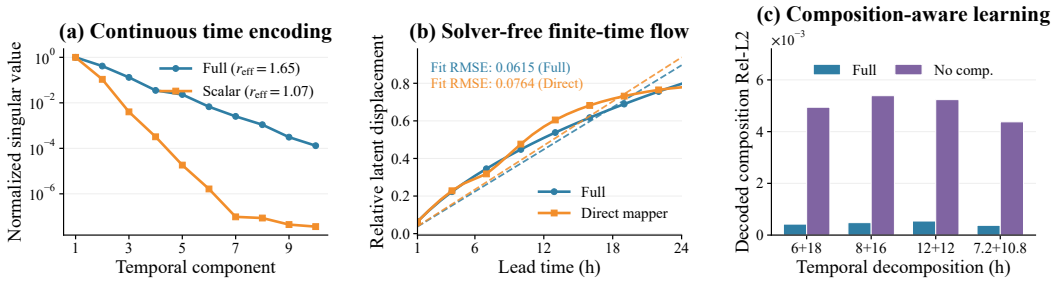

Figure 5: Ablations of PhiCast's three core mechanisms. (a) Normalized singular-value spectrum of the time representation; r_{eff} is its effective rank and *Scalar* uses scalar-time conditioning. (b) Relative latent displacement for PhiCast and the *Direct mapper*, which retains the conditioned blocks but removes the identity-preserving gate and initial-state velocity. Dashed lines are through-origin least-squares fits, with fitting RMSE shown. (c) Direct-versus-composed decoded Rel-L2; *No comp.* removes the latent path-consistency and composed-forecast objectives. Lower is better in (b)–(c).

strongest in both NRMSE and ACC throughout all three intervals. Its advantage therefore persists across the interior of each supervision gap rather than being concentrated near the observed state or a single supervised endpoint. The stable ranking as the intervals move later also indicates that the result is not driven only by local smoothness around the forecast origin. The agreement between NRMSE and ACC further shows that the improvement covers both physical-amplitude error and non-climatological spatial structure. These trends support the intended role of the shared finite-time evolution: sparse supervision anchors a common temporal representation from which intermediate horizons remain directly and reliably queryable.

Table 2: Forecast accuracy strictly between supervised leads on the complete 2018 test set, reported as mean ± sample standard deviation over seeds 42–44. Within each interval, NRMSE and ACC average the five integer leads and four variables after variable-wise RMSE normalization.

	7–11 h		13–17 h		19–23 h	
Model	NRMSE ↓	ACC ↑	NRMSE ↓	ACC ↑	NRMSE ↓	ACC ↑
PhiCast	0.1562±0.0005	0.9423±0.0004	0.1767±0.0009	0.9270±0.0006	0.1958±0.0013	0.9145±0.0009
ClimODE	0.1887±0.0007	0.8956±0.0023	0.2094±0.0006	0.8791±0.0007	0.2267±0.0007	0.8736±0.0010
ClimaX	0.1706±0.0004	0.9221±0.0005	0.1865±0.0003	0.9105±0.0005	0.2004±0.0001	0.9030±0.0002
ResNet	0.1656±0.0003	0.9267±0.0006	0.1876±0.0002	0.9076±0.00005	0.2052±0.0003	0.8982±0.0001
U-Net	0.1807±0.0002	0.9137±0.0005	0.2066±0.0002	0.8900±0.0004	0.2279±0.0002	0.8755±0.0002
FNO	0.1932±0.0001	0.9085±0.0005	0.2226±0.0001	0.8804±0.0002	0.2464±0.0004	0.8609±0.0003
SFNO	0.3190±0.0054	0.7746±0.0082	0.3518±0.0029	0.7332±0.0065	0.3805±0.0003	0.7033±0.0028
WeatherGFT	0.2047±0.0008	0.8969±0.0014	0.2380±0.0013	0.8624±0.0021	0.2609±0.0016	0.8441±0.0024

5 Conclusion

We introduced PhiCast to recast weather forecasting as continuous querying over one coherent latent atmosphere. Its central contribution is representational: the observed multivariate state is encoded once, that same spatial latent is evolved by a shared finite-time flow, and every requested horizon is decoded from the resulting trajectory. Continuous time is therefore not an auxiliary condition attached to otherwise independent predictors, but the coordinate along which a common atmospheric representation changes. The time-compositional objectives further make this interpretation operational by encouraging direct and decomposed queries to describe the same future state. Together, these choices separate forecast query cadence from the discrete lead-time supervision used to train the model, while avoiding interpolation, autoregressive feedback, repeated encoding, and solver-dependent time stepping. Beyond the reported accuracy improvements, this work establishes coherent continuous atmospheric representation as a viable modeling principle for weather prediction: forecasts at different horizons can be learned as related states of one evolving atmosphere rather than as a collection of discrete tasks. This perspective opens a route toward temporally adaptive forecasting systems that can align their output cadence with downstream scientific and operational needs while preserving cross-horizon dynamical coherence.

REFERENCES

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 26392–26416, 2025.

Omri Azencot, N Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent Koopman autoencoders. In *International conference on machine learning*, pp. 475–485. PMLR, 2020.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

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

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural ODEs. *Advances in neural information processing systems*, 34:21325–21337, 2021.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 1(8), 2024.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pp. 2806–2823. PMLR, 2023.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Junfeng Chen and Kailiang Wu. Deep-OSG: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493:112498, 2023.

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

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36:47143–47175, 2023b.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):5145, 2022.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xianglong Hou, Xinquan Huang, and Paris Perdikaris. CFO: Learning continuous-time PDE dynamics via flow-matched neural operators. In *International Conference on Learning Representations*, volume 2026, pp. 86968–86994, 2026.

Ruoxi Jiang, Peter Y Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. *Advances in Neural Information Processing Systems*, 36: 27645–27669, 2023.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Vladimir Kostic, Pietro Novelli, Riccardo Grazzi, Karim Lounici, et al. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *International Conference on Learning Representations*, volume 2024, pp. 54329–54341, 2024.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Kai Lagemann, Christian Lagemann, and Sach Mukherjee. Invariance-based learning of latent dynamics. In *International Conference on Learning Representations*, volume 2024, pp. 38843–38887, 2024.

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

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

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Xin Man, Chenghong Zhang, Jin Feng, Changyu Li, and Jie Shao. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*, 2023.

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

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. *Advances in Neural Information Processing Systems*, 37:68740–68771, 2024.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

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

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Samuel H Rudy, J Nathan Kutz, and Steven L Brunton. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *Journal of Computational Physics*, 396:483–506, 2019.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *International conference on machine learning*, pp. 19388–19405. PMLR, 2022.

Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and Weather Forecasting with Physics-informed Neural ODEs. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 8408–8430, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/21f3c05fc5afcaebd8a71d7e98a1b171-Paper-Conference.pdf.

Zhong Yi Wan, Leonardo Zepeda-Nunez, Anudhyan Boral, and Fei Sha. Evolve smoothly, fit consistently: Learning smooth latent dynamics for advection-dominated systems. *arXiv preprint arXiv:2301.10391*, 2023.

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. *Advances in Neural Information Processing Systems*, 37:23325–23351, 2024.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous pde dynamics forecasting with implicit neural representations. *arXiv preprint arXiv:2209.14855*, 2022.

Xuan Zhang, Jacob Helwig, Yuchao Lin, Yaochen Xie, Cong Fu, Stephan Wojtowytsch, and Shuiwang Ji. Sinenet: Learning temporal dynamics in time-dependent partial differential equations. In *International Conference on Learning Representations*, volume 2024, pp. 11612–11653, 2024.

A Expanded Related Works

Data-driven global weather forecasting. WeatherBench and WeatherBench 2 established standardized data processing, evaluation metrics, and reference systems for data-driven global weather forecasting (Rasp et al., 2020; 2024). To model the spatial structure of global atmospheric fields, FourCastNet introduced efficient high-resolution forecasting with Fourier operators (Kurth et al., 2023), while Spherical FNO extended spectral operator learning to spherical geometry (Bonev et al., 2023). For multivariate representation learning, ClimaX developed a foundation model with cross-variable interactions (Nguyen et al., 2023), whereas Earthformer and Stormer used space–time attention and scalable Transformers to model atmospheric dynamics (Gao et al., 2022; Nguyen et al., 2024). Pangu-Weather, GraphCast, and FuXi further improved medium-range forecasting through three-dimensional Earth-specific networks, multiscale graph message passing, and cascaded prediction systems (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023a). AIFS explored an operational data-driven forecasting architecture, while NeuralGCM coupled learned components with a structured atmospheric circulation model (Lang et al., 2024; Kochkov et al., 2024). W-MAE and Aurora demonstrated the value of pretraining and foundation modeling across variables and tasks (Man et al., 2023; Bodnar et al., 2024), and GenCast extended data-driven forecasting to probabilistic weather ensembles (Price et al., 2025). Overall, most existing systems still formulate atmospheric

evolution as discrete transitions at predefined intervals, so predictions at unsupervised intermediate lead times require interpolation, recursive forecasting, or additional horizon-specific training and are not explicitly tied to a unified atmospheric trajectory. In contrast, PhiCast encodes the initial atmosphere once and directly queries future states through a shared continuous latent evolution, enabling supervised and intermediate lead times to be represented within the same coherent trajectory.

Neural operators and spatiotemporal dynamics learning. Neural operators learn mappings between function spaces and provide an alternative to conventional numerical solvers for high-dimensional physical forecasting. The Fourier Neural Operator captures global spatial dependencies in the spectral domain (Li et al., 2020), MeshGraphNets represent local physical interactions on mesh graphs (Pfaff et al., 2020), and Message Passing Neural PDE Solvers extend graph-based evolution learning across spatial resolutions (Brandstetter et al., 2022). SineNet uses multiscale spatial processing to reduce information loss and phase errors in time-dependent systems (Zhang et al., 2024). Moving beyond explicit grid states, DINo combines implicit neural representations with latent dynamics to provide continuous spatial decoding (Yin et al., 2022), while CFO learns continuous-time PDE evolution operators through flow matching (Hou et al., 2026). Other studies focus on long-term dynamical properties, such as preserving the invariant measures of chaotic attractors during neural-operator training (Jiang et al., 2023). Overall, most neural-operator approaches emphasize spatial field mappings, discretization flexibility, or particular dynamical properties, while spatial encoding, continuous lead-time querying, and cross-horizon consistency remain separate components rather than a single coherent evolution representation. In contrast, PhiCast preserves the encoded atmospheric field as a spatial latent state and evolves that same representation through a shared finite-time flow, jointly connecting spatial structure, continuous lead time, and coherent cross-horizon dynamics.

Continuous-time modeling and consistent evolution. Neural ODEs represent hidden-state evolution as the continuous solution of a differential equation (Chen et al., 2018), and Latent ODEs use a shared latent trajectory to explain irregularly sampled sequences (Rubanova et al., 2019). GRU-ODE-Bayes, Neural CDEs, Continuous Recurrent Units, and ContiFormer address irregular observations through continuous hidden-state updates, controlled dynamics, closed-form recurrent transitions, and continuous-time attention, respectively (De Brouwer et al., 2019; Kidger et al., 2020; Schirmer et al., 2022; Chen et al., 2023b). In contrast to methods that rely on numerical integration, Neural Flows directly parameterize finite-time transformations and can evaluate a requested state in a single query (Biloš et al., 2021). For atmospheric prediction, ClimODE uses physically informed neural ODEs to model continuous weather dynamics (Verma et al., 2024), WeatherGFT studies generalization to fine-grained forecast intervals (Xu et al., 2024), and continuous ensemble forecasting constructs temporally continuous probabilistic predictions (Andrae et al., 2025). Beyond queryability, continuous dynamics should remain consistent under different temporal decompositions: Deep-OSG imposes semigroup structure on learned evolution operators (Chen & Wu, 2023), consistent Koopman autoencoders reduce latent drift through forward–backward agreement (Azencot et al., 2020), and related methods constrain latent evolution through dynamical invariance, trajectory smoothness, time-homogeneous stochastic representations, and time-stepping consistency (Lagemann et al., 2024; Wan et al., 2023; Kostic et al., 2024; Rudy et al., 2019). Overall, most continuous-time approaches either depend on numerical integration, target irregular-sequence interpolation, or impose individual consistency properties without jointly maintaining a shared spatial atmosphere and a directly queryable compositional trajectory. In contrast, PhiCast combines one-time atmospheric encoding, a solver-free continuous lead-time flow, and time-compositional consistency, so forecasts at different horizons are generated as coherent states of the same latent atmospheric evolution rather than independent predictions.

B Complete Model and Learning Details

The main text follows the two modules shown in Figure 2. This appendix records the surrounding spatial interfaces, the complete learning objective, and the end-to-end algorithm used to train and query PhiCast.

B.1 Spatial Atmospheric Interfaces

The input contains near-surface temperature, two horizontal wind components, and 500 hPa geopotential. Each variable is standardized with training-set statistics. Longitude uses periodic padding and latitude uses replicated boundary padding. Let $P \in \mathbb{R}^{B\times C_P\times H\times W}$ collect $C_P = 4$ sine–cosine latitude and longitude channels. The one-time spatial encoding is

$$\widetilde{X}(t_0) = \mathrm{Concat}[\mathcal{N}(X(t_0)), P]\,, \qquad z_0 = E\Big(\widetilde{X}(t_0)\Big) \in \mathbb{R}^{B\times d\times H\times W}. \tag{7}$$

Here, $X(t_0) \in \mathbb{R}^{B\times C\times H\times W}$ is the physical atmospheric field at initial time t_0; $\mathcal{N}$ is channel-wise standardization; P is the coordinate tensor with $C_P = 4$ channels; Concat is channel-wise concatenation; $\widetilde{X}(t_0)$ is the normalized coordinate-augmented input; E is the encoder; and z_0 is the initial spatial latent with width $d = 128$. The dimensions B, C, H, and W denote batch, variable, latitude, and longitude. Coordinate-aware cylindrical convolutions and residual blocks preserve the spatial alignment of z_0 with the atmospheric grid instead of compressing the full field into one vector. These 3×3 convolutions use periodic padding in longitude and replicated boundary padding in latitude; they are not rotation-equivariant spherical convolutions.

The shared decoder $D : \mathbb{R}^{B\times d\times H\times W} \to \mathbb{R}^{B\times C\times H\times W}$ first reconstructs common features and then uses separate heads for T_{2m}, u_{10}, v_{10}, and Z_{500}. Its output is used through the residual anchor in Equation 6; the same D and the cached reference $D(z_0)$ are reused for every lead-time query.

For reproducibility, all hidden and latent features have width 128 with eight GroupNorm groups. The encoder consists of a 3×3 cylindrical stem, a 1×1 lift to 128 channels, and four residual blocks, each containing two GroupNorm–GELU–3×3 cylindrical-convolution stages. The state MLP in each finite-time block is $128 \to 128 \to 128$ with SiLU, and its affine head is a linear $128 \to 256$ map producing the FiLM scale and shift. The three branch outputs are concatenated and fused by 1×1 convolutions of widths $384 \to 256 \to 128$ with GELU. The velocity map V applies GroupNorm, parallel 16-mode spectral and 3×3 cylindrical convolutions, GELU, and a 1×1 projection to 128 channels; its scalar gain logit is initialized to $\alpha = -3$. The decoder uses a 1×1 lift, two 16-mode FNO blocks at width 128, and three heads: one-channel temperature and geopotential heads and a two-channel wind head, each formed by a 3×3 cylindrical convolution, GELU, and a 1×1 projection. The *Direct mapper* ablation retains the same four multiscale, time–state-conditioned blocks and widths but removes the velocity branch and replaces the identity-preserving gate by unit scaling; it is trained with the same data and optimization budget.

B.2 Time-Compositional Learning Objective

Let $\bar{X} = \mathcal{N}(X)$ and $\bar{Y} = \mathcal{N}(Y)$ be standardized input and target fields, and let $\hat{\bar{X}}$ denote the standardized prediction inside $\mathcal{N}^{-1}$ in Equation 6. For a target $Y(t_0 + \tau_{b,\mathrm{h}})$ at physical lead $\tau_{b,\mathrm{h}}$, define $\bar{\tau}_b = \tau_{b,\mathrm{h}}/T$. The direct forecast term is

$$\mathcal{L}_{\mathrm{forecast}} = \frac{1}{BCHW}\sum_{b=1}^{B}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} r_b w_i \left(\hat{\bar{X}}_{bcij}(t_0+\tau_{b,\mathrm{h}}) - \bar{Y}_{bcij}(t_0+\tau_{b,\mathrm{h}})\right)^2. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{forecast}}$ is the latitude- and horizon-weighted forecast loss; b, c, i, and j index the B, C, H, and W dimensions; $\tau_{b,\mathrm{h}}$ and $\bar{\tau}_b$ are the physical and normalized sampled leads of example b; $w_i = \cos(\varphi_i)/\overline{w}$ is the normalized area weight at latitude φ_i; and $r_b = (0.8 + 0.6\bar{\tau}_b)/\overline{r}$ is the normalized horizon weight. The scalars $\overline{w}$ and $\overline{r}$ are the corresponding spatial and mini-batch means. We set $T = 24$ hours because it is the largest directly supervised duration in this bounded experiment; this normalization choice does not impose an architectural upper limit on the query interval.

For latent time composition, sample $u, v \sim \mathcal{U}(0, 1)$ and set $\tau_1 = u\bar{\tau}_b$ and $\tau_2 = v(\bar{\tau}_b - \tau_1)$, so both are normalized non-negative subdurations with $\tau_1 + \tau_2 \le \bar{\tau}_b$. We then define

$$z^{\mathrm{sg}}_{\mathrm{dir}} = \phi(z_0, \tau_1 + \tau_2), \qquad z^{\mathrm{sg}}_{\mathrm{cmp}} = \phi(\phi(z_0, \tau_1), \tau_2)\,. \tag{9}$$

Here, z_0 is the encoded initial atmosphere, ϕ is the complete finite-time map, $z^{\mathrm{sg}}_{\mathrm{dir}}$ is the latent reached by one direct query, $z^{\mathrm{sg}}_{\mathrm{cmp}}$ is reached by composing two queries, u and v are independent

Table 3: Algorithmic summary of PhiCast training and direct continuous lead-time inference. The composed paths regularize training only; inference always uses one encoding and one direct finite-time query per requested horizon.

Inputs: training pairs $(X(t_0), Y(t_0+\tau_{\mathrm{h}}))$, supervised physical-lead set $\mathcal{S}_{\mathrm{h}}$, normalization scale T, encoder E, finite-time map ϕ, decoder D.	
Training	
1	Sample $(X(t_0), Y(t_0+\tau_{b,\mathrm{h}}))$ with $\tau_{b,\mathrm{h}} \in \mathcal{S}_{\mathrm{h}}$ and set $\bar{\tau}_b = \tau_{b,\mathrm{h}}/T$; normalize and append spherical coordinates using Equation 7.
2	Encode once: $z_0 \leftarrow E(\widetilde{X}(t_0))$; cache $D(z_0)$.
3	Direct query: $z(\bar{\tau}_b) \leftarrow \phi(z_0, \bar{\tau}_b)$ using Equations 1–5; decode $\hat{X}(t_0+\tau_{b,\mathrm{h}})$ with Equation 6.
4	Compute the direct forecast loss $\mathcal{L}_{\mathrm{forecast}}$ from Equation 8.
5	Sample (u, v), form $z^{\mathrm{sg}}_{\mathrm{dir}}$ and $z^{\mathrm{sg}}_{\mathrm{cmp}}$ with Equation 9, and compute $\mathcal{L}_{\mathrm{sg}}$.
6	Sample ρ, decode the composed path with Equation 11, and compute $\mathcal{L}_{\mathrm{comp}}$.
7	Update all parameters by minimizing Equation 12.
Inference for one origin and a set of physical leads $\mathcal{Q}_{\mathrm{h}} \subseteq [0, T]$	
1	Compute $z_0 \leftarrow E(\widetilde{X}(t_0))$ and $D(z_0)$ once.
2	For each $\tau_{\mathrm{h}} \in \mathcal{Q}_{\mathrm{h}}$, set $\bar{\tau} = \tau_{\mathrm{h}}/T$, evaluate $z(\bar{\tau}) \leftarrow \phi(z_0, \bar{\tau})$ directly, and decode with Equation 6.
3	Return $\{\hat{X}(t_0+\tau_{\mathrm{h}}) : \tau_{\mathrm{h}} \in \mathcal{Q}_{\mathrm{h}}\}$; do not integrate, feed predictions back, or re-encode intermediate fields.

unit-uniform random variables, τ_1 and τ_2 are the resulting normalized subdurations, and $\mathcal{U}(p, q)$ is the uniform distribution on $[p, q]$. Their relative discrepancy is

$$\mathcal{L}_{\mathrm{sg}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\left\| z^{\mathrm{sg},(b)}_{\mathrm{dir}} - z^{\mathrm{sg},(b)}_{\mathrm{cmp}} \right\|_2^2}{\left\| \mathrm{sg}\left(z^{\mathrm{sg},(b)}_{\mathrm{dir}} \right) \right\|_2^2 + \varepsilon_{\mathrm{stab}}}. \tag{10}$$

Here, $\mathcal{L}_{\mathrm{sg}}$ is the semigroup-consistency loss; b indexes a sample; $\|\cdot\|_2$ is the Euclidean norm over latent channels and locations; sg is stop-gradient in the denominator; and $\varepsilon_{\mathrm{stab}} > 0$ prevents division by zero. This objective encourages $\phi_{\tau_1+\tau_2} \approx \phi_{\tau_2} \circ \phi_{\tau_1}$ but does not impose an exact semigroup.

For composition in decoded weather space, draw $\rho \sim \mathcal{U}(0.2, 0.8)$, set $\tau_1 = \rho\bar{\tau}_b$ and $\tau_2 = (1-\rho)\bar{\tau}_b$, and let $z^{\mathrm{pred}}_{\mathrm{cmp}} = \phi(\phi(z_0, \tau_1), \tau_2)$. The composed prediction remains referenced to the original atmosphere,

$$\hat{\hat{X}}_{\mathrm{cmp}}(t_0+\tau_{b,\mathrm{h}}) = \mathcal{N}(X(t_0)) + D(z^{\mathrm{pred}}_{\mathrm{cmp}}) - D(z_0). \tag{11}$$

Here, $\hat{\hat{X}}_{\mathrm{cmp}}$ is the standardized composed forecast, $X(t_0)$ is the observed physical field, $\mathcal{N}$ is channel-wise standardization, ρ is the random split fraction, τ_1 and τ_2 are its two subintervals, $z^{\mathrm{pred}}_{\mathrm{cmp}}$ is the composed latent, and $D(z^{\mathrm{pred}}_{\mathrm{cmp}}) - D(z_0)$ is the decoded increment relative to the original latent. Applying Equation 8 to this path gives $\mathcal{L}_{\mathrm{comp}}$. Unlike $\mathcal{L}_{\mathrm{sg}}$, this term does not directly penalize the distance between direct and composed decoded forecasts; it supervises the composed forecast against the common verification target and thereby encourages decoded agreement indirectly. The complete loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{forecast}} + \lambda_{\mathrm{sg}}\mathcal{L}_{\mathrm{sg}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}}. \tag{12}$$

Here, $\mathcal{L}$ is the total objective; $\mathcal{L}_{\mathrm{forecast}}$, $\mathcal{L}_{\mathrm{sg}}$, and $\mathcal{L}_{\mathrm{comp}}$ are the direct, latent-composition, and decoded-composition terms; and $\lambda_{\mathrm{sg}}, \lambda_{\mathrm{comp}} \geq 0$ are their weights. The composed paths are used only during training.

B.3 Algorithmic Summary

Table 3 summarizes the complete PhiCast training procedure and its direct continuous lead-time inference workflow.

C Experimental Details

Dataset and preprocessing. We use WeatherBench, a standardized benchmark derived from ERA5 reanalysis for data-driven global weather forecasting (Rasp et al., 2020; 2024). All experi-

ments use the 5.625° grid (32×64 spatial locations) and predict four prognostic fields: 2-m temperature T_{2m} in kelvin, 10-m zonal and meridional winds u_{10} and v_{10} in meters per second, and 500-hPa geopotential Z_{500}. We use 2000–2015 for training, 2016–2017 for validation, and 2018 for testing. Each field is standardized independently with the mean and standard deviation computed only from the training period. The forecast loss is latitude weighted using $w(\varphi) = \cos(\varphi)/\operatorname{mean}_\varphi \cos(\varphi)$ to account for the changing physical area of regular latitude–longitude cells. During training, a forecast lead is sampled uniformly from $\{6, 12, 18, 24\}$ hours. The complete test evaluates every integer lead from 1 to 24 hours at 1,456 six-hourly forecast origins in 2018; thus, 20 of the 24 evaluated lead times are absent from direct forecast supervision. Every prediction is an independent direct query from the observed origin, and no reported result uses autoregressive rollout. Confidence intervals are computed with 10,000 seven-day block-bootstrap draws, preserving short-range temporal dependence between neighboring origins.

Comparison models. We compare with seven models: ClimODE (Verma et al., 2024), ClimaX (Nguyen et al., 2023), WeatherGFT (Xu et al., 2024), SFNO (Bonev et al., 2023), FNO (Li et al., 2020), U-Net (Ronneberger et al., 2015), and ResNet (He et al., 2016). Each receives the same four normalized fields, training samples, forecast origins, and supervised leads as PhiCast. ClimODE retains its augmented velocity–weather state, advection-aware derivative networks, uncertainty-correction head, and explicit Euler integration; its initial velocity is inferred from the observed state alone to preserve the single-origin protocol without target fitting. ClimaX is initialized from its released pretrained weights and retains its direct prediction interface. The other deterministic direct models receive spherical coordinate channels and the normalized lead-time condition; ResNet, U-Net, FNO, SFNO, and WeatherGFT predict an increment added to the normalized input field. Every model is evaluated independently from the same observed origin at each target time. Architectures requiring a different native input interface are adapted to the common four-variable 32×64 grid; these are not presented as unchanged pretrained checkpoints. No capacity-expanded baseline variant is used.

Optimization and model selection. All models are trained for at most 200 epochs with 100 optimizer updates per epoch and effective batch size four. PhiCast and ClimODE use micro-batches of two with two-step gradient accumulation; the remaining baselines use an effective batch size of four, either directly or through accumulation. We optimize with AdamW, weight decay 10^{-5}, cosine learning-rate annealing, and gradient-norm clipping at 1.0. The initial learning rate is 2×10^{-4} for PhiCast, 10^{-5} for ClimaX, and 5×10^{-4} for ClimODE, ResNet, U-Net, FNO, SFNO, and WeatherGFT. PhiCast uses $\lambda_{\mathrm{sg}} = 0.08$ and $\lambda_{\mathrm{comp}} = 0.25$ in Equation 12. Validation is performed every 10 epochs on 32 fixed forecast origins at every integer lead from 1 to 24 hours. Training stops after six validation checks without improvement, and the checkpoint with the lowest mean validation Rel-L2 is retained. The same validation rule is applied to every model and ablation, and no test-set metric is used for checkpoint selection. Table 1 reports three independent training runs with seeds 42, 43, and 44 for every model; Appendix Tables 9 and 10 give the individual runs. Unless explicitly stated otherwise, mechanism probes, qualitative figures, efficiency measurements, ablations, and spatial-resolution transfer use the seed-42 checkpoint. Training and inference are performed on one NVIDIA RTX 4070 Laptop GPU with 8 GB memory.

Architecture summary. We use the original model configurations adopted in the reported comparison, without capacity-expanded baseline variants. PhiCast has latent width 128 and four flow blocks with 16 spectral modes per spatial axis; each block combines full- and coarse-resolution spectral branches with time–state FiLM conditioning, GroupNorm, and SiLU activations. ClimODE uses the official residual and attention derivative networks with 128- and 64-channel hidden stages, an augmented velocity–weather state, an uncertainty-correction network, and explicit Euler integration on a one-hour internal grid. ClimaX uses 2×2 patches, width 1024, eight Transformer blocks, 16 attention heads, an MLP expansion ratio of four, a two-layer prediction head, and dropout and drop-path rates of 0.1. ResNet uses width 96 and eight residual blocks, each containing two 3×3 circular convolutions with GroupNorm(8) and GELU. U-Net uses base width 48, three encoder–decoder levels with channel ratio $1{:}2{:}4{:}8$, average-pooling downsampling, bilinear upsampling, GroupNorm, and GELU. FNO uses width 64, four operator blocks, and 12 modes per spatial axis, with each block combining a spectral path and a local 1×1 path followed by GroupNorm(8) and GELU. SFNO uses width 96 and four spherical-operator layers. WeatherGFT uses a learned 1×1 lift

into its 69-channel native interface and a 64-channel hybrid PDE–attention core. All models use AdamW, cosine scheduling, at most 200 epochs of 100 updates, effective batch size four, and gradient clipping at 1.0; supervision is provided at 6, 12, 18, and 24 hours, validation is performed every 10 epochs, and early stopping uses patience six. PhiCast additionally uses the semigroup and composition-forecast objectives with $\lambda_{\mathrm{sg}} = 0.08$ and $\lambda_{\mathrm{comp}} = 0.25$.

D NOTATION

Atmospheric data, dimensions, and time

t_0	Initial forecast time or forecast origin.
τ_{h}	Physical forecast duration in hours.
$\tau, s, a, b, \tau_1, \tau_2$	Normalized non-negative durations; $\tau = \tau_{\mathrm{h}}/T$, while the remaining symbols are auxiliary durations used in continuity or composition statements.
$T, \bar{\tau}$	Duration normalization scale and normalized duration $\bar{\tau} = \tau_{\mathrm{h}}/T$; $T = 24$ hours in the experiments.
$T_{\max}$	Upper endpoint of the bounded normalized-duration query interval used in the appendix analysis.
$X(t_0), Y(t_0 + \tau_{\mathrm{h}})$	Observed atmospheric input field and target future atmospheric field in physical units.
$\hat{X}(t_0 + \tau_{\mathrm{h}})$	Predicted future atmospheric field in physical units; a hat denotes a model prediction.
$\bar{X}, \bar{Y}, \hat{\bar{X}}$	Standardized input, target, and prediction, respectively.
$\Delta\hat{\bar{X}}$	Standardized forecast increment produced by the residual decoder anchor.
$T_{2m}, u_{10}, v_{10}, Z_{500}$	2-m temperature, 10-m zonal wind, 10-m meridional wind, and 500-hPa geopotential.
$\widetilde{X}(t_0)$	Coordinate-augmented standardized input passed to the encoder.
P, C_P	Coordinate tensor and its number of channels; $C_P = 4$.
B, C, H, W	Batch size, number of physical variables, latitude count, and longitude count.
d, L	Latent channel width and number of finite-time flow blocks.
b, c, i, j, ℓ, k	In indexed training sums, indices of sample, variable, latitude, longitude, flow block, and temporal frequency; standalone a and b in the theory denote durations.
$\mathbb{R}$	Set of real numbers.

Latent states and learned mappings

E, D	Atmospheric encoder and shared variable-aware decoder.
$z_0, z(\bar{\tau}), z, z'$	Encoded initial spatial latent, latent queried at normalized duration $\bar{\tau}$, and generic admissible latent states used in the block and continuity analyses.
$z_\ell, \Delta z_\ell$	Latent entering block ℓ and displacement proposed by that block.
ϕ, ϕ_τ	Duration-conditioned finite-time latent map and shorthand $\phi_\tau(z) = \phi(z, \tau)$.
$\phi_{\mathrm{transient}}$	Transient component of the finite-time latent map.
h, h_{cond}	Normalized current block feature and its time–state-conditioned form.
$h^{(\ell)}_{\mathrm{full}}, h^{(\ell)}_{\mathrm{coarse}}, h^{(\ell)}_{\mathrm{cond}}$	Full-resolution, coarse-resolution, and conditioned branch features in block ℓ.
F	Concatenation-and-1×1-fusion mapping used to propose a block displacement.
V, α	Initial-state velocity mapping and its learned scalar gain logit.
$\bar{\tau}\sigma(\alpha)V(z_0)$	Initial-state tendency obtained by scaling the initial-state velocity by normalized duration and the learned gain.
$\psi(\bar{\tau}), e(\bar{\tau})$	Twelve-dimensional continuous duration-feature vector and its 128-dimensional learned embedding.
f_k	Fixed temporal frequency with the deterministic assignment $(f_1, f_2, f_3, f_4) = (0.25, 0.5, 1, 2)$; it is not sampled or learned.
$c(h), q(h, \bar{\tau})$	Pooled state context and joint time–state conditioning vector $q = e + c$.
A, γ, β	Affine conditioning head and its channel-wise scale and shift outputs.
$g(\bar{\tau})$	Identity-preserving duration gate $\bar{\tau}/\sqrt{1 + \bar{\tau}^2}$ for normalized duration $\bar{\tau}$.

Losses, weights, and composition variables

$\mathcal{L}_{\mathrm{forecast}}$	Latitude- and horizon-weighted forecast loss.
$\mathcal{L}_{\mathrm{sg}}$	Normalized latent semigroup-consistency loss.

$\mathcal{L}_{\mathrm{comp}}$	Forecast loss applied to a temporally composed latent path.
$\mathcal{L}$	Complete training objective.
$\lambda_{\mathrm{sg}}, \lambda_{\mathrm{comp}}$	Non-negative weights of the two consistency terms.
$w(\varphi), w_i, \varphi, \varphi_i, \overline{w}$	Normalized cosine-latitude weighting function, its value at row i, generic latitude, latitude of row i, and the spatial mean used for normalization.
$r_b, \overline{r}, \tau_{b,\mathrm{h}}, \bar{\tau}_b$	Horizon weight, its mini-batch normalizer, and the physical and normalized lead times of sample b.
$\mathcal{S}_{\mathrm{h}}, \mathcal{Q}_{\mathrm{h}}$	Sets of supervised and requested physical lead times, respectively.
u, v	Independent unit-uniform random variables used to construct normalized subdurations for the semigroup objective.
ρ	Random split fraction used to form $\tau_1 = \rho\bar{\tau}_b$ and $\tau_2 = (1-\rho)\bar{\tau}_b$.
$z^{\mathrm{sg}}_{\mathrm{dir}}, z^{\mathrm{sg}}_{\mathrm{cmp}}$	Direct and composed latents used by the semigroup-consistency objective.
$z^{\mathrm{pred}}_{\mathrm{cmp}}$	Composed latent decoded for target-supervised forecast composition.
$\hat{X}_{\mathrm{cmp}}$	Standardized weather prediction decoded from the composed path.
sg	Stop-gradient operator.
$\varepsilon_{\mathrm{stab}}$	Positive numerical stabilizer in the normalized semigroup loss.
ε	Uniform upper bound on the composition defect in the theoretical analysis.

Operators and theoretical quantities

$\mathcal{N}, \mathcal{N}^{-1}$	Channel-wise standardization and inverse standardization.
Concat	Concatenation along the channel dimension.
$\odot$	Channel-wise multiplication with spatial broadcasting.
$\mathrm{GroupNorm}$	Group normalization.
$\mathrm{SiLU}, \mathrm{GELU}$	Sigmoid linear unit and Gaussian error linear unit.
$\sigma, \tanh$	Logistic sigmoid and hyperbolic tangent.
mean	Arithmetic mean over the subscripted dimension or collection.
$\log, \sin, \cos, \pi$	Natural logarithm, sine, cosine, and the circle constant.
$\lVert \cdot \rVert_2$	Euclidean norm over all indicated latent or field entries.
$\circ$	Function composition.
$\mathcal{U}(p, q)$	Uniform distribution on the real interval $[p, q]$.
$\mathcal{Z}$	Compact latent-state region considered in the continuity analysis.
L_z, L_τ, L_D	Lipschitz constants with respect to latent state, query duration, and decoding.
s	Comparison duration in the continuity bound.
$\delta(z; a, b)$	Defect between one direct query of duration $a+b$ and the composition of durations a and b.
m, Δ_k, p	Number of temporal subintervals, duration of the kth subinterval, and index of the geometric sum in a decomposition of τ.
R^2, r_{eff}	Coefficient of determination for the affine latent-displacement fit and effective rank of the time-representation spectrum.
$\mathrm{RMSE}, \mathrm{NRMSE}$	Root mean square error and normalized RMSE.
RMS	Root mean square magnitude.
$\mathrm{ACC}, \mathrm{Rel\text{-}L2}$	Anomaly correlation coefficient and relative L_2 error.

E Continuity and Time-Composition Properties

This section records the structural properties that follow from the finite-time parameterization used by PhiCast. The statements concern continuity and internal consistency of the learned query map; they do not claim that continuity alone guarantees forecast accuracy at unsupervised or extrapolated lead times. For compactness, let $\phi_\tau(z) \equiv \phi(z, \tau)$, let $\mathcal{Z} \subset \mathbb{R}^{d\times H\times W}$ be a compact forward-invariant region containing the reachable closure of admissible direct and composed latent trajectories, and assume $\phi_\tau(\mathcal{Z}) \subseteq \mathcal{Z}$ for $\tau \in [0, T_{\max}]$. The duration τ lies in this bounded normalized query interval.

Continuity across queried lead times. All components of the duration feature map in Equation 1 are continuously differentiable on $[0, T_{\max}]$. SiLU, sigmoid, and hyperbolic tangent are smooth; linear maps, finite-dimensional spectral convolutions, pooling, bilinear upsampling, cylindrical convolutions, and concatenation are locally Lipschitz; and the normalization layers use a positive numerical stabilizer. Therefore, on the bounded domain $\mathcal{Z} \times [0, T_{\max}]$, each flow block and their finite

composition are locally Lipschitz in both latent state and query duration. Consequently, there exist finite constants L_z and L_τ such that

$$\|\phi_\tau(z) - \phi_s(z')\|_2 \le L_z \|z - z'\|_2 + L_\tau|\tau - s|, \qquad z, z' \in \mathcal{Z}, \quad \tau, s \in [0, T_{\max}]. \tag{13}$$

Here, $z, z' \in \mathcal{Z}$ are two latent states; $\tau, s \in [0, T_{\max}]$ are two query durations; ϕ_τ and ϕ_s are their finite-time maps; L_z and L_τ are finite Lipschitz constants with respect to latent input and duration; and $\|\cdot\|_2$ is the Euclidean norm. In particular, fixing the encoded atmosphere z_0 gives

$$\|\phi_\tau(z_0) - \phi_s(z_0)\|_2 \le L_\tau|\tau - s|, \tag{14}$$

Here, z_0 is one fixed encoded initial atmosphere, τ and s are two lead times, and L_τ is the duration-Lipschitz constant from Equation 13. Thus, nearby lead times produce nearby latent states. If the shared decoder is Lipschitz with constant L_D on the same latent range, the corresponding standardized forecasts satisfy

$$\left\|\hat{\hat{X}}(t_0 + T\tau) - \hat{\hat{X}}(t_0 + Ts)\right\|_2 \le L_D L_\tau|\tau - s|. \tag{15}$$

Here, $\hat{\hat{X}}(t_0 + T\tau)$ and $\hat{\hat{X}}(t_0 + Ts)$ are standardized predictions from the same origin t_0 at normalized durations τ and s, $T\tau$ and Ts are the corresponding physical durations, L_D is the decoder Lipschitz constant, and L_τ is the latent duration-Lipschitz constant. This establishes continuity of the model output with respect to requested lead time without introducing numerical time steps. The constants are data- and parameter-dependent and are not asserted to be small; the empirical smoothness analysis in Section 4.2 measures the realized behavior of the trained model.

Exact origin consistency. At zero duration, $\psi(0)$ is finite and the identity gate satisfies $g(0) = 0$. Every transient block in Equation 3 therefore reduces exactly to the identity. The initial-state tendency in Equation 5 evaluates to zero because $\bar{\tau} = 0$, yielding

$$\phi_0(z) = z. \tag{16}$$

Here, ϕ_0 is the zero-duration finite-time map and z is any admissible latent state. Substituting Equation 16 into the residual decoder in Equation 6 gives $\hat{X}(t_0) = X(t_0)$. Thus, the continuous trajectory begins exactly at the observed atmosphere rather than learning the zero-time boundary only from samples.

Approximate time composition. Define the latent composition defect for durations $a, b \ge 0$ by

$$\delta(z; a, b) = \|\phi_{a+b}(z) - \phi_b(\phi_a(z))\|_2 . \tag{17}$$

Here, z is the reference latent, $a, b \ge 0$ are two durations, $\phi_{a+b}(z)$ is the direct query, $\phi_b(\phi_a(z))$ is the two-part composed query, and $\delta(z; a, b)$ is their Euclidean discrepancy. An exact autonomous flow has $\delta = 0$, but sub-daily atmospheric evolution is externally forced and PhiCast does not assert this identity as a physical law. Instead, Equation 10 acts as an empirical path-consistency regularizer over trajectories represented in the training distribution. Exact closure is also not built into the architecture: the second subinterval recomputes the state-dependent velocity from $\phi_a(z)$, and the nonlinear duration gate is not additive. Suppose that $\delta(z; a, b) \le \varepsilon$ on $\mathcal{Z}$ for the fixed macroscopic partitions of interest and that every duration map is L_z-Lipschitz in its latent input. For a decomposition $\tau = \sum_{k=1}^{m} \Delta_k$, repeated application of the triangle inequality gives

$$\|\phi_\tau(z) - \phi_{\Delta_m} \circ \cdots \circ \phi_{\Delta_1}(z)\|_2 \le \varepsilon \sum_{p=0}^{m-2} L_z^p. \tag{18}$$

Here, τ is the total duration; $m \ge 2$ is the number of subintervals; Δ_k is the duration of subinterval k; $\phi_{\Delta_m} \circ \cdots \circ \phi_{\Delta_1}$ is the corresponding m-part composition; ε is a uniform upper bound on each two-part defect; L_z is the latent-input Lipschitz constant; and p indexes the geometric sum. Exact origin consistency implies $L_z \ge 1$ for a constant uniform over intervals containing zero. For $L_z = 1$, the right-hand side is $(m-1)\varepsilon$; for $L_z > 1$, it is $\varepsilon(L_z^{m-1} - 1)/(L_z - 1)$ and may grow rapidly with partition depth. The bound is intended for the finite two- and three-part macroscopic decompositions studied here, not an asymptotic limit under arbitrarily fine partitioning, for which a step-size-dependent defect analysis would be required. Equation 18 therefore motivates reducing empirical defect while making clear that the unconstrained network has no global exact-semigroup guarantee. At inference, the model uses the direct map ϕ_τ once; composition is a training regularizer and trajectory diagnostic.

F EXPANDED RESULTS

F.1 FINE-GRAINED COMPONENT ABLATIONS

Figure 6 uses mechanism-specific diagnostics to test the intended role of each component. All comparisons use fixed checkpoints trained with the same data, optimization budget, forecast supervision, and validation rule. The consistency, coarse-input, and short-time diagnostics use the same 32 fixed 2018 forecast origins.

The semigroup objective substantially improves agreement between a direct latent query and a composition of shorter queries. Averaged over the $6{+}18$, $8{+}16$, $12{+}12$, and $7.2{+}10.8$ hour decompositions, the latent consistency error decreases from 0.2507 without the semigroup term to 0.0757 for the complete model, corresponding to a 69.8% reduction. Panel (a) thus verifies the role of latent semigroup supervision: it makes the learned state transition approximately independent of whether a duration is queried in one step or split into subintervals. Panel (b) evaluates the complementary composition-forecast objective in decoded space. For each of the 32 origins, we predict the same 24-hour endpoint using a direct query and the $12{+}12$, $6{+}18$, $3{+}21$, and $8{+}8{+}8$ hour decompositions. We compute the root mean square deviation of these five forecasts from their per-origin mean after normalizing each variable by its training-set standard deviation. With latent semigroup supervision retained, removing only the composition-forecast objective increases this normalized across-decomposition spread from 0.0183 to 0.0303; equivalently, including the objective reduces it by 39.6%. The result therefore shows that the decoded 24-hour forecast is less sensitive to how the total duration is partitioned. It provides empirical evidence of approximate composition consistency for the tested decompositions, rather than a claim of exact path independence for arbitrary times.

We next examine the representational roles of continuous time encoding and multiscale spatial processing. For lead times sampled densely over $[0, 24]$ hours, we form the matrix of learned time embeddings and analyze its singular-value spectrum. The complete continuous encoding retains eight numerical temporal components above a relative threshold of 10^{-3}, compared with three for scalar-time conditioning; its effective rank also increases from 1.07 to 1.65. The slower spectral decay in panel (c) shows that the proposed time code preserves multiple temporal frequencies and nonlinear duration trends instead of collapsing lead time onto an almost one-dimensional direction. This richer coordinate system is important because hours without direct forecast supervision must be represented between supervised horizons without introducing horizon-specific heads. To test the coarse spatial branch at its native scale, we apply the same 2×2 average pooling used by that branch to the input field and then restore the original grid before forecasting. The complete model obtains lower Rel-L2 at every evaluated horizon: 0.2603, 0.2611, 0.2683, and 0.2815 at 6, 12, 18, and 24 hours, respectively, compared with 0.2623, 0.2633, 0.2708, and 0.2829 after removing the multiscale branch. Panel (d) therefore indicates that the coarse branch contributes useful large-scale atmospheric context that the full-resolution path alone does not recover from spatially degraded inputs. Finally, the duration gate is essential near the temporal origin. The complete model has an effectively exact zero-time error of 3.5×10^{-8} and reduces the one-hour error from 0.1049 to 0.0515. Panel (e) directly tests the boundary condition of a continuous flow: without the gate, the model can introduce a finite artificial jump even when almost no time has elapsed; with the gate, the initial state is preserved exactly and the short-time trajectory starts from the correct atmosphere. This verifies that enforcing $\phi(z_0, 0) = z_0$ is a structural requirement for faithful continuous lead-time queries, rather than a cosmetic parameterization choice.

Table 8 tests whether composition-aware learning also improves pointwise forecasts. Removing both objectives yields small, mixed RMSE/ACC changes; its clearer effect is the lower direct-versus-composed disagreement in Figure 5(c) and Figure 6(a)–(b).

F.2 MACRO-AVERAGE CONTINUOUS LEAD-TIME FORECAST CURVE

Figure 7 reports the macro-average continuous lead-time curve corresponding to the main forecast table and complements the variable-wise physical-unit RMSE curves in Figure 3.

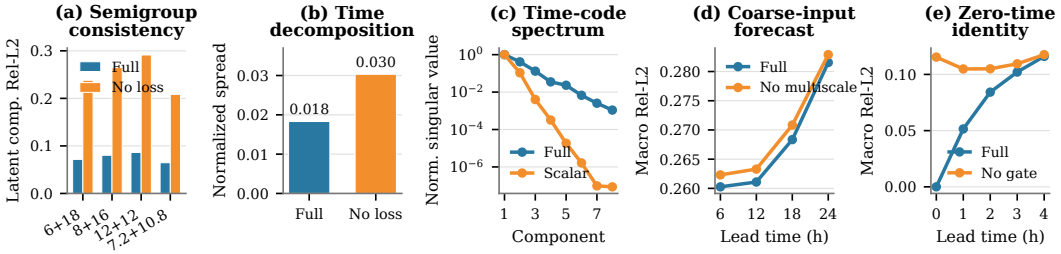

Figure 6: Mechanism-aligned ablations of PhiCast. (a) Latent consistency between direct and composed finite-time queries with and without the semigroup objective. (b) Normalized across-decomposition spread of decoded forecasts for the same 24-hour endpoint, comparing the complete model with an ablation trained without the composition-forecast objective while retaining latent semigroup supervision. The five evaluated paths are the direct 24-hour query and the 12+12, 6+18, 3+21, and 8+8+8 hour decompositions. (c) Singular-value spectrum of the learned continuous time embedding compared with scalar-time conditioning. (d) Forecast error from inputs coarsened by the 2×2 pooling scale used in the multiscale branch. (e) Zero- and short-time forecast errors with and without the identity-preserving duration gate. Lower is better in (a), (b), (d), and (e); the slower spectral decay in (c) indicates a richer temporal representation.

Table 8: Pointwise forecast accuracy for PhiCast and the No comp. ablation, trained without both composition-aware objectives. Both use seed-42 checkpoints and all 1,456 forecast origins in 2018. Seen leads are 6, 12, 18, and 24 hours; unseen leads are the other integer hours in 1–24. Latitude-weighted RMSE is in native physical units (K for T_{2m}, m s^{-1} for winds, and m^2 s^{-2} for Z_{500}); ACC uses the 2000–2015 calendar-hour climatology. Lower RMSE and higher ACC are better.

		RMSE ↓				ACC ↑			
Leads	Model	T_{2m}	u_{10}	v_{10}	Z_{500}	T_{2m}	u_{10}	v_{10}	Z_{500}
Seen	PhiCast	1.2475	1.5270	1.5578	147.49	0.88959	0.92663	0.92629	0.98403
	No comp.	1.2514	1.5222	1.5540	146.86	0.88897	0.92710	0.92666	0.98417
Unseen	PhiCast	1.2241	1.4101	1.4423	132.76	0.89521	0.93783	0.93721	0.98710
	No comp.	1.2342	1.4080	1.4414	133.08	0.89368	0.93803	0.93729	0.98704

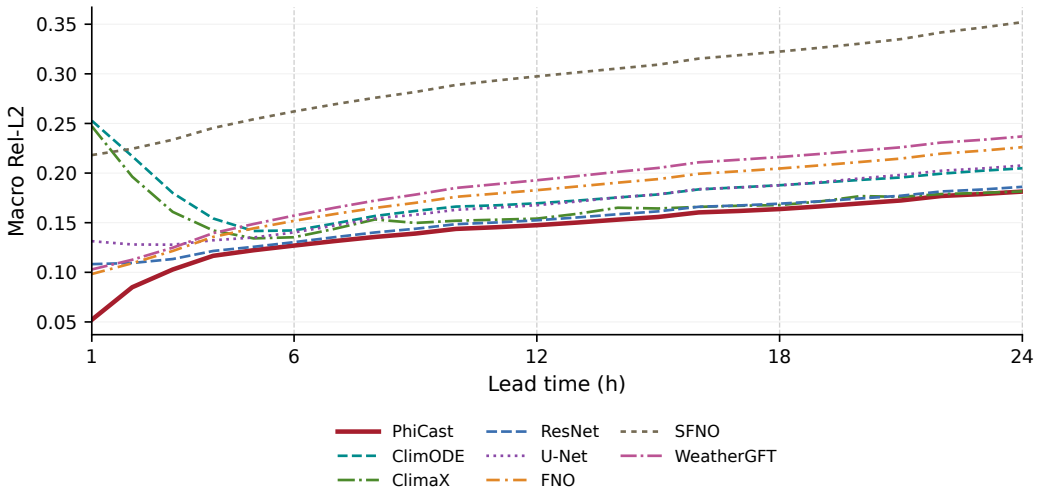

Figure 7: Macro Rel-L2 for PhiCast and seven baseline seed-42 checkpoints at every integer lead from 1 to 24 hours. Vertical dashed lines mark direct forecast supervision at 6, 12, 18, and 24 hours. All forecasts start from the same observed origin. ClimODE integrates its ODE, and the remaining models use direct mappings. No temporal interpolation or autoregressive rollout is used.

F.3 INDIVIDUAL RANDOM-SEED RESULTS

Every model in the main comparison was trained independently with seeds 42, 43, and 44 under the identical data split, update budget, validation schedule, early-stopping rule, and complete-2018 evaluation protocol. Table 1 reports the mean and sample standard deviation across these three runs. Tables 9 and 10 list the individual results, confirming that the aggregate ranking is not caused by one favorable initialization.

Table 9: Individual-seed results at the forecast-supervised leads (6, 12, 18, and 24 hours). RMSE is reported in physical units and ACC uses the training-period calendar-hour climatology.

		RMSE ↓				ACC ↑			
Model	Seed	T_{2m}	u_{10}	v_{10}	Z_{500}	T_{2m}	u_{10}	v_{10}	Z_{500}
PhiCast	42	1.248	1.527	1.558	147.49	0.88959	0.92663	0.92629	0.98403
PhiCast	43	1.238	1.527	1.557	147.16	0.89088	0.92667	0.92634	0.98410
PhiCast	44	1.253	1.542	1.573	149.48	0.88837	0.92516	0.92485	0.98359
ClimODE	42	1.803	1.742	1.761	198.69	0.78674	0.90338	0.90477	0.97087
ClimODE	43	1.835	1.737	1.756	197.65	0.78182	0.90391	0.90528	0.97120
ClimODE	44	1.825	1.739	1.757	200.64	0.78292	0.90365	0.90513	0.97035
ClimaX	42	1.419	1.574	1.599	139.79	0.85328	0.92181	0.92220	0.98569
ClimaX	43	1.420	1.575	1.599	139.56	0.85319	0.92177	0.92220	0.98575
ClimaX	44	1.423	1.577	1.601	139.65	0.85252	0.92159	0.92205	0.98573
ResNet	42	1.492	1.575	1.599	173.11	0.84730	0.92174	0.92213	0.97799
ResNet	43	1.493	1.581	1.604	172.87	0.84696	0.92109	0.92164	0.97804
ResNet	44	1.488	1.579	1.602	173.90	0.84757	0.92129	0.92186	0.97781
U-Net	42	1.565	1.746	1.761	191.51	0.82937	0.90292	0.90471	0.97294
U-Net	43	1.583	1.747	1.760	190.91	0.82655	0.90282	0.90482	0.97311
U-Net	44	1.592	1.742	1.757	192.78	0.82447	0.90336	0.90516	0.97257
FNO	42	1.622	1.886	1.929	215.95	0.81735	0.88564	0.88444	0.96548
FNO	43	1.629	1.886	1.926	222.70	0.81702	0.88575	0.88483	0.96324
FNO	44	1.621	1.883	1.922	217.36	0.81670	0.88611	0.88533	0.96500
SFNO	42	2.670	2.933	3.180	336.90	0.61507	0.72294	0.68098	0.91400
SFNO	43	2.613	2.904	3.146	328.59	0.63563	0.73075	0.69072	0.91876
SFNO	44	2.621	2.927	3.179	336.33	0.62699	0.72641	0.68338	0.91476
WeatherGFT	42	1.807	2.057	1.959	253.15	0.79168	0.86491	0.88095	0.95298
WeatherGFT	43	1.765	2.042	1.952	245.63	0.79935	0.86717	0.88201	0.95567
WeatherGFT	44	1.754	2.027	1.960	241.47	0.80139	0.86905	0.88119	0.95706

Table 10: Individual-seed results at the 20 integer leads without direct forecast supervision in 1–24 hours. RMSE is reported in physical units and ACC uses the training-period calendar-hour climatology.

		RMSE ↓				ACC ↑			
Model	Seed	T_{2m}	u_{10}	v_{10}	Z_{500}	T_{2m}	u_{10}	v_{10}	Z_{500}
PhiCast	42	1.224	1.410	1.442	132.76	0.89521	0.93783	0.93721	0.98710
PhiCast	43	1.215	1.410	1.442	132.26	0.89631	0.93786	0.93727	0.98718
PhiCast	44	1.226	1.422	1.454	133.78	0.89469	0.93680	0.93620	0.98688
ClimODE	42	1.881	1.773	1.834	194.00	0.77786	0.89988	0.89655	0.97221
ClimODE	43	1.888	1.775	1.840	193.92	0.77719	0.89966	0.89583	0.97227
ClimODE	44	1.907	1.768	1.832	195.37	0.77330	0.90047	0.89668	0.97184
ClimaX	42	1.668	1.614	1.709	160.53	0.80371	0.91770	0.91088	0.98111
ClimaX	43	1.662	1.603	1.685	155.82	0.80532	0.91890	0.91340	0.98224
ClimaX	44	1.660	1.606	1.688	156.35	0.80544	0.91859	0.91316	0.98209
ResNet	42	1.535	1.479	1.509	160.81	0.84297	0.93137	0.93100	0.98106
ResNet	43	1.525	1.487	1.517	159.93	0.84432	0.93062	0.93029	0.98125
ResNet	44	1.545	1.482	1.512	161.12	0.84080	0.93109	0.93076	0.98099
U-Net	42	1.728	1.646	1.674	179.09	0.80341	0.91427	0.91436	0.97640
U-Net	43	1.691	1.629	1.647	176.37	0.81079	0.91602	0.91726	0.97710
U-Net	44	1.716	1.638	1.661	179.18	0.80576	0.91513	0.91577	0.97637
FNO	42	1.603	1.744	1.788	195.55	0.82481	0.90319	0.90160	0.97178
FNO	43	1.601	1.740	1.782	200.20	0.82638	0.90367	0.90229	0.97039
FNO	44	1.603	1.739	1.780	196.78	0.82434	0.90372	0.90258	0.97141
SFNO	42	2.590	2.746	3.030	319.36	0.63718	0.75750	0.71087	0.92287
SFNO	43	2.531	2.709	2.974	306.82	0.65753	0.76603	0.72382	0.92926
SFNO	44	2.540	2.738	3.030	317.69	0.64887	0.76089	0.71274	0.92403
WeatherGFT	42	1.792	1.898	1.822	227.35	0.80009	0.88606	0.89792	0.96219
WeatherGFT	43	1.793	1.886	1.816	220.86	0.79999	0.88764	0.89865	0.96424
WeatherGFT	44	1.790	1.874	1.823	217.94	0.80069	0.88907	0.89802	0.96512

F.4 Mechanistic Role of the Initial-State Tendency

We next examine how the two terms in Equation 5 participate in the queried latent evolution. At five representative lead times, Figure 8(a) reports the root-mean-square magnitude of the transient displacement $\phi_{\mathrm{transient}}(z_0, \bar{\tau}) - z_0$ and the initial-state tendency $\bar{\tau}\sigma(\alpha)V(z_0)$, each normalized by the RMS magnitude of z_0. The displayed percentage is the latter magnitude divided by the sum of the two branch magnitudes; it is a scale diagnostic rather than a decomposition of forecast error. The initial-state-tendency share grows from 6.6% at 1 hour to 18.2% at 24 hours, while the transient path remains the dominant nonlinear component throughout the interval. This behavior matches the intended separation: the transient flow captures bounded lead-dependent adjustment, whereas the initial-state tendency supplies the directional component derived from $V(z_0)$.

To test whether that growing term affects the decoded forecast, we perform a frozen-checkpoint intervention on the same 32 fixed 2018 origins. We evaluate the trained model once in full and once with only $\bar{\tau}\sigma(\alpha)V(z_0)$ removed; all learned weights, inputs, transient states, and decoder operations remain unchanged. Figure 8(b) shows that the intervention increases mean Rel-L2 by 0.17% at 1 hour, 2.36% at 12 hours, and 4.11% at 24 hours, with a 2.54% increase averaged over all 1–24 hour queries. Because the intervened network is not retrained, this is a mechanism diagnostic rather than a capacity-matched ablation. Nevertheless, the horizon-dependent degradation shows that the initial-state tendency contributes increasingly as the queried interval grows.

F.5 Inference Efficiency

We measure batch-one FP32 inference on the same NVIDIA RTX 4070 Laptop GPU after five warm-up iterations. Single-query latency is averaged over 20 repetitions and includes the mapping from an input weather field to one forecast; the 1–24 hour suite is averaged over three repetitions. Each model is loaded and measured independently. The suite queries all 24 horizons from the same origin. PhiCast caches one encoding, while ClimODE integrates once to 24 hours, retains the complete augmented state at each hour, and applies its output correction at each requested time. We also report the less efficient protocol that restarts ClimODE independently for every horizon. The other direct baselines evaluate their predictor at each horizon. Table 11 and Figure 9 show that compact

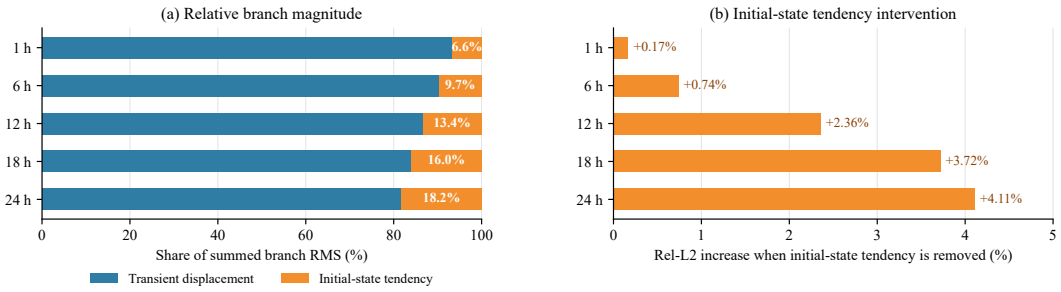

Figure 8: Initial-state-tendency mechanism analysis using the trained PhiCast checkpoint and 32 fixed 2018 forecast origins. Horizontal bars report five representative lead times. Panel (a) compares the transient displacement with the initial-state tendency $\bar{\tau}\sigma(\alpha)V(z_0)$ as shares of their summed latent RMS magnitudes. Panel (b) reports the percentage increase in macro Rel-L2 when the initial-state tendency is removed at inference while every learned parameter remains frozen. The first panel describes branch scale rather than additive forecast contribution. The second is a frozen-checkpoint intervention rather than a separately trained ablation.

convolutional and operator baselines have the lowest absolute latency. PhiCast is not the fastest model, but its single-query latency remains approximately constant with forecast duration. Cached ClimODE narrows the full-suite latency difference considerably relative to repeated integration, so the efficiency advantage claimed here is direct, solver-free access to an individual requested horizon rather than a large universal speedup for dense query suites.

Table 11: End-to-end GPU inference latency in milliseconds. The 1–24 hour suite reports 24 forecasts from one origin; lower is better. ClimODE's main row uses one integration that stores all hourly augmented states; its independent-query suite is shown separately. The two ClimODE suites produce identical outputs.

Model	1 h	6 h	12 h	24 h	1–24 h suite
PhiCast	25.93	26.15	23.92	24.14	398.41
ClimODE	26.42	90.39	176.01	345.02	504.96
ClimODE (independent suite)	–	–	–	–	4514.57
ClimaX	10.98	10.79	10.96	11.02	262.93
ResNet	6.20	7.10	8.61	7.42	160.36
U-Net	7.02	6.90	7.06	6.86	161.98
FNO	4.59	4.79	5.17	4.83	99.02
SFNO	5.56	5.36	5.49	5.58	131.09
WeatherGFT	78.43	77.39	78.50	75.65	1868.83

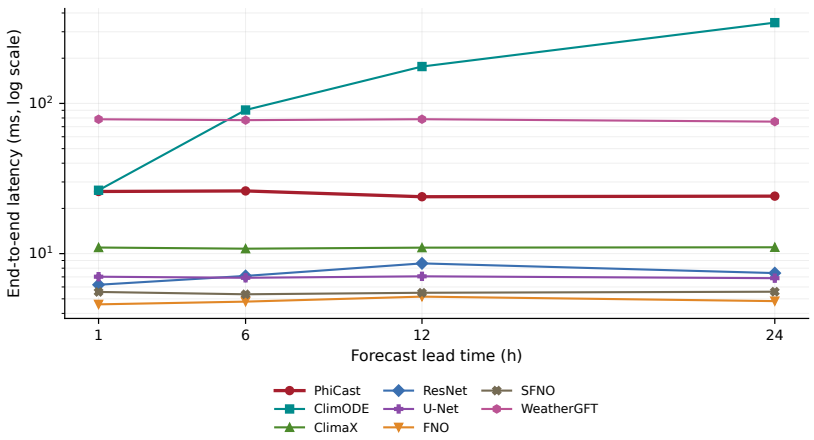

Figure 9: End-to-end single-query latency for PhiCast and seven baselines. The logarithmic vertical axis accommodates the inexpensive direct operators and the duration-dependent ODE integration of ClimODE. Each model is timed independently under the protocol of Table 11.

F.6 Bootstrap Confidence Intervals

Tables 12 and 13 provide supplementary scale-normalized Rel-L2 uncertainty estimates. These complement the physical-unit RMSE and ACC results in Table 1. Intervals are obtained from 10,000 seven-day block-bootstrap draws over the 1,456 six-hourly forecast origins in 2018.

Table 12: Mean Rel-L2 and 95% confidence intervals for temperature and wind variables.

Model	T_{2m}	u_{10}	v_{10}
PhiCast	**0.004176 [0.004142, 0.004211]**	**0.25236 [0.24980, 0.25484]**	**0.31431 [0.31115, 0.31749]**
ClimODE	0.006409 [0.006358, 0.006462]	0.31631 [0.31316, 0.31937]	0.39514 [0.39140, 0.39890]
ResNet	0.005244 [0.005203, 0.005287]	0.26632 [0.26368, 0.26889]	0.33083 [0.32757, 0.33409]
ClimaX	0.005567 [0.005520, 0.005615]	0.28822 [0.28562, 0.29075]	0.36615 [0.36303, 0.36931]
U-Net	0.005844 [0.005803, 0.005888]	0.29611 [0.29303, 0.29912]	0.36670 [0.36319, 0.37024]
FNO	0.005510 [0.005464, 0.005558]	0.31255 [0.30918, 0.31585]	0.39035 [0.38654, 0.39416]
SFNO	0.008895 [0.008818, 0.008971]	0.49458 [0.48929, 0.49971]	0.66569 [0.65964, 0.67183]
WeatherGFT	0.006116 [0.006067, 0.006168]	0.33919 [0.33554, 0.34273]	0.39758 [0.39389, 0.40137]

Table 13: Mean Rel-L2 and 95% confidence intervals for geopotential and the four-variable macro average.

Model	Z_{500}	Macro
PhiCast	**0.002305 [0.002273, 0.002339]**	**0.14329 [0.14211, 0.14448]**
ClimODE	0.003419 [0.003374, 0.003465]	0.18032 [0.17891, 0.18176]
ResNet	0.002835 [0.002798, 0.002874]	0.15131 [0.15010, 0.15253]
ClimaX	0.002706 [0.002677, 0.002735]	0.16566 [0.16454, 0.16683]
U-Net	0.003145 [0.003103, 0.003188]	0.16795 [0.16659, 0.16935]
FNO	0.003418 [0.003366, 0.003473]	0.17796 [0.17647, 0.17950]
SFNO	0.005718 [0.005637, 0.005803]	0.29372 [0.29135, 0.29616]
WeatherGFT	0.003959 [0.003903, 0.004017]	0.18671 [0.18520, 0.18825]

F.7 Lagged-Forecast Fusion under Different Initialization Sets

We additionally probe sensitivity to the available initialization set using the existing single-state checkpoints. For each of 256 target times, forecasts initialized from the listed historical ages are propagated directly to a common target and combined by an unweighted arithmetic mean. This is a lagged-forecast late-fusion test rather than native variable-length or asynchronous observation encoding. Table 14 shows that PhiCast obtains the lowest Rel-L2 in all four tested averaging configurations. Because changing the available ages also changes the relative weight assigned to the most

recent, usually most accurate initialization, differences between columns should not be interpreted as an isolated measure of missing-observation robustness.

Table 14: Rel-L2 for lagged-forecast fusion under four initialization sets. Column headers list the ages, in hours before the common target, of the single-state forecasts included in the unweighted mean. The same 256 target times are used in every column. This experiment does not use native asynchronous or missing-observation encoding. Lower is better.

Model	Regular: 24/18/12/6	Irregular: 24/18/6	Missing middle: 24/6	Missing latest: 24/18/12
PhiCast	**0.1408**	**0.1432**	**0.1393**	**0.1540**
ClimODE	0.1607	0.1629	0.1570	0.1766
ResNet	0.1443	0.1467	0.1425	0.1583
ClimaX	0.1471	0.1489	0.1452	0.1588
U-Net	0.1602	0.1628	0.1569	0.1767
FNO	0.1747	0.1773	0.1705	0.1930
SFNO	0.2605	0.2631	0.2541	0.2910
WeatherGFT	0.1766	0.1797	0.1730	0.1975

F.8 Qualitative Forecast Fields

Figure 10 compares a fixed 2018 case at the 15-hour lead time, which receives no direct forecast supervision. All methods start from the same observed atmospheric state and are evaluated independently at the requested horizon. ClimODE uses its internal ODE solver, and the remaining models use a direct mapping. Each variable column has a shared color scale across models, and the per-panel Rel-L2 values complement visual inspection of spatial structures and amplitudes.

F.9 Complete 1–24 Hour Forecast Sequence

Figure 11 visualizes all 24 direct hourly queries from the same encoded 2018 atmospheric state. Each variable is displayed as a 4×6 grid ordered from 1 to 24 hours. The number above each field gives the lead time and its latitude-weighted Rel-L2 for this case. Red titles and borders identify the four supervised horizons at 6, 12, 18, and 24 hours; the remaining 20 fields are queried at lead times absent from direct forecast supervision. The gradual spatial evolution across adjacent supervised and unseen times provides a qualitative view of the continuously queryable trajectory learned by PhiCast.

F.10 Zero-Shot Spatial-Resolution Generalization

We evaluate whether the learned temporal dynamics transfer to a finer spatial grid without fine-tuning. The seed-42 checkpoints trained at 5.625° are applied directly to native 2.8125° WeatherBench fields. The evaluation covers T_{2m}, u_{10}, v_{10}, and Z_{500} over the complete 2018 test year, with 1,456 six-hourly forecast origins and independent queries at every integer lead time from 1 to 24 hours. Normalization remains fixed to the training statistics, while ACC uses the native-2.8125° 2000–2015 calendar-hour climatology. For ClimaX, the learned positional embedding is bicubically interpolated from the training patch grid to the target patch grid using its released resolution-adaptation routine, while all other learned weights remain fixed. This supplemental comparison focuses on PhiCast, ClimaX, and ClimODE because they represent the proposed coherent finite-time formulation, the strongest weather-specific direct-query baseline, and the weather-specific continuous-time baseline, respectively. The experiment is therefore designed to test whether the temporal modeling advantage survives a spatial-resolution shift, rather than to repeat the full main-table comparison with generic convolutional and operator predictors. Table 15 reports the resulting four-variable evaluation. PhiCast gives the best macro normalized RMSE and macro ACC, with the largest gains on temperature and geopotential; ClimaX remains stronger for v_{10} RMSE. Thus, the transfer advantage is broad but not uniform across every variable.

Figure 12 resolves the aggregate result across lead time. PhiCast degrades smoothly as the requested horizon increases and retains the lowest macro Rel-L2 over most of the interval. The variable-wise panels expose the complementary behavior hidden by the macro average: PhiCast is consistently strongest for T_{2m}, becomes preferable for u_{10} at longer leads, and is competitive for Z_{500}, whereas

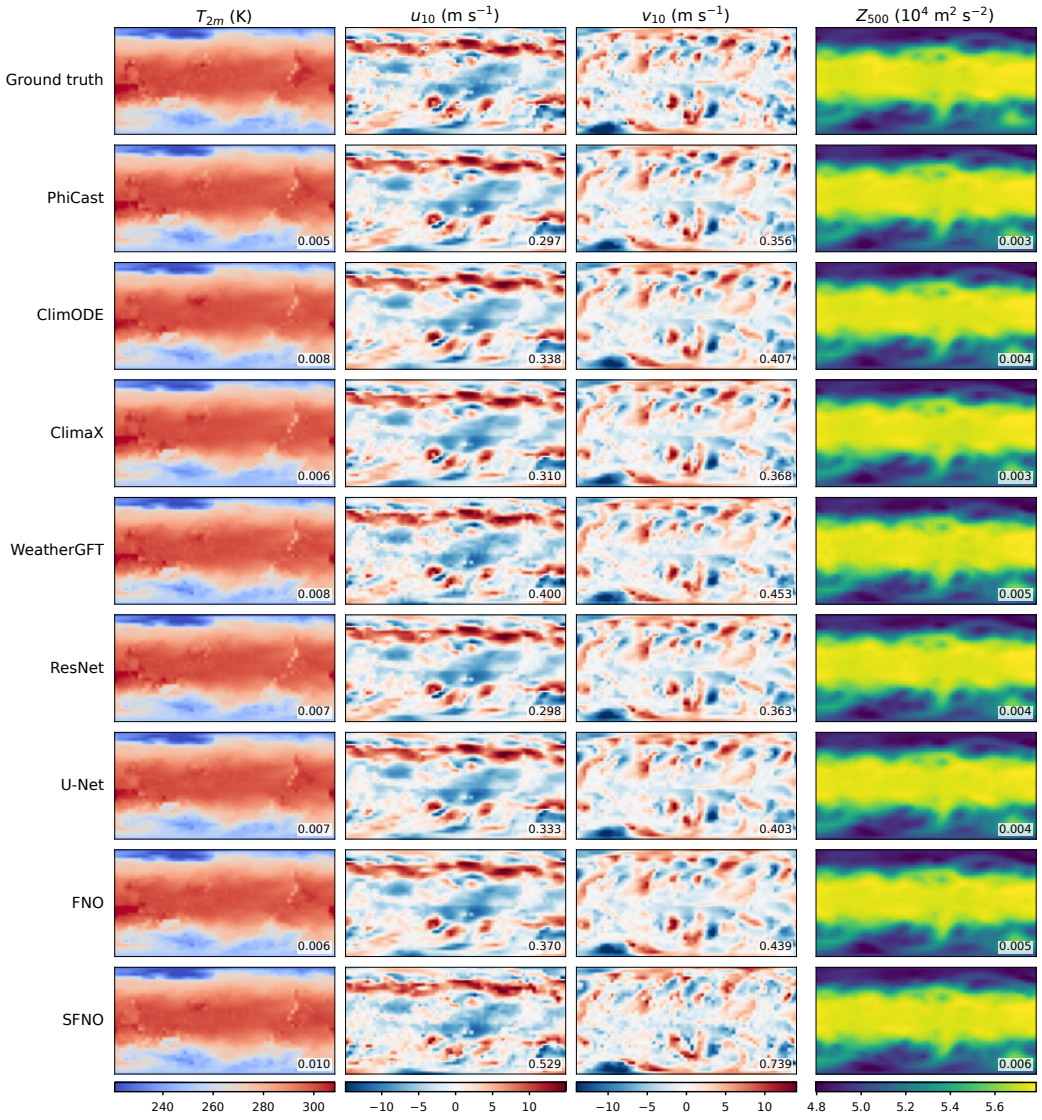

Figure 10: Forecast fields for a fixed 2018 test case at the unsupervised 15-hour lead. Columns show T_{2m}, u_{10}, v_{10}, and Z_{500}; rows show ground truth followed by PhiCast and all seven baselines. Each variable uses one common physical-unit color scale across rows. The Rel-L2 for the displayed sample and variable is shown in each predicted panel.

Table 15: Zero-shot transfer from $5.625°$ training to native $2.8125°$ evaluation on the complete 2018 test set over 1–24 hour queries. RMSE is reported in physical units (K for T_{2m}, m s^{-1} for wind, and m^2 s^{-2} for Z_{500}). NRMSE is the macro average after division by each variable's training-set standard deviation. Higher ACC and lower RMSE/NRMSE are better. Best values are bold.

Model	T_{2m} RMSE	T_{2m} ACC	u_{10} RMSE	u_{10} ACC	v_{10} RMSE	v_{10} ACC	Z_{500} RMSE	Z_{500} ACC	Macro NRMSE	Macro ACC
PhiCast	**1.866**	**0.785**	**2.413**	**0.813**	2.672	0.774	**352.23**	**0.911**	**0.2957**	**0.8206**
ClimaX	3.207	0.531	2.465	0.801	**2.461**	**0.809**	380.08	0.892	0.3050	0.7584
ClimODE	2.521	0.666	2.641	0.770	3.079	0.688	406.70	0.878	0.3390	0.7505

ClimaX is stronger for v_{10} in both aggregate RMSE and ACC. This per-lead evaluation prevents a favorable result at one horizon from dominating the conclusion.

Figure 13 provides a spatial comparison at the unseen 15-hour lead. All models start from the same 2018 atmospheric state and each row shares one physical-unit color scale. PhiCast preserves the

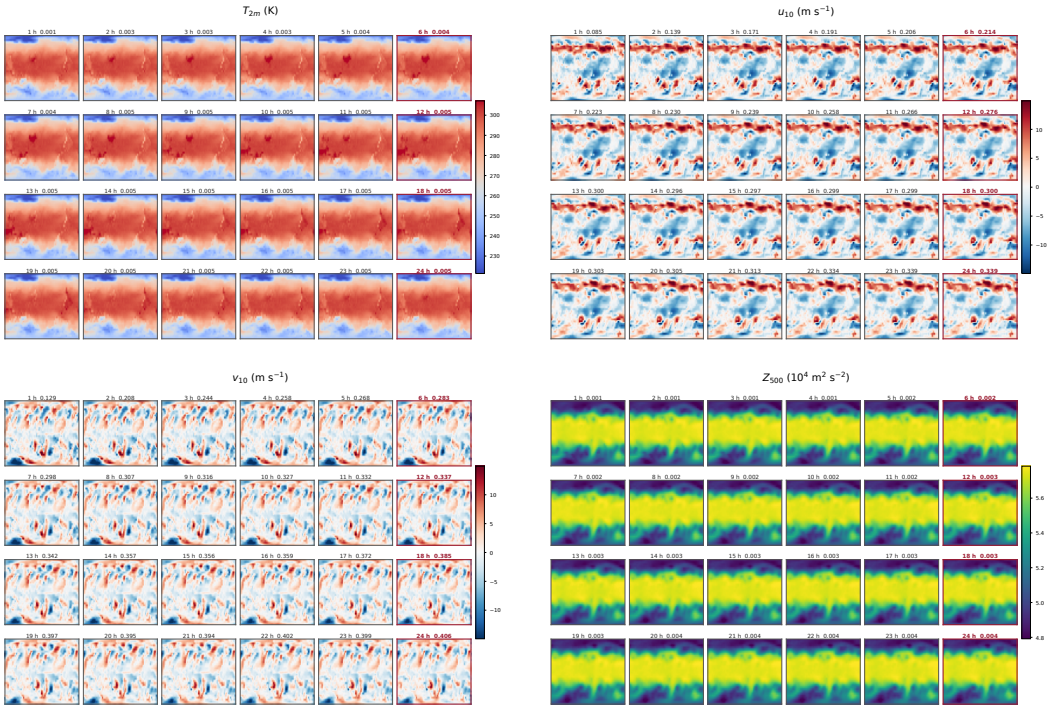

Figure 11: Complete PhiCast prediction sequence for one fixed 2018 forecast origin. The four panels show all direct 1–24 hour queries for T_{2m}, u_{10}, v_{10}, and Z_{500}. Each panel uses a common physical-unit color scale across its 24 fields. Titles report lead time and case-specific Rel-L2; directly supervised horizons are highlighted in red, while the remaining horizons receive no direct forecast supervision. No intermediate prediction is fed back to the model.

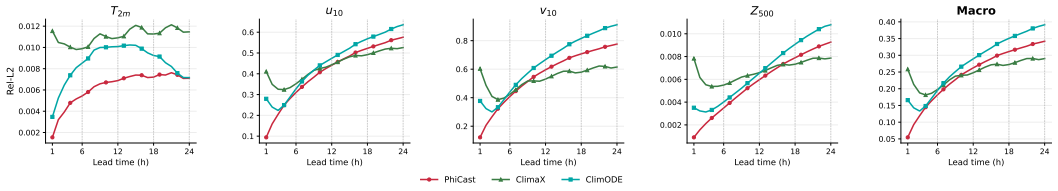

Figure 12: Zero-shot 2.8125° Rel-L2 as a function of forecast lead time for the four variables and their macro average. All checkpoints are trained only at 5.625° and evaluated without fine-tuning. Vertical dashed lines mark the directly supervised lead times at 6, 12, 18, and 24 hours; the remaining integer lead times receive no direct forecast supervision.

large-scale temperature and geopotential structure, while the wind panels reflect the quantitative trade-off in Table 15.

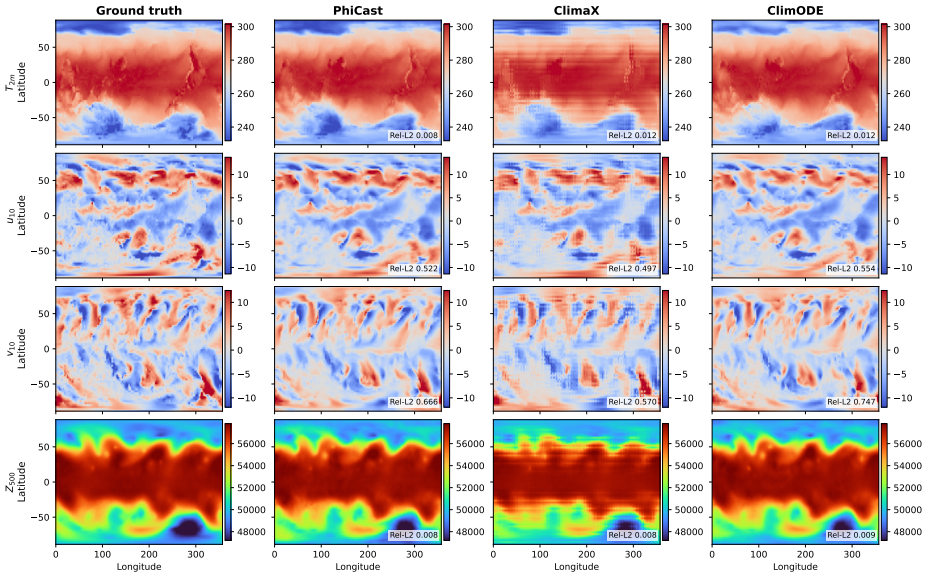

Figure 13: Native-2.8125° forecast fields for a fixed 2018 case at the unseen 15-hour lead. Rows show T_{2m}, u_{10}, v_{10}, and Z_{500}; columns show ground truth, PhiCast, ClimaX, and ClimODE. Each row uses a shared physical-unit color scale, and each predicted panel reports its case-specific latitude-weighted Rel-L2. No model is fine-tuned at 2.8125°.

G Limitations and Future Work

The experiments are intentionally centered on direct 1–24 hour queries at 5.625° resolution with four representative atmospheric variables. This setting makes it possible to evaluate every integer lead time over the complete 2018 test year and to isolate the behavior of the continuous-query mechanism under a consistent protocol. Given the available computational resources, we prioritize this dense temporal evaluation and three-seed main comparison over a substantially larger collection of pressure levels, forecast ranges, and high-resolution training runs. The resulting conclusions are therefore scoped to continuously queried short-range forecasts within the modeled interval, while the zero-shot 2.8125° experiment provides an initial indication of transfer beyond the training grid.

These choices leave several natural opportunities for extension rather than changing the central formulation. The present map conditions on elapsed duration and the observed state rather than explicit solar-forcing metadata; absolute-time or insolation features can be incorporated when scaling to longer query intervals. Future work can also extend the same one-encoding continuous evolution to multilevel atmospheric states and higher-resolution training as additional computational resources become available. The continuously queried latent trajectory can further be combined with probabilistic or ensemble decoding and observation-assimilation interfaces, extending the present deterministic formulation toward uncertainty-aware forecasts and more diverse atmospheric inputs.

H AI Use Statement

In this work, we used generative AI tools solely for language polishing and improving the clarity and fluency of the manuscript. The scientific ideas, methodology, experiments, analysis, and conclusions were developed entirely by the authors. All AI-assisted revisions were carefully reviewed and approved by the authors. We take full responsibility for the final content of this work.