WAVEGSSM: GRAPH WAVE STATE SPACE MODELS FOR PROPAGATING SPATIO-TEMPORAL PATTERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal graph models typically encode each snapshot with a GNN and then connect the resulting representations through a temporal module. This space-then-time design is effective, yet it does not explicitly represent how a pattern moves across the graph. We show empirically that, for a propagating process, the same present field can lead to different futures when its recent rate of change differs, motivating an explicit representation of motion in the predictive state. We introduce WaveGSSM, a second-order graph state-space model that maintains two coupled latent states at each node, one for the current pattern and one for its temporal rate of change. A graph-wave transition updates the motion state through graph interactions and uses it to advance the pattern state, coupling spatial propagation and temporal evolution within a single rollout. We evaluate WaveGSSM on four temporal-graph benchmarks and global weather forecasting. It consistently achieves the best mean performance across the temporal-graph benchmarks and reduces the geopotential RMSE by 20.2% on average for 1- to 5-day weather forecasts relative to a backbone-matched snapshot model, while better preserving large-scale atmospheric patterns.

1 INTRODUCTION

Spatio-temporal graph learning commonly combines graph-based spatial encoding with recurrence (Cini et al., 2023), attention (Karmim et al., 2024), or state-space models (Li et al., 2024) for temporal modeling. Most methods therefore first encode each graph snapshot and then connect the resulting representations across time (Wang et al., 2025b). However, this space-then-time factorization does not explicitly represent how a pattern moves through the graph. For propagating processes, the same spatial distribution of activation can lead to different futures when its rate of change, or velocity, differs. Existing models may infer this velocity from temporal history, but it remains implicit rather than being represented in the graph state. Errors in the inferred velocity can accumulate over recursive forecasts and shift the predicted pattern away from its correct location. This raises a natural question. Can the activation pattern and its velocity be represented jointly, so that propagation through graph edges and evolution through time are advanced within the same transition?

Answering this question requires a predictive representation that captures both the present pattern and how it is moving. Figure 1 shows why this information matters in a simple synthetic forecasting setting. At time t, the two trajectories in panel (a) have the same graph and identical node patterns, but their recent histories indicate motion in opposite directions. They therefore evolve to different future locations, as shown in panels (b) and (c). For each trajectory, all history-aware models receive the same three-frame history ending at time t. We ask whether they can use this history to predict the two different futures. We compare two space-then-time baselines, GCN-LSTM and GCN-SSM, against WaveGSSM and a variant without the motion state. GCN-LSTM and GCN-SSM combine snapshot-level graph encoding with a temporal module. They capture part of the motion information, while the full WaveGSSM most accurately predicts both motion-dependent futures. This motivates representing pattern and motion jointly within the evolving graph state rather than leaving motion implicit in temporal memory. We also provide a theoretical analysis of this state requirement in Observation 1.

The same state-representation problem also appears in real medium-range weather forecasting, where coherent atmospheric patterns move over several days while changing their strength and

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.

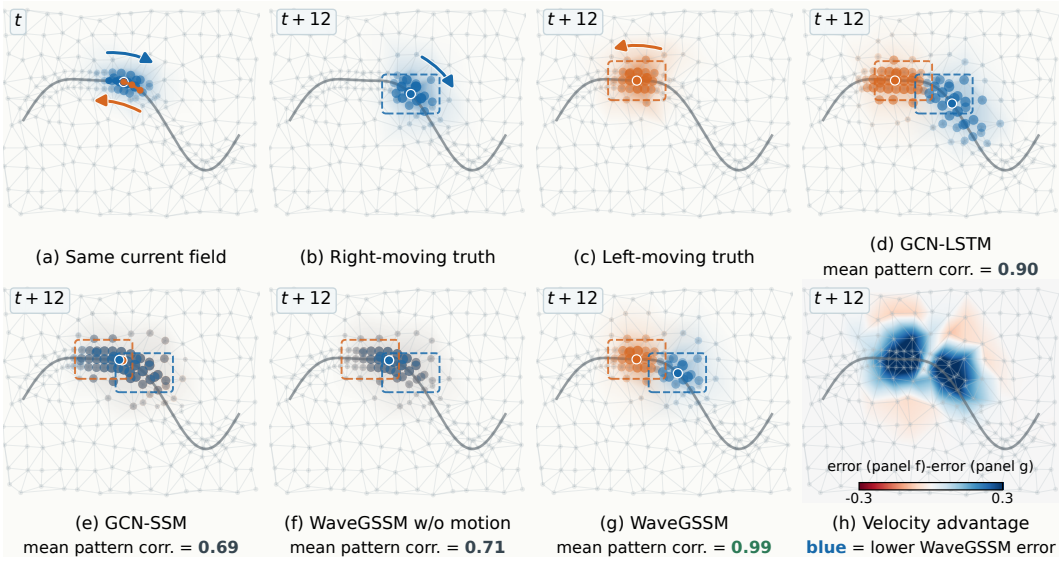


Figure 1: A synthetic experiment shows that the present graph field alone does not determine the future of a propagating pattern. Paired trajectories, shown in blue and orange, share the same graph and present field at time t but arrive with opposite recent motion (a), leading to different future locations at $t + 12$ (b, c). Panels (d)–(g) compare history-aware baselines, WaveGSSM without the motion state, and the full WaveGSSM. The values below panels (d)–(g) report the mean Pearson pattern correlation over the two motion directions at $t + 12$ for the displayed held-out pair. Panel (h) visualizes the error reduction from carrying the motion state. WaveGSSM best preserves the two motion-dependent futures, while aggregate future-separation and centroid-error results are reported in Appendix Figure 6.

shape (Kornhuber et al., 2020; Wang et al., 2025a). A forecast can therefore produce a plausible large-scale pattern while placing it at the wrong location. Figure 2 illustrates this failure mode in a five-day forecast of 500 hPa geopotential height, or z500, a standard field for describing large-scale atmospheric circulation. A backbone-matched GraphCast-style predictor (Lam et al., 2023) captures the broad Atlantic-to-Europe pattern but displaces the ridge–trough structure relative to ERA5, while WaveGSSM keeps the predicted structure more closely aligned. Hence, this example further motivates modeling not only the present pattern, but also its motion.

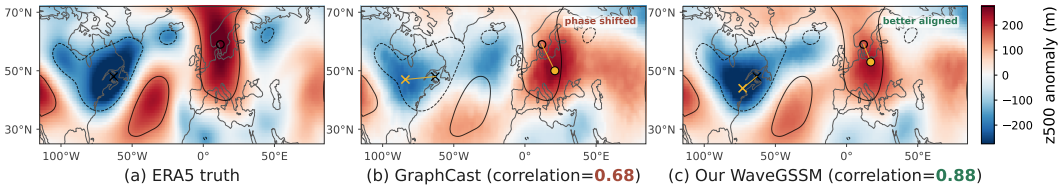


Figure 2: A plausible weather forecast can still place a coherent pattern at the wrong location. (a) The heatmap shows the ERA5 reanalysis as the evaluation ground truth of predicting five-day 500 hPa geopotential-height anomaly, or z500. (b) Prediction by GraphCast. (c) Prediction by our WaveGSSM. Yellow markers show the predicted centers and displacement to the ground-truth (marked with black contours and markers). GraphCast* captures the broad pattern but shifts it relative to ERA5, whereas WaveGSSM keeps the structure better aligned, with correlation increasing from 0.68 to 0.88.

We therefore introduce WaveGSSM, a wave-inspired Graph State Space Model for propagating spatio-temporal signals. Following second-order wave dynamics, each node maintains two coupled latent states. The pattern state represents the current latent pattern, while the motion state represents its temporal rate of change. A graph-wave transition updates the motion state through graph interactions and then uses the updated motion to advance the pattern state. Spatial propagation and temporal

evolution are therefore coupled within a single state transition. WaveGSSM explicitly incorporates motion into the evolving graph state rather than leaving it to be inferred implicitly by a separate temporal module.

Our main contributions can be summarized as follows:

- We show that the present field can be an incomplete predictive state for propagating graph signals. Paired synthetic trajectories and a theoretical modal analysis show how different recent motion can produce different futures from the same present field, motivating the explicit representation of motion in the graph state.
- We introduce WaveGSSM, a wave-inspired graph state-space model with coupled pattern and motion states. Its second-order graph-wave transition advances the two states jointly, coupling spatial propagation and temporal evolution in a single transition.
- We demonstrate strong performance across four temporal-graph benchmarks and global weather forecasting. WaveGSSM achieves the best mean performance on all four temporal-graph benchmarks and reduces the geopotential RMSE by 20.2% on average for 1- to 5-day forecasts relative to a backbone-matched snapshot model, with larger pattern gains when propagation becomes more important.

2 Related Work

Spatio-temporal graph learning. Many spatio-temporal graph models factor spatial and temporal modeling by first encoding each graph snapshot and then connecting the resulting representations through recurrence Yu et al. (2018); Gao et al. (2023); Cini et al. (2023), attention Karmim et al. (2024), or state-space sequence models (Li et al., 2024). Recent work has also explored generative pre-training (Li et al., 2023b), continuous operators on product graphs (Einizade et al., 2024), and continual adaptation to streaming graphs (Chen & Liang, 2025). Across these formulations, motion is typically inferred from successive representations rather than represented as part of the evolving graph state. WaveGSSM instead couples the current latent pattern and its recent motion within a single dynamical state.

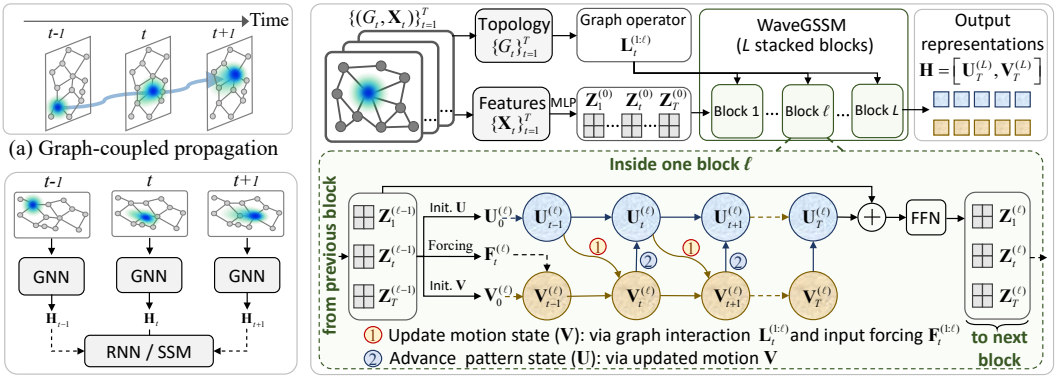
Graph state-space models. GraphSSM is the closest prior work in this family (Li et al., 2024). It brings state-space sequence modeling to temporal graphs by incorporating evolving graph structure into the temporal processor. Related approaches include GraphMamba for graph representation learning (Behrouz & Hashemi, 2024) and message-passing state-space models that combine sequence modeling with graph propagation (Ceni et al., 2025). Our new WaveGSSM differs in the structure of the state itself. Its latent state explicitly contains coupled pattern and motion states whose joint evolution determines the next graph transition.

Physics-inspired graph dynamics. Differential equations have been used to structure graph representation learning. GraphCON builds GNNs from graph-coupled oscillator dynamics (Rusch et al., 2022), Graph Wave Networks formulates message passing as wave propagation (Yue et al., 2025), and TDE-GNN models temporal dependencies with differential-equation-inspired graph dynamics (Eliasof et al., 2024). However, these approaches mainly use the dynamics to organize representation propagation within the graph network, rather than to model the motion state of an observed temporal graph process. In this paper, WaveGSSM instead uses graph-wave dynamics as the recurrent state transition along the observed sequence to solve the spatio-temporal pattern learning problem. Specifically, its pattern and motion states jointly encode the current latent pattern and its temporal change, allowing spatial propagation and temporal evolution to be advanced in the same transition.

3 Method

We consider an observed temporal graph sequence $\{(G_t, \mathbf{X}_t)\}_{t=1}^{T}$, where $G_t = (\mathcal{V}, E_t)$, $N = |\mathcal{V}|$, and $\mathbf{X}_t \in \mathbb{R}^{N \times F_{\text{in}}}$. The goal is to predict future node fields or labels from the observed sequence.

Overview. WaveGSSM first maps each snapshot to a hidden representation $\mathbf{Z}_t^{(0)} \in \mathbb{R}^{N \times d}$ and then processes the complete hidden sequence with L stacked graph-wave blocks. Block ℓ takes



(a) Graph-coupled propagation

(b) Space-first snapshot baseline

(c) Our WaveGSSM: Coupled Space-Time State Evolution

Figure 3: Motivation and architecture of WaveGSSM. (a) Spatio-temporal graph signals contain patterns that propagate through graph edges over time rather than evolving as independent snapshots. (b) Snapshot baselines first encode each snapshot with a GNN and then connect the representations with a temporal module such as an RNN or SSM, yielding a space-then-time pipeline. (c) WaveGSSM instead rolls forward coupled pattern and motion states within each graph-wave block. The pattern state $\mathbf{U}_t$ represents the current latent pattern, while the motion state $\mathbf{V}_t$ represents its temporal rate of change. Graph interaction and an input-dependent forcing update $\mathbf{V}_t$, which then advances $\mathbf{U}_t$, coupling spatial propagation and temporal evolution within a single rollout.

$\{\mathbf{Z}_t^{(\ell-1)}\}_{t=1}^T$ as input, performs a recurrent graph-coupled rollout over time, and returns $\{\mathbf{Z}_t^{(\ell)}\}_{t=1}^T$ to the next block. After the final block, a task-specific readout produces the prediction. Figure 3 (c) illustrates this overall architecture, and Algorithm 1 summarizes the forward pass.

The key difference from a space-then-time architecture (Figure 3 (b)) is the state evolution within each block. WaveGSSM maintains a *pattern state* $\mathbf{U}_t^{(\ell)}$, representing the current latent pattern, and a *motion state* $\mathbf{V}_t^{(\ell)}$, representing its temporal rate of change. Both states are initialized from the incoming hidden sequence $\{\mathbf{Z}_t^{(\ell-1)}\}_{t=1}^T$. At each temporal step, graph interactions and the current hidden input update the motion state, which then advances the pattern state. The evolved pattern is injected back into the hidden stream before being passed to the next block. Thus, $\mathbf{Z}_{1:T}^{(\ell)}$ propagates information across network depth, while $(\mathbf{U}_t^{(\ell)}, \mathbf{V}_t^{(\ell)})$ form the recurrent state evolving over time within a block.

We next define the graph propagation operator (Section 3.1), the coupled pattern–motion evolution (Section 3.2), and its realization as a residual WaveGSSM block with task readout (Section 3.3).

3.1 Graph Propagation Operator

The propagation operator acts on a support graph $G_\star = (\mathcal{V}, E_\star)$, which specifies the candidate edges for latent interaction. For fixed-mesh tasks, $G_\star$ is the prescribed mesh. For evolving graphs, it is constructed once from the observed structural sequence as detailed in Appendix B.2.

Let $\mathbf{B}_\star \in \mathbb{R}^{|E_\star| \times N}$ be an incidence matrix of the support graph, and let $\mathbf{a}_t \in \{0,1\}^{|E_\star|}$ indicate which support edges are active at step t. Block ℓ learns the positive base edge strengths $\bar{\mathbf{w}}^{(\ell)} \in \mathbb{R}^{|E_\star|}$ and defines the weighted graph Laplacian

$$\mathbf{L}_t^{(\ell)} = \mathbf{B}_\star^\top \operatorname{diag}\left(\mathbf{a}_t \odot \bar{\mathbf{w}}^{(\ell)}\right) \mathbf{B}_\star \in \mathbb{R}^{N \times N}. \tag{1}$$

The resulting operator is sparse, symmetric, and positive semidefinite. It becomes time-dependent when the graph evolves. For a fixed graph, all support edges remain active and the learned operator is shared across time. The operator therefore determines where latent motion can propagate and how strongly neighboring nodes interact.

3.2 COUPLED PATTERN–MOTION STATE EVOLUTION

Within block ℓ, we consider continuous pattern and motion states $\mathbf{U}^{(\ell)}(\tau), \mathbf{V}^{(\ell)}(\tau) \in \mathbb{R}^{N\times d}$, where τ denotes continuous latent time. The pattern state represents the current latent pattern, while the motion state represents its rate of change,

$$\mathbf{V}^{(\ell)}(\tau) = \frac{d}{d\tau}\mathbf{U}^{(\ell)}(\tau). \tag{2}$$

Under the graph-wave formulation, $\mathbf{U}^{(\ell)}$ and $\mathbf{V}^{(\ell)}$ are interpreted as displacement and velocity, respectively. For the continuous-time formulation, we denote the graph operator and forcing by $\mathbf{L}^{(\ell)}(\tau)$ and $\mathbf{F}^{(\ell)}(\tau)$, whose discrete counterparts are $\mathbf{L}_t^{(\ell)}$ and $\mathbf{F}_t^{(\ell)}$.

We motivate their evolution with a forced and damped graph-wave equation

$$\frac{d^2}{d\tau^2}\mathbf{U}^{(\ell)}(\tau) + \frac{d}{d\tau}\mathbf{U}^{(\ell)}(\tau)\mathbf{\Gamma}^{(\ell)} + \mathbf{L}^{(\ell)}(\tau)\mathbf{U}^{(\ell)}(\tau)\mathbf{K}^{(\ell)} = \mathbf{F}^{(\ell)}(\tau). \tag{3}$$

Here, $\mathbf{K}^{(\ell)} \in \mathbb{R}^{d\times d}$ and $\mathbf{\Gamma}^{(\ell)} \in \mathbb{R}^{d\times d}$ are learnable positive diagonal matrices. The former controls graph-interaction strength across latent channels, while the latter controls damping and motion persistence. The forcing $\mathbf{F}^{(\ell)}(\tau) \in \mathbb{R}^{N\times d}$ introduces information from the input sequence, and the graph force $-\mathbf{L}^{(\ell)}(\tau)\mathbf{U}^{(\ell)}(\tau)\mathbf{K}^{(\ell)}$ couples neighboring nodes.

Using the motion state defined above, Eq. (3) can equivalently be written as the first-order system

$$\frac{d}{d\tau}\mathbf{V}^{(\ell)}(\tau) = -\mathbf{V}^{(\ell)}(\tau)\mathbf{\Gamma}^{(\ell)} - \mathbf{L}^{(\ell)}(\tau)\mathbf{U}^{(\ell)}(\tau)\mathbf{K}^{(\ell)} + \mathbf{F}^{(\ell)}(\tau), \tag{4}$$

$$\frac{d}{d\tau}\mathbf{U}^{(\ell)}(\tau) = \mathbf{V}^{(\ell)}(\tau). \tag{5}$$

To realize these dynamics as a recurrent state transition within each WaveGSSM block, we discretize the system using a semi-implicit Euler scheme (Cromer, 1981). For block ℓ, let $\Delta t^{(\ell)} > 0$ be a learnable step size and define $\tau_t^{(\ell)} = t\,\Delta t^{(\ell)}$ for $t = 0, \ldots, T$. We denote the corresponding discrete states by $\mathbf{U}_t^{(\ell)} \approx \mathbf{U}^{(\ell)}(\tau_t^{(\ell)})$ and $\mathbf{V}_t^{(\ell)} \approx \mathbf{V}^{(\ell)}(\tau_t^{(\ell)})$. At rollout step t, $\mathbf{L}_t^{(\ell)}$ is the graph operator defined in Section 3.1, while $\mathbf{F}_t^{(\ell)}$ is generated from the incoming hidden representation as specified in Section 3.3. The discrete WaveGSSM transition is

$$\mathbf{V}_t^{(\ell)} = \left[\mathbf{V}_{t-1}^{(\ell)} + \Delta t^{(\ell)}\left(-\mathbf{L}_t^{(\ell)}\mathbf{U}_{t-1}^{(\ell)}\mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)}\right)\right]\left(\mathbf{I}_d + \Delta t^{(\ell)}\mathbf{\Gamma}^{(\ell)}\right)^{-1}, \tag{6}$$

$$\mathbf{U}_t^{(\ell)} = \mathbf{U}_{t-1}^{(\ell)} + \Delta t^{(\ell)}\mathbf{V}_t^{(\ell)}. \tag{7}$$

A derivation of Eqs. (6)–(7) is given in Appendix A.2. The motion state is updated first with damping treated implicitly, and the updated motion then advances the pattern state. Here, $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix. Because $\mathbf{\Gamma}^{(\ell)}$ is positive and diagonal, the inverse in Eq. (6) is implemented by elementwise division. Through this two-stage update, spatial propagation and temporal evolution are coupled within the same transition. Graph interactions update the motion state, and the updated motion advances the pattern, enabling oscillatory and propagating dynamics. Observation 1 illustrates this standard second-order state dependence in the graph-wave setting.

Observation 1 (Pattern state alone does not determine graph-wave evolution)**.** *Let* $\mathbf{u}(\tau) \in \mathbb{R}^N$ *satisfy*

$$\ddot{\mathbf{u}}(\tau) + \gamma\dot{\mathbf{u}}(\tau) + \kappa\mathbf{L}\mathbf{u}(\tau) = \mathbf{0},$$

where $\mathbf{L} \succeq 0$*,* $\kappa > 0$*, and* $\gamma \geq 0$*. For an eigenpair* $(\lambda_i, \mathbf{q}_i)$ *of* $\mathbf{L}$ *with* $\lambda_i > 0$*, define* $a_i(\tau) = \mathbf{q}_i^\top\mathbf{u}(\tau)$*. If* $\kappa\lambda_i > \gamma^2/4$ *and* $\omega_i = (\kappa\lambda_i - \gamma^2/4)^{1/2}$*, then*

$$a_i(\tau) = e^{-\gamma\tau/2}\left[a_i(0)\cos(\omega_i\tau) + \frac{\dot{a}_i(0) + \frac{\gamma}{2}a_i(0)}{\omega_i}\sin(\omega_i\tau)\right]. \tag{8}$$

The future graph-wave state, therefore, depends on both the displacement $a_i(0)$ *and velocity* $\dot{a}_i(0)$*, which correspond to modal components of the pattern and motion states in WaveGSSM.*

The proof is given in Appendix A.1. Observation 1 explains the phenomenon illustrated in Figure 1. Two trajectories can share the same pattern state while evolving differently when their motion states differ. WaveGSSM represents both components explicitly through $(\mathbf{U}_t, \mathbf{V}_t)$.

For a fixed graph operator, the semi-implicit rollout is non-amplifying for oscillatory graph modes under suitable conditions, as shown in Appendix A.3.

3.3 From Graph-Wave Dynamics to WaveGSSM

Next, we embed the graph-wave transition in a residual neural block. Each observed snapshot is first mapped to hidden width d by a shared node-wise input projection $\phi_{\text{in}} : \mathbb{R}^{N\times F_{\text{in}}} \to \mathbb{R}^{N\times d}$,

$$\mathbf{Z}_t^{(0)} = \phi_{\text{in}}(\mathbf{X}_t) \in \mathbb{R}^{N\times d}, \qquad t = 1, \dots, T. \tag{9}$$

Block ℓ receives the hidden sequence $\{\mathbf{Z}_t^{(\ell-1)}\}_{t=1}^T$ and yields an updated sequence of the same shape.

The first hidden frame initializes the pattern state. When at least two frames are available, their hidden difference initializes the motion state

$$\mathbf{U}_0^{(\ell)} = \psi_U^{(\ell)}\left(\mathrm{LN}\left(\mathbf{Z}_1^{(\ell-1)}\right)\right), \tag{10}$$

$$\mathbf{V}_0^{(\ell)} = \psi_V^{(\ell)}\left(\mathrm{LN}\left(\mathbf{Z}_2^{(\ell-1)}\right) - \mathrm{LN}\left(\mathbf{Z}_1^{(\ell-1)}\right)\right). \tag{11}$$

Here, LN denotes the layer normalization, and $\psi_U^{(\ell)}, \psi_V^{(\ell)} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ are lightweight node-wise maps with independent parameters. For a single-frame input, we set $\mathbf{V}_0^{(\ell)} = \mathbf{0}$. The initialization, therefore, provides each block with both a latent pattern and an input-dependent motion cue.

Each hidden snapshot also produces a forcing term

$$\mathbf{F}_t^{(\ell)} = g^{(\ell)}\left(\mathbf{Z}_t^{(\ell-1)}\right) \in \mathbb{R}^{N\times d}, \tag{12}$$

where $g^{(\ell)} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ is a lightweight node-wise forcing map. Starting from $(\mathbf{U}_0^{(\ell)}, \mathbf{V}_0^{(\ell)})$, the block applies Eqs. (6) and (7) along the input sequence using $\{\mathbf{L}_t^{(\ell)}\}_{t=1}^T$ and $\{\mathbf{F}_t^{(\ell)}\}_{t=1}^T$.

The evolved pattern state is returned to the hidden stream through a residual update

$$\hat{\mathbf{Z}}_t^{(\ell)} = \mathbf{Z}_t^{(\ell-1)} + \mathrm{Drop}\left(\mathrm{LN}\left(\mathbf{U}_t^{(\ell)}\right)\right), \tag{13}$$

$$\mathbf{Z}_t^{(\ell)} = \hat{\mathbf{Z}}_t^{(\ell)} + \mathrm{Drop}\left(\mathrm{FFN}\left(\mathrm{LN}\left(\hat{\mathbf{Z}}_t^{(\ell)}\right)\right)\right). \tag{14}$$

Here, Drop denotes dropout, and FFN is a standard node-wise feed-forward network. The motion state remains part of the internal recurrent state and is passed between consecutive graph-wave transitions. Stacking L blocks alternates motion-carrying state evolution with feature refinement. Full definitions of the projection, initialization, forcing, and residual maps are given in Appendix B.

Task readout. After L WaveGSSM blocks, the final block produces the hidden representation $\mathbf{Z}_T^{(L)}$ and the terminal pattern–motion state $(\mathbf{U}_T^{(L)}, \mathbf{V}_T^{(L)})$. For the state readout, we form

$$\mathbf{H} = \left[\mathbf{U}_T^{(L)}, \mathbf{V}_T^{(L)}\right] \in \mathbb{R}^{N\times 2d}. \tag{15}$$

A task-specific decoder maps the selected representation to node labels or future fields. Temporal node classification uses either $\mathbf{H}$ or $\mathbf{Z}_T^{(L)}$, as detailed in Appendix B.5. For mesh-based forecasting, $\mathbf{H}$ is projected to the decoder width and mapped back to the target grid, conditioning the forecast on both terminal pattern and motion states.

4 Experiments

4.1 Experimental Setup

We evaluate WaveGSSM on temporal node classification and global weather forecasting. The former tests whether the proposed state evolution generalizes beyond explicitly physical processes, while the latter evaluates the motivating setting of multi-step spatial propagation.

Temporal graph node classification. We evaluate WaveGSSM on DBLP-3, Brain, Reddit, and DBLP-10 from the GraphSSM benchmark (Li et al., 2024), reporting Micro-F1 and Macro-F1. Baselines include graph embedding methods, discrete-time temporal GNNs, and GraphSSM. Full dataset, split, model, and implementation details are provided in Appendix D.1.

Global weather forecasting. We construct a daily global forecasting task from ERA5 (Hersbach et al., 2020; Lam et al., 2023) on a 1° latitude–longitude grid with a global icosahedral mesh. The model jointly forecasts geopotential at six pressure levels, daily-maximum 2 m temperature, and level-1 soil moisture. We train on 1979–2023, validate on 2024, and evaluate on 354 initializations from 2025 using latitude-weighted RMSE and centred anomaly correlation.

Because the original GraphCast setting differs in forecast interval, resolution, and evaluation period, we train GraphCast* under the same daily ERA5 setting as WaveGSSM. We also include GraphSSM (Li et al., 2024) and GNN-LSTM as state-space and recurrent baselines. Full experimental details are provided in Appendix D.2, with additional motion-state analysis in Appendix F.3.

4.2 Generalization to Temporal Graph Benchmarks

We first examine whether the motion-carrying state is also useful for general temporal graph tasks. As shown in Table 1, WaveGSSM obtains the highest mean result on all eight reported metrics across DBLP-3, Brain, Reddit, and DBLP-10. The gains are most pronounced on DBLP-3 and Brain, while the margins are smaller on Reddit and DBLP-10. These results show that the proposed state evolution also provides an effective representation mechanism for discrete-time temporal graphs.

Table 1: Node-classification performance (%) on four temporal graph benchmarks. Mean ± standard deviation over five runs where available. Higher is better. Best and second-best results are highlighted in **green** and **blue**, respectively.

	DBLP-3		Brain		Reddit		DBLP-10	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
DeepWalk	$47.53_{\pm 0.4}$	$47.21_{\pm 0.2}$	$51.45_{\pm 0.6}$	$51.03_{\pm 0.8}$	$26.82_{\pm 0.6}$	$26.75_{\pm 0.4}$	66.38	67.12
Node2Vec	$48.79_{\pm 0.3}$	$48.42_{\pm 0.4}$	$53.51_{\pm 0.5}$	$52.95_{\pm 0.6}$	$25.47_{\pm 0.6}$	$25.44_{\pm 0.5}$	67.31	66.93
HTNE	$48.98_{\pm 0.2}$	$48.74_{\pm 0.3}$	$22.31_{\pm 0.8}$	$22.12_{\pm 0.5}$	$26.96_{\pm 0.5}$	$26.80_{\pm 0.7}$	68.79	68.36
M^2DNE	$49.12_{\pm 0.5}$	$48.87_{\pm 0.4}$	$23.79_{\pm 0.5}$	$23.54_{\pm 0.4}$	$25.79_{\pm 0.6}$	$25.61_{\pm 0.4}$	69.71	69.75
DynamicTriad	$48.78_{\pm 0.5}$	$48.63_{\pm 0.6}$	$21.71_{\pm 0.7}$	$21.94_{\pm 0.7}$	$28.76_{\pm 0.5}$	$28.51_{\pm 0.5}$	66.95	66.42
MPNN	$81.78_{\pm 0.6}$	$81.46_{\pm 1.2}$	$90.97_{\pm 1.4}$	$91.01_{\pm 1.5}$	$40.85_{\pm 1.3}$	$40.64_{\pm 1.2}$	$67.74_{\pm 0.3}$	$65.05_{\pm 0.5}$
STAR	$84.74_{\pm 1.0}$	$84.20_{\pm 1.2}$	$92.08_{\pm 1.3}$	$92.23_{\pm 1.3}$	$43.42_{\pm 2.3}$	$43.43_{\pm 2.4}$	$72.98_{\pm 1.5}$	$72.03_{\pm 1.2}$
tNodeEmbed	$84.51_{\pm 1.2}$	$83.57_{\pm 1.1}$	$92.35_{\pm 0.8}$	$92.30_{\pm 1.0}$	$42.11_{\pm 1.8}$	$42.06_{\pm 1.3}$	$74.19_{\pm 1.8}$	$74.23_{\pm 2.2}$
EvolveGCN	$84.01_{\pm 1.5}$	$83.12_{\pm 1.5}$	$92.20_{\pm 1.3}$	$92.00_{\pm 1.0}$	$41.24_{\pm 1.3}$	$41.11_{\pm 1.5}$	$71.32_{\pm 0.5}$	$71.20_{\pm 0.7}$
SpikeNet	$83.92_{\pm 1.5}$	$83.04_{\pm 1.1}$	$92.00_{\pm 1.2}$	$91.97_{\pm 1.2}$	$40.42_{\pm 2.0}$	$40.20_{\pm 2.1}$	$74.86_{\pm 0.5}$	$74.65_{\pm 0.5}$
ROLAND	$84.21_{\pm 1.4}$	$84.06_{\pm 1.5}$	$92.14_{\pm 1.2}$	$91.85_{\pm 1.1}$	$44.22_{\pm 2.2}$	$44.25_{\pm 1.9}$	$75.01_{\pm 1.1}$	$74.98_{\pm 1.0}$
GraphSSM	$83.52_{\pm 1.6}$	$82.98_{\pm 1.5}$	$92.36_{\pm 0.8}$	$92.75_{\pm 0.8}$	$46.79_{\pm 1.9}$	$46.65_{\pm 2.1}$	$76.29_{\pm 0.7}$	$75.51_{\pm 0.8}$
WaveGSSM (Ours)	$\mathbf{86.34}_{\pm 1.5}$	$\mathbf{86.26}_{\pm 1.6}$	$\mathbf{93.36}_{\pm 0.7}$	$\mathbf{93.64}_{\pm 0.5}$	$\mathbf{47.47}_{\pm 1.7}$	$\mathbf{47.51}_{\pm 1.7}$	$\mathbf{76.38}_{\pm 0.4}$	$\mathbf{76.12}_{\pm 0.5}$

4.3 Global Weather Forecasting

We next evaluate WaveGSSM on global weather forecasting, a real-world setting with propagating atmospheric patterns. The model forecasts eight ERA5 target fields. Figure 4 highlights four variables: the geopotential height at 500 and 200 hPa, the daily-maximum 2 m temperature, and the level-1 soil moisture. RMSE measures pointwise error, while anomaly correlation (AC) measures the large-scale pattern agreement. Our results for the remaining pressure levels are given in Appendix F.1.

WaveGSSM achieves the lowest RMSE and highest AC for every displayed variable and lead time. Relative to the backbone-matched GraphCast* baseline, it reduces geopotential RMSE by 20.2% on average over all six pressure levels and forecast days 1–5. The lead-time dependence is especially revealing. For z500, the RMSE reduction decreases from 38.1% at day 1 to 7.0% at day 5. In contrast, the AC advantage becomes more pronounced toward the medium range, reaching its strongest level around one week. Thus, while the pointwise-error margin narrows with lead, WaveGSSM better maintains large-scale pattern agreement over medium-range forecasts. For propagating patterns, this distinction matters because a forecast can remain locally plausible even as spatial displacement accumulates over time.

Importantly, WaveGSSM also outperforms GraphSSM and GNN-LSTM across the displayed forecasts. The advantage, therefore, extends beyond the backbone-matched snapshot comparison and remains present against both graph state-space and recurrent temporal models. The controlled motion-state ablations in Section 4.4 next examine whether this advantage depends on carrying the proposed motion state.

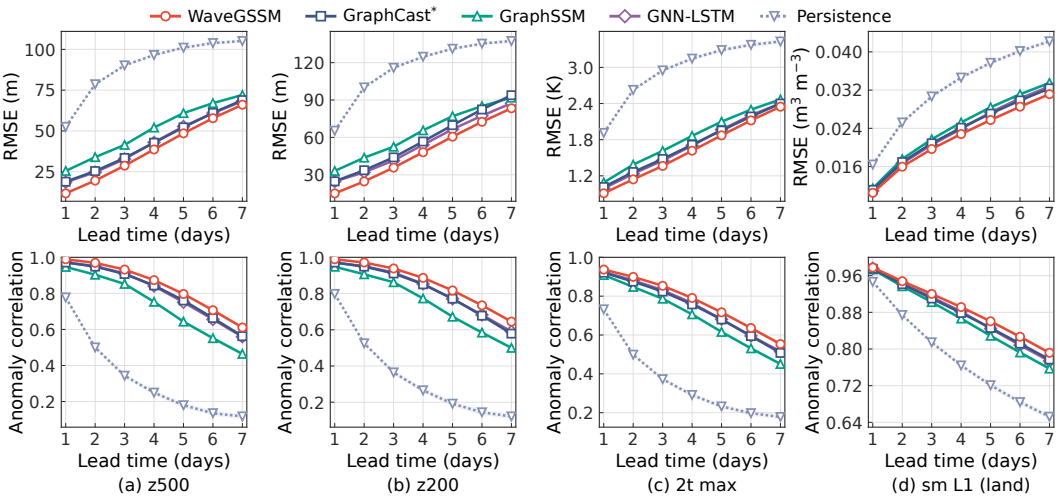


Figure 4: Forecast skill on the 2025 ERA5 test set. Latitude-weighted RMSE and centred anomaly correlation over forecast days 1–7 for (a) 500 hPa geopotential height, (b) 200 hPa geopotential height, (c) daily-maximum 2 m temperature, and (d) level-1 soil moisture over land. WaveGSSM achieves the lowest RMSE and highest anomaly correlation across all displayed fields and leads.

Table 2: Retrained ablation of persistent motion-state carry. Both variants use the same dataset-specific architecture, parameter count, training protocol, and validation rule. Values are mean ± standard deviation over five seeds. Higher is better.

	DBLP-3		Brain		Reddit	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
WaveGSSM	**86.34**$_{\pm 1.54}$	**86.26**$_{\pm 1.58}$	**93.36**$_{\pm 0.66}$	**93.64**$_{\pm 0.47}$	**47.47**$_{\pm 1.68}$	**47.51**$_{\pm 1.73}$
WaveGSSM w/o motion-state carry	84.51$_{\pm 1.41}$	84.45$_{\pm 1.55}$	90.00$_{\pm 0.66}$	90.65$_{\pm 0.92}$	45.40$_{\pm 0.59}$	45.27$_{\pm 0.44}$

4.4 Ablating Motion-State Carry

The previous comparisons involve different model families. Here, we retrain WaveGSSM variants to test whether carrying the motion state contributes to predictive performance.

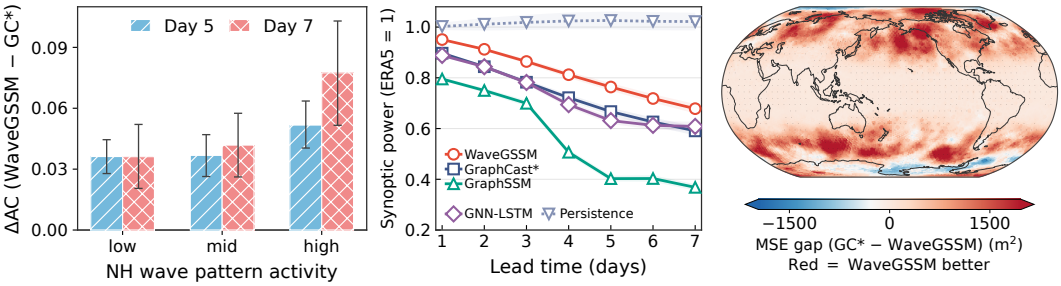
For a temporal node classification, the no-carry variant preserves the input-derived initial motion state and the instantaneous graph response, but discards the motion state produced by each transition before the next. This removes persistent motion-state carry, while keeping the rest of the model and training procedure unchanged. Implementation details are provided in Appendix E.1.

On the three temporal-graph datasets included in this ablation, removing persistent motion-state carry lowers both Micro-F1 and Macro-F1 by 1.81–3.36 percentage points. The ablated model remains competitive on DBLP-3, but falls below GraphSSM on Brain and Reddit. Since it is retrained independently, the remaining gap indicates that persistent motion-state carry contributes to the performance gains.

The same trend holds for weather forecasting. The parameter-matched variant resets the incoming motion state before every internal transition, including the first, while retaining the instantaneous response within each transition. After independent retraining, mean geopotential RMSE over all six pressure levels and forecast days 1–5 increases from 30.72 to 31.53 m, a 2.63% increase. The ablated model also has a higher z200 RMSE at every reported lead from day 1 to day 10. The complete results are given in Appendix F.2.

4.5 When Does the Motion-Carrying State Help?

The ablations above show that carrying the motion state substantially contributes to the predictive performance. We now ask when the weather-forecasting advantage becomes strongest and whether



(a) Wave-activity gain (b) Large-scale pattern strength (c) Spatial MSE gap, day 5

Figure 5: When and where WaveGSSM improves, and how well synoptic structure is preserved. Panel (a) shows the z500 anomaly-correlation gain $\Delta\mathrm{AC} = \mathrm{AC}_{\mathrm{WaveGSSM}} - \mathrm{AC}_{\mathrm{GC}^*}$ relative to GraphCast*. Panel (a) groups days 5 and 7 by Northern-Hemisphere extratropical wave activity. Panel (b) reports z500 forecast power relative to ERA5 over zonal wavenumbers 4–9. Panel (c) shows the day-5 spatial MSE gap, where red favors WaveGSSM. The advantage grows with wave activity, while WaveGSSM preserves more power over the synoptic band.

it arises from a better propagation of coherent structure rather than stronger spatial smoothing. If motion matters, the gain should increase, when the large-scale wave activity is stronger and as the displacement errors accumulate with the forecast lead.

The gain grows when propagation matters. We stratify the z500 anomaly-correlation advantage by wave activity (Figure 5a). The advantage increases from low- to high-activity cases and is larger at day 7 than at day 5 in the most active regime. At day 7, using $(1 - \mathrm{AC})$ as pattern error, WaveGSSM removes 18% of the error remaining in GraphCast* for the most wave-active third of forecasts, compared with 9% for the least active third. The day-5 spatial error map further shows that the improvement is geographically broad rather than confined to a small number of regions (Figure 5c). The supporting seasonal and regional breakdowns show larger gains during boreal winter and spring and across the extratropics (Appendix Figure 10). These results emphasize that the advantage becomes stronger when large-scale spatial propagation plays a larger role.

Large-scale spatial variability is better preserved. A deterministic forecast can reduce the squared error by smoothing spatial variability toward the mean. Hence, we examine the power of z500 over zonal wavenumbers 4–9, which capture large-scale spatial patterns associated with evolving weather systems (Figure 5b). At day 7, WaveGSSM retains 68% of the ERA5 variability in this band, compared with 59% for GraphCast* and 37% for GraphSSM. Its day-7 spectrum also remains closest to ERA5 across the selected band (Appendix Figure 10c). Persistence retains much of the initial spatial variability, but forecasts poorly as the atmosphere evolves. Preserving variability alone is therefore insufficient. The corresponding patterns must also evolve toward the correct future states.

These results show that WaveGSSM's advantage grows in regimes where spatial propagation matters and is not explained by stronger spatial smoothing. The retrained ablations in Section 4.4 test the role of motion-state carry directly, while Appendix F.3 examines how the trained weather model uses its internal motion-state pathway. Additional spatial, vertical, bias, and case-level analyses are provided in Appendices G and H.

5 Conclusion

We studied spatio-temporal graph prediction from the perspective of state representation, showing that the present field alone can be insufficient when future evolution depends on recent motion. The new WaveGSSM addresses this limitation with coupled pattern and motion states that evolve through a second-order graph-wave transition, allowing spatial propagation and temporal evolution to occur within the same state update. Across temporal-graph benchmarks and daily global weather forecasting, WaveGSSM consistently improves the predictive performance, with particularly strong gains in maintaining large-scale pattern agreement over longer rollouts. These results support explicit motion representation as a useful state-space principle for learning propagating dynamics on graphs.

AI USE STATEMENT

Generative AI tools were used to assist with language editing of the manuscript, identifying potentially relevant related work, checking theoretical analyses and mathematical derivations, providing suggestions on figure presentation such as color choices, and assisting the first author with coding. All figures were produced from our own data, and AI tools were not used to generate scientific data or figures. Relevant literature identified with AI assistance was independently verified, and all AI-assisted theoretical content, derivations, and code were checked by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The appendix provides the information needed to reproduce our experiments. Model architecture and implementation details are given in Appendix B, while dataset preprocessing, training, and evaluation protocols are provided in Appendices D.1 and D.2. Ablation procedures are detailed in Appendices E.1, F.2, and F.3. Theoretical derivations and stability analysis are provided in Appendix A.

REFERENCES

Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 119–130, 2024.

Andrea Ceni, Alessio Gravina, Claudio Gallicchio, Davide Bacciu, Carola-Bibiane Schonlieb, and Moshe Eliasof. Message-passing state-space models: Improving graph learning with modern sequence modeling. *arXiv preprint arXiv:2505.18728*, 2025.

Wei Chen and Yuxuan Liang. Expand and compress: Exploring tuning principles for continual spatio-temporal graph forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 81631–81656, 2025.

Andrea Cini, Ivan Marisca, Daniele Zambon, and Cesare Alippi. Taming local effects in graph-based spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 36:55375–55393, 2023.

Alan Cromer. Stable solutions using the euler approximation. *American Journal of Physics*, 49(5): 455–459, 1981.

Aref Einizade, Fragkiskos D Malliaros, and Jhony H Giraldo. Continuous product graph neural networks. *Advances in Neural Information Processing Systems*, 37:90226–90252, 2024.

Moshe Eliasof, Eldad Haber, Eran Treister, and Carola-Bibiane B Schönlieb. On the temporal domain of differential equation inspired graph neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1792–1800. PMLR, 2024.

Chao Gao, Junyou Zhu, Fan Zhang, Zhen Wang, and Xuelong Li. A novel representation learning for dynamic graphs based on graph convolutional networks. *IEEE Transactions on Cybernetics*, 53(6): 3599–3612, 2023.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

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

Yannis Karmim, Marc Lafon, Raphaël F S'niehotta, and Nicolas Thome. Supra-laplacian encoding for transformer on dynamic graphs. *Advances in Neural Information Processing Systems*, 37: 17215–17246, 2024.

Kai Kornhuber, Dim Coumou, Elisabeth Vogel, Corey Lesk, Jonathan F Donges, Jascha Lehmann, and Radley M Horton. Amplified rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nature Climate Change*, 10(1):48–53, 2020.

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.

Jintang Li, Zhouxin Yu, Zulun Zhu, Liang Chen, Qi Yu, Zibin Zheng, Sheng Tian, Ruofan Wu, and Changhua Meng. Scaling up dynamic graph representation learning via spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8588–8596, 2023a.

Jintang Li, Ruofan Wu, Xinzhou Jin, Boqun Ma, Liang Chen, and Zibin Zheng. State space models on temporal graphs: A first-principles study. *Advances in Neural Information Processing Systems*, 37:127030–127058, 2024.

Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. Gpt-st: generative pre-training of spatio-temporal graph neural networks. *Advances in neural information processing systems*, 36: 70229–70246, 2023b.

Yuanfu Lu, Xiao Wang, Chuan Shi, Philip S Yu, and Yanfang Ye. Temporal network embedding with micro-and macro-dynamics. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 469–478, 2019.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4838–4845, 2021.

Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5363–5370, 2020.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.

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.

T Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael Bronstein. Graph-coupled oscillator networks. In *International conference on machine learning*, pp. 18888–18909. PMLR, 2022.

Uriel Singer, Ido Guy, and Kira Radinsky. Node embedding over temporal graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 4605–4612, 2019.

Mingzhao Wang, Yu Huang, Christian LE Franzke, Naiming Yuan, Zuntao Fu, and Niklas Boers. Evidence for preferred propagating terrestrial heatwave pathways due to rossby wave activity. *Nature Communications*, 16(1):4742, 2025a.

Yili Wang, Tairan Huang, Changlong He, Qiutong Li, and Jianliang Gao. Simple and efficient heterogeneous temporal graph neural network. *Advances in Neural Information Processing Systems*, 38:110346–110372, 2025b.

Dongkuan Xu, Wei Cheng, Dongsheng Luo, Xiao Liu, and Xiang Zhang. Spatio-temporal attentive rnn for node classification in temporal attributed graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3947–3953, 2019.

Jiaxuan You, Tianyu Du, and Jure Leskovec. Roland: graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2358–2366, 2022.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018.

Juwei Yue, Haikuo Li, Jiawei Sheng, Yihan Guo, Xinghua Zhang, Chuan Zhou, Tingwen Liu, and Li Guo. Graph wave networks. In *Proceedings of the ACM on Web Conference 2025*, pp. 1365–1379, 2025.

Lekui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. Dynamic network embedding by modeling triadic closure process. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yuan Zuo, Guannan Liu, Hao Lin, Jia Guo, Xiaoqian Hu, and Junjie Wu. Embedding temporal network via neighborhood formation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2857–2866, 2018.

A Theoretical Analysis

This section provides the theoretical details supporting WaveGSSM. We first prove that the present graph field alone does not determine the future of an oscillatory graph mode when its recent motion is unknown. We then derive the semi-implicit transition used by WaveGSSM and analyze its modal stability.

A.1 Why the Pattern State Alone Is Not a Complete State

Proof of Observation 1. Since $\mathbf{L}$ is symmetric, let $(\lambda_i, \mathbf{q}_i)$ be an eigenpair satisfying

$$\mathbf{L}\mathbf{q}_i = \lambda_i \mathbf{q}_i.$$

Define the projected modal amplitude

$$a_i(\tau) = \mathbf{q}_i^\top \mathbf{u}(\tau).$$

Projecting the graph-wave equation onto this mode gives

$$\begin{aligned}
0 &= \mathbf{q}_i^\top \left(\ddot{\mathbf{u}}(\tau) + \gamma \dot{\mathbf{u}}(\tau) + \kappa \mathbf{L}\mathbf{u}(\tau)\right) && (16)\\
&= \mathbf{q}_i^\top \ddot{\mathbf{u}}(\tau) + \gamma \mathbf{q}_i^\top \dot{\mathbf{u}}(\tau) + \kappa \mathbf{q}_i^\top \mathbf{L}\mathbf{u}(\tau) && (17)\\
&= \ddot{a}_i(\tau) + \gamma \dot{a}_i(\tau) + \kappa (\mathbf{L}\mathbf{q}_i)^\top \mathbf{u}(\tau) && (18)\\
&= \ddot{a}_i(\tau) + \gamma \dot{a}_i(\tau) + \kappa \lambda_i \mathbf{q}_i^\top \mathbf{u}(\tau) && (19)\\
&= \ddot{a}_i(\tau) + \gamma \dot{a}_i(\tau) + \kappa \lambda_i a_i(\tau). && (20)
\end{aligned}$$

Thus, each graph mode follows a damped oscillator. Set

$$a_i(\tau) = e^{-\gamma\tau/2} b_i(\tau).$$

Then

$$\begin{aligned}
\dot{a}_i(\tau) &= e^{-\gamma\tau/2}\left(\dot{b}_i(\tau) - \frac{\gamma}{2} b_i(\tau)\right), && (21)\\
\ddot{a}_i(\tau) &= e^{-\gamma\tau/2}\left(\ddot{b}_i(\tau) - \gamma \dot{b}_i(\tau) + \frac{\gamma^2}{4} b_i(\tau)\right). && (22)
\end{aligned}$$

Substituting these expressions into Eq. (20) yields

$$\begin{aligned}
0 &= \ddot{a}_i(\tau) + \gamma \dot{a}_i(\tau) + \kappa \lambda_i a_i(\tau) && (23)\\
&= e^{-\gamma\tau/2}\left(\ddot{b}_i(\tau) - \gamma \dot{b}_i(\tau) + \frac{\gamma^2}{4} b_i(\tau)\right) && (24)\\
&\quad + \gamma e^{-\gamma\tau/2}\left(\dot{b}_i(\tau) - \frac{\gamma}{2} b_i(\tau)\right) + \kappa \lambda_i e^{-\gamma\tau/2} b_i(\tau) && (25)\\
&= e^{-\gamma\tau/2}\left[\ddot{b}_i(\tau) - \gamma \dot{b}_i(\tau) + \frac{\gamma^2}{4} b_i(\tau) + \gamma \dot{b}_i(\tau) - \frac{\gamma^2}{2} b_i(\tau) + \kappa \lambda_i b_i(\tau)\right] && (26)\\
&= e^{-\gamma\tau/2}\left[\ddot{b}_i(\tau) + \left(\kappa \lambda_i - \frac{\gamma^2}{4}\right) b_i(\tau)\right] && (27)\\
&= e^{-\gamma\tau/2}\left[\ddot{b}_i(\tau) + \omega_i^2 b_i(\tau)\right]. && (28)
\end{aligned}$$

Because $e^{-\gamma\tau/2} > 0$, Eq. (28) implies

$$\ddot{b}_i(\tau) + \omega_i^2 b_i(\tau) = 0.$$

Therefore,

$$\begin{aligned}
b_i(\tau) &= c_1 \cos(\omega_i \tau) + c_2 \sin(\omega_i \tau), && (29)\\
a_i(\tau) &= e^{-\gamma\tau/2}\left[c_1 \cos(\omega_i \tau) + c_2 \sin(\omega_i \tau)\right]. && (30)
\end{aligned}$$

The constants are determined by the initial displacement and velocity:

$$a_i(0) = c_1, \tag{31}$$

$$\dot{a}_i(0) = \dot{b}_i(0) - \frac{\gamma}{2} b_i(0) \tag{32}$$

$$= \omega_i c_2 - \frac{\gamma}{2} c_1 \tag{33}$$

$$= \omega_i c_2 - \frac{\gamma}{2} a_i(0), \tag{34}$$

$$c_2 = \frac{\dot{a}_i(0) + \frac{\gamma}{2} a_i(0)}{\omega_i}. \tag{35}$$

Substituting Eq. (35) into Eq. (30) gives Eq. (8).

It remains to show that equal displacement does not imply equal future evolution. Consider two trajectories satisfying

$$\mathbf{u}^{(1)}(0) = \mathbf{u}^{(2)}(0), \qquad \mathbf{q}_i^\top \dot{\mathbf{u}}^{(1)}(0) \neq \mathbf{q}_i^\top \dot{\mathbf{u}}^{(2)}(0).$$

Using Eq. (8), their modal difference is

$$a_i^{(1)}(\tau) - a_i^{(2)}(\tau) = e^{-\gamma\tau/2} \left[\left(a_i^{(1)}(0) - a_i^{(2)}(0) \right) \cos(\omega_i \tau) \right] \tag{36}$$

$$+ e^{-\gamma\tau/2} \left[\frac{\dot{a}_i^{(1)}(0) - \dot{a}_i^{(2)}(0) + \frac{\gamma}{2} \left(a_i^{(1)}(0) - a_i^{(2)}(0) \right)}{\omega_i} \sin(\omega_i \tau) \right] \tag{37}$$

$$= e^{-\gamma\tau/2} \frac{\dot{a}_i^{(1)}(0) - \dot{a}_i^{(2)}(0)}{\omega_i} \sin(\omega_i \tau) \tag{38}$$

$$= e^{-\gamma\tau/2} \frac{\mathbf{q}_i^\top \dot{\mathbf{u}}^{(1)}(0) - \mathbf{q}_i^\top \dot{\mathbf{u}}^{(2)}(0)}{\omega_i} \sin(\omega_i \tau). \tag{39}$$

For any time τ satisfying $\sin(\omega_i \tau) \neq 0$, the right-hand side is nonzero. Two graph fields can therefore share the same current displacement and still evolve into different future states. A displacement-only representation is not a complete Markov state for the oscillatory graph-wave dynamics considered here. □

Generally, the theorem does not claim that a recurrent model cannot estimate motion from an observed history. It shows that a state intended to determine the future of a second-order propagating process must contain information about both the present field and its recent change. WaveGSSM represents these components explicitly through the pattern and motion states $(\mathbf{U}_t, \mathbf{V}_t)$, as examined empirically in Figure 1.

A.2 Derivation of the Semi-Implicit Rollout

We derive Eqs. (6)–(7) from the first-order system in Section 3.2. At rollout step t, the graph interaction and forcing are evaluated explicitly using $\mathbf{U}_{t-1}^{(\ell)}$, $\mathbf{L}_t^{(\ell)}$, and $\mathbf{F}_t^{(\ell)}$, while the damping term is evaluated implicitly at $\mathbf{V}_t^{(\ell)}$. The semi-implicit Euler update therefore satisfies

$$\frac{\mathbf{V}_t^{(\ell)} - \mathbf{V}_{t-1}^{(\ell)}}{\Delta t^{(\ell)}} = -\mathbf{V}_t^{(\ell)} \mathbf{\Gamma}^{(\ell)} - \mathbf{L}_t^{(\ell)} \mathbf{U}_{t-1}^{(\ell)} \mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)}, \tag{40}$$

$$\frac{\mathbf{U}_t^{(\ell)} - \mathbf{U}_{t-1}^{(\ell)}}{\Delta t^{(\ell)}} = \mathbf{V}_t^{(\ell)}. \tag{41}$$

Rearranging Eq. (40) gives

$$\mathbf{V}_t^{(\ell)} \left(\mathbf{I}_d + \Delta t^{(\ell)} \mathbf{\Gamma}^{(\ell)} \right) = \mathbf{V}_{t-1}^{(\ell)} + \Delta t^{(\ell)} \left(-\mathbf{L}_t^{(\ell)} \mathbf{U}_{t-1}^{(\ell)} \mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)} \right), \tag{42}$$

$$\mathbf{V}_t^{(\ell)} = \left[\mathbf{V}_{t-1}^{(\ell)} + \Delta t^{(\ell)} \left(-\mathbf{L}_t^{(\ell)} \mathbf{U}_{t-1}^{(\ell)} \mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)} \right) \right] \left(\mathbf{I}_d + \Delta t^{(\ell)} \mathbf{\Gamma}^{(\ell)} \right)^{-1}. \tag{43}$$

Similarly, Eq. (41) gives

$$\mathbf{U}_t^{(\ell)} = \mathbf{U}_{t-1}^{(\ell)} + \Delta t^{(\ell)} \mathbf{V}_t^{(\ell)}. \tag{44}$$

These are Eqs. (6) and (7) in the main text.

A.3 Modal Stability of the Semi-Implicit Rollout

We next analyze the numerical behavior of the semi-implicit rollout under a fixed graph operator. The analysis isolates the integrator by removing the forcing term and holding the graph operator constant. Since $\mathbf{K}$ and $\mathbf{\Gamma}$ are positive diagonal matrices, each graph eigenmode and latent channel can be studied independently. We fix one WaveGSSM block and suppress the block index ℓ for clarity.

Proposition 1 (Modal stability of the semi-implicit rollout)**.** *Let* $\mathbf{L} \succeq 0$ *be fixed,*

$$\mathbf{K} = \mathrm{diag}(\kappa_1, \dots, \kappa_d), \qquad \kappa_r > 0,$$

and

$$\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_d), \qquad \gamma_r \geq 0.$$

For a graph eigenvalue $\lambda_i > 0$ *and latent channel* r*, define*

$$\omega_{ir}^2 = \kappa_r \lambda_i,$$

and let $\Delta t > 0$ *denote the step size. The unforced semi-implicit rollout has the modal transition*

$$\begin{bmatrix} a_t \\ b_t \end{bmatrix} = \mathbf{A}_{ir} \begin{bmatrix} a_{t-1} \\ b_{t-1} \end{bmatrix}, \qquad \mathbf{A}_{ir} = \frac{1}{1 + \Delta t \gamma_r} \begin{bmatrix} 1 + \Delta t \gamma_r - (\Delta t)^2 \omega_{ir}^2 & \Delta t \\ -\Delta t \omega_{ir}^2 & 1 \end{bmatrix}. \tag{45}$$

If $\gamma_r = 0$ *and* $0 < \Delta t \omega_{ir} < 2$*, then*

$$\rho(\mathbf{A}_{ir}) = 1.$$

If $\gamma_r > 0$ *and*

$$\Delta t^2 \omega_{ir}^2 < 4 + 2\Delta t \gamma_r,$$

then

$$\rho(\mathbf{A}_{ir}) < 1.$$

In contrast, undamped explicit Euler satisfies

$$\rho(\mathbf{A}_{ir}^{\mathrm{EE}}) = \sqrt{1 + (\Delta t)^2 \omega_{ir}^2} > 1$$

for every $\Delta t > 0$*.*

Proof. Let $\mathbf{q}_i$ be a normalized eigenvector satisfying

$$\mathbf{L}\mathbf{q}_i = \lambda_i \mathbf{q}_i,$$

and let $\mathbf{u}_{r,t}$ and $\mathbf{v}_{r,t}$ denote the r-th columns of the pattern and motion states. Define

$$a_t = \mathbf{q}_i^\top \mathbf{u}_{r,t}, \qquad b_t = \mathbf{q}_i^\top \mathbf{v}_{r,t}.$$

Projecting the unforced update in Eqs. (6) and (7) onto graph mode i and latent channel r gives

$$\begin{aligned} b_t &= \frac{b_{t-1} - h\kappa_r \lambda_i a_{t-1}}{1 + h\gamma_r} && (46)\\ &= \frac{b_{t-1} - h\omega_{ir}^2 a_{t-1}}{1 + h\gamma_r}, && (47)\\ a_t &= a_{t-1} + h b_t && (48)\\ &= a_{t-1} + h\frac{b_{t-1} - h\omega_{ir}^2 a_{t-1}}{1 + h\gamma_r} && (49)\\ &= \frac{\left(1 + h\gamma_r - h^2\omega_{ir}^2\right) a_{t-1} + h b_{t-1}}{1 + h\gamma_r}. && (50) \end{aligned}$$

Therefore,

$$\begin{aligned} \begin{bmatrix} a_t \\ b_t \end{bmatrix} &= \frac{1}{1 + \Delta t \gamma_r} \begin{bmatrix} 1 + h\gamma_r - h^2\omega_{ir}^2 & h \\ -h\omega_{ir}^2 & 1 \end{bmatrix} \begin{bmatrix} a_{t-1} \\ b_{t-1} \end{bmatrix} && (51)\\ &= \mathbf{A}_{ir} \begin{bmatrix} a_{t-1} \\ b_{t-1} \end{bmatrix}. && (52) \end{aligned}$$

The determinant and trace of $\mathbf{A}_{ir}$ are

$$\det(\mathbf{A}_{ir}) = \frac{1 + h\gamma_r - h^2\omega_{ir}^2 + h^2\omega_{ir}^2}{(1 + h\gamma_r)^2} \tag{53}$$

$$= \frac{1}{1 + \Delta t\gamma_r}, \tag{54}$$

$$\mathrm{tr}(\mathbf{A}_{ir}) = \frac{2 + h\gamma_r - h^2\omega_{ir}^2}{1 + h\gamma_r}. \tag{55}$$

Let

$$\delta_{ir} = \det(\mathbf{A}_{ir}), \qquad \theta_{ir} = \mathrm{tr}(\mathbf{A}_{ir}).$$

The characteristic polynomial is

$$p_{ir}(z) = z^2 - \theta_{ir} z + \delta_{ir}. \tag{56}$$

For $\gamma_r > 0$, the second-order Jury conditions satisfy

$$1 - \delta_{ir} = 1 - \frac{1}{1 + \Delta t\gamma_r} \tag{57}$$

$$= \frac{h\gamma_r}{1 + h\gamma_r} > 0, \tag{58}$$

$$1 - \theta_{ir} + \delta_{ir} = 1 - \frac{2 + h\gamma_r - h^2\omega_{ir}^2}{1 + h\gamma_r} + \frac{1}{1 + \Delta t\gamma_r} \tag{59}$$

$$= \frac{h^2\omega_{ir}^2}{1 + h\gamma_r} > 0, \tag{60}$$

$$1 + \theta_{ir} + \delta_{ir} = 1 + \frac{2 + h\gamma_r - h^2\omega_{ir}^2}{1 + h\gamma_r} + \frac{1}{1 + \Delta t\gamma_r} \tag{61}$$

$$= \frac{4 + 2h\gamma_r - h^2\omega_{ir}^2}{1 + h\gamma_r}. \tag{62}$$

If

$$\Delta t^2\omega_{ir}^2 < 4 + 2\Delta t\gamma_r,$$

all three quantities are positive. Both roots of Eq. (56) therefore lie strictly inside the unit disk:

$$\rho(\mathbf{A}_{ir}) < 1.$$

For $\gamma_r = 0$, Eqs. (54) and (55) reduce to

$$\det(\mathbf{A}_{ir}) = 1, \tag{63}$$

$$\mathrm{tr}(\mathbf{A}_{ir}) = 2 - h^2\omega_{ir}^2. \tag{64}$$

Under $0 < h\omega_{ir} < 2$,

$$-2 < 2 - h^2\omega_{ir}^2 < 2.$$

The eigenvalues are therefore a complex-conjugate pair with product one, which gives

$$\rho(\mathbf{A}_{ir}) = 1.$$

For comparison, undamped explicit Euler gives

$$\begin{bmatrix} a_t \\ b_t \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & h \\ -h\omega_{ir}^2 & 1 \end{bmatrix}}_{\mathbf{A}_{ir}^{\mathrm{EE}}} \begin{bmatrix} a_{t-1} \\ b_{t-1} \end{bmatrix}, \tag{65}$$

$$\lambda_{\pm}\left(\mathbf{A}_{ir}^{\mathrm{EE}}\right) = 1 \pm \mathrm{i}h\omega_{ir}, \tag{66}$$

$$\left|\lambda_{\pm}\left(\mathbf{A}_{ir}^{\mathrm{EE}}\right)\right| = \sqrt{1 + h^2\omega_{ir}^2} > 1. \tag{67}$$

Thus, explicit Euler amplifies every nonzero undamped oscillatory mode, whereas the semi-implicit rollout admits a non-amplifying conservative regime and becomes contractive under positive damping and the stated step-size condition. □

Algorithm 1 WaveGSSM forward pass

Require: Observed temporal graph sequence $\{(G_t, \mathbf{X}_t)\}_{t=1}^{T}$

```
Encode each snapshot Z_t^(0) ← φ_in(X_t)
for ℓ = 1, ..., L do
    Initialize pattern state U_0^(ℓ) from Z_1^(ℓ−1)
    Initialize motion state V_0^(ℓ) from the difference between the first hidden frames
    for t = 1, ..., T do
        Obtain graph operator L_t^(ℓ) and forcing F_t^(ℓ)
        Update motion state V_t^(ℓ) using Eq. (6)
        Update pattern state U_t^(ℓ) using Eq. (7)
        Update hidden representation Z_t^(ℓ) using Eqs. (13)–(14)
    end for
end for
Form the terminal representation from Z_T^(L) or [U_T^(L), V_T^(L)]
return Task-specific prediction
```

For a frozen graph operator, a sufficient condition covering all nonconstant graph modes and latent channels is

$$(\Delta t)^2 \kappa_{\max} \lambda_{\max}(\mathbf{L}) < 4 + 2\Delta t \gamma_{\min}. \tag{68}$$

where

$$\kappa_{\max} = \max_r \kappa_r, \qquad \gamma_{\min} = \min_r \gamma_r.$$

If $\gamma_{\min} > 0$, this condition makes every nonconstant frozen mode contractive. If some channels have zero damping, the corresponding conservative modes are non-amplifying under

$$(\Delta t)^2 \kappa_{\max} \lambda_{\max}(\mathbf{L}) < 4.$$

This result concerns the graph-wave integrator under a fixed graph operator and without forcing, rather than the full nonlinear forecasting network. It shows that the proposed semi-implicit discretization avoids the systematic amplification of explicit Euler and admits a stable oscillatory regime for the graph-wave component.

B WaveGSSM Architecture and Implementation Details

This section first summarizes the complete WaveGSSM forward pass and then provides the implementation details for the graph operator, input and forcing maps, wave-state rollout, and task-specific readout.

B.1 Forward Pass Overview

Algorithm 1 summarizes how the components introduced in Section 3 form a complete forward pass. Each input snapshot is first mapped to a hidden representation. Within each WaveGSSM block, the first hidden frames initialize the pattern and motion states. These states are then advanced sequentially over the observed graph snapshots, after which the evolved pattern state updates the hidden representation passed to the next block.

The inner loop is the motion-carrying graph rollout across the observed sequence, while the outer loop stacks multiple WaveGSSM blocks. The motion state is carried between consecutive transitions within a block, whereas the updated hidden sequence is passed between blocks. The following subsections specify each operation in detail.

B.2 Support Graph Construction

The support graph $G_\star = (\mathcal{V}, E_\star)$ defines the candidate edges for latent interaction. For each snapshot, we convert the raw edge set into a simple undirected graph by removing self-loops, ignoring edge direction, and merging duplicated edges.

For an evolving graph, we keep edges that occur in at least k_{sup} snapshots

$$E_\star = \left\{ e \in \bigcup_{t=1}^{T} E_t \;\middle|\; \sum_{t=1}^{T} \mathbf{1}[e \in E_t] \geq k_{\mathrm{sup}} \right\}. \tag{69}$$

Here, k_{sup} is a dataset-specific preprocessing threshold. Setting $k_{\mathrm{sup}} = 1$ gives the union of all observed edge sets, while larger values remove rarely observed edges.

The support graph is constructed once and shared across WaveGSSM blocks. Under the transductive node-classification protocol, the complete structural sequence is observed during training and inference, while node labels are not used to construct the support graph.

After assigning an arbitrary orientation to each support edge, we construct the incidence matrix $\mathbf{B}_\star \in \mathbb{R}^{|E_\star| \times N}$. At snapshot t, $\mathbf{a}_t \in \{0,1\}^{|E_\star|}$ indicates which support edges are active. Block ℓ learns positive base edge strengths $\bar{\mathbf{w}}^{(\ell)} \in \mathbb{R}^{|E_\star|}$ and forms the weighted graph Laplacian $\mathbf{L}_t^{(\ell)}$ defined in Eq. (1). The support topology is shared across blocks, while the learned edge strengths remain block-specific.

For fixed-mesh tasks, no support-frequency threshold is applied. The prescribed mesh directly defines $G_\star$, and all mesh edges remain available throughout the rollout. This is the setting used for ERA5 forecasting.

B.3 Input and Forcing Maps

Input projection. Each observed snapshot is projected to hidden width d as

$$\mathbf{Z}_t^{(0)} = \phi_{\mathrm{in}}(\mathbf{X}_t) = \mathrm{Drop}\left(\mathrm{LN}\left(\mathrm{GELU}\left(\mathbf{X}_t\mathbf{W}_{\mathrm{in}} + \mathbf{1}_N\mathbf{b}_{\mathrm{in}}^\top\right)\right)\right) \in \mathbb{R}^{N\times d}, \tag{70}$$

where $\mathbf{W}_{\mathrm{in}} \in \mathbb{R}^{F_{\mathrm{in}}\times d}$ and $\mathbf{b}_{\mathrm{in}} \in \mathbb{R}^d$. For block $\ell > 1$, the input is the hidden sequence $\{\mathbf{Z}_t^{(\ell-1)}\}_{t=1}^T$ produced by the preceding block.

Gated linear map. The forcing and initialization modules use the gated linear map

$$\mathrm{GL}(\mathbf{Z}) = \left(\mathbf{Z}\mathbf{W}^{(v)} + \mathbf{1}_N(\mathbf{b}^{(v)})^\top\right) \odot \sigma\left(\mathbf{Z}\mathbf{W}^{(g)} + \mathbf{1}_N(\mathbf{b}^{(g)})^\top\right). \tag{71}$$

For $\mathbf{Z} \in \mathbb{R}^{N\times d_{\mathrm{in}}}$, the weights satisfy $\mathbf{W}^{(v)}, \mathbf{W}^{(g)} \in \mathbb{R}^{d_{\mathrm{in}}\times d_{\mathrm{out}}}$ and the biases satisfy $\mathbf{b}^{(v)}, \mathbf{b}^{(g)} \in \mathbb{R}^{d_{\mathrm{out}}}$. Each occurrence of GL uses independent parameters.

Forcing. The node-wise forcing map $g^{(\ell)}$ in Eq. (12) is implemented as

$$\mathbf{F}_t^{(\ell)} = \mathrm{GL}_{\mathrm{force}}^{(\ell)}\left(\mathrm{LN}\left(\mathbf{Z}_t^{(\ell-1)}\right)\right) \in \mathbb{R}^{N\times d}. \tag{72}$$

For ERA5 forecasting, the forcing additionally uses the hidden temporal difference

$$\Delta\mathbf{Z}_t^{(\ell-1)} = \begin{cases} \mathbf{0}_{N\times d}, & t = 1, \\ \mathrm{LN}\left(\mathbf{Z}_t^{(\ell-1)}\right) - \mathrm{LN}\left(\mathbf{Z}_{t-1}^{(\ell-1)}\right), & t > 1, \end{cases} \tag{73}$$

through

$$\mathbf{F}_t^{(\ell)} \leftarrow \mathbf{F}_t^{(\ell)} + \mathrm{GL}_{\mathrm{imp}}^{(\ell)}\left(\Delta\mathbf{Z}_t^{(\ell-1)}\right). \tag{74}$$

Temporal node classification uses the base forcing map.

B.4 Wave-State Initialization and Rollout

Initial state. Each WaveGSSM block initializes its wave state from the incoming hidden sequence. The node-wise initialization map $\psi_U^{(\ell)}$ in the main text is implemented by $\mathrm{GL}_{\mathrm{init},u}^{(\ell)}$, giving

$$\mathbf{U}_0^{(\ell)} = \psi_U^{(\ell)}\left(\mathrm{LN}\left(\mathbf{Z}_1^{(\ell-1)}\right)\right) = \mathrm{GL}_{\mathrm{init},u}^{(\ell)}\left(\mathrm{LN}\left(\mathbf{Z}_1^{(\ell-1)}\right)\right) \in \mathbb{R}^{N\times d}. \tag{75}$$

When at least two input snapshots are available, the initial motion state is obtained from their hidden difference

$$\mathbf{Z}_{\mathrm{vel}}^{(\ell)} = \mathrm{LN}\left(\mathbf{Z}_{2}^{(\ell-1)}\right) - \mathrm{LN}\left(\mathbf{Z}_{1}^{(\ell-1)}\right), \tag{76}$$

$$\mathbf{V}_{0}^{(\ell)} = \psi_{V}^{(\ell)}\left(\mathbf{Z}_{\mathrm{vel}}^{(\ell)}\right) = \mathrm{GL}_{\mathrm{init},v}^{(\ell)}\left(\mathbf{Z}_{\mathrm{vel}}^{(\ell)}\right) \in \mathbb{R}^{N\times d}. \tag{77}$$

For a single-frame input, $\mathbf{V}_{0}^{(\ell)} = \mathbf{0}_{N\times d}$. The pattern-state and motion-state initialization maps use independent parameters.

Within-block rollout. Starting from $(\mathbf{U}_{0}^{(\ell)}, \mathbf{V}_{0}^{(\ell)})$, the block processes $\{\mathbf{L}_{t}^{(\ell)}\}_{t=1}^{T}$ and $\{\mathbf{F}_{t}^{(\ell)}\}_{t=1}^{T}$ using the updates in Eqs. (6) and (7). At each transition, the updated motion state advances the pattern state and is passed directly to the following transition. The motion state therefore acts as a recurrent state within the block.

The resulting pattern-state trajectory is injected into the hidden stream through Eqs. (13) and (14). The motion state remains internal to the block. Across network depth, the updated hidden sequence $\{\mathbf{Z}_{t}^{(\ell)}\}_{t=1}^{T}$ is passed to the next block, which initializes and evolves its own wave state.

Autoregressive weather rollout. For weather forecasting, the explicit pattern and motion states persist only within each graph-wave processor call. Across forecast days, predicted prognostic fields form the rolling input history for the next call together with known forcing and static features. Each new call reconstructs its wave state from this history rather than receiving an explicit $(\mathbf{U}, \mathbf{V})$ pair from the preceding forecast day. No future ERA5 target is used during evaluation.

B.5 Task-Specific Readouts

The final WaveGSSM block provides the terminal hidden representation $\mathbf{Z}_{T}^{(L)}$ and the state readout $\mathbf{H}$ defined in Eq. (15). The implementation supports either the `state` readout $\mathbf{H}$ or the `final` readout $\mathbf{Z}_{T}^{(L)}$.

For temporal node classification, the selected readout for each dataset is listed in Table 4. The selected representation is passed through a learned projection followed by a linear node classifier.

ERA5 forecasting uses the `state` readout. A learned projection maps $\mathbf{H}$ to the mesh-decoder width before the mesh-to-grid decoder predicts the target fields. The frozen-checkpoint readout intervention in Appendix F.3 scales only the terminal motion-state component $\mathbf{V}_{T}^{(L)}$ while leaving $\mathbf{U}_{T}^{(L)}$, the decoder parameters, and all tensor dimensions unchanged.

C Paired Motion Diagnostic

The diagnostic in Figure 1 isolates a basic property of propagating graph signals. Two trajectories are generated on the same irregular graph using the same symmetric propagation operator. They have identical fields at the final observed step but arrive there with opposite recent motion. Their future trajectories therefore differ even though the graph and present field are the same. This construction directly tests whether a model can distinguish futures that cannot be determined from the current snapshot alone.

Compared models. We compare a current-only GCN, GCN-LSTM, GCN-SSM, WaveGSSM without the motion state, and the full WaveGSSM. The temporal models receive the same three observed fields, while the current-only GCN receives only the shared final field. The WaveGSSM variant without the motion state retains the graph architecture and hidden width of the full model but removes the motion-state input and its recurrent propagation. All models are trained with the same free-running multi-step objective and have closely matched parameter counts. Training, validation, and testing use disjoint graphs, and each result is averaged over five random seeds.

Evaluation. The visualized example is selected using only properties of the ground-truth trajectories rather than model performance. Forecast position difference measures the separation between the

predicted right-moving and left-moving patterns along the transport path. A value near zero indicates that the model produces nearly the same future for the two histories. Centroid error measures the distance between the predicted and true pattern locations. Panel (h) in Figure 1 directly compares the rollout errors of WaveGSSM with and without the motion state, while Figure 6 reports aggregate results over all held-out pairs and seeds.

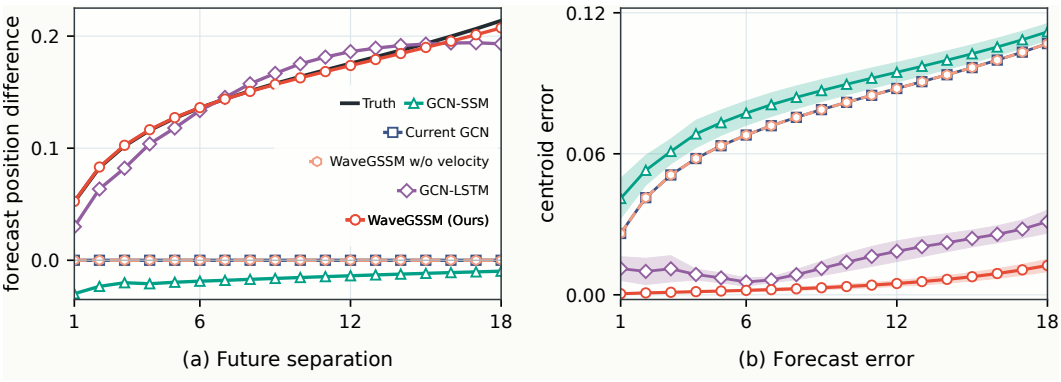


Figure 6: Aggregate paired-motion diagnostic on the held-out graph. Panel (a) reports the signed separation between the predicted right-moving and left-moving pattern centres; the black curve is the ground-truth mean over 48 held-out trajectory pairs. Each model curve first averages over the 48 pairs within a training seed and then over five seeds. Panel (b) reports mean centroid error over the corresponding 96 directional trajectories.

D Experimental Setup

We evaluate WaveGSSM in two complementary settings. Temporal node classification examines whether the proposed state evolution is useful beyond explicitly physical forecasting, while global ERA5 forecasting evaluates the motivating setting in which coherent spatial patterns must be advanced over multiple prediction steps. This section gives the datasets, evaluation protocols, model configurations, and metrics used in both settings.

D.1 Temporal Node Classification

Task and datasets. We evaluate WaveGSSM on the four discrete-time temporal graph benchmarks adopted by GraphSSM (Li et al., 2024), including DBLP-3, Brain, Reddit, and DBLP-10. Each dataset contains a sequence of graph snapshots on a shared node set together with node features and node-level class labels. The task is transductive node classification after processing the observed graph sequence.

Table 3 summarizes the dataset statistics. The structural temporal-continuity score $\mathrm{TC_s}$ is the mean Jaccard overlap between consecutive edge sets, while $\mathrm{TC_f}$ is the mean cosine similarity between consecutive node-feature matrices (Li et al., 2024). The datasets span different temporal regimes, from rapidly changing connectivity in Brain to substantially more persistent structure in DBLP-10.

Table 3: Statistics of the temporal node-classification benchmarks (Li et al., 2024). $\mathrm{TC_s}$ and $\mathrm{TC_f}$ measure temporal continuity of graph structure and node features. The final column gives the train, validation, and test proportions used for WaveGSSM.

Dataset	#Nodes	#Edges	#Feat.	#Cls.	#Steps	$\mathrm{TC_s}$	$\mathrm{TC_f}$	Split
DBLP-3	4,257	23,540	100	3	10	0.139	0.468	81/9/10
Brain	5,000	1,955,488	20	10	12	0.024	0.070	81/9/10
Reddit	8,291	264,050	20	4	10	0.030	0.556	81/9/10
DBLP-10	28,085	236,894	128	10	27	0.823	0.823	75/5/20

Protocol and evaluation. We follow the transductive setting of Li et al. (2024). Graph structure and node features from all snapshots are available during training and inference, while supervision is restricted to the training nodes. The support graph can therefore be constructed from the observed structural sequence without using validation or test labels.

For WaveGSSM, we use the train, validation, and test proportions listed in Table 3. Validation nodes are used only for model selection and early stopping. WaveGSSM is trained with cross-entropy loss on the training nodes. For each seed, we select the checkpoint with the highest validation Micro-F1 and report the corresponding test Micro-F1 and Macro-F1. Results are averaged over five seeds from 0 to 4, with population standard deviation.

Baselines and result provenance. We compare against the baseline families reported by GraphSSM (Li et al., 2024). Static graph embedding methods include DeepWalk (Perozzi et al., 2014) and Node2Vec (Grover & Leskovec, 2016). Temporal graph embedding methods include HTNE (Zuo et al., 2018), M^{2}DNE (Lu et al., 2019), and DynamicTriad (Zhou et al., 2018). Discrete-time temporal GNN baselines include MPNN (Panagopoulos et al., 2021), STAR (Xu et al., 2019), tNodeEmbed (Singer et al., 2019), EvolveGCN (Pareja et al., 2020), SpikeNet (Li et al., 2023a), and ROLAND (You et al., 2022). GraphSSM (Li et al., 2024) is the closest state-space comparison.

The results for these legacy baselines in Table 1 are taken from the GraphSSM benchmark. The GraphSSM row reports results from our rerun of the baseline.

WaveGSSM configuration. For temporal node classification, the observed graph at snapshot t determines the activity vector $\mathbf{a}_t$ and the corresponding operator $\mathbf{L}_t^{(\ell)}$. We set $k_{\sup} = 2$ for DBLP-3 and Brain, $k_{\sup} = 1$ for Reddit, and $k_{\sup} = 4$ for DBLP-10. These thresholds retain edges observed in at least the specified number of snapshots.

Each node snapshot is first encoded by a node-wise MLP and then processed by L stacked WaveGSSM blocks with a learned initial motion state. No auxiliary spatial mixer is used. All four datasets use an FFN expansion ratio of 2 and the `state` readout, which concatenates the terminal pattern and motion states, as described in Section B.5. The selected representation is passed through a learned projection followed by a linear node classifier.

Training uses AdamW, gradient clipping at 1.0, a REDUCELRONPLATEAU learning-rate schedule, and validation-based early stopping. Architecture and optimization hyperparameters are selected separately for each dataset using only its validation split. The resulting configurations are listed in Table 4.

Table 4: WaveGSSM configurations for temporal node classification. Here, d denotes hidden width and L denotes the number of WaveGSSM blocks. All runs use AdamW, gradient clipping at 1.0, validation-based early stopping, and a learned initial motion state.

Dataset	d	L	Epochs	Learning rate	Weight decay	Dropout	Readout
DBLP-3	88	3	200	4.0×10^{-3}	5×10^{-4}	0.12	state
Brain	224	2	220	4.8×10^{-3}	8×10^{-4}	0.20	state
Reddit	144	6	400	1.2×10^{-3}	5×10^{-5}	0.05	state
DBLP-10	280	2	480	4.5×10^{-3}	5×10^{-4}	0.13	state

D.2 ERA5 GLOBAL WEATHER FORECASTING

The ERA5 experiment evaluates the motivating forecasting setting of WaveGSSM, where a multivariate spatial field must be advanced over several autoregressive steps. We evaluate both pointwise forecast accuracy and agreement in large-scale anomaly patterns over a ten-day rollout.

Data and target variables. We use daily ERA5 reanalysis (Hersbach et al., 2020) regridded to a regular 1° latitude–longitude grid with 360×181 grid points. The prognostic state contains eight target channels. Six are geopotential at 150, 200, 300, 500, 850, and 1000 hPa. The remaining two are daily-maximum 2 m temperature and level-1 volumetric soil moisture.

At each forecast step, the model also receives top-of-atmosphere incident solar radiation and sine/cosine encodings of the annual and daily cycles. Static inputs consist of surface geopotential and the land–sea mask. Dynamic channels are standardized using statistics from the training period and transformed back to physical units for evaluation.

Forecast protocol. We train on 1979-01-01 through 2023-12-31, use 2024 for validation, and reserve 2025 for testing. Each forecast starts from two consecutive daily states and advances the prognostic state by one day per autoregressive step. During rollout, the predicted prognostic state is fed back into the next input together with known forcing and static features. No future ERA5 target is used after initialization.

Training uses a three-step free-running objective. Evaluation uses a ten-day autoregressive rollout. Requiring a complete verification window gives 354 test initializations in 2025. Atmospheric variables are evaluated globally, while soil moisture is evaluated only over land.

Graph and model configuration. All weather models use the same 1° grid and global level-5 icosahedral mesh following the graph construction of GraphCast (Lam et al., 2023). The mesh contains $10{,}242$ nodes. The grid-to-mesh encoder contains $101{,}892$ directed edges, the mesh processor contains $81{,}900$ directed edges, and the mesh-to-grid decoder contains $195{,}480$ directed edges. Each graph edge carries four geometric features describing relative spherical displacement and distance, while each mesh node receives three geometric features.

For ERA5, $G_\star$ is the prescribed icosahedral mesh and all mesh edges remain available throughout the rollout. WaveGSSM uses hidden width $d = 128$ and three processor blocks. Node and edge transformations use single-hidden-layer MLPs. The two input states initialize the pattern and motion states as described in Section B.4, and the forcing uses the temporal-difference term defined in Section B.3.

Baselines and controlled comparisons. GraphCast* is our central backbone-matched snapshot baseline. It uses the same grid-to-mesh encoder, mesh-to-grid decoder, hidden width, processor depth, target variables, two-frame history, training horizon, loss, and optimization protocol as WaveGSSM. It predicts future representations without the graph-wave pattern–motion transition. The asterisk denotes our daily ERA5 implementation rather than the published GraphCast checkpoint or its original operational configuration. We therefore refer to it as a GraphCast-style backbone-matched baseline.

GraphSSM (Li et al., 2024) is adapted to the same forecasting input and output interface using two graph state-space blocks with state dimension 16. GNN-LSTM combines per-frame graph processing with a node-shared single-layer LSTM. Both use the same forecast variables, data split, mesh interface, training targets, and evaluation protocol as WaveGSSM. Because they retain their native temporal processors, we treat them as cross-model comparisons rather than controlled component ablations. Persistence and climatology are included as non-learned references. Persistence keeps the initialized prognostic state fixed, while climatology uses the ERA5 day-of-year mean computed from the 1979–2023 training period.

The retrained motion-state-carry ablation and frozen-checkpoint interventions are detailed in Appendices F.2 and F.3.

Training objective and optimization. All matched weather runs use a latitude-weighted multivariate mean-squared-error objective. Grid-point weights are proportional to $\cos\varphi$ and normalized to have unit mean. Each target channel is weighted inversely by the variance of its daily increments over the training period. Geopotential levels additionally use pressure-dependent weights, while the two surface channels use fixed task weights. The loss is averaged over target variables, valid grid points, and the three free-running training steps.

We optimize with AdamW using an initial learning rate of 6×10^{-4}. Training runs for 300 epochs on 4 H100 GPUs with a batch size of 4 per device. Gradient checkpointing is used inside the WaveGSSM processor. The full WaveGSSM and retrained motion-state-carry ablation are selected independently on the 2024 validation set using the same checkpoint-selection rule. All frozen interventions use the selected full WaveGSSM checkpoint without additional training or model selection.

Evaluation metrics. For forecast field $\hat{y}$ and ERA5 target y, the latitude-weighted RMSE is

$$\mathrm{RMSE} = \left[\frac{\sum_i w_i \left(\hat{y}_i - y_i\right)^2}{\sum_i w_i}\right]^{1/2}, \qquad w_i = \cos\varphi_i. \tag{78}$$

The sum is taken over valid grid points. Ocean points are excluded for soil moisture. RMSE is computed separately for each target variable and forecast lead.

We compute anomaly correlation relative to the 1979–2023 ERA5 day-of-year climatology. Let $\hat{a}_i$ and a_i denote forecast and observed anomalies, with latitude-weighted spatial means

$$\overline{\hat{a}}_w = \frac{\sum_i w_i \hat{a}_i}{\sum_i w_i}, \qquad \overline{a}_w = \frac{\sum_i w_i a_i}{\sum_i w_i}. \tag{79}$$

The centred anomaly correlation is

$$\mathrm{AC} = \frac{\sum_i w_i \left(\hat{a}_i - \overline{\hat{a}}_w\right)\left(a_i - \overline{a}_w\right)}{\sqrt{\sum_i w_i \left(\hat{a}_i - \overline{\hat{a}}_w\right)^2}\sqrt{\sum_i w_i \left(a_i - \overline{a}_w\right)^2}}. \tag{80}$$

AC measures agreement in the spatial anomaly pattern after removing the weighted spatial mean. We use it together with RMSE and do not interpret it as a direct measure of physical phase.

E Additional Temporal-Graph Details

This section provides the implementation details of the matched no-carry variant used in Section 4.4.

E.1 Matched No-Carry Variant

Matched no-carry variant. We construct a controlled variant, denoted WaveGSSM w/o motion-state carry. It retains the support graph, learned graph operator, forcing maps, stiffness, damping, initial-motion-state map, hidden width, number of blocks, task readout, classifier, loss, and optimization protocol of the full model. The only change is the state passed from one rollout step to the next.

Let the motion state supplied to transition t be

$$\overline{\mathbf{V}}_{t-1}^{(\ell)} = \begin{cases} \mathbf{V}_0^{(\ell)}, & t = 1, \\ \mathbf{0}_{N\times d}, & t > 1. \end{cases} \tag{81}$$

The input-derived initial motion cue is therefore preserved, but a motion state produced during the rollout is not carried into the following transition. The instantaneous response at step t is

$$\widetilde{\mathbf{V}}_t^{(\ell)} = \left[\overline{\mathbf{V}}_{t-1}^{(\ell)} + \Delta t^{(\ell)}\left(-\mathbf{L}_t^{(\ell)}\mathbf{U}_{t-1}^{(\ell)}\mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)}\right)\right]\left(\mathbf{I}_d + \Delta t^{(\ell)}\mathbf{\Gamma}^{(\ell)}\right)^{-1}, \tag{82}$$

followed by the pattern-state update

$$\mathbf{U}_t^{(\ell)} = \mathbf{U}_{t-1}^{(\ell)} + \Delta t^{(\ell)}\widetilde{\mathbf{V}}_t^{(\ell)}. \tag{83}$$

For every transition after initialization, $\overline{\mathbf{V}}_{t-1}^{(\ell)} = \mathbf{0}$. Eliminating the instantaneous auxiliary variable then gives

$$\mathbf{U}_t^{(\ell)} = \mathbf{U}_{t-1}^{(\ell)} + \left(\Delta t^{(\ell)}\right)^2\left(-\mathbf{L}_t^{(\ell)}\mathbf{U}_{t-1}^{(\ell)}\mathbf{K}^{(\ell)} + \mathbf{F}_t^{(\ell)}\right)\left(\mathbf{I}_d + \Delta t^{(\ell)}\mathbf{\Gamma}^{(\ell)}\right)^{-1}, \qquad t > 1. \tag{84}$$

The recurrent state in Eq. (84) is first-order. The next pattern state depends on the preceding pattern state and the current graph-conditioned forcing, but not on a motion state carried from the preceding step.

The ablated model preserves the same external block interface as the full model. Its terminal instantaneous response $\widetilde{\mathbf{V}}_T^{(\ell)}$ occupies the motion-state slot wherever that tensor is required by the implementation, and the task-specific readout is otherwise unchanged. The comparison therefore removes persistent motion-state recurrence without changing the model width, decoder interface, or number of learnable parameters. The corresponding results are reported in Table 2.

F Additional Weather Forecasting Results and Ablations

This section reports the complete ERA5 forecast results and evaluates the role of the motion state. We first provide lead-wise skill for all predicted fields. We then evaluate a separately retrained model without motion-state carry and use frozen-checkpoint interventions to examine how the trained full model uses its internal motion-state pathway and terminal motion-state readout. All experiments use the 2025 ERA5 test set and the evaluation protocol in Section D.2.

F.1 Lead-Wise Forecast Skill

Table 5 reports the four fields highlighted in the main text, while Table 6 gives the remaining geopotential levels. For RMSE, we report the relative reduction of WaveGSSM over GraphCast*

$$\mathrm{Reduction}_{\mathrm{RMSE}} = \frac{\mathrm{RMSE}_{\mathrm{GC}^*} - \mathrm{RMSE}_{\mathrm{WaveGSSM}}}{\mathrm{RMSE}_{\mathrm{GC}^*}} \times 100\%. \tag{85}$$

For anomaly correlation, we report the absolute gain

$$\Delta\mathrm{AC} = \mathrm{AC}_{\mathrm{WaveGSSM}} - \mathrm{AC}_{\mathrm{GC}^*}. \tag{86}$$

The absolute AC difference avoids inflating the apparent improvement as the reference correlation decreases at longer leads.

The two metrics reveal different lead-time trends. RMSE reductions are largest at short range and generally narrow with lead, while the AC advantage grows into the medium range and remains positive through day 10. WaveGSSM therefore retains better agreement with the observed anomaly pattern even as the pointwise RMSE margin becomes smaller.

WaveGSSM has lower RMSE for every reported geopotential and soil-moisture entry. For 2 m daily maximum temperature, the models are effectively tied at day 10, where the rounded RMSE values differ by only 0.001 K. WaveGSSM has higher anomaly correlation for every reported field and lead.

F.2 Retrained Ablation of Persistent Motion-State Carry

We first test whether carrying the motion state remains useful when the ablated model is allowed to retrain and adapt. The parameter-matched variant is trained independently under the same data, objective, optimizer, rollout horizon, and validation-selection protocol as WaveGSSM. Before every internal graph-wave transition, including the first, its incoming motion state is set to zero. Each transition still computes an instantaneous motion response and uses it to advance the pattern state. The terminal instantaneous motion state also remains available to the decoder.

We use mean geopotential RMSE over all six pressure levels and forecast days 1–5 as the aggregate endpoint

$$\overline{\mathrm{RMSE}}_{\mathrm{geo},1\text{–}5} = \frac{1}{30} \sum_{p\in\mathcal{P}} \sum_{h=1}^{5} \mathrm{RMSE}(p,h), \qquad \mathcal{P} = \{150, 200, 300, 500, 850, 1000\}\ \mathrm{hPa}. \tag{87}$$

We also report z200 at regularly sampled forecast leads because it is already used as the representative upper-tropospheric field in the main text.

Removing motion-state carry increases the mean geopotential RMSE from 30.72 to 31.53 m over all six pressure levels and forecast days 1–5, corresponding to a 2.63% increase. The same direction holds at every reported z200 lead, with relative increases between 2.46% and 3.90%.

The difference is modest but directionally consistent across the reported leads. Because the ablated model is retrained from initialization, its remaining pathways can adapt to the removed motion-state carry. The remaining performance gap indicates that carrying the motion state provides additional predictive value within the proposed architecture.

F.3 Frozen-Checkpoint Motion-State Interventions

The retrained comparison in Table 7 evaluates motion-state carry after the ablated model has adapted through training. We next examine how the selected full predictor uses this pathway at inference time. We freeze the validation-selected WaveGSSM checkpoint and separately intervene on the motion state entering the internal graph-wave transitions and the terminal motion state supplied to the decoder.

Table 5: Lead-wise forecast skill for the four fields highlighted in the main text on the 2025 ERA5 test set. The upper block reports latitude-weighted RMSE and its relative reduction under WaveGSSM. The lower block reports centred anomaly correlation and the absolute gain $\Delta\mathrm{AC} = \mathrm{AC}_{\mathrm{WaveGSSM}} - \mathrm{AC}_{\mathrm{GC}^*}$. RMSE units are metres for $z500$ and $z200$, kelvin for 2 m daily maximum temperature, and $10^{-3}\,\mathrm{m}^3\mathrm{m}^{-3}$ for soil moisture layer 1.

Field	Model	Lead time (days)									
		1	2	3	4	5	6	7	8	9	10
RMSE ↓											
	WaveGSSM	11.7	19.6	28.7	38.6	48.5	57.8	66.1	73.3	79.3	84.1
$z500$ (m)	GraphCast*	18.9	25.4	33.4	42.8	52.2	61.1	68.9	75.5	80.7	85.0
	Reduction	**38.1%**	**22.9%**	**14.1%**	**9.6%**	**7.0%**	**5.3%**	**4.1%**	**2.8%**	**1.7%**	**1.1%**
	WaveGSSM	15.1	24.7	35.8	48.1	60.6	72.5	83.4	93.0	101.4	108.2
$z200$ (m)	GraphCast*	24.7	33.6	44.0	56.7	69.8	82.4	93.9	103.6	111.6	118.3
	Reduction	**39.1%**	**26.6%**	**18.6%**	**15.2%**	**13.2%**	**12.0%**	**11.2%**	**10.2%**	**9.2%**	**8.5%**
	WaveGSSM	0.910	1.142	1.364	1.616	1.872	2.120	2.348	2.551	2.740	2.903
2t max (K)	GraphCast*	1.014	1.263	1.480	1.715	1.952	2.184	2.400	2.594	2.762	2.902
	Reduction	**10.2%**	**9.6%**	**7.8%**	**5.8%**	**4.1%**	**2.9%**	**2.2%**	**1.7%**	**0.8%**	0.0%
	WaveGSSM	10.50	15.98	19.67	22.81	25.73	28.53	31.15	33.59	35.87	37.97
sm L1 (10^{-3})	GraphCast*	11.12	16.94	20.86	24.13	27.08	29.83	32.36	34.72	36.97	39.05
	Reduction	**5.5%**	**5.7%**	**5.7%**	**5.5%**	**5.0%**	**4.4%**	**3.7%**	**3.3%**	**3.0%**	**2.8%**
Centred anomaly correlation ↑											
	WaveGSSM	0.989	0.969	0.932	0.874	0.797	0.707	0.611	0.515	0.427	0.353
$z500$	GraphCast*	0.971	0.949	0.908	0.844	0.761	0.663	0.562	0.467	0.384	0.313
	ΔAC	**+0.018**	**+0.020**	**+0.024**	**+0.030**	**+0.036**	**+0.044**	**+0.049**	**+0.048**	**+0.043**	**+0.040**
	WaveGSSM	0.989	0.971	0.938	0.887	0.817	0.735	0.645	0.554	0.469	0.395
$z200$	GraphCast*	0.971	0.949	0.911	0.851	0.772	0.678	0.579	0.486	0.405	0.336
	ΔAC	**+0.018**	**+0.022**	**+0.027**	**+0.036**	**+0.045**	**+0.057**	**+0.066**	**+0.068**	**+0.064**	**+0.059**
	WaveGSSM	0.937	0.899	0.853	0.791	0.717	0.636	0.553	0.473	0.392	0.321
2t max	GraphCast*	0.921	0.875	0.824	0.758	0.680	0.594	0.508	0.424	0.350	0.287
	ΔAC	**+0.016**	**+0.024**	**+0.029**	**+0.033**	**+0.037**	**+0.042**	**+0.045**	**+0.049**	**+0.042**	**+0.034**
	WaveGSSM	0.978	0.948	0.920	0.891	0.860	0.827	0.792	0.756	0.720	0.685
sm L1	GraphCast*	0.975	0.941	0.910	0.879	0.846	0.812	0.777	0.742	0.707	0.673
	ΔAC	**+0.003**	**+0.007**	**+0.010**	**+0.012**	**+0.014**	**+0.015**	**+0.015**	**+0.014**	**+0.013**	**+0.012**

Intervention semantics. Let α_{c} scale the motion state entering the internal graph-wave transitions and α_{r} scale the terminal motion state supplied to the decoder. Before every internal transition, including the first, the incoming motion state is replaced by

$$\widetilde{\mathbf{V}}_{t-1}^{(\ell)} = \alpha_{\mathrm{c}} \mathbf{V}_{t-1}^{(\ell)}. \tag{88}$$

The transition then computes its instantaneous motion response and advances the pattern state as usual. At the decoder, the state readout becomes

$$\widetilde{\mathbf{H}}_{\mathrm{state}} = \left[\mathbf{U}_T^{(L)}, \alpha_{\mathrm{r}} \mathbf{V}_T^{(L)}\right]. \tag{89}$$

Setting $\alpha_{\mathrm{c}} = 0$ suppresses both the initialized motion cue and the motion state carried between subsequent internal transitions, while preserving the instantaneous response computed by each transition. Setting $\alpha_{\mathrm{r}} = 0$ removes only the terminal motion state from the decoder input. Neither intervention changes the model parameters, graph structure, forcing, or tensor dimensions.

The explicit wave state is propagated only within each graph-wave processor call. Across forecast days, predicted prognostic fields form the next rolling input history, from which a new wave state is reconstructed. An explicit $(\mathbf{U}, \mathbf{V})$ pair is not passed directly between forecast days.

Design and endpoint. All conditions use the same checkpoint selected only on 2024 validation data. They are evaluated on the same ordered 354 initializations from the held-out year 2025, all eight output channels, and lead days 1–10. The frozen 2×2 factorial evaluates

$$(\alpha_{\mathrm{c}}, \alpha_{\mathrm{r}}) \in \{(1,1), (0,1), (1,0), (0,0)\}. \tag{90}$$

Table 6: Lead-wise forecast skill for the remaining geopotential levels. The upper block reports latitude-weighted RMSE in metres and its relative reduction under WaveGSSM. The lower block reports centred anomaly correlation and the absolute gain ΔAC. WaveGSSM improves both metrics at every pressure level and lead shown.

Level	Model	Lead time (days)									
		1	2	3	4	5	6	7	8	9	10
RMSE (m) ↓											
	WaveGSSM	9.7	15.6	22.3	29.3	36.1	42.4	47.8	52.5	56.2	59.2
1000 hPa	GraphCast*	15.6	20.1	25.9	32.4	39.0	45.0	50.1	54.4	57.9	60.7
	Reduction	**37.7%**	**22.1%**	**14.0%**	**9.8%**	**7.4%**	**5.7%**	**4.7%**	**3.6%**	**2.8%**	**2.5%**
	WaveGSSM	9.1	14.9	21.4	28.1	34.8	41.0	46.4	51.1	54.8	57.7
850 hPa	GraphCast*	15.1	19.3	24.9	31.2	37.6	43.4	48.5	52.7	56.1	58.9
	Reduction	**39.7%**	**22.7%**	**14.2%**	**9.9%**	**7.3%**	**5.6%**	**4.3%**	**3.1%**	**2.2%**	**1.9%**
	WaveGSSM	16.4	26.7	39.0	52.6	66.1	78.9	90.4	100.4	109.0	115.9
300 hPa	GraphCast*	25.8	34.9	45.9	59.0	72.2	84.8	96.1	105.5	113.1	119.3
	Reduction	**36.2%**	**23.4%**	**14.9%**	**10.9%**	**8.5%**	**7.0%**	**5.9%**	**4.8%**	**3.7%**	**2.9%**
	WaveGSSM	14.2	23.5	33.0	43.3	53.9	64.1	73.5	82.1	89.5	95.7
150 hPa	GraphCast*	23.4	33.0	42.2	53.0	64.4	75.5	85.8	94.8	102.3	108.6
	Reduction	**39.2%**	**28.8%**	**21.7%**	**18.2%**	**16.3%**	**15.0%**	**14.2%**	**13.4%**	**12.5%**	**11.9%**
Centred anomaly correlation ↑											
	WaveGSSM	0.984	0.959	0.914	0.848	0.763	0.667	0.569	0.475	0.390	0.319
1000 hPa	GraphCast*	0.959	0.932	0.882	0.810	0.715	0.609	0.503	0.407	0.323	0.252
	ΔAC	**+0.025**	**+0.027**	**+0.032**	**+0.038**	**+0.048**	**+0.058**	**+0.066**	**+0.068**	**+0.067**	**+0.067**
	WaveGSSM	0.986	0.962	0.920	0.856	0.775	0.680	0.584	0.489	0.405	0.333
850 hPa	GraphCast*	0.961	0.936	0.889	0.820	0.729	0.625	0.521	0.425	0.341	0.270
	ΔAC	**+0.025**	**+0.026**	**+0.031**	**+0.036**	**+0.046**	**+0.055**	**+0.063**	**+0.064**	**+0.064**	**+0.063**
	WaveGSSM	0.988	0.968	0.931	0.873	0.797	0.708	0.613	0.518	0.429	0.355
300 hPa	GraphCast*	0.971	0.947	0.906	0.843	0.759	0.662	0.561	0.466	0.385	0.316
	ΔAC	**+0.017**	**+0.021**	**+0.025**	**+0.030**	**+0.038**	**+0.046**	**+0.052**	**+0.052**	**+0.044**	**+0.039**
	WaveGSSM	0.989	0.969	0.938	0.890	0.827	0.751	0.669	0.584	0.504	0.433
150 hPa	GraphCast*	0.970	0.944	0.907	0.851	0.777	0.690	0.596	0.507	0.428	0.361
	ΔAC	**+0.019**	**+0.025**	**+0.031**	**+0.039**	**+0.050**	**+0.061**	**+0.073**	**+0.077**	**+0.076**	**+0.072**

Table 7: **Retrained ablation of motion-state carry on ERA5.** Both variants are independently trained under the same protocol with identical parameter counts. The aggregate geopotential RMSE averages all six pressure levels over forecast days 1–5. All values are in metres and lower is better.

Model	Geopotential RMSE 6 levels, D1–5	z200 RMSE				
		D1	D3	D5	D7	D10
WaveGSSM	**30.72**	**15.08**	**35.77**	**60.59**	**83.36**	**108.17**
WaveGSSM w/o motion-state carry	31.53	15.59	37.16	62.49	85.58	110.84
Relative increase	+2.63%	+3.40%	+3.90%	+3.13%	+2.66%	+2.46%

The continuous analysis fixes one scale at one while sweeping the other over $\{1, .75, .5, .25, 0\}$. The scale grid was fixed before inspecting the test results, and its endpoints coincide with the corresponding factorial conditions.

For intervention condition m, initialization i, lead h, and output channel q, let $\mathrm{wMSE}_{m,i,h,q}$ be the cosine-latitude-weighted mean squared error and let s_q be the fixed training scale of channel q stored with the checkpoint runtime. The primary dimensionless endpoint is

$$E_m = \sqrt{\frac{1}{NHQ}\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{q=1}^{Q}\frac{\mathrm{wMSE}_{m,i,h,q}}{s_q^2}}, \qquad \Delta_m = 100\left(\frac{E_m}{E_{\mathrm{Full}}} - 1\right), \tag{91}$$

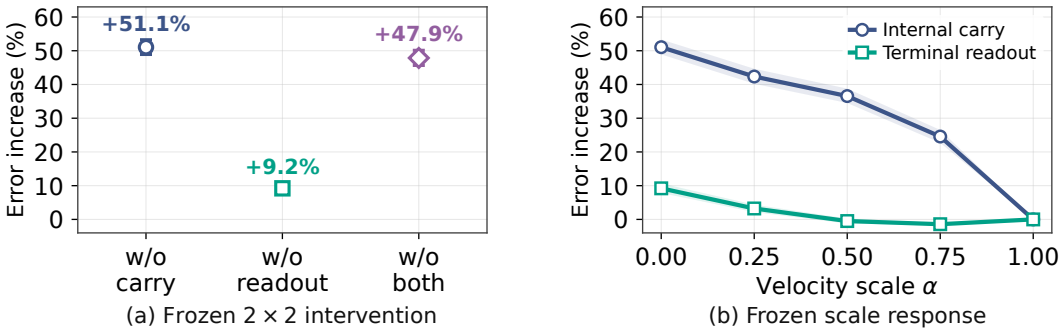


Figure 7: Frozen-checkpoint functional reliance on the motion-state pathway. All conditions use the same validation-selected checkpoint and 354 held-out 2025 ERA5 initializations without retraining. Vertical axes report percentage changes in aggregate standardized error relative to Full over eight fields and forecast days 1–10. (a) Non-Full conditions in the internal-pathway × readout design. (b) Error response to scaling the internal motion-state pathway or terminal motion-state readout while keeping the other pathway unchanged. The plotted "w/o carry" condition corresponds to $\alpha_c = 0$ and therefore suppresses the motion state entering all internal transitions, including the initialized motion state.

Table 8: Frozen-checkpoint motion-state intervention on ERA5. Values report the percentage change in aggregate standardized error relative to Full.

Condition	Internal path	Readout	Error change (%)
Full	On	On	0.0
Internal path off	Off	On	+51.1
Readout off	On	Off	+9.2
Both off	Off	Off	+47.9

where $N = 354$, $H = 10$, and $Q = 8$. Soil moisture is evaluated over land. Positive Δ_m means that the intervention increases forecast error.

Factorial results. Table 8 reports the complete 2×2 design. Suppressing the internal motion-state pathway produces the largest degradation. Removing terminal readout alone has a smaller effect, while joint removal is less harmful than the sum of the two individual effects, indicating that the two pathways are not simply additive.

Continuous scale response. Table 9 shows a graded response to the internal motion-state scale. Reducing α_c consistently increases the aggregate endpoint. The degradation at $\alpha_c = 0$ is therefore part of a continuous sensitivity pattern rather than an isolated effect of complete removal.

Terminal readout behaves differently. A scale of .75 slightly improves the frozen endpoint, .5 remains close to Full, and stronger attenuation becomes harmful. This non-monotonic response suggests that the terminal motion state acts partly through feature calibration rather than simple dependence on its magnitude. Because the scales are test-time interventions on the held-out year, the favorable .75 condition is treated only as a diagnostic.

Structure of the intervention effect. Figure 8 shows that suppressing the internal motion-state pathway is harmful across the full ten-day horizon and across geopotential, temperature, and soil moisture. The relative effect is largest at short lead, where the Full forecast error is smallest. This trend should not be interpreted as accumulation of an explicit motion state across forecast days, since each processor call reconstructs its wave state from the rolling weather history.

For z500, suppressing the internal pathway degrades both phase coherence and amplitude fidelity over days 3–7 and zonal wavenumbers 4–9. The larger spectral contribution comes from amplitude error, so we do not attribute the intervention effect primarily to phase.

Table 9: Motion-state scale sweep on the frozen ERA5 checkpoint. Values report percentage changes in aggregate standardized error relative to Full, with the other pathway kept at full scale. Lower is better.

Scale α	Internal path error change (%)	Terminal readout error change (%)
1.00	0.0	0.0
0.75	+24.6	−1.4
0.50	+36.6	−0.5
0.25	+42.4	+3.2
0.00	+51.1	+9.2

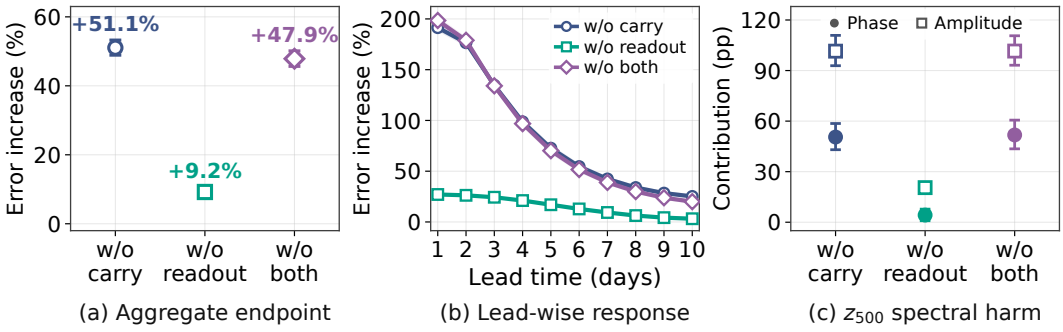


Figure 8: Frozen-checkpoint motion-state diagnostics. (a) Aggregate standardized error increase relative to Full for the three non-Full interventions over eight fields and forecast days 1–10. (b) Lead-wise error increase relative to Full. (c) Phase and amplitude contributions to z500 spectral error over days 3–7 and zonal wavenumbers 4–9, normalized by the Full spectral error and reported in percentage points. Suppressing the internal motion-state pathway increases error across forecast leads and affects both spectral components, with the larger contribution coming from amplitude.

The phase-specific damage under $\alpha_c = 0$ increases with initialization wave activity (Figure 9a). A season-adjusted linear fit gives a slope of 10.37 percentage points per activity standard deviation, consistent with a stronger phase effect in more wave-active conditions. This relationship is specific to the phase component and does not imply the same activity dependence for total forecast error.

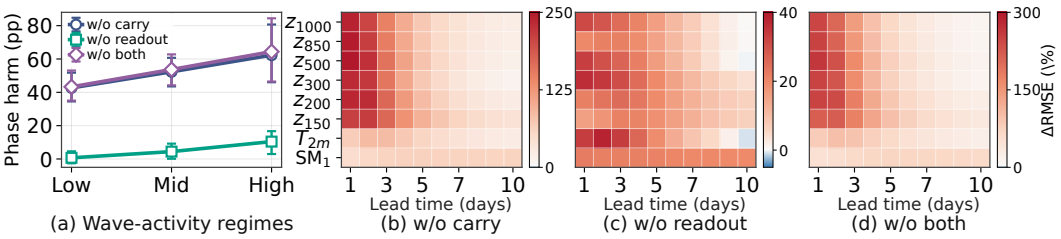


Figure 9: Regime and variable detail for the frozen factorial. (a) Phase-harm contribution over days 3–7 and zonal wavenumbers 4–9, normalized by Full total spectral error and expressed in percentage points, stratified by initialization wave activity using thresholds fixed from the 2024 validation set. (b)–(d) Physical-RMSE increase relative to Full by target field and lead time under suppression of the internal motion-state pathway, removal of terminal readout, and joint removal. Each heatmap uses its own colorbar to show within-condition structure. Conditions without the internal motion-state pathway show broader degradation, while readout-only effects are smaller.

G Atmospheric Forecast Diagnostics

We next examine the spatial and structural character of the ERA5 forecast improvement. All diagnostics use the same 354 held-out 2025 initializations and the evaluation protocol in Section D.2. We examine whether the gain is geographically broad, whether it extends across atmospheric levels, whether it can be explained by mean-bias correction, and whether the forecasts retain synoptic-scale variability.

G.1 Additional Regime and Spectral Detail

Figure 10 provides the supporting breakdowns omitted from the focused main-text analysis. The seasonal and regional views show that the z500 anomaly-correlation advantage is not confined to one initialization season or geographical region, while the day-7 spectrum resolves how power is distributed within and around the selected synoptic band.

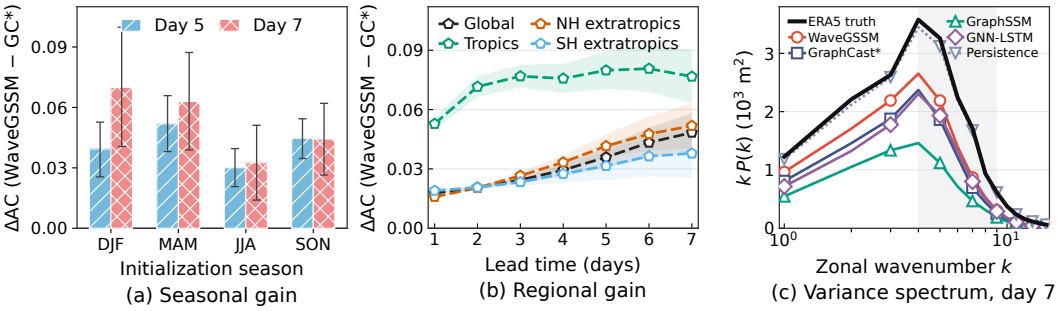


Figure 10: Additional seasonal, regional, and spectral detail for the z500 forecast advantage. Panels (a) and (b) report $\Delta\mathrm{AC} = \mathrm{AC}_{\mathrm{WaveGSSM}} - \mathrm{AC}_{\mathrm{GC}^*}$ relative to GraphCast*. Panel (a) groups the Northern-Hemisphere extratropical gain at days 5 and 7 by initialization season, where DJF, MAM, JJA, and SON denote December–February, March–May, June–August, and September–November. Panel (b) reports the gain by region and forecast lead. Panel (c) shows the day-7 area-weighted z500 spectrum, with zonal wavenumbers 4–9 shaded.

G.2 Horizontal Spatial Skill

Global RMSE does not reveal whether a forecast improvement is spatially broad. We therefore evaluate local skill at every valid grid point.

For variable v, lead h, and grid point $x = (\lambda, \varphi)$, let $\mathrm{RMSE}_M(v, h, x)$ denote the RMSE of model M over the 354 test initializations. We define

$$S_{v,h}(x) = \frac{\mathrm{RMSE}_{\mathrm{WaveGSSM}}(v, h, x) - \mathrm{RMSE}_{\mathrm{GC}^*}(v, h, x)}{\mathrm{RMSE}_{\mathrm{GC}^*}(v, h, x)}. \tag{92}$$

Negative values favor WaveGSSM and positive values favor GraphCast*. The maps use blue and red accordingly. We also report the fraction of valid grid points with $S_{v,h}(x) < 0$ as a measure of the spatial coverage of the improvement.

The geopotential advantage is geographically broad. At day 1, WaveGSSM reduces global RMSE by 39.1% for $z200$ and 38.1% for $z500$, while achieving lower local RMSE at more than 99% of valid grid points for both fields. The advantage remains widespread as forecast lead increases. At day 7, WaveGSSM has lower local RMSE at 94.6% of valid $z200$ grid points and 81.8% of valid $z500$ grid points.

The surface variables show a weaker and less uniform spatial advantage. WaveGSSM improves 2 m daily maximum temperature and soil moisture layer 1 over much of the domain, but the gains are less extensive than for upper-air geopotential. Together, the maps show that the geopotential improvement is spatially broad rather than concentrated in a small number of favorable regions.

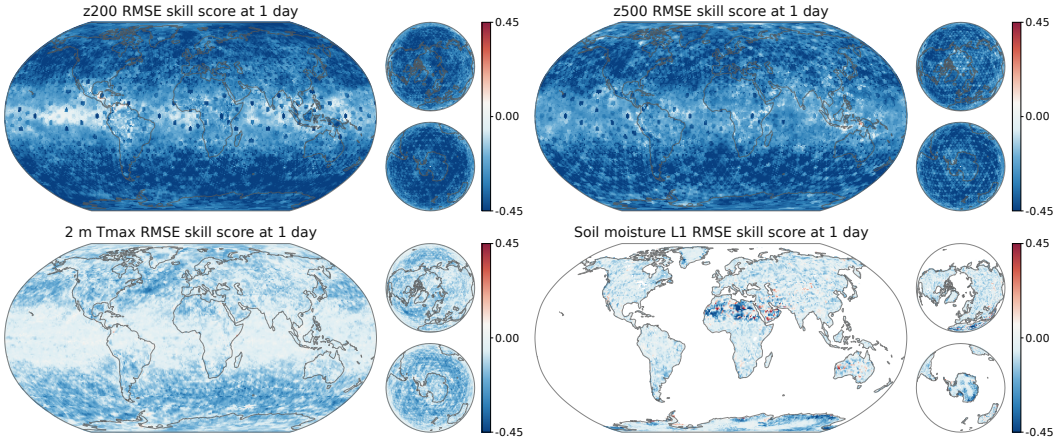


Figure 11: Spatial RMSE skill relative to GraphCast* at lead day 1. Each panel shows the normalized local RMSE difference in Eq. (92) over the 2025 test initializations. Blue indicates lower WaveGSSM RMSE, while red indicates lower GraphCast* RMSE. The variables are $z200$, $z500$, 2 m daily maximum temperature, and soil moisture layer 1. Polar insets display the high-latitude regions. The short-range geopotential improvement is broad rather than concentrated in a small number of locations.

G.3 Vertical Structure of Geopotential Skill

The horizontal maps show that the geopotential improvement is geographically broad. We next examine how this advantage varies across latitude and pressure level.

For model M, lead h, latitude φ, and pressure level p, we compute RMSE over longitude and the test initializations

$$\mathrm{RMSE}_M(h, \varphi, p) = \left[\frac{1}{N_{\mathrm{test}} I} \sum_{n=1}^{N_{\mathrm{test}}} \sum_{i=1}^{I} \left(\hat{z}_{M,n}(h, \lambda_i, \varphi, p) - z_n(h, \lambda_i, \varphi, p) \right)^2 \right]^{1/2}, \tag{93}$$

where $N_{\mathrm{test}} = 354$ and $I = 360$. We then define the normalized skill difference

$$S_h(\varphi, p) = \frac{\mathrm{RMSE}_{\mathrm{WaveGSSM}}(h, \varphi, p) - \mathrm{RMSE}_{\mathrm{GC}^*}(h, \varphi, p)}{\mathrm{RMSE}_{\mathrm{GC}^*}(h, \varphi, p)}. \tag{94}$$

Negative values favor WaveGSSM.

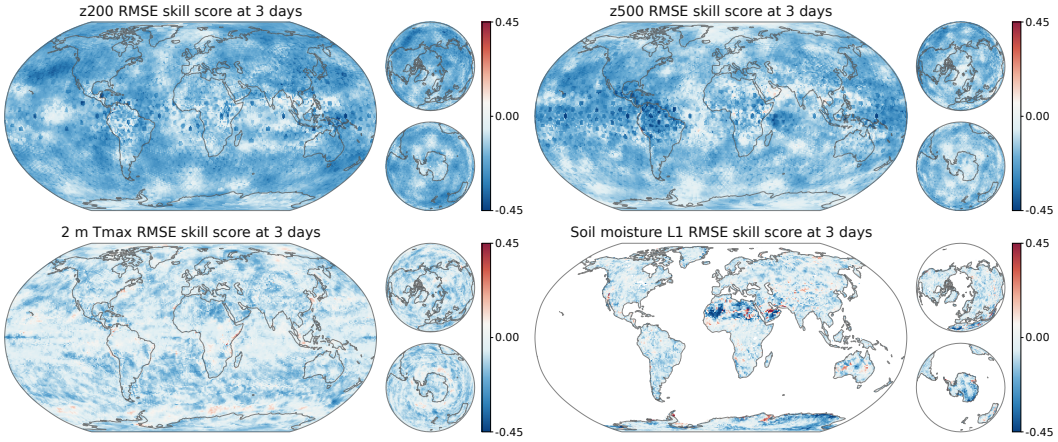


Figure 12: Spatial RMSE skill relative to GraphCast* at lead day 3. The score and color convention follow Figure 11. WaveGSSM reduces global RMSE by 18.6% for $z200$ and 14.1% for $z500$, and the corresponding local improvement remains geographically widespread.

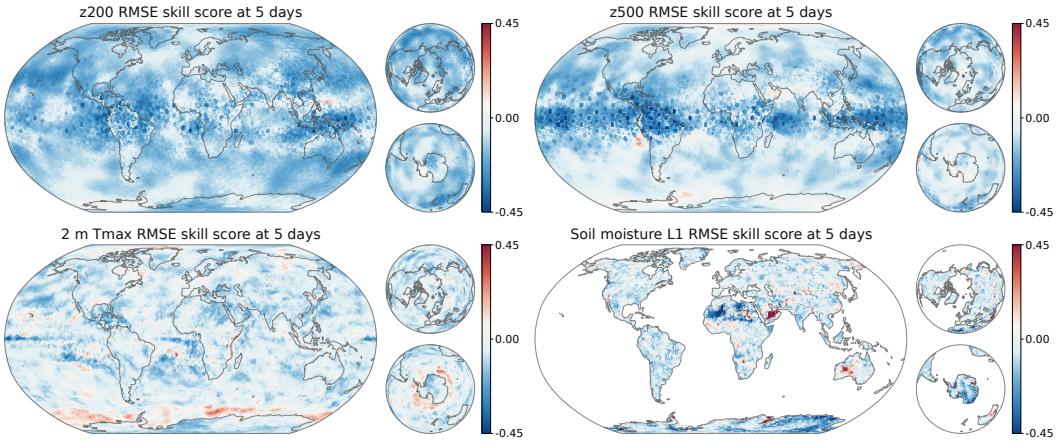


Figure 13: Spatial RMSE skill relative to GraphCast* at lead day 5. Blue denotes lower local RMSE for WaveGSSM under the normalized score in Eq. (92). The strongest and most continuous improvements occur in the geopotential fields, while the surface-variable gains are more regionally heterogeneous.

Figure 15 shows that the improvement extends across the predicted pressure levels rather than being confined to a single field. The strongest and most persistent gains appear in the upper-level geopotential fields, particularly around 200 and 300 hPa. This pattern is consistent with the lead-wise results, where the upper-level fields retain some of the largest medium-range advantages.

The latitude–pressure structure is also coherent across broad regions rather than appearing as isolated cells. Together with the horizontal maps, this shows that the geopotential improvement extends across both geographical location and atmospheric level.

G.4 Systematic Bias Analysis

Lower RMSE can reflect better transient forecasts, smaller systematic bias, or both. We therefore examine the mean forecast bias separately from the total error.

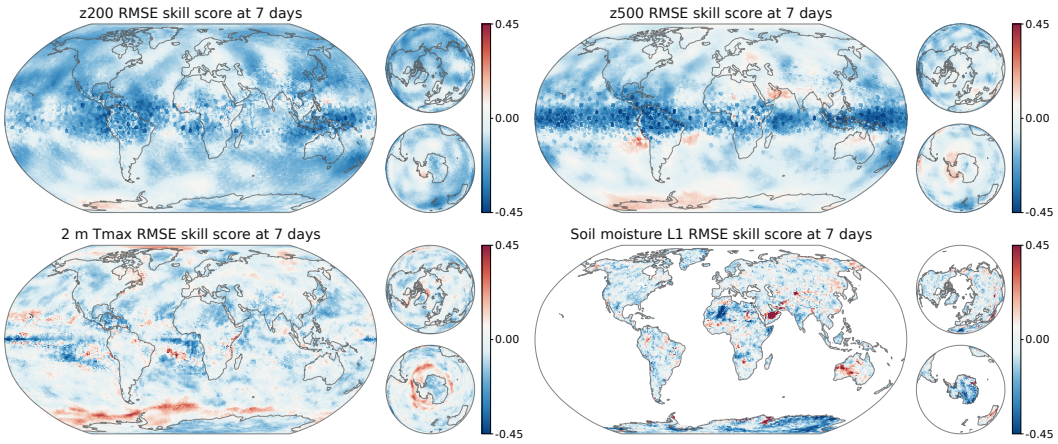


Figure 14: Spatial RMSE skill relative to GraphCast* at lead day 7. The score and color convention follow the earlier lead-time maps. WaveGSSM has lower local RMSE at 94.6% of valid $z200$ grid points and 81.8% of valid $z500$ grid points. The pointwise margin is smaller than at short range, but the upper-air advantage remains spatially broad after one week.

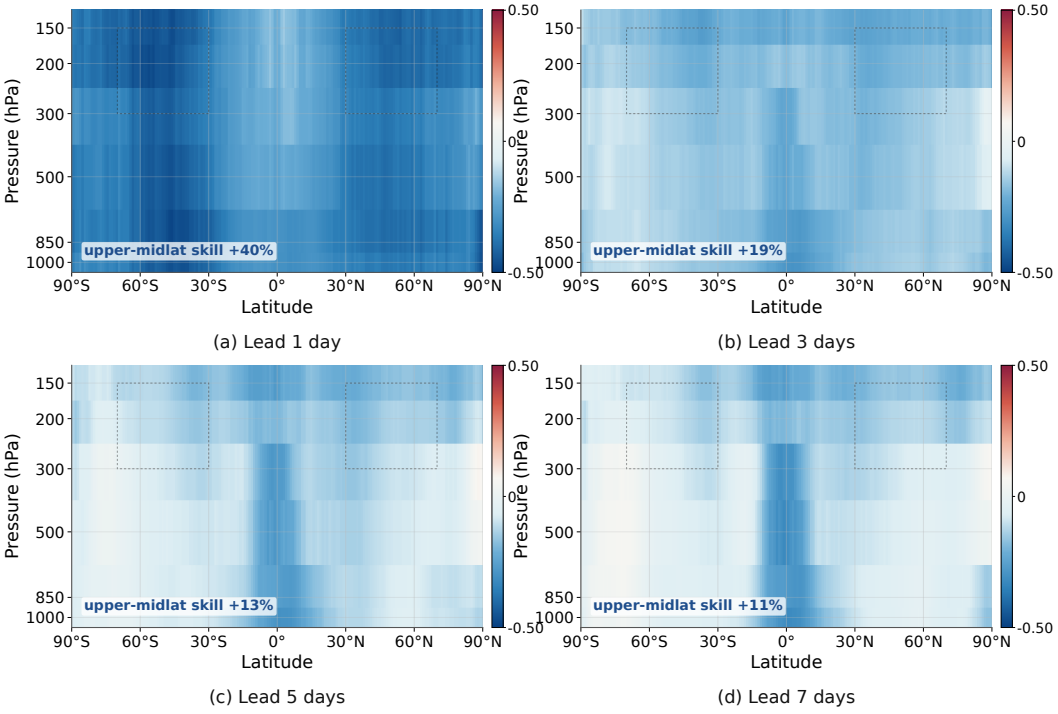


Figure 15: Latitude–pressure structure of geopotential skill. Each panel shows the normalized RMSE difference in Eq. (94) over longitude and the 2025 test initializations. Blue indicates lower WaveGSSM RMSE and red indicates lower GraphCast* RMSE. The advantage extends across multiple pressure levels and remains strongest in the upper-level geopotential fields.

For model M, variable v, lead h, and grid point $x = (\lambda, \varphi)$, we define

$$B_M(v,h,x) = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \left[\hat{y}_{M,n}(v,h,x) - y_n(v,h,x)\right], \qquad N_{\text{test}} = 354. \tag{95}$$

Positive values indicate forecasts above ERA5 on average, while negative values indicate forecasts below ERA5.

To compare bias magnitude between models, we use

$$D_{\mathrm{bias}}(v,h,x) = |B_{\mathrm{WaveGSSM}}(v,h,x)| - |B_{\mathrm{GC}^*}(v,h,x)|\,. \tag{96}$$

Negative values favor WaveGSSM. We summarize each map using the latitude-weighted absolute bias

$$\mathrm{ABias}_M(v,h) = \frac{\sum_x w_x\,|B_M(v,h,x)|}{\sum_x w_x}, \qquad w_x = \cos\varphi_x, \tag{97}$$

$$\mathrm{Reduction}_{\mathrm{bias}} = \frac{\mathrm{ABias}_{\mathrm{GC}^*} - \mathrm{ABias}_{\mathrm{WaveGSSM}}}{\mathrm{ABias}_{\mathrm{GC}^*}} \times 100\%. \tag{98}$$

For soil moisture, the sums include land points only.

WaveGSSM substantially reduces systematic bias in the geopotential fields. For $z200$, the latitude-weighted absolute bias is reduced by 61.6%, 62.4%, and 59.3% at days 3, 5, and 7. The corresponding reductions for $z500$ are 35.0%, 30.9%, and 27.3%. The improvement therefore persists across the medium-range forecast window.

The surface fields show a weaker effect. At day 7, absolute bias is reduced by 12.1% for 2 m daily maximum temperature, while soil moisture changes little. The strongest bias reductions therefore occur in the geopotential fields, where the overall forecast advantage is also most pronounced.

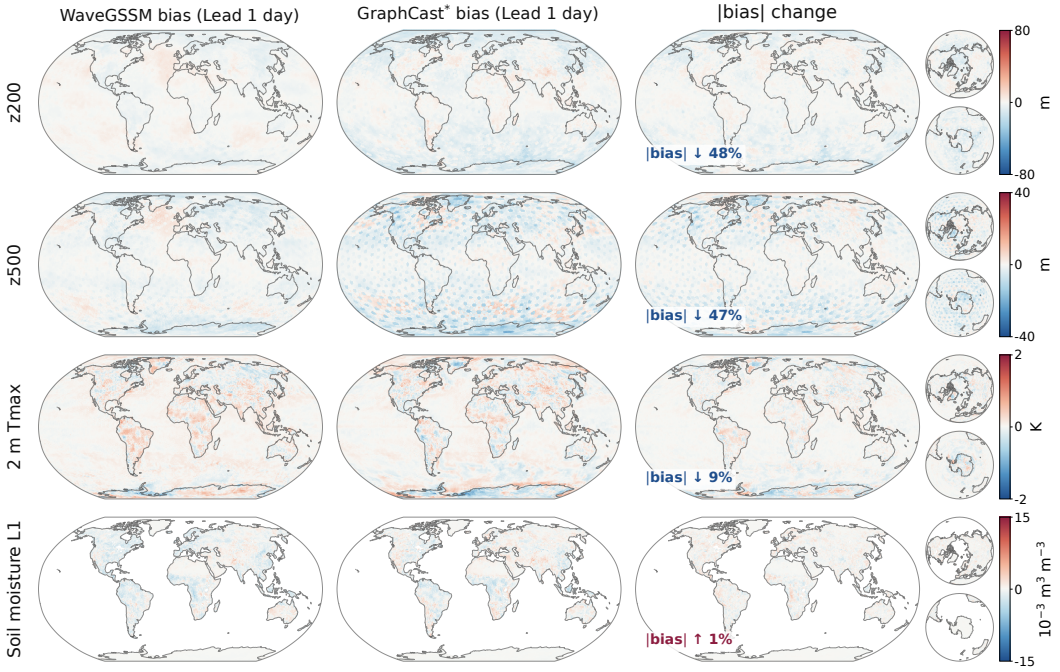


Figure 16: Systematic forecast bias at lead day 1. The first two columns show signed mean bias for WaveGSSM and GraphCast*. The third column shows the bias-magnitude difference in Eq. (96). Blue indicates smaller bias magnitude for WaveGSSM and red indicates smaller bias magnitude for GraphCast*.

G.5 Synoptic-Scale Spectral Retention

Lower squared error can also be obtained by damping spatial variability toward a smoother forecast (Lam et al., 2023; Price et al., 2025). We therefore examine whether WaveGSSM preserves the spatial scales associated with evolving weather systems rather than obtaining lower error through stronger smoothing.

For each initialization and forecast lead, we compute geopotential anomalies relative to the 1979–2023 ERA5 day-of-year climatology

$$a(\varphi_j,\lambda_i) = f(\varphi_j,\lambda_i) - c(\varphi_j,\lambda_i), \tag{99}$$

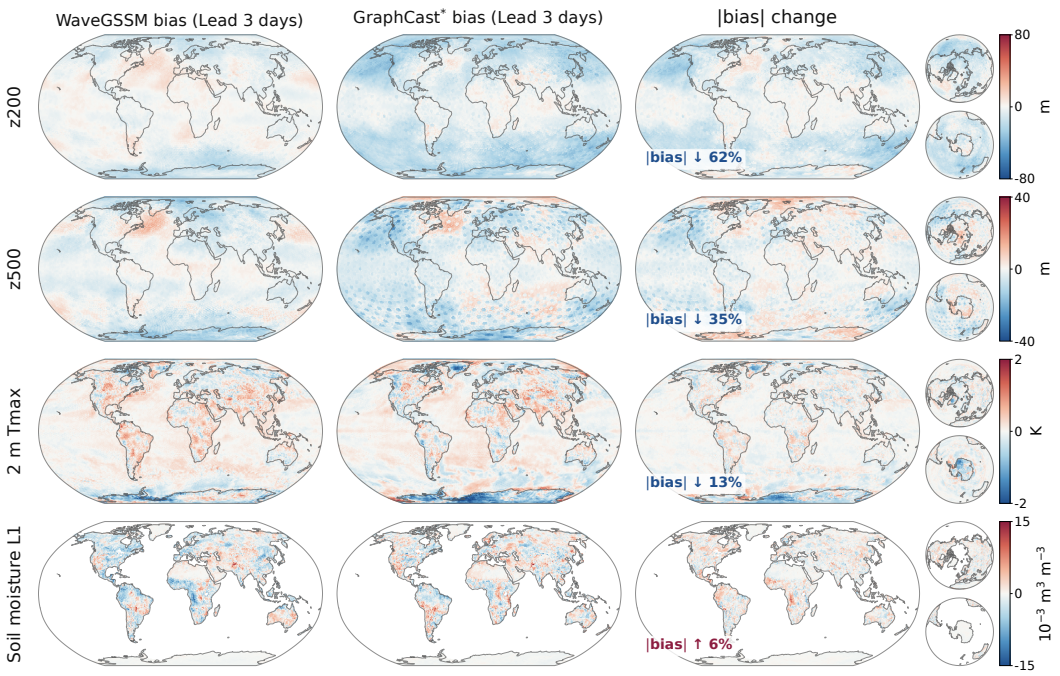


Figure 17: Systematic forecast bias at lead day 3. WaveGSSM reduces latitude-weighted absolute bias by 61.6% for $z200$ and 35.0% for $z500$. The surface-field changes are smaller.

where f denotes either a forecast or the corresponding ERA5 field. Geopotential is converted to geopotential height before spectral analysis.

At each latitude φ_j, we compute the discrete Fourier transform along the $I = 360$ longitude points

$$\widehat{a}_j(k) = \sum_{i=0}^{I-1} a(\varphi_j, \lambda_i) \exp\left(-\frac{2\pi \mathrm{i} k i}{I}\right), \qquad k = 0, \ldots, \left\lfloor \frac{I}{2} \right\rfloor. \tag{100}$$

The corresponding one-sided Parseval-normalized power is

$$P_j(k) = \frac{m(k)}{I^2} \left|\widehat{a}_j(k)\right|^2, \tag{101}$$

where $m(k) = 2$ for positive non-Nyquist wavenumbers and $m(k) = 1$ for the zero and Nyquist components. This normalization makes the summed spectrum equal to the zonal anomaly variance.

For initialization n, we aggregate the spectra across latitude using cosine weights

$$P_n(k) = \frac{\sum_j \cos\varphi_j P_{n,j}(k)}{\sum_j \cos\varphi_j}. \tag{102}$$

The spectrum shown in Figure 10c is obtained by averaging $P_n(k)$ over the 354 test initializations.

We use zonal wavenumbers 4–9 as an operational synoptic-scale band. For each initialization, the retained variance ratio is

$$r_n = \frac{\sum_{k=4}^{9} P_n^{\mathrm{model}}(k)}{\sum_{k=4}^{9} P_n^{\mathrm{ERA5}}(k)}, \qquad R_{\mathcal{B}} = \frac{1}{N_{\mathrm{test}}} \sum_{n=1}^{N_{\mathrm{test}}} r_n. \tag{103}$$

A value of one indicates that the forecast retains the same amount of variance as ERA5 in this band. The retention ratio measures amplitude at the selected scales, while anomaly correlation measures agreement in the spatial anomaly pattern.

Figure 5b shows $R_{\mathcal{B}}$ across forecast lead for z500. At day 7, WaveGSSM retains 68% of the ERA5 synoptic-scale variance, compared with 59% for GraphCast* and 37% for GraphSSM. Figure 10c

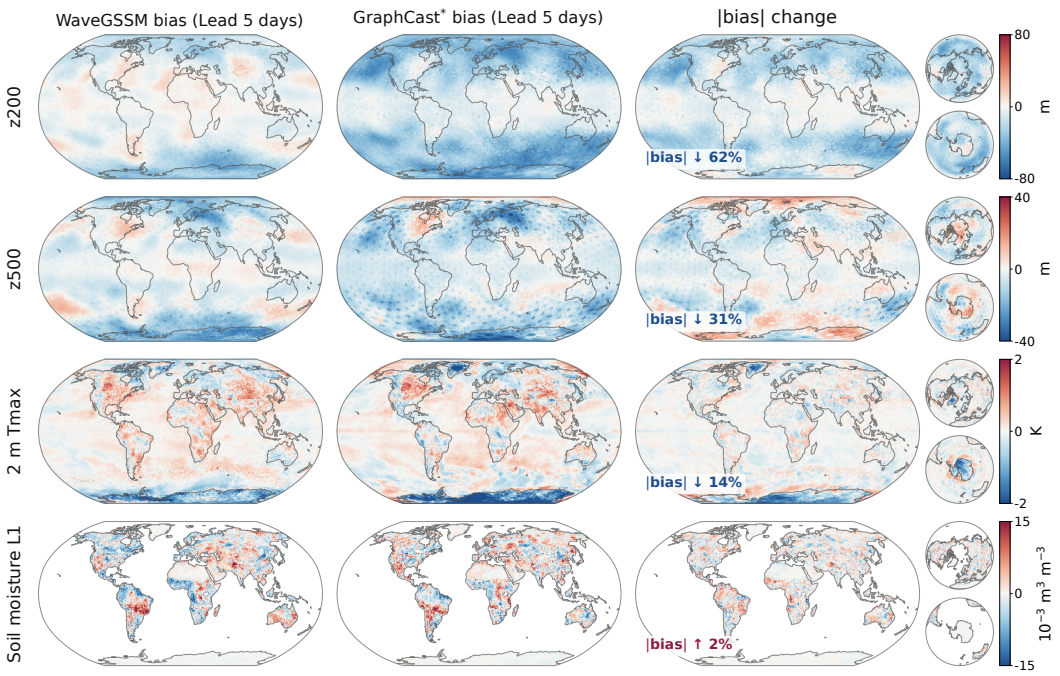


Figure 18: Systematic forecast bias at lead day 5. WaveGSSM reduces latitude-weighted absolute bias by 62.4% for $z200$ and 30.9% for $z500$. The geopotential bias reduction remains broad at medium range.

shows the corresponding day-7 spectrum, where WaveGSSM remains closest to ERA5 across much of the selected band.

Persistence retains nearly all of its initial spatial variance but has substantially lower forecast skill as weather systems evolve. Variance retention alone is therefore not sufficient. Together with the anomaly-correlation results, the spectral analysis shows that WaveGSSM retains more synoptic-scale variability while maintaining better agreement with the observed atmospheric pattern.

H Additional Forecast Analysis

H.1 Auditable Forecast Case

Aggregate scores establish that WaveGSSM improves pattern placement on average, but a case study makes the failure mode visually concrete. The case in Figure 20 is selected by an auditable rule rather than by hand. Among 2025 Northern-Hemisphere extratropical z500 forecasts, the selected initialization must have strong ERA5 wave amplitude, a positive day 7 anomaly-correlation gain for WaveGSSM relative to GraphCast*, and a lower WaveGSSM RMSE. This ranking rule avoids a common pitfall in qualitative forecast figures, where an example can look convincing but be difficult to reproduce.

The plotted quantity is the z500 zonal anomaly. This removes the dominant meridional background gradient and exposes the ridge and trough phase. The selected forecast is initialized on 2025-03-29 and verified at day 7 on 2025-04-05. The numerical evolution is consistent with the visual story. At day 1, both learned forecasts are accurate, but WaveGSSM already reduces RMSE by 37.9% relative to GraphCast*. At day 3 the RMSE reduction is 41.7%. At day 5 it is 30.6%. By day 7, GraphCast* has anomaly correlation 0.21, while WaveGSSM reaches 0.62 and lowers RMSE by 27.5%. This is the failure mode from the introduction in a real forecast. A snapshot model can retain a plausible-looking pattern but place it poorly. The motion-aware rollout keeps the same large-scale wave structure more closely aligned with ERA5.

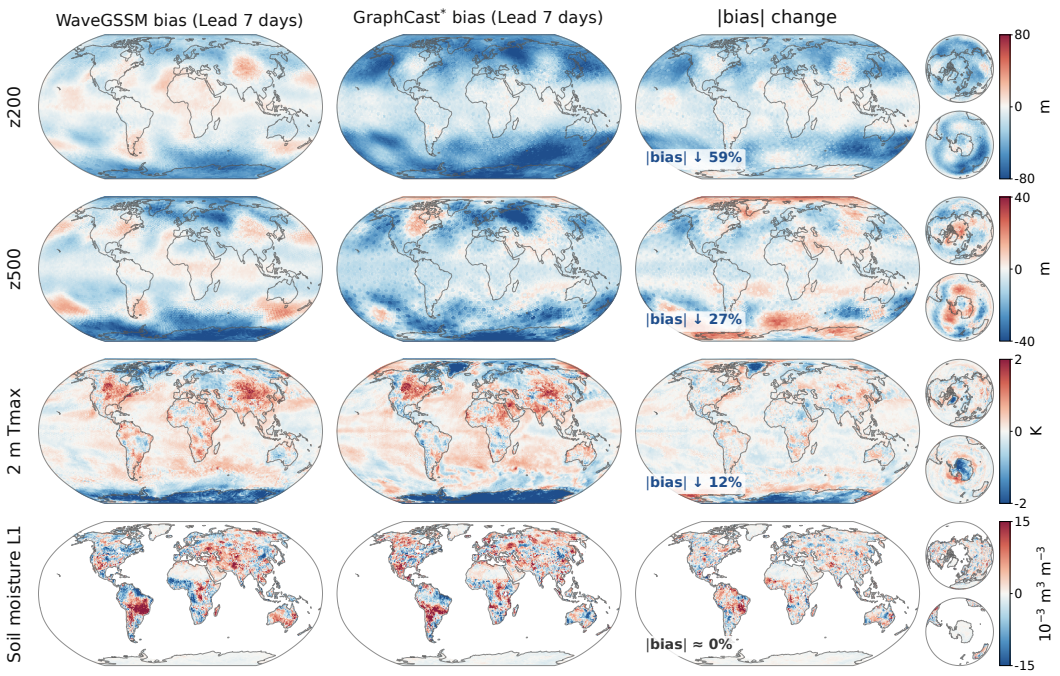


Figure 19: Systematic forecast bias at lead day 7. WaveGSSM reduces latitude-weighted absolute bias by 59.3% for $z200$, 27.3% for $z500$, and 12.1% for 2 m daily maximum temperature. Soil-moisture bias changes little.

H.2 Extratropical Skill Across Model Families

The preceding results compare WaveGSSM primarily with the backbone-matched GraphCast* model over the global domain. We next ask whether the advantage remains visible in the extratropics and against other temporal model families.

We define the evaluation region as

$$\mathcal{E} = \{(\lambda, \varphi) \mid 30^\circ \leq |\varphi| \leq 70^\circ\}, \tag{104}$$

including all longitudes in both hemispheres. This fixed latitude band contains much of the midlatitude circulation in which synoptic ridges and troughs evolve. It is used as an operational extratropical diagnostic rather than as a dynamically identified atmospheric waveguide, whose position varies between cases and seasons.

RMSE and centred anomaly correlation are computed using the definitions in Section D.2, restricted to grid points in $\mathcal{E}$ and retaining the cosine-latitude weights. The comparison includes WaveGSSM, GraphCast*, GraphSSM, and GNN-LSTM. GraphCast* is the backbone-matched snapshot predictor. GraphSSM provides a graph state-space comparison, while GNN-LSTM represents a conventional recurrent temporal model. The latter two retain their own temporal processors and should therefore be interpreted as cross-model baselines rather than controlled architectural ablations.

Across the lead days shown in Figure 21, WaveGSSM achieves the lowest RMSE and highest anomaly correlation for both $z200$ and $z500$. At day 7, it reduces $z200$ RMSE by 11.2% relative to GraphCast*, 9.4% relative to GraphSSM, and 4.1% relative to GNN-LSTM. For $z500$, the corresponding reductions are 4.5%, 8.6%, and 3.4%.

The anomaly-correlation curves show a clearer separation than the RMSE curves at the later leads. This is consistent with the global result that WaveGSSM retains better agreement with the observed anomaly pattern as the rollout progresses. The comparison shows that the improvement is not restricted to one matched snapshot baseline and remains present in a region where evolving synoptic structures are prominent. It does not, by itself, identify which architectural component causes the difference between WaveGSSM and the two external temporal baselines.

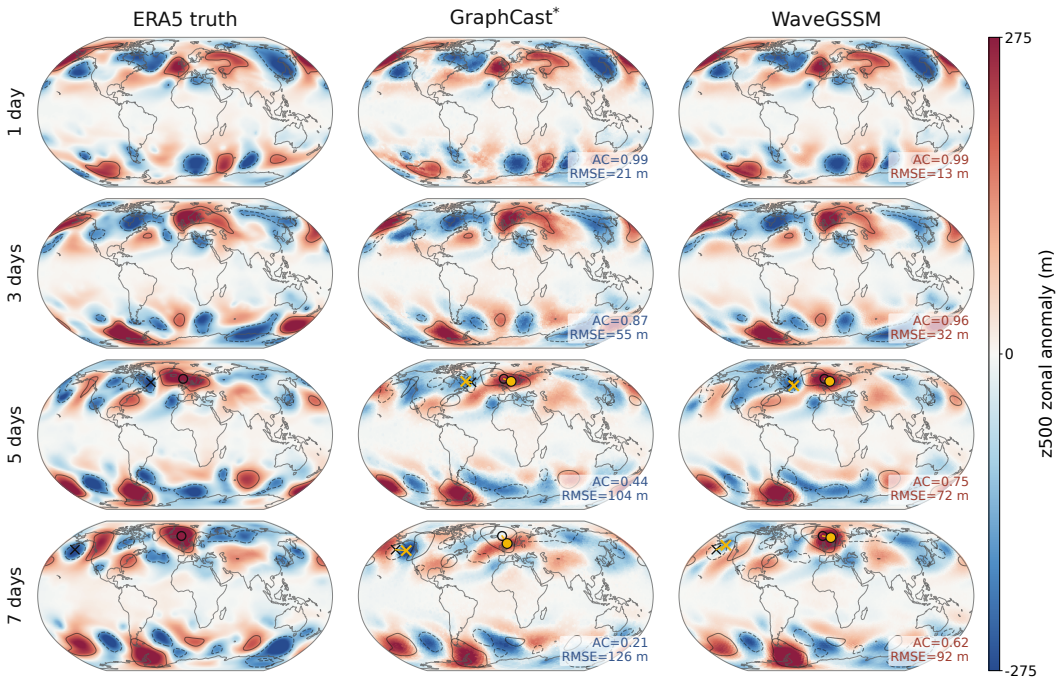


Figure 20: Auditable z500 forecast case showing phase alignment over a one-week rollout. Rows are lead days 1, 3, 5, and 7. Columns are ERA5 truth, GraphCast*, and WaveGSSM. Shading shows the 500 hPa geopotential-height zonal anomaly (m). Black contours and markers trace the ERA5 ridge and trough reference, while yellow markers and segments show predicted centres and displacement. The case is initialized on 2025-03-29 and selected by the ranking rule described in the text. At day 7, WaveGSSM raises anomaly correlation from 0.21 to 0.62 and reduces RMSE by 27.5% relative to GraphCast*, illustrating the same phase-preservation effect quantified in the aggregate diagnostics.

H.3 Spatial Comparison with GraphSSM

The previous subsection shows that WaveGSSM outperforms GraphSSM in aggregate extratropical scores. We now examine whether this difference is geographically widespread or concentrated in a small number of locations.

For variable v, lead h, and grid point x, we define

$$S_{v,h}^{\text{GraphSSM}}(x) = \frac{\text{RMSE}_{\text{WaveGSSM}}(v,h,x) - \text{RMSE}_{\text{GraphSSM}}(v,h,x)}{\text{RMSE}_{\text{GraphSSM}}(v,h,x)}. \tag{105}$$

Negative values indicate lower local RMSE for WaveGSSM. As in Section G.2, the local RMSE is computed over the 354 test initializations independently at every grid point.

WaveGSSM has lower latitude-weighted global RMSE than GraphSSM for both upper-air fields at all three displayed leads. The relative reductions for $z200$ and $z500$ are 31.9% and 30.7% at day 3, 21.3% and 20.3% at day 5, and 9.2% and 8.3% at day 7. The reduction narrows with lead, but the spatial advantage remains broad. At day 7, WaveGSSM has lower local RMSE at 92.8% of valid $z200$ grid points and 89.4% of valid $z500$ grid points.

The maps therefore show that the aggregate improvement over GraphSSM is not driven by a small number of favorable regions. The strongest margins occur at days 3 and 5, while the day-7 comparison remains positive across most grid points even as the global percentage difference becomes smaller.

GraphSSM and WaveGSSM differ in their state parameterization, processor design, and model configuration. This comparison consequently does not constitute an ablation of motion-state carry and cannot attribute the spatial difference to a single component. Its role is to show that WaveGSSM's advantage extends beyond the backbone-matched GraphCast* comparison and remains geographically broad against an independently designed graph state-space model.

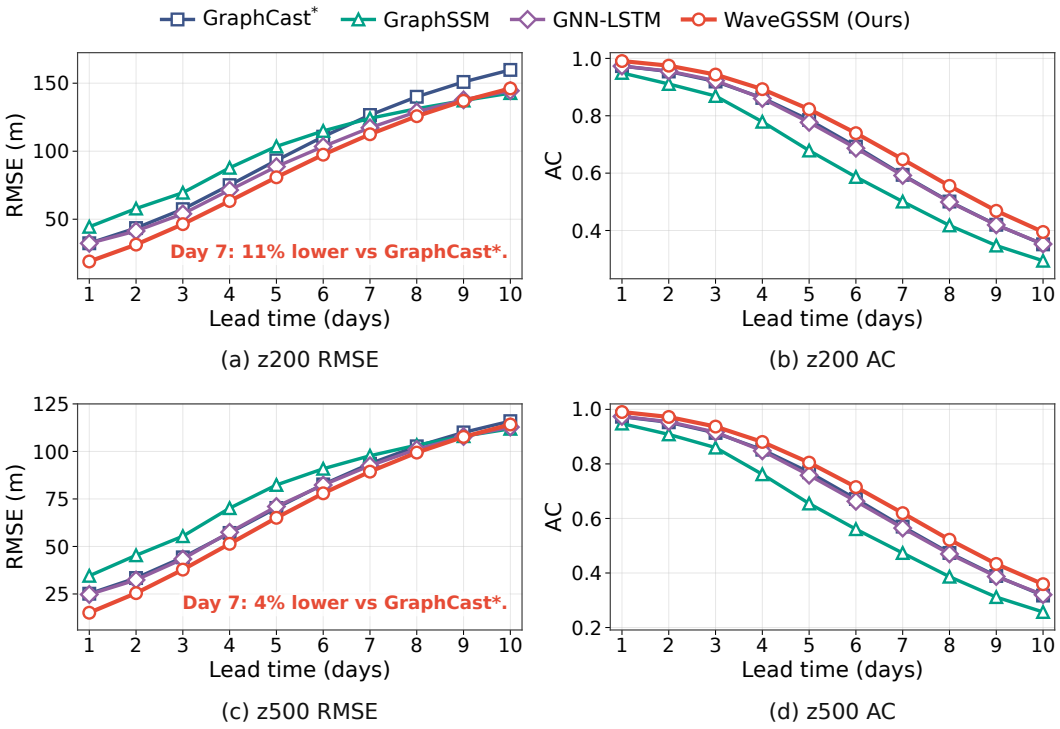


Figure 21: Forecast skill across temporal model families in the extratropics. RMSE and centred anomaly correlation are computed for $z200$ and $z500$ over all longitudes between 30° and 70° absolute latitude. The evaluation uses the 354 test initializations from 2025 and the same cosine-latitude weighting as the global metrics. Curves compare WaveGSSM, the backbone-matched GraphCast* model, GraphSSM, and GNN-LSTM. WaveGSSM has the lowest RMSE and highest anomaly correlation throughout the displayed day-1 to day-7 horizon.

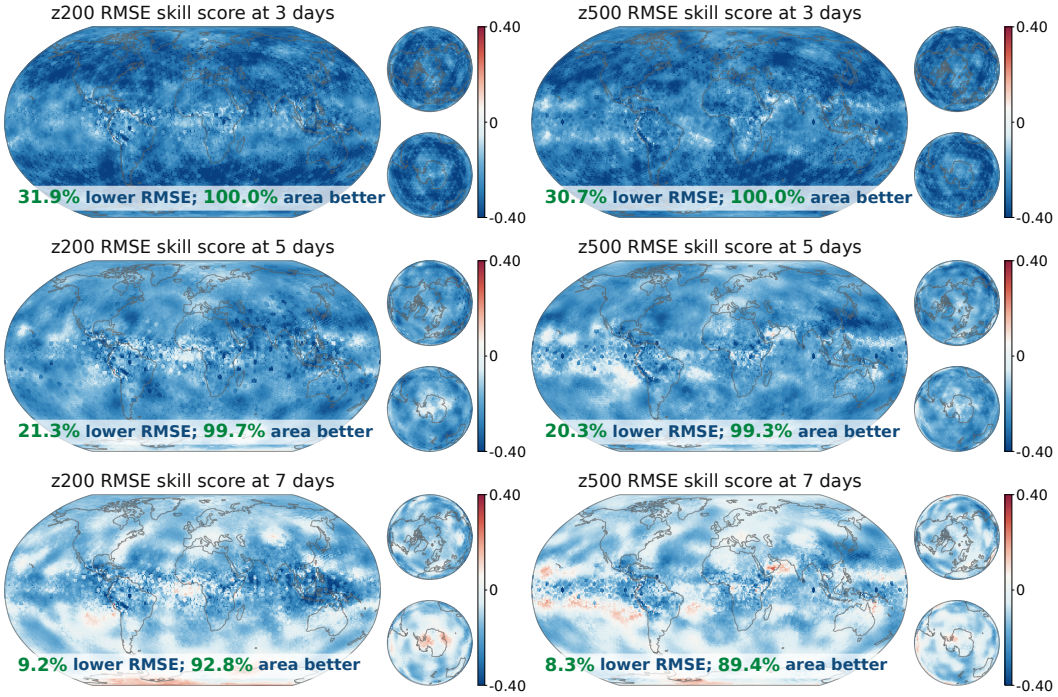


Figure 22: Spatial forecast skill relative to GraphSSM. Each panel shows the normalized local RMSE difference in Eq. (105) for $z200$ or $z500$ at lead days 3, 5, and 7. Blue indicates lower WaveGSSM RMSE, while red indicates lower GraphSSM RMSE. Polar insets display the high-latitude regions. The badges report the latitude-weighted global RMSE reduction and the fraction of valid grid points at which WaveGSSM has lower RMSE. The comparison is between two distinct graph state-space architectures and is not a component ablation.