STBRIDGE: A SPATIO-TEMPORAL NEURAL BRIDGE FOR GLOBAL WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medium-range weather forecasting requires preserving atmospheric structure over long lead times while controlling cross-scale error accumulation, maintaining spatial transport, and keeping meteorological variables coherently coupled. Existing models often rely on spherical graphs, multi-scale operators, temporal cascades, broad receptive fields, or variable-aware representations, but they typically model scale transfer, spatial transport, temporal evolution, and cross-variable interactions in separate pathways. To address these limitations, we propose STBridge, a Spatio-Temporal Neural Bridge that organizes global weather forecasting around a shared grid state. The model aggregates the encoded atmospheric history across multiple resolutions and returns the fused multi-grid state to the forecast grid, then learns directional spatial neighborhoods and temporal states for forecast evolution. The shared state supports spatial-semantic message exchange, followed by grid-wise coupling of the six atmospheric variables within each cell; a variable-conditioned decoder produces the future fields. Experiments on the ERA5 dataset show that STBridge achieves the best average ACC and the lowest average RMSE among the compared foundation and baseline models, while ablation and visualization experiments confirm the effectiveness and accuracy of its core modules. These results demonstrate that a shared grid state can coordinate multi-scale context, spatio-temporal relations, spatial-semantic routing, and cross-variable coupling for global medium-range forecasting.

1 INTRODUCTION

Global medium-range weather forecasting estimates the evolution of temperature, humidity, wind, pressure, and geopotential height over lead times from several hours to several days, given the recent atmospheric state on a global latitude–longitude grid. This task must represent interactions between large-scale circulation and local weather systems and preserve spatial transport across multiple scales (Keisler, 2022; Bonev et al., 2023; Tang et al., 2026). It must also maintain the coupled evolution of atmospheric variables across regions, vertical levels, and weather regimes (Nguyen et al., 2023; Gan et al., 2025; Chen et al., 2025a). Long rollouts accumulate forecast errors, and forecast skill degrades with lead time as fine-scale structures are smoothed and spatial transport becomes harder to represent. This work focuses on deterministic and distributional forecasting for six atmospheric variables over the 6–168 h medium-range horizon.

Existing data-driven systems address these difficulties with several complementary strategies. Spherical graphs and geometry-aware operators connect distant regions while respecting Earth geometry (Keisler, 2022; Lam et al., 2023; Oskarsson et al., 2024). Hierarchical temporal aggregation reduces the number of autoregressive steps over long horizons (Bi et al., 2023). Horizon-cascaded systems assign successive forecast stages to specialized dynamics (Chen et al., 2023). Long-horizon Transformer models extend temporal context, and high-resolution forecasting models preserve fine-scale structure over longer forecasts (Nguyen et al., 2024; Han et al., 2024). Weather and climate foundation models provide transferable representations for multivariate forecasting (Schmude et al., 2024). Observation-driven systems incorporate external observations into the forecasting process (Chen et al., 2025c; Sun et al., 2025). Generative models represent forecast uncertainty through probabilistic or multiple plausible trajectories (Price et al., 2025; Shi et al., 2025; Cachay et al., 2025). These methods improve different parts of the forecasting problem, yet multi-scale context, spatial transport, temporal evolution, and cross-variable exchange remain confined to separate representations or fixed pathways. Information from different grid scales, locations, lead times, and variables therefore cannot be associated and transferred within a unified forecast state.

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.

To address this limitation, we propose **STBridge**, a **spatio-temporal neural bridge** that uses the forecast grid as a shared intermediate representation that links multi-scale, spatial, temporal, semantic, and cross-variable information. The process begins with Adaptive Multi-Grid Bridging, which aggregates the encoded atmospheric history into a multi-scale grid hierarchy and returns a fused state to the forecast grid. Spatio-Temporal Learnable Neighborhood then learns directional spatial neighbors and temporal states on the fused grid state, allowing broad circulation, local transport, and lead-dependent dynamics to pass through a common grid representation. Building on this bridge, Dynamic Spatial Graph Routing exchanges spatial-semantic messages along the learned neighborhoods. Grid-Wise Multi-Element Coupling then couples the atmospheric variables within each grid cell. A variable-conditioned decoder maps the coupled state to the forecast fields. A compound geometric objective supervises field accuracy together with spatial structure, cross-scale agreement, and coupling consistency. Our contributions are summarized as follows:

- We establish a multi-scale spatio-temporal grid bridge in which the forecast grid links broad circulation, fine-scale structure, local transport, and lead-dependent temporal dynamics. Adaptive Multi-Grid Bridging and Spatio-Temporal Learnable Neighborhood provide the cross-scale and temporal states that allow these signals to interact on the shared grid.
- We place spatial-semantic routing and cross-variable interaction on the shared grid state. Dynamic Spatial Graph Routing and Grid-Wise Multi-Element Coupling coordinate spatial-semantic message exchange with atmospheric-variable coupling before decoding, representing both forms of interaction within a common forecasting state.
- Experiments on the ERA5 dataset show that STBridge achieves the best average ACC and the lowest average RMSE across the six atmospheric variables and forecast leads. Ablation and sensitivity studies further verify the contribution of each proposed module.

2 Related Works

Geometry-aware Graph Modeling. Global weather models represent spatial relations through grid nodes, spherical meshes, or continuous fields. Early graph networks established multiscale message passing, and later grid-to-graph encoders and hierarchical meshes strengthened spherical connectivity and long-range propagation (Keisler, 2022; Lam et al., 2023; Chantry et al., 2025; Oskarsson et al., 2024). Nested global and regional graphs extend this exchange across geographic domains and boundary conditions (Gao et al., 2025; Du et al., 2025). Other models learn global mappings via Fourier and spherical harmonic operators or continuous attention, retaining a global receptive field over the sphere (Kurth et al., 2023; Bonev et al., 2023; Anonymous, 2026; Li et al., 2024). Spherical multigrid and hybrid operators further tie different resolutions to Earth geometry (Hu et al., 2025; 2026; Tang et al., 2026).

Hierarchical Multi-scale Modeling. Hierarchies organize local, mesoscale, planetary, and vertical processes across forecast leads. Horizon-cascaded systems train distinct sub-models for the short, medium, and long ranges, each specializing in a separate forecast regime (Chen et al., 2023). Hierarchical temporal-aggregation approaches train separate models for several lead intervals (1, 3, 6, and 24 hours) under a shared architectural design, reducing the number of autoregressive steps over long horizons (Bi et al., 2023). Vertical interaction, large kernels, cross-resolution transfer, and hourly prediction further retain fine spatial structure and temporal continuity (Couairon et al., 2026; Zhu et al., 2024; Han et al., 2024; Zhong et al., 2024). Multi-interval training, hybrid dynamics, and adaptive rollout schedules coordinate successive predictions over longer horizons (Nguyen et al., 2024; Kochkov et al., 2024; Tian et al., 2026). Mixed-scale conditioning and efficient multiscale processing combine broad circulation with local atmospheric variation in the forecast state (Chen et al., 2026a; Yu et al., 2026; Chen et al., 2026b).

Multivariate Coupling Modeling. Atmospheric variables, vertical levels, and observations form an interdependent forecast state. Heterogeneous pretraining and foundation models share information across variables, resolutions, and tasks (Nguyen et al., 2023; Lessig et al., 2023; Schmude et al., 2024; Bodnar et al., 2025). Variable-aware experts and task weights preserve the identity of each meteorological element and coordinate predictions spanning different units and dynamics (Gan et al., 2025; Chen et al., 2025a;b). Observation-driven forecasting and differentiable assimilation use satellite and sparse observations to constrain multivariate evolution (Sun et al., 2025; Wang et al., 2026; Chen et al., 2025c; Xu et al., 2026; Li et al., 2026).

3 PRELIMINARIES

We consider medium-range weather forecasting from an L-step atmospheric history $\mathbf{X}_{t-L+1:t}$ over N grid cells, with consecutive states six hours apart. Given the history and static geographic descriptors, STBridge predicts variables in $\mathcal{V}$ at leads in $\mathcal{T}$:

$$\widetilde{\mathbf{X}}_{t-L+1:t} \in \mathbb{R}^{L\times N\times C_x}, \qquad \mathbf{E}_{\text{geo}} \in \mathbb{R}^{N\times C_g},$$
$$\widehat{\widetilde{\mathbf{Y}}}_{t+\mathcal{T}} = \mathcal{F}_\Theta\left(\widetilde{\mathbf{X}}_{t-L+1:t}, \mathbf{E}_{\text{geo}}, \{\mathbf{e}_\tau\}_{\tau\in\mathcal{T}}, \{\mathbf{e}_v\}_{v\in\mathcal{V}}\right) \in \mathbb{R}^{|\mathcal{T}|\times N\times|\mathcal{V}|}. \tag{1}$$

where C_x and C_g count dynamic channels and static descriptors. The current implementation uses $L = 2$ and $C_x = C_g = 6$; Appendix C.1 specifies the channels and normalization. The vectors $\mathbf{e}_\tau$ and $\mathbf{e}_v$ encode lead and variable identity, and Θ denotes model parameters. For variable v, $\widetilde{X}^v = (X^v - \mu_v)/s_v$ uses fixed training-history statistics. STBridge maps the history and static descriptors to the complete forecast sequence, with physical units restored by $\widehat{Y}^v_{i,t+\tau} = \mu_v + s_v\widehat{\widetilde{Y}}^v_{i,t+\tau}$.

To represent atmospheric states at multiple spatial scales, STBridge organizes the forecast grid and its coarser representations into a grid hierarchy $\mathcal{G} = \{G^{(1)}, G^{(2)}, G^{(4)}, G^{(8)}\}$. Here $G^{(1)}$ is the forecast grid, and each coarse-grid superscript gives its pooling factor. For a state $\mathbf{H} \in \mathbb{R}^{N\times d}$ with feature dimension d, paired restriction and prolongation operators define the cross-scale interface

$$[\mathcal{R}_s(\mathbf{H})]_j = \frac{\sum_{i\in\mathcal{C}_s(j)} a_i\mathbf{h}_i}{\sum_{i\in\mathcal{C}_s(j)} a_i}, \qquad [\mathcal{U}_s(\mathbf{H}^{(s)})]_i = \sum_{j\in\mathcal{B}_s(i)} b^{(s)}_{ij}\mathbf{h}^{(s)}_j, \qquad s \in \{2, 4, 8\}. \tag{2}$$

where $\mathcal{C}_s(j)$ contains the cells pooled into coarse cell j, a_i is the area weight of cell i, and $\mathcal{B}_s(i)$ contains the coarse cells used to interpolate cell i. Area-weighted average pooling implements $\mathcal{R}_s$; interpolation supplies nonnegative weights $b^{(s)}_{ij}$ with $\sum_{j\in\mathcal{B}_s(i)} b^{(s)}_{ij} = 1$. Each $G^{(s)} = (\mathcal{N}^{(s)}, \mathcal{E}^{(s)})$ contains grid cells $\mathcal{N}^{(s)}$ and their spatial relations $\mathcal{E}^{(s)}$ at scale s. As illustrated in Figure 1, STBridge uses the forecast grid to connect representations across scales and link spatial, temporal, and meteorological semantic information.

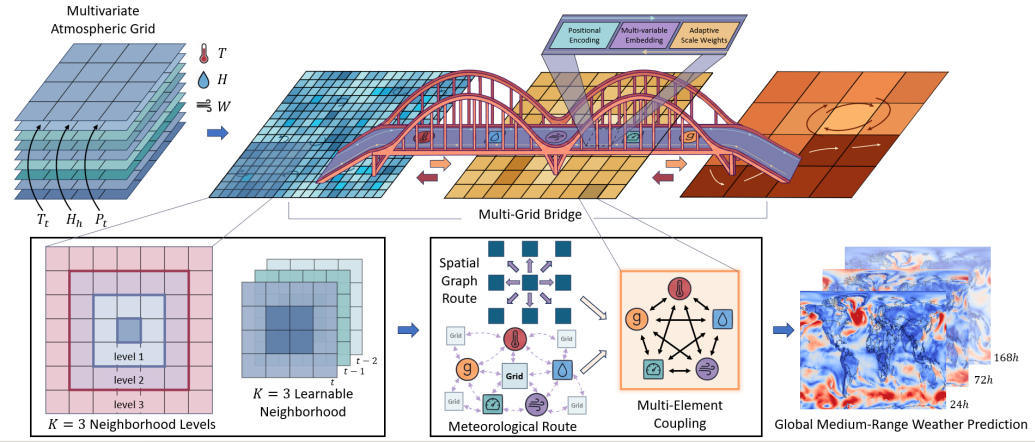

Figure 1: Conceptual illustration of the multi-scale spatio-temporal bridge in STBridge.

4 METHOD

4.1 OVERALL ARCHITECTURE

Building on the grid representation defined in Section 3, STBridge combines four interacting modules to transform atmospheric history into future meteorological fields, as shown in Figure 2. A shared encoder combines normalized atmospheric history, static geographic descriptors, lead encodings, and variable tokens into grid representations $\mathbf{H}^v_{0,\tau}$. Adaptive Multi-Grid Bridging integrates information across resolutions, and Spatio-Temporal Learnable Neighborhood learns spatial affinities and temporal states from the fused representation. Dynamic Spatial Graph Routing combines messages from neighboring cells with meteorological semantic context. The resulting routed state feeds Grid-Wise Multi-Element Coupling, which learns interactions among variables within each cell. A variable-conditioned decoder and target head map the coupled states to future meteorological fields.

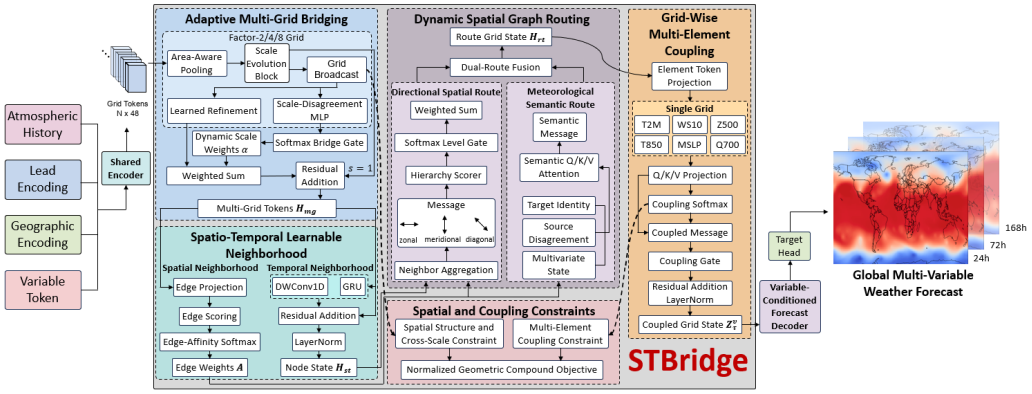

Figure 2: STBridge architecture and its four interacting modules

4.2 Adaptive Multi-Grid Bridging

Adaptive Multi-Grid Bridging connects coarse-scale circulation and fine-scale atmospheric structure in STBridge by transferring features across grid resolutions and fusing them on the forecast grid.

Scale Evolution Block. Area-Aware Pooling first aggregates the encoded state at each scale $s \in \mathcal{S}_c = \{2, 4, 8\}$. A scale embedding $\mathbf{e}_s$ is added to each pooled state, which the Scale Evolution Block evolves conditioned on the forecast lead. Grid Broadcast then returns the evolved features to the forecast grid:

$$\mathbf{U}_{\tau}^{v,(s)} = \mathcal{U}_s\big(\Phi_s\big(\mathcal{R}_s(\mathbf{H}_{0,\tau}^{v}) + \mathbf{e}_s, \mathbf{e}_{\tau}\big)\big), \qquad s \in \mathcal{S}_c. \tag{3}$$

Here $\mathcal{R}_s$ performs area-weighted pooling, Φ_s evolves the coarse state, and $\mathcal{U}_s$ interpolates it onto the forecast grid. Learned Refinement combines each returned state $\mathbf{U}_{\tau}^{v,(s)}$, the encoded state $\mathbf{H}_{0,\tau}^{v}$, and their absolute difference to produce the bridge message $\boldsymbol{\delta}_{\tau}^{v,(s)}$.

Softmax Bridge Gate. The Scale-Disagreement MLP scores each branch from the encoded and returned states, their absolute difference, and the lead embedding. The Softmax Bridge Gate normalizes these scores across scales to obtain the dynamic weights $\alpha_{i,\tau}^{v,(s)}$. A weighted sum of the bridge messages is added to the encoded state:

$$\alpha_{i,\tau}^{v,(s)} \geq 0, \qquad \sum_{s \in \mathcal{S}_c} \alpha_{i,\tau}^{v,(s)} = 1, \qquad \widetilde{\mathbf{h}}_{\mathrm{mg},i,\tau}^{v} = \mathbf{h}_{0,i,\tau}^{v} + \sum_{s \in \mathcal{S}_c} \alpha_{i,\tau}^{v,(s)} \boldsymbol{\delta}_{i,\tau}^{v,(s)}. \tag{4}$$

The encoded forecast-grid state supplies the residual base, while the three coarse scales contribute location- and lead-dependent messages. Feature normalization yields the fused state $\mathbf{H}_{\mathrm{mg},\tau}^{v}$. This state combines coarse circulation context with local features and provides the input for spatio-temporal neighborhood learning.

4.3 Spatio-Temporal Learnable Neighborhood

From the fused multi-grid state, STBridge learns directional spatial affinities and lead-dependent temporal states through Spatio-Temporal Learnable Neighborhood. The spatial branch produces edge weights at grid distances $\mathcal{D}_3 = \{1, 2, 4\}$, while the temporal branch updates the state across forecast leads.

Spatial Neighborhood. Within each distance level k, the module assigns state-dependent weights to valid neighbors in the zonal, meridional, and diagonal directions. Edge Projection maps the fused endpoint states into edge features. Edge Scoring combines these features with direction- and distance-specific parameters and target and lead embeddings, and Edge-Affinity Softmax normalizes the scores over the candidate set $\mathcal{N}_k(i)$ within each distance level. These weights determine directional messages from the temporal node states $\mathbf{h}_{\mathrm{st,j},\tau}^{v}$ defined by the temporal branch below:

$$a_{ij,\tau}^{v,(k)} \geq 0, \qquad \sum_{j \in \mathcal{N}_k(i)} a_{ij,\tau}^{v,(k)} = 1, \qquad \mathbf{m}_{i,\tau}^{v,(k)} = \sum_{j \in \mathcal{N}_k(i)} a_{ij,\tau}^{v,(k)} W_{r_{ij},k} \mathbf{h}_{\mathrm{st,j},\tau}^{v}. \tag{5}$$

Here r_{ij} identifies the edge direction, and $W_{r_{ij},k}$ transforms messages for that direction and distance. The normalized weights $a_{ij,\tau}^{v,(k)}$ describe the relative influence of neighboring cells and supply the spatial affinities used by graph routing.

Temporal Neighborhood. Alongside the spatial affinities, DWConv1D captures changes across three consecutive forecast leads, and a GRU carries the recurrent latent state. Their projected responses update the multi-grid state through residual addition:

$$\begin{aligned} \mathbf{T}^{v}_{\tau_\ell} = F_{\mathrm{temp},3}\Big(\mathbf{H}^{v}_{\mathrm{mg},\tau_{\ell-2:\ell}}\Big), \qquad \mathbf{s}^{v}_{\tau_\ell} = F_{\mathrm{rec}}\Big(\mathbf{H}^{v}_{\mathrm{mg},\tau_\ell}, \mathbf{s}^{v}_{\tau_{\ell-1}}\Big), \\ \widetilde{\mathbf{H}}^{v}_{\mathrm{st},\tau_\ell} = \mathbf{H}^{v}_{\mathrm{mg},\tau_\ell} + W_t \mathbf{T}^{v}_{\tau_\ell} + W_h \mathbf{s}^{v}_{\tau_\ell}. \end{aligned} \tag{6}$$

Here $F_{\mathrm{temp},3}$ is the depthwise temporal filter, F_{rec} is the GRU update, and W_t and W_h align their feature dimensions. The recurrence proceeds in increasing lead order, and LayerNorm converts the residual state into $\mathbf{H}^{v}_{\mathrm{st},\tau_\ell}$. The resulting edge weights and temporal node states jointly determine how STBridge routes information between grid cells at each forecast lead.

4.4 Dynamic Spatial Graph Routing

Dynamic Spatial Graph Routing connects different cells on the forecast grid and integrates their spatial messages with meteorological semantic context in STBridge.

Directional Spatial Route. Neighbor Aggregation uses the learned affinities to form the directional messages $\mathbf{m}^{v,(k)}_{i,\tau}$ in Equation 5. A Hierarchy Scorer evaluates each distance level using the node state, its aggregated message, and the forecast lead. The Softmax Level Gate converts these scores into weights $\pi^{v,(k)}_{i,\tau}$ for the weighted sum:

$$\mathbf{h}^{v}_{\mathrm{sp,i},\tau} = \sum_{k=1}^{3} \pi^{v,(k)}_{i,\tau} \mathbf{m}^{v,(k)}_{i,\tau}, \qquad \pi^{v,(k)}_{i,\tau} \geq 0, \qquad \sum_{k=1}^{3} \pi^{v,(k)}_{i,\tau} = 1. \tag{7}$$

The directional transforms $W_{r,k}$ distinguish transport direction and distance, while the level weights adapt the spatial context supplied to each cell.

Meteorological Semantic Route. The temporal node state supplies the query for Semantic Q/K/V Attention. Three descriptor tokens $\mathbf{p}^{v,(p)}_{i,\tau}$ encode target identity, internal cross-scale disagreement, and the multivariate latent state, supplying keys and values. Attention aggregates the descriptor values into a semantic message. The semantic update and its gated fusion with the spatial state are:

$$\mathbf{h}^{v}_{\mathrm{sem,i},\tau} = \mathbf{h}^{v}_{\mathrm{st,i},\tau} + \sum_{p=1}^{3} \omega^{v,(p)}_{i,\tau} W_V \mathbf{p}^{v,(p)}_{i,\tau}, \qquad \mathbf{h}^{v}_{\mathrm{rt,i},\tau} = \boldsymbol{\eta}^{v}_{i,\tau} \odot \mathbf{h}^{v}_{\mathrm{sp,i},\tau} + (1 - \boldsymbol{\eta}^{v}_{i,\tau}) \odot \mathbf{h}^{v}_{\mathrm{sem,i},\tau}. \tag{8}$$

Here $\omega^{v,(p)}_{i,\tau} \geq 0$ and $\sum_p \omega^{v,(p)}_{i,\tau} = 1$, and W_V projects descriptor values into the node-feature space. Dual-Route Fusion computes an element-wise gate $\boldsymbol{\eta}^{v}_{i,\tau} \in [0,1]^d$ from the spatial and semantic states, combining them into the routed state for within-cell variable coupling.

4.5 Grid-Wise Multi-Element Coupling

Using the routed grid state, STBridge applies Grid-Wise Multi-Element Coupling to learn dependencies among meteorological variables within each grid cell.

Coupling Softmax. Element Token Projection combines each variable's routed state with its variable embedding to form tokens $\mathbf{p}^{u}_{i,\tau}$. Q/K/V Projection uses the target token as the query and source-variable tokens as keys and values. Coupling Softmax normalizes their compatibility scores across source variables, producing the weights $c^{u\to v}_{i,\tau}$ for the coupled message. Coupling, gated residual addition, and output decoding follow

$$\begin{aligned} c^{u\to v}_{i,\tau} \geq 0, \qquad \sum_{u\in\mathcal{V}} c^{u\to v}_{i,\tau} = 1, \\ \mathbf{c}^{v}_{i,\tau} = \sum_{u\in\mathcal{V}} c^{u\to v}_{i,\tau} V_u \mathbf{p}^{u}_{i,\tau}, \qquad \widetilde{\mathbf{z}}^{v}_{i,\tau} = \mathbf{h}^{v}_{\mathrm{rt,i},\tau} + \gamma^{v}_{i,\tau} \mathbf{c}^{v}_{i,\tau}, \\ \widehat{Y}^{v}_{i,t+\tau} = \mu_v + s_v P^{v}_{\theta}\left(D^{v}_{\theta}[\mathbf{z}^{v}_{i,\tau}, \mathbf{e}_v, \mathbf{e}_\tau]\right). \end{aligned} \tag{9}$$

Here V_u projects source tokens into the message space, $\gamma^{v}_{i,\tau}$ gates their contribution, and LayerNorm yields $\mathbf{z}^{v}_{i,\tau}$ from the residual state. The Variable-Conditioned Forecast Decoder D^{v}_{θ} and target head P^{v}_{θ} use this state and identity embeddings to predict the normalized field; μ_v and s_v restore physical units.

Spatial and Coupling Constraints. Training combines field error with four structural terms $\mathcal{Q} = \{\nabla, \mathrm{spec}, \mathrm{bridge}, \mathrm{coup}\}$:

$$\mathcal{L}_{\mathrm{lat}} = \frac{1}{|\mathcal{T}||\mathcal{V}|N} \sum_{\tau,v,i} w_i \left(\widehat{\widetilde{Y}}^{v}_{i,t+\tau} - \widetilde{Y}^{v}_{i,t+\tau} \right)^2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{lat}} \left[\prod_{q \in \mathcal{Q}} (1 + \lambda_q \widetilde{\mathcal{L}}_q) \right]^{1/4}. \tag{10}$$

Here w_i denotes area weights normalized to unit mean, λ_q weights each structural term, and $\widetilde{\mathcal{L}}_q$ is its loss normalized by a detached moving reference. The structural terms constrain spatial gradients and spectra of the predicted fields, agreement between grid scales, and consistency among variable representations in latent space. Detailed definitions of the loss terms and their normalization are provided in Appendix C.6. Joint optimization trains the grid-transfer, routing, and coupling modules to produce the multivariable forecast sequence.

5 Experiments

Datasets. We use ERA5 fields distributed through WeatherBench 2 (Hersbach et al., 2020; Rasp et al., 2024). STBridge uses 2019–2020 for training and 2021 for validation. The test set contains 256 matched initializations from 2022, with forecasts evaluated every six hours through 168 hours. We predict three surface fields, 2-m temperature (T2M), 10-m wind speed (WS10), and mean sea-level pressure (MSLP), together with temperature at 850 hPa (T850), specific humidity at 700 hPa (Q700), and geopotential at 500 hPa (Z500). All models are evaluated on a common 1.5° grid, using identical test initializations, ERA5 verification fields, target variables, forecast leads, and evaluation metrics. Appendix B.1 gives the data and grid protocol.

Metrics. Latitude-weighted root mean squared error (RMSE) measures physical-field error, and the anomaly correlation coefficient (ACC) measures agreement between forecast and observed climatological anomalies. Lower RMSE and higher ACC indicate better deterministic forecasts. Ensemble forecasts are evaluated with the continuous ranked probability score (CRPS), spread–mean error (SME), and relative quantile error (RQE). Appendix B.3 defines the metrics, latitude weights, climatological anomalies, and ensemble notation.

Baselines. We compare STBridge with Pangu-Weather (Bi et al., 2023), FuXi (Chen et al., 2023), OneForecast (Gao et al., 2025), and Aurora (Bodnar et al., 2025). Pangu-Weather is abbreviated as Pangu in the figures and tables. Appendix B.2 describes the baseline configurations and forecast alignment.

5.1 Main Results

Medium-range forecasting. Figures 3 and 4 compare the ACC and RMSE of STBridge, Pangu-Weather, FuXi, OneForecast, and Aurora for six meteorological variables at six-hour intervals through 168 hours.

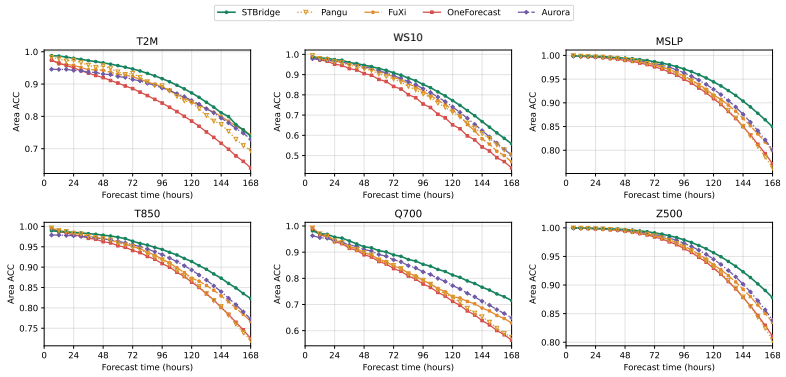

Figure 3: Medium-range ACC of all evaluated models for six meteorological variables.

STBridge achieves the strongest overall forecasting performance across the six variables, with the best average ACC and the lowest average RMSE. Its ACC advantage becomes particularly clear at longer leads for WS10, MSLP, T850, Q700, and Z500. The corresponding RMSE curves show lower errors for these fields toward the end of the forecast horizon, demonstrating STBridge's ability to maintain prediction accuracy across surface and upper-air variables through seven days.

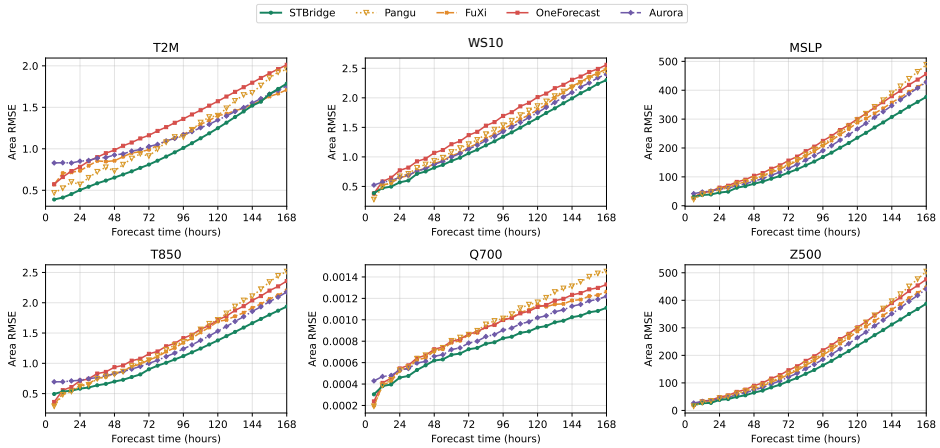

Figure 4: Medium-range RMSE of all evaluated models for six meteorological variables.

Ensemble forecasting. Table 1 compares four-member ensembles from STBridge, Pangu-Weather, and FuXi for T2M, T850, WS10, and Z500 at 24, 72, 120, and 168 hours, using the same 32 test initializations. CRPS measures distributional forecast error, SME measures aggregate spread–error mismatch, and RQE measures relative deviations between predicted quantiles and observations. STBridge achieves the lowest scores in 45 of the 48 variable–lead–metric combinations, including all RQE entries and all CRPS entries for T2M, T850, and Z500. STBridge also obtains the lowest SME throughout the four forecast leads for T2M, T850, and WS10, showing consistent agreement between ensemble spread and mean error for these variables.

Table 1: Four-member ensemble forecasting results for four meteorological variables.

Variable	Model	CRPS (↓)				SME (↓)				RQE (↓)			
		24	72	120	168	24	72	120	168	24	72	120	168
T2M	STBridge	**0.3208**	**0.5107**	**0.7416**	**1.0381**	**0.2988**	**0.3430**	**0.4351**	**0.5337**	**0.00159**	**0.00229**	**0.00331**	**0.00485**
	Pangu	0.9010	0.8660	0.9731	1.1138	0.3882	0.4308	0.5190	0.5409	0.00427	0.00406	0.00457	0.00533
	FuXi	1.2280	1.8624	2.3431	2.6657	0.3023	0.8132	1.1174	1.4916	0.00619	0.00889	0.01100	0.01200
T850	STBridge	**0.4101**	**0.6502**	**1.0021**	**1.4760**	**0.4056**	**0.5288**	**0.6144**	**0.6829**	**0.00176**	**0.00270**	**0.00412**	**0.00592**
	Pangu	0.7019	0.9341	1.2468	1.5538	0.4121	0.5760	0.7074	0.7551	0.00325	0.00428	0.00584	0.00750
	FuXi	1.0000	1.8086	2.5838	3.1355	0.4997	1.2339	1.7386	2.2470	0.00474	0.00795	0.01100	0.01400
WS10	STBridge	**0.4922**	**0.7713**	**1.1760**	1.5776	**0.2811**	**0.3163**	**0.3167**	**0.4977**	**0.15800**	**0.25300**	**0.37700**	**0.49300**
	Pangu	0.6915	0.9410	1.2121	**1.3651**	0.3565	0.4727	0.5759	0.5144	0.26200	0.34100	0.42900	0.50000
	FuXi	1.8587	2.4576	2.5833	2.9923	1.3264	1.8653	1.8586	2.4324	0.45900	0.56300	0.61500	0.63000
Z500	STBridge	**30.105**	**76.626**	**158.03**	**266.26**	39.873	**64.516**	**93.188**	138.02	**0.00100**	**0.00165**	**0.00316**	**0.00577**
	Pangu	62.066	121.40	209.44	282.97	**20.833**	74.148	119.35	**128.85**	0.00159	0.00288	0.00507	0.00712
	FuXi	74.096	189.24	305.66	395.01	26.997	120.14	180.07	249.77	0.00187	0.00441	0.00730	0.00933

5.2 Model Analysis

Spatial and temporal visualization. Figures 5 and 6 compare STBridge forecasts with ERA5 fields and show signed errors for WS10 and Q700 at 24, 72, and 168 hours. STBridge reproduces the broad wind bands and tropical moisture distribution at the first two leads, with errors concentrated near sharp gradients and localized weather structures. At 168 hours, STBridge retains the main circulation and humidity patterns, while larger signed errors mark differences in the position and intensity of finer structures. Together, the two fields show accurate large-scale spatial evolution alongside the growth of localized errors over the forecast horizon.

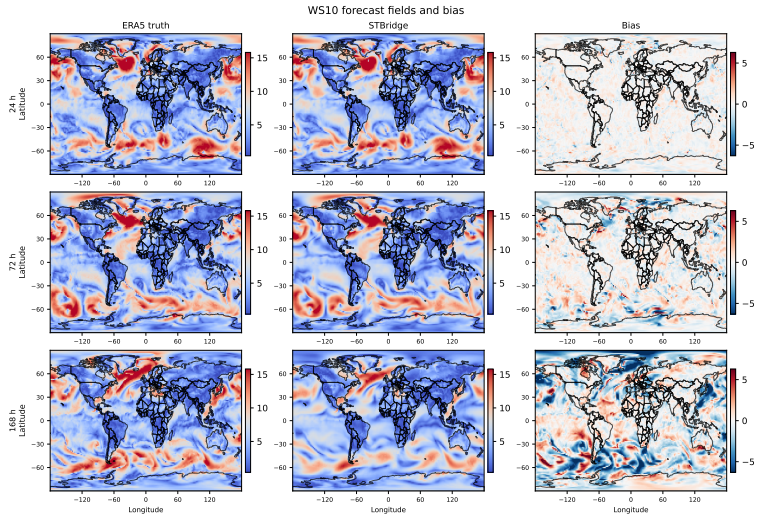

Figure 5: STBridge forecasts, ERA5 fields, and signed errors for WS10 at 24, 72, and 168 hours.

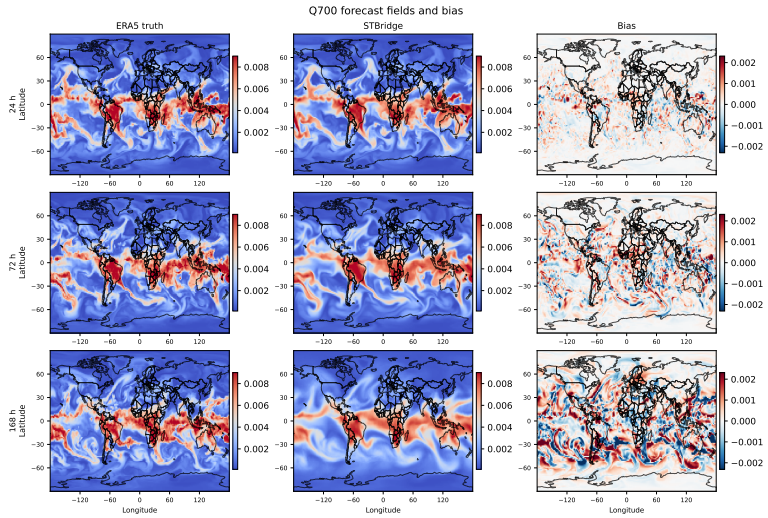

Figure 6: STBridge forecasts, ERA5 fields, and signed errors for Q700 at 24, 72, and 168 hours.

Fixed-point forecasting. Figure 7 compares fixed-point forecasts of WS10, Q700, and Z500 from STBridge, Pangu-Weather, FuXi, OneForecast, and Aurora against ERA5. At the selected locations, STBridge achieves the lowest MAE and RMSE among the evaluated models, with RMSE values of 0.938, 0.000762, and 44.6 for WS10, Q700, and Z500, respectively. STBridge follows the later WS10 fluctuations more closely, reduces the sustained positive Q700 bias seen in the baseline trajectories, and tracks the Z500 peaks and troughs with smaller late-lead deviations. These results demonstrate STBridge's accuracy in tracking local atmospheric evolution at the displayed locations.

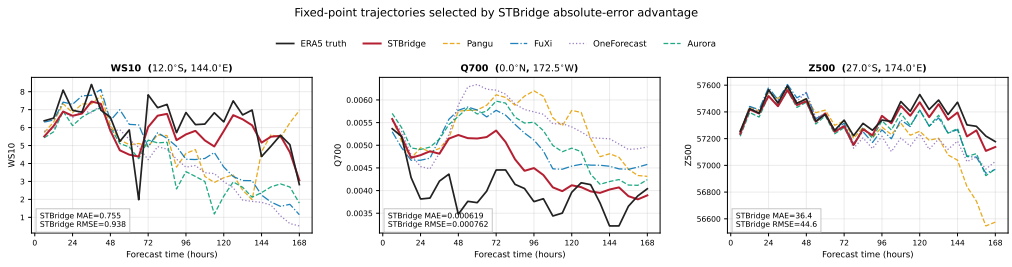

Figure 7: Fixed-point forecasts of WS10, Q700, and Z500 from all evaluated models against ERA5.

Ablations. Table 2 reports the effects of removing each STBridge module on WS10 and Z500 forecasts. Removing Adaptive Multi-Grid Bridging raises their 24-hour RMSE from 0.569 and 38.2 to 0.697 and 45.7, supporting the contribution of cross-scale information transfer. Removing Spatio-Temporal Learnable Neighborhood increases these errors to 0.700 and 46.0, demonstrating the value of learned spatial affinities and temporal states. Removing Dynamic Spatial Graph Routing produces the largest deterioration, to 0.726 and 47.2, identifying spatial-semantic routing as the strongest contributor in this comparison. Removing Grid-Wise Multi-Element Coupling increases RMSE to 0.636 and 43.2, confirming the additional benefit of within-cell variable interactions. Every removal raises RMSE and lowers ACC across the reported leads; the complete STBridge achieves the best results, supporting the effectiveness of all four modules.

Table 2: Ablation study of STBridge on WS10 and Z500.

		RMSE (↓)							ACC (↑)						
Variable	Model	24	48	72	96	120	144	168	24	48	72	96	120	144	168
WS10	w/o Multi-Grid Bridging	0.697	0.948	1.191	1.478	1.782	2.117	2.425	0.964	0.932	0.883	0.820	0.752	0.642	0.530
	w/o Learnable Neighborhood	0.700	0.952	1.195	1.481	1.784	2.119	2.429	0.959	0.930	0.880	0.817	0.750	0.640	0.527
	w/o Graph Routing	0.726	0.976	1.222	1.509	1.811	2.145	2.453	0.956	0.927	0.876	0.803	0.738	0.632	0.519
	w/o Multi-Element Coupling	0.636	0.890	1.132	1.414	1.720	2.056	2.366	0.967	0.938	0.901	0.840	0.767	0.660	0.551
	STBridge	**0.569**	**0.817**	**1.060**	**1.342**	**1.657**	**1.990**	**2.301**	**0.974**	**0.946**	**0.908**	**0.852**	**0.772**	**0.668**	**0.558**
Z500	w/o Multi-Grid Bridging	45.7	72.9	115.3	173.7	245.9	323.2	399.6	0.978	0.973	0.965	0.950	0.922	0.886	0.841
	w/o Learnable Neighborhood	46.0	73.2	115.5	174.0	246.2	323.6	400.1	0.968	0.966	0.957	0.942	0.916	0.874	0.833
	w/o Graph Routing	47.2	74.5	117.0	175.6	248.0	325.6	402.1	0.959	0.956	0.945	0.932	0.903	0.865	0.824
	w/o Multi-Element Coupling	43.2	70.1	112.2	170.4	242.3	319.6	395.6	0.982	0.975	0.972	0.958	0.933	0.896	0.850
	STBridge	**38.2**	**64.6**	**106.2**	**163.9**	**235.3**	**312.1**	**387.7**	**0.999**	**0.997**	**0.991**	**0.979**	**0.957**	**0.923**	**0.878**

Neighborhood hierarchy. Table 3 shows that $K = 3$ achieves the lowest RMSE and highest ACC for WS10 and Z500 at every reported lead. With $K = 2$, the neighborhood covers distances $\{1, 2\}$ and omits distance 4; its higher errors are consistent with insufficient spatial coverage for broader interactions. Increasing K from 3 to 4 adds distance 8 but raises the 24-hour RMSE from 0.569 to 0.699 for WS10 and from 38.2 to 45.8 for Z500, with lower ACC across the reported leads. The additional distance level therefore reduces forecasting performance, establishing $K = 3$ as the most effective neighborhood configuration in this comparison.

Table 3: Effect of neighborhood depth K on WS10 and Z500.

		RMSE (↓)							ACC (↑)						
Variable	Model	24	48	72	96	120	144	168	24	48	72	96	120	144	168
WS10	$K = 1$	0.704	0.954	1.199	1.486	1.788	2.124	2.432	0.959	0.925	0.872	0.803	0.744	0.629	0.518
	$K = 2$	0.701	0.951	1.197	1.482	1.784	2.120	2.431	0.960	0.929	0.877	0.807	0.748	0.637	0.523
	$K = 4$	0.699	0.944	1.195	1.480	1.782	2.118	2.429	0.963	0.932	0.883	0.816	0.747	0.636	0.530
	STBridge ($K = 3$)	**0.569**	**0.817**	**1.060**	**1.342**	**1.657**	**1.990**	**2.301**	**0.974**	**0.946**	**0.908**	**0.852**	**0.772**	**0.668**	**0.558**
Z500	$K = 1$	46.4	74.5	116.9	174.6	247.5	324.9	402.1	0.966	0.960	0.948	0.929	0.902	0.871	0.831
	$K = 2$	45.5	73.6	116.0	174.5	246.7	324.2	401.5	0.970	0.968	0.951	0.936	0.912	0.875	0.830
	$K = 4$	45.8	73.0	115.4	173.8	246.0	323.4	400.5	0.983	0.978	0.968	0.953	0.926	0.890	0.844
	STBridge ($K = 3$)	**38.2**	**64.6**	**106.2**	**163.9**	**235.3**	**312.1**	**387.7**	**0.999**	**0.997**	**0.991**	**0.979**	**0.957**	**0.923**	**0.878**

6 Conclusion and Future Work

In this paper, we presented STBridge, a spatio-temporal neural bridge for global medium-range weather forecasting. STBridge transfers information across grid scales, learns spatial and temporal relations, and combines spatial-semantic routing with variable coupling within each grid cell. Experiments on ERA5 show the best average ACC and lowest average RMSE among the evaluated models across six atmospheric variables and forecasts through 168 hours. Ablation and neighborhood studies establish the contribution of each module and the effectiveness of the three-level neighborhood, while spatial fields and fixed-point trajectories demonstrate accurate atmospheric structure and local evolution. Future work will extend STBridge to higher-resolution forecasting and larger probabilistic ensembles.

7 AI USE STATEMENT

In this study, we employed a Large Language Model (LLM) solely as a writing-assistance tool to improve the readability and overall presentation of the manuscript. All substantive research, data analysis, and intellectual reasoning were conducted exclusively by the authors. Every suggestion generated by the LLM was carefully examined, adapted where necessary, and validated by us to guarantee its consistency with our research objectives. The authors take full and exclusive responsibility for all content presented herein. The LLM contributed to the following specific tasks:
• **Text Refinement:** LLMs helped correct grammatical issues, refine sentence structures, and maintain uniform terminology and stylistic tone throughout the paper.
• **Coherence and Structure:** LLMs assisted in restructuring the flow of arguments, improving the coherence between sections and paragraphs, and ensuring smoother transitions that enhance the overall narrative clarity.

8 ETHICS STATEMENT

Our work only focuses on the scientific problem, so there is no potential ethical risk.

9 REPRODUCIBILITY STATEMENT

Implementation details are provided in Appendices B and C, including model configuration, optimization, dataset descriptions, and metric calculation. The source code is provided in supplementary materials for reproducibility.

REFERENCES

Anonymous. STC-vit: Spatio temporal continuous vision transformer for medium-range global weather forecasting. *Submitted to Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=yTeXmoeZ79. Under review.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3. URL https://doi.org/10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Salva Rühling Cachay, Miika Aittala, Karsten Kreis, Noah D Brenowitz, Arash Vahdat, Morteza Mardani, and Rose Yu. Elucidated rolling diffusion models for probabilistic forecasting of complex dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NnRQOPzL9P.

Matthew Chantry, Simon Lang, Mihai Alexe, Jesper Dramsch, Baudouin Raoult, Mariana Clare, Mario Santa Cruz, Sara Hahner, Rilwan Adewoyin, Florian Pinault, Ana Prieto Nemesio, Gabriel Moldovan, Linus Magnusson, Zied Ben Bouallegue, Steffen Tietsche, Ewan Mark Pinnington, Jakob Schloer, Andy Brown, Florian Pappenberger, and Florence Rabier. AIFS - ECMWF's Data-Driven Forecasting System. In *105th Annual AMS Meeting 2025*, volume 105 of *American Meteorological Society Meeting Abstracts*, pp. 449087, January 2025.

Hao Chen, Tao Han, Song Guo, Jie Zhang, Yonghan Dong, Yunlong Yu, and Lei Bai. Va-moe: Variables-adaptive mixture of experts for incremental weather forecasting. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7915–7924, 2025a. URL https://api.semanticscholar.org/CorpusID:274445800.

Hao Chen, Tao Han, Jie Zhang, Song Guo, Fenghua Ling, and Lei Bai. EMFormer: Efficient multi-scale transformer for accumulative context weather forecasting. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=cKztWFFGNE.

Jun Chen, Yan Fang, Minghui Qiu, Yueran Qiu, Lin Chen, Shuxin Zhong, Yu Zhang, and Kaishun Wu. Storminsight: Hierarchical environmental forcing and vertical coupling for weather system evolution. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=fYLR5f1Xq2.

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

Kun Chen, Lei Bai, Fenghua Ling, Peng Ye, Tao Chen, Hang Fan, Hao Chen, Yi Xiao, Kang Chen, Tao Han, Jing-Jia Luo, and Wanli Ouyang. Towards an end-to-end artificial intelligence driven global weather forecasting system, 2025c. URL https://arxiv.org/abs/2312.12462.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 2023. doi: 10.1038/s41612-023-00512-1. URL https://doi.org/10.1038/s41612-023-00512-1.

Guillaume Couairon, Renu Singh, Anastase Charantonis, Christian Lessig, and Claire Monteleoni. Archesweathergen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Science Advances*, 12(17):eadx2372, 2026. doi: 10.1126/sciadv.adx2372. URL https://www.science.org/doi/abs/10.1126/sciadv.adx2372.

Haoxing Du, Lyna Kim, Joan Creus-Costa, Jack Michaels, Anuj Shetty, Todd Hutchinson, Christopher Riedel, and John Dean. Weathermesh-3: Fast and accurate operational global weather forecasting, 2025. URL https://arxiv.org/abs/2503.22235.

Lihao Gan, Xin Man, Chenghong Zhang, and Jie Shao. EWMoE: An effective model for global weather forecasting with mixture-of-experts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):210–218, April 2025. doi: 10.1609/aaai.v39i1.31997. URL https://doi.org/10.1609/aaai.v39i1.31997.

Yuan Gao, Hao Wu, Ruiqi Shu, huanshuo dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, Jiahao Wu, Li Qing, Hui Xiong, and Xiaomeng Huang. Oneforecast: A universal framework for global and regional weather forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9xGSeVolcN.

Tao Han, Song Guo, Fenghua Ling, Kang Chen, Junchao Gong, Jingjia Luo, Junxia Gu, Kan Dai, Wanli Ouyang, and Lei Bai. Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting, 2024. URL https://arxiv.org/abs/2402.00059.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Yifan Hu, Fukang Yin, Weimin Zhang, Kaijun Ren, Junqiang Song, Kefeng Deng, and Di Zhang. Spherical multigrid neural operator for improving autoregressive global weather forecasting. *Scientific Reports*, 15, 2025. URL https://api.semanticscholar.org/CorpusID:277548592.

Yifan Hu, Fukang Yin, Weimin Zhang, Kaijun Ren, Junqiang Song, and Kefeng Deng. A hybrid framework for global weather forecasting via low-resolution dynamical core and multigrid neural operator. *npj Climate and Atmospheric Science*, 9(1):112, 2026. ISSN 2397-3722. doi: 10.1038/s41612-026-01374-z. URL https://doi.org/10.1038/s41612-026-01374-z.

Ryan Keisler. Forecasting global weather with graph neural networks, 2022. URL https://arxiv.org/abs/2202.07575.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07744-y. URL https://doi.org/10.1038/s41586-024-07744-y.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701900. doi: 10.1145/3592979.3593412. URL https://doi.org/10.1145/3592979.3593412.

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

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning, 2023. URL https://arxiv.org/abs/2308.13280.

Guowen Li, Xintong Liu, Yang Liu, Mengxuan Chen, Shilei Cao, Xuehe Wang, Juepeng Zheng, Jinxiao Zhang, Haoyuan Liang, Lixian Zhang, Jiuke Wang, Meng Jin, Hong Cheng, and Haohuan Fu. Tianquan-s2s: A subseasonal-to-seasonal global weather model via incorporate climatology state. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7Dvmq7MhwU.

Zijie Li, Anthony Zhou, Saurabh Patil, and Amir Barati Farimani. Cafa: Global weather forecasting with factorized attention on sphere, 2024. URL https://arxiv.org/abs/2405.07395.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=aBP01akha9.

Joel Oskarsson, Tomas Landelius, Marc Peter Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wTIzpqX121.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam,

and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637 (8044):84–90, 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

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

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate, 2024. URL https://arxiv.org/abs/2409.13598.

Jimeng Shi, Bowen Jin, Jiawei Han, Sundararaman Gopalakrishnan, and Giri Narasimhan. Codicast: Conditional diffusion model for global weather forecasting with uncertainty quantification. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 9853–9861. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1095. URL https://doi.org/10.24963/ijcai.2025/1095. AI and Social Good.

Xiuyu Sun, Xiaohui Zhong, Xiaoze Xu, Yuanqing Huang, Hao Li, J. David Neelin, Deliang Chen, Jie Feng, Wei Han, Libo Wu, and Yuan Qi. A data-to-forecast machine learning system for global weather. *Nature Communications*, 16(1):6658, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62024-1. URL https://doi.org/10.1038/s41467-025-62024-1.

Hao Tang, Hao Chen, and Chao Li. Generalized spherical neural operators: Green's function formulation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 11180–11201, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/13079a5dd5c994a7e5f0cc691e0338e8-Paper-Conference.pdf.

Jindong Tian, Yifei Ding, Ronghui Xu, Hao Miao, Guo Guo, and Bin Yang. Arrow: An adaptive rollout and routing method for global weather forecasting. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 54823–54847, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/59fb5cf4683eaec4c56918a69637c9d2-Paper-Conference.pdf.

Wuxin Wang, Weicheng Ni, Lilan Huang, Tao Hao, Ben Fei, Shuo Ma, Taikang Yuan, Yanlai Zhao, Kefeng Deng, Xiaoyong Li, Hongze Leng, Boheng Duan, Lei Bai, Weimin Zhang, Junqiang Song, and Kaijun Ren. Xichen: A global weather observation-to-forecast machine learning system via four-dimensional variational gradient-guided flexible assimilation, 2026. URL https://arxiv.org/abs/2507.09202.

WeatherBench 2 Team. Weatherbench 2 data guide. https://weatherbench2.readthedocs.io/en/latest/data-guide.html, 2024. Accessed 2026-08-22.

Xiaoze Xu, Xiuyu Sun, Songling Zhu, Xiaohui Zhong, Yuanqing Huang, Zijian Zhu, Jun Liu, and Hao Li. Fuxiweather2: Learning accurate atmospheric state estimation for operational global weather forecasting, 2026. URL https://arxiv.org/abs/2603.15358.

Yue Yu, Weiqi Chen, Binqing Wu, Wanyi Jiang, Dongliang Cui, Zongjiang Shang, Bo Wu, Liang Sun, and Ling Chen. ClimateAR: Multi-scale autoregressive generative modeling for climate forecasting, 2026. URL https://openreview.net/forum?id=MMcyzQUPqb.

Xiaohui Zhong, Lei Chen, Xu Fan, Wenxu Qian, Jun Liu, and Hao Li. Fuxi-2.0: Advancing machine learning weather forecasting model for practical applications. *ArXiv*, abs/2409.07188, 2024. URL https://api.semanticscholar.org/CorpusID:272593408.

Shengchen Zhu, Yiming Chen, Peiying Yu, Xiang Qu, Yuxiao Zhou, Yiming Ma, Zhizhan Zhao, Yukai Liu, Hao Mi, and Bin Wang. Puyun: Medium-range global weather forecasting using large kernel attention convolutional networks, 2024. URL https://arxiv.org/abs/2409.02123.

APPENDIX

A NOTATION

Table 4 summarizes the grid states, messages, gates, and loss terms used by STBridge.

Table 4: Key STBridge notation.

Symbol	Description
$\mathbf{X}, \mathbf{Y}, \widehat{\mathbf{Y}}$	Atmospheric history, verifying field, and forecast.
$\mathcal{R}_s, \mathcal{U}_s$	Area-weighted restriction to scale s and bilinear return to the grid.
$\mathbf{H}_{0,\tau}, \mathbf{H}_{\mathrm{mg},\tau}$	Encoded state and its fused multi-grid state.
$\mathbf{H}_{\mathrm{st},\tau}, \mathbf{A}_\tau$	State and edge affinities from neighborhood learning.
$\mathbf{H}_{\mathrm{sp},\tau}, \mathbf{H}_{\mathrm{sem},\tau}, \mathbf{H}_{\mathrm{rt},\tau}$	Spatial, semantic, and fused route states.
$\mathbf{Z}_\tau$	Coupled state supplied to the forecast decoder.
$\alpha_{i,\tau}^{(s)}, \boldsymbol{\delta}_{i,\tau}^{(s)}$	Adaptive scale weight and refined message returned to cell i.
$a_{ij,\tau}, \mathbf{m}_{i,\tau}^{(k)}$	Edge affinity and directional message at neighborhood level k.
$c_{i,\tau}^{u\to v}, \gamma_{i,\tau}^{v}$	Influence from source element u to target v and coupling gate.
$\mathcal{L}_{\mathrm{lat}}, \widetilde{\mathcal{L}}_q, \lambda_q$	Field risk, normalized structural risk, and its positive weight.

B EXPERIMENTAL DETAILS

B.1 DATASET DETAILS

The deterministic and ensemble evaluations in Section 5 use the same spatial grid and six-hourly forecast leads.

STBridge uses 2019–2020 ERA5 fields for training and 2021 for validation. All models undergo deterministic evaluation on the same 256 initializations from 2022, with forecasts scored every six hours from 6 to 168 hours. Metric curves connect all 28 leads directly, without interpolation or smoothing.

We use the 1.5° evaluation grid adopted in the WeatherBench 2 benchmark, comprising 240 longitude columns and 121 latitude rows including the poles (Rasp et al., 2024; WeatherBench 2 Team, 2024). This choice aligns the experiment with the benchmark's evaluation resolution and gives all models a consistent spatial basis for comparison. Forecasts and verification fields are aligned using first-order conservative regridding, with weights proportional to the overlap area between source and target cells. The longitude spacing is $360^\circ/240 = 1.5^\circ$, and the latitude spacing is $180^\circ/(121-1) = 1.5^\circ$.

Ensemble evaluation protocol. The four-member ensemble comparison uses the same 32 season-balanced initializations for STBridge, Pangu-Weather, and FuXi. STBridge's correlated input perturbations and spread scale are selected through four-fold out-of-fold calibration. Each fold uses 24 fitting and 8 held-out initializations, keeping the fitting cases separate from the held-out verification cases.

Ablation evaluation protocol. The module-removal variants start from independent module-neutral random initializations, while the complete STBridge row uses the locked main-results system. Alternative neighborhood configurations use the same initialization protocol as the reduced variants. Neighborhood level k adds a grid distance of 2^{k-1}; the $K = 2$, $K = 3$, and $K = 4$ configurations therefore use distances $\{1, 2\}$, $\{1, 2, 4\}$, and $\{1, 2, 4, 8\}$, respectively. All variants are scored on WS10 and Z500 using the reported forecast leads.

B.2 BASELINE DETAILS

Pangu-Weather uses a three-dimensional Earth-specific Transformer and hierarchical temporal aggregation for global medium-range forecasting (Bi et al., 2023). FuXi uses a cascade of models for short-, medium-, and long-range forecast windows (Chen et al., 2023). OneForecast combines multi-scale information propagation with nested grids for global and regional forecasting (Gao et al.,

2025). Aurora is an Earth-system foundation model that learns representations from diverse geophysical data and adapts them to downstream forecasting tasks (Bodnar et al., 2025). The evaluated models follow the common verification protocol specified in Appendix B.1.

All deterministic forecasts follow the shared 256-case manifest, retain all 28 leads, and use the same ERA5 targets and climatological anomalies. Alignment checks cover channel order, normalization, grid coordinates, periodic boundaries, pressure levels, derived surface variables, regridding, and physical units.

B.3 Metric Details

Deterministic metrics. For an error field on the grid, latitude-weighted RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i \left(\widehat{Y}_i - Y_i\right)^2}. \tag{11}$$

where $N = HW$ is the node count, w_i is the normalized latitude weight, and $\widehat{Y}_i$ and Y_i are forecast and truth at node i. Lower RMSE indicates smaller physical-field error.

Let $Y' = Y - \overline{Y}_{\mathrm{clim}}$ and $\widehat{Y}' = \widehat{Y} - \overline{Y}_{\mathrm{clim}}$ be climatological anomalies. The anomaly correlation coefficient is

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N} w_i \widehat{Y}_i' Y_i'}{\sqrt{\sum_{i=1}^{N} w_i (\widehat{Y}_i')^2}\sqrt{\sum_{i=1}^{N} w_i (Y_i')^2}}. \tag{12}$$

where $\overline{Y}_{\mathrm{clim}}$ is the matched climatological mean and primes denote anomaly fields. Higher ACC indicates better spatial anomaly agreement. Bias is the latitude-weighted mean signed error.

Ensemble metrics. For ensemble members $\{x_m\}_{m=1}^{M}$ and observation y, empirical CRPS is

$$\mathrm{CRPS} = \frac{1}{M}\sum_{m=1}^{M} |x_m - y| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} |x_m - x_n|. \tag{13}$$

where M is the member count, x_m is member m, and y is the verifying observation. CRPS jointly evaluates ensemble location and dispersion.

For ensemble mean μ_b, spread σ_b, and N_e verification entries, spread–mean error is

$$\mathrm{SME} = \left| \frac{1}{N_e}\sum_{b=1}^{N_e} (\sigma_b - |y_b - \mu_b|) \right|. \tag{14}$$

where N_e is the number of ensemble verification entries, μ_b and σ_b are ensemble mean and standard deviation, and y_b is truth. Lower SME indicates closer agreement between the mean spread and the mean absolute error of the ensemble mean.

For empirical quantiles, relative quantile error is

$$\mathrm{RQE} = \frac{1}{N_e K_{\mathrm{qtl}}}\sum_{b=1}^{N_e}\sum_{\kappa=1}^{K_{\mathrm{qtl}}} \nu_\kappa \frac{|\,\mathrm{Quant}_{\alpha_\kappa,b} - y_b|}{\max(|y_b|, \epsilon)}. \tag{15}$$

where K_{qtl} is the number of quantile levels, b indexes the N_e verification entries, $\mathrm{Quant}_{\alpha_\kappa,b}$ is the empirical quantile at level α_κ, $\nu_\kappa \geq 0$ is its weight normalized by $\sum_\kappa \nu_\kappa = K_{\mathrm{qtl}}$, and $\epsilon > 0$ protects the relative denominator. Lower RQE indicates smaller weighted relative deviations of the predicted quantiles from the observations.

B.4 Implementation Details

Architecture configuration. The evaluated STBridge configuration has hidden width 48, two graph blocks, and coupling width 32. It uses grid factors 2, 4, and 8, three spatial neighborhood levels with offsets 1, 2, and 4, and a temporal convolution with kernel size three. Its content-router, local-adapter, and lead-shape widths are 24, 48, and 32. The reported scores come from the specialist checkpoints described below.

Let N_s be the node count at scale s. The multi-grid blocks require

$$\mathcal{O}\left(d^2 \sum_{s\in\mathcal{S}_{\mathrm{grid}}} N_s\right) \quad \text{with} \quad N_s \approx N/s^2. \tag{16}$$

where d is the hidden width, N is the node count, and $\mathcal{S}_{\text{grid}} = \{1, 2, 4, 8\}$ is the scale set. The geometric sum keeps coarse-grid computation bounded by a constant fraction of the grid block.
For the subsequent routing stage, the directional graph contains at most $8KN$ local edges and has complexity $\mathcal{O}(KNd^2)$. Coupling at each grid location has complexity $\mathcal{O}(N|\mathcal{V}|^2 d_c)$ because each node exchanges information across all element pairs. Hidden states dominate activation memory, while pooled grids and sparse local relations add bounded auxiliary storage.

Optimization settings. The specialists use AdamW with initial learning rate 1.8×10^{-6} and minimum learning rate 1.8×10^{-7}, eight epochs with one warmup epoch, batch size one, and AMP with TF32 numerics. Cosine decay sets the learning-rate schedule, and an RMSE/ACC frontier score determines checkpoint selection. These are the training settings for the checkpoints that generated the reported results. STBridge training was conducted on eight GPUs with 32 GB of memory each (8×32 GB).

Inference. At inference, specialist parameters and moving references are fixed. The encoder processes atmospheric history once, and the latent recurrence advances through all 28 six-hour forecast leads. The decoder returns the full multivariable forecast sequence in physical units.

C Model Details

C.1 Inputs, Encoding, and Direct Output

Input configuration. The current implementation uses $L = 2$, $C_x = 6$, and $C_g = 6$. The two atmospheric states are six hours apart:

$$\mathbf{X}_t^{\text{hist}} = [\mathbf{X}_{t-6\text{h}}, \mathbf{X}_t].$$

The encoder uses both frames, providing twelve historical values per grid cell. In Equation 1, consecutive history indices denote six-hourly samples. The temporal kernel of size three acts on forecast latent states, separately from the two-frame historical input.
The dynamic channels are ordered as T2M (2-m temperature, K), WS10 (10-m wind speed, $\mathrm{m\,s^{-1}}$), MSLP (mean sea-level pressure, Pa), T850 (850-hPa temperature, K), Q700 (700-hPa specific humidity, $\mathrm{kg\,kg^{-1}}$), and Z500 (500-hPa geopotential, $\mathrm{m^2\,s^{-2}}$). All six variables supply historical input. WS10 is scalar wind speed, and Z500 is geopotential. The history axis is separate from the six-channel variable axis.
At cell i, geographic descriptors are the six dimensionless coordinate encodings

$$\mathbf{E}_{\text{geo},i} = [\sin\phi_i, \cos\phi_i, \sin\lambda_i, \cos\lambda_i, \sin(2\lambda_i), \cos(2\lambda_i)],$$

where latitude ϕ_i and longitude λ_i are in radians. These descriptors comprise the full static input; raw coordinates, terrain elevation, surface geopotential, land–sea masks, and calendar or diurnal features are excluded. Lead embeddings and variable tokens enter separately and are not counted in C_g.
On the 121×240 grid, $N = 29{,}040$ and the tensor dimensions, excluding the batch axis, are

$$\mathbf{X}_t^{\text{hist}} \in \mathbb{R}^{2\times 29040\times 6}, \qquad \mathbf{E}_{\text{geo}} \in \mathbb{R}^{29040\times 6}, \qquad \widehat{\mathbf{Y}}_{t+\mathcal{T}} \in \mathbb{R}^{28\times 29040\times 6},$$

with $\mathcal{T} = \{6, 12, \ldots, 168\}$ h.

Training-history normalization. For each dynamic variable v, the implementation computes a mean and population standard deviation over all spatial cells and both history frames in the training windows. If M is the number of values counted for that variable,

$$\mu_v = \frac{1}{M}\sum_{m=1}^{M} x_m^v, \qquad s_v = \max\left(\sqrt{\frac{1}{M}\sum_{m=1}^{M}(x_m^v - \mu_v)^2}, 10^{-8}\right).$$

Overlapping windows count repeated timestamps once per occurrence. Inputs and corresponding targets use the same pair (μ_v, s_v), fixed during validation and testing. The six geographic encodings are used directly. The implementation saves fitted statistics in `normalization.json`; variable order, physical units, training-data range, and the window-counting convention define their interpretation. A separately computed statistics file can be loaded for a protocol that counts each unique training timestamp once.

Encoding and output. The shared encoder combines normalized atmospheric history $\widetilde{\mathbf{X}}_{t-L+1:t}$, static geographic descriptors $\mathbf{E}_{\text{geo}}$, target identity, and forecast lead to form $\mathbf{H}_{0,\tau}^v \in \mathbb{R}^{N\times d}$. Each

coarse branch carries $N_s \times d$ states and returns an $N \times d$ message. Neighborhood learning supplies an $N \times d$ temporal state and sparse edge weights; routing and coupling preserve the grid-cell order. The encoder and direct output map are

$$\begin{aligned}\mathbf{H}^v_{0,\tau} &= E_\theta([F_{\text{hist}}(\widetilde{\mathbf{X}}_{t-L+1:t}), \mathbf{E}_{\text{geo}}]) + \mathbf{e}_v + \mathbf{e}_\tau,\\ \widehat{\mathbf{Y}}^v_{t+\tau} &= \mu_v \mathbf{1} + s_v P^v_\theta(D^v_\theta([\mathbf{Z}^v_\tau, \mathbf{e}_v, \mathbf{e}_\tau]))\,.\end{aligned} \tag{17}$$

where F_{hist} maps the L input states to a grid history representation and E_θ combines it with geographic descriptors; target and lead embeddings are then added. The Variable-Conditioned Forecast Decoder D^v_θ and target head P^v_θ produce the normalized field, and the training-split mean μ_v and standard deviation s_v restore physical units. Input dimensions follow Equation 1; the hidden widths are specified in Appendix B.4.

C.2 Adaptive Multi-Grid Transfer

For each $s \in \{2, 4, 8\}$, the area-aware restriction in the multi-grid weather space creates a coarse state, a scale block conditioned on the forecast lead evolves it, and bilinear prolongation returns the evolved states to their originating forecast-grid cells. A learned refinement and scale score then provide the message and gate used by the adaptive bridge:

$$\begin{aligned}\widetilde{\mathbf{H}}^{v,(s)}_\tau &= \Phi_s(\mathcal{R}_s(\mathbf{H}^v_{0,\tau}) + \mathbf{e}_s, \mathbf{e}_\tau), & \mathbf{U}^{v,(s)}_\tau &= \mathcal{U}_s(\widetilde{\mathbf{H}}^{v,(s)}_\tau),\\ \boldsymbol{\delta}^{v,(s)}_\tau &= \Psi_s([\mathbf{U}^{v,(s)}_\tau, \mathbf{H}^v_{0,\tau}, |\mathbf{U}^{v,(s)}_\tau - \mathbf{H}^v_{0,\tau}|]),\\ r^{v,(s)}_{i,\tau} &= f_g([\mathbf{h}^v_{0,i,\tau}, \mathbf{u}^{v,(s)}_{i,\tau}, |\mathbf{u}^{v,(s)}_{i,\tau} - \mathbf{h}^v_{0,i,\tau}|, \mathbf{e}_\tau]).\end{aligned} \tag{18}$$

For a latitude row centered at ϕ_i, its cell boundaries are the midpoints to adjacent rows, clipped at the poles; angles are expressed in radians. The area weight is the cell solid angle, so the common squared Earth radius cancels in normalized pooling. To define restriction on the odd latitude dimension, use non-overlapping blocks of up to s rows and s columns, retaining the final partial row block. This gives $(H_s, W_s) = (61, 120), (31, 60), (16, 30)$ for $s = 2, 4, 8$. Each block pools the area-weighted valid cells. Prolongation interpolates using the area-weighted coarse-cell coordinates, wraps longitude periodically, and clamps latitude outside the coarse-center range. Its weights are nonnegative and sum to one.

The finite-cell area weight is $a_i = \Delta\lambda_i[\sin(\phi_i^+) - \sin(\phi_i^-)]$. The scale embedding $\mathbf{e}_s$ and evolution block Φ_s follow Equation 3; Ψ_s implements Learned Refinement and f_g implements the Scale-Disagreement MLP. The Softmax Bridge Gate gives $\alpha^{v,(s)}_{i,\tau} = \exp(r^{v,(s)}_{i,\tau}) / \sum_{q \in \mathcal{S}_c} \exp(r^{v,(q)}_{i,\tau})$. These weights and refined messages enter Equation 4, followed by feature normalization.

C.3 Periodic Geometry and Spatio-Temporal State

Longitude indices wrap periodically so east–west messages cross the dateline continuously. Pole-adjacent cells retain valid latitude neighbors, and incoming affinities are normalized over this valid support. Direction- and distance-specific parameters distinguish zonal, meridional, and diagonal transport at offsets 1, 2, and 4. Projected endpoint states and target and lead embeddings form the edge score

$$e^{v,(k)}_{ij,\tau} = \frac{(W^e_Q \mathbf{h}^v_{\text{mg,i},\tau})^\top (W^e_K \mathbf{h}^v_{\text{mg,j},\tau})}{\sqrt{d_e}} + b_{r_{ij},k} + \mathbf{u}^\top_{r_{ij},k}[\mathbf{e}_v, \mathbf{e}_\tau]. \tag{19}$$

where r_{ij} and k identify direction and distance level. Edge-Affinity Softmax gives $a^{v,(k)}_{ij,\tau} = \exp(e^{v,(k)}_{ij,\tau}) / \sum_{q \in \mathcal{N}_k(i)} \exp(e^{v,(k)}_{iq,\tau})$, normalized within each distance level as in Equation 5. The hierarchy scorer $b^{v,(k)}_{i,\tau} = f_k([\mathbf{h}^v_{\text{st,i},\tau}, \mathbf{m}^{v,(k)}_{i,\tau}, \mathbf{e}_\tau])$ sets the level weights through $\pi^{v,(k)}_{i,\tau} = \exp(b^{v,(k)}_{i,\tau}) / \sum_{q=1}^{3} \exp(b^{v,(q)}_{i,\tau})$ in Equation 7.

Alongside these spatial affinities, a depthwise temporal convolution at lead $\tau_j = 6j$ captures short-term variations, while a GRU carries the recurrent state across forecast leads. Their projected responses enter the temporal state update in the Method section:

$$\mathbf{T}^v_{\tau_j} = F_{\text{temp},3}(\mathbf{H}^v_{\text{mg},\tau_{\text{j}-2:\text{j}}}), \qquad \mathbf{s}^v_{\tau_j} = F_{\text{rec}}(\mathbf{H}^v_{\text{mg},\tau_{\text{j}}}, \mathbf{s}^v_{\tau_{j-1}}). \tag{20}$$

The states remain aligned with cells; $F_{\text{temp},3}$ is a depthwise temporal filter over three leads, and F_{rec} is the recurrent update. Channel projections W_t and W_h align the responses before residual addition and layer normalization. The recurrence initializes $\mathbf{s}^v_{\tau_0} = \mathbf{0}$ and extends the temporal-

convolution input to nonpositive lead indices by copying $\mathbf{H}^{v}_{\mathrm{mg},\tau_1}$. Both temporal operators process the lead-conditioned latent sequence in increasing lead order.

C.4 Semantic Routing and Output Gates

The directional message uses the temporal node state together with affinities learned from the multi-grid state:

$$\mathbf{m}^{v,(k)}_{i,\tau} = \sum_{j\in\mathcal{N}_k(i)} a^{v,(k)}_{ij,\tau} W_{r_{ij},k} \mathbf{h}^{v}_{\mathrm{st,j},\tau}. \tag{21}$$

The edge weights are computed from the multi-grid state, and message aggregation uses the temporal node state.

Semantic routing queries three descriptors from the spatio-temporal state: target identity, internal cross-scale disagreement, and the multivariate latent state. The descriptors supply keys and values for the attention operation:

$$\begin{aligned} \mathbf{d}^{v}_{i,\tau} &= \sum_{s\in\mathcal{S}_c} \alpha^{v,(s)}_{i,\tau} |\mathbf{u}^{v,(s)}_{i,\tau} - \mathbf{h}^{v}_{0,i,\tau}|, \\ \mathbf{p}^{v,(1)}_{i,\tau} &= \mathbf{e}_v, \qquad \mathbf{p}^{v,(2)}_{i,\tau} = \mathbf{d}^{v}_{i,\tau}, \qquad \mathbf{p}^{v,(3)}_{i,\tau} = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} \mathbf{h}^{u}_{\mathrm{st,i},\tau}. \end{aligned} \tag{22}$$

All three descriptors have width d. Their compatibility scores are $\ell^{v,(p)}_{i,\tau} = (W_Q \mathbf{h}^{v}_{\mathrm{st,i},\tau})^\top W_K \mathbf{p}^{v,(p)}_{i,\tau}/\sqrt{d_a}$, where d_a is the query–key width. Semantic attention assigns $\omega^{v,(p)}_{i,\tau} = \exp(\ell^{v,(p)}_{i,\tau})/\sum_{r=1}^{3}\exp(\ell^{v,(r)}_{i,\tau})$ and forms the residual semantic state as in Equation 8:

$$\begin{aligned} \omega^{v,(p)}_{i,\tau} \ge 0, \qquad \sum_{p=1}^{3} \omega^{v,(p)}_{i,\tau} = 1, \\ \mathbf{c}^{v}_{\mathrm{sem,i},\tau} = \sum_{p=1}^{3} \omega^{v,(p)}_{i,\tau} W_V \mathbf{p}^{v,(p)}_{i,\tau}, \\ \mathbf{H}^{v}_{\mathrm{sem},\tau} = \mathbf{H}^{v}_{\mathrm{st},\tau} + \mathbf{C}^{v}_{\mathrm{sem},\tau}. \end{aligned} \tag{23}$$

Here $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, and $\mathbf{p}^{(3)}$ are the target, disagreement, and multivariate descriptors. Their compatibility scores determine nonnegative weights $\omega^{v,(p)}_{i,\tau}$ that sum to one; $\mathbf{C}^{v}_{\mathrm{sem},\tau}$ stacks the weighted messages. The spatial state stacks the weighted directional messages from the graph route. The gate $\boldsymbol{\eta}^{v}_{i,\tau} = \sigma(f_\eta([\mathbf{h}^{v}_{\mathrm{sp,i},\tau}, \mathbf{h}^{v}_{\mathrm{sem,i},\tau}])) \in (0,1)^d$ fuses the two route states.

C.5 Multi-Element Coupling

At cell i and lead τ, the routed state for variable u and embedding $\mathbf{e}_u$ form its token. Query and key projections have width d_c; value projections return width d so messages can be added to the routed state. The message and its injection gate are

$$\begin{aligned} \mathbf{p}^{u}_{i,\tau} &= W_u[\mathbf{h}^{u}_{\mathrm{rt,i},\tau}, \mathbf{e}_u], & \mathbf{c}^{v}_{i,\tau} &= \sum_{u\in\mathcal{V}} c^{u\to v}_{i,\tau} V_u \mathbf{p}^{u}_{i,\tau}, \\ \gamma^{v}_{i,\tau} &= \sigma\big(f_c([\mathbf{h}^{v}_{\mathrm{rt,i},\tau}, \mathbf{c}^{v}_{i,\tau}, \mathbf{e}_\tau])\big), & \widetilde{\mathbf{z}}^{v}_{i,\tau} &= \mathbf{h}^{v}_{\mathrm{rt,i},\tau} + \gamma^{v}_{i,\tau}\mathbf{c}^{v}_{i,\tau}. \end{aligned} \tag{24}$$

For target v and source u, the compatibility score is $\ell^{u\to v}_{i,\tau} = (Q_v \mathbf{p}^{v}_{i,\tau})^\top K_u \mathbf{p}^{u}_{i,\tau}/\sqrt{d_c}$. Coupling Softmax gives $c^{u\to v}_{i,\tau} = \exp(\ell^{u\to v}_{i,\tau})/\sum_{q\in\mathcal{V}}\exp(\ell^{q\to v}_{i,\tau})$, using the same target query for every source. LayerNorm maps the residual state $\widetilde{\mathbf{z}}^{v}_{i,\tau}$ to $\mathbf{z}^{v}_{i,\tau}$ for direct decoding in Equation 17.

C.6 Objective Definitions

The primary risk measures variable-normalized physical-field error over leads and grid cells with latitude weights normalized to unit mean:

$$w_i = \frac{a_i}{N^{-1}\sum_{j=1}^{N} a_j},$$

$$\mathcal{L}_{\text{lat}} = \frac{1}{|\mathcal{T}||\mathcal{V}|N} \sum_{\tau,v,i} w_i \left(\frac{\widehat{Y}^v_{i,t+\tau} - Y^v_{i,t+\tau}}{s_v} \right)^2. \tag{25}$$

where a_i is the finite-cell area weight and $s_v > 0$ is the training-split standard deviation of target variable v, consistent with the normalization in Equation 1.

The field risk is combined with gradient, spectral, bridge, and coupling risks. With $\mathcal{P}$ denoting the zonal power spectrum operator, these terms are:

$$\mathcal{L}_{\nabla} = \frac{1}{|\mathcal{V}||\mathcal{T}|} \sum_{v,\tau} \|\nabla_s \widehat{\widetilde{\mathbf{Y}}}^v_{t+\tau} - \nabla_s \widetilde{\mathbf{Y}}^v_{t+\tau}\|_{w,1}, \tag{26}$$

$$\mathcal{L}_{\text{spec}} = \frac{1}{|\mathcal{V}||\mathcal{T}|} \sum_{v,\tau} \|\log(\epsilon + \mathcal{P}(\widehat{\widetilde{\mathbf{Y}}}^v_{t+\tau})) - \log(\epsilon + \mathcal{P}(\widetilde{\mathbf{Y}}^v_{t+\tau}))\|_1, \tag{27}$$

$$\mathcal{L}_{\text{bridge}} = \frac{1}{|\mathcal{V}||\mathcal{S}_c||\mathcal{T}|Nd} \sum_{v,s,\tau} \|\mathbf{U}^{v,(s)}_{\tau} - \text{sg}(\mathbf{H}^v_{0,\tau})\|^2_{\text{F}}, \tag{28}$$

$$\mathcal{L}_{\text{coup}} = \frac{1}{N|\mathcal{T}||\mathcal{V}|d} \sum_{i,\tau,v} \|\mathbf{c}^v_{i,\tau} - V_v \mathbf{p}^v_{i,\tau}\|^2_2. \tag{29}$$

where ∇_s is the latitude–longitude finite-difference operator, $\|\cdot\|_{w,1}$ is the latitude-weighted absolute norm, $\mathcal{S}_c = \{2, 4, 8\}$ is the set of coarse context scales, and sg stops gradients. The coupling risk compares messages and target-variable representations in the dimensionless latent space. To place these structural terms on a common optimization scale, an exponential moving reference is maintained for each $q \in \{\nabla, \text{spec}, \text{bridge}, \text{coup}\}$:

$$m^{(\ell)}_q = \beta m^{(\ell-1)}_q + (1-\beta)\,\text{sg}(\mathcal{L}^{(\ell)}_q), \qquad \widetilde{\mathcal{L}}^{(\ell)}_q = \frac{\mathcal{L}^{(\ell)}_q}{\text{sg}(m^{(\ell)}_q) + \epsilon}. \tag{30}$$

where β is the moving-average decay and sg stops gradients through the reference. The references persist across optimization steps and are initialized from the first detached loss values. The normalized risks enter the multiplicative objective in Equation 10.

C.7 Training and Forecasting Algorithms

Algorithm 1 produces the full forecast sequence and intermediate states; Algorithm 2 evaluates the objective and updates the model parameters.

Algorithm 1 STBridge sequence forecast

Require: Normalized history $\widetilde{\mathbf{X}}_{t-L+1:t}$, geographic descriptors, lead and target embeddings, parameters Θ
Ensure: Forecast sequence $\{\widehat{\mathbf{Y}}_{t+\tau_j}\}_{j=1}^{28}$ and intermediate states for the structural risks
1: Encode the normalized history and initialize the recurrent latent states
2: **for** $\ell = 1, \ldots, 28$ in increasing lead order **do**
3: Set $\tau = \tau_\ell = 6\ell$ hours and form grid tokens using geography, target, and lead embeddings
4: **for** $s \in \{2, 4, 8\}$ **do**
5: Restrict the state with spherical area weights, evolve with Φ_s, and restore by bilinear interpolation
6: **end for**
7: Compute the bridge mixture $\mathbf{H}_{\mathrm{mg},\tau}$ using gates conditioned on node and lead
8: Build valid periodic directional neighborhoods at distances 1, 2, and 4 and compute their edge weights
9: Form the temporal-filter and recurrent states at cells, then obtain $\mathbf{H}_{\mathrm{st},\tau}$
10: Aggregate weighted directional messages, mix distance levels, and fuse the spatial and semantic routes into $\mathbf{H}_{\mathrm{rt},\tau}$
11: Mix element tokens with target-conditioned coupling weights and an injection gate to obtain $\mathbf{Z}_\tau$
12: Decode all six normalized future fields and restore their physical units
13: **end for**
14: **return** Forecast sequence and intermediate bridge states, variable tokens, and coupling messages

The returned sequence and intermediate bridge states and coupling messages enter the risks defined in Appendix C.6.

Algorithm 2 6–168h rollout and compound optimization

Require: Parameters Θ, persistent moving references $\{m_q\}$, training batch with normalized history $\widetilde{\mathbf{X}}_{t-L+1:t}$ and targets $\{\mathbf{Y}_{t+\tau_j}\}_{j=1}^{28}$
Ensure: Updated parameters Θ and moving references $\{m_q\}$
1: Initialize the recurrent state at the start of the sequence
2: Run Algorithm 1 once on the full 28-lead sequence
3: Compute $\mathcal{L}_{\mathrm{lat}}$, $\mathcal{L}_{\nabla}$, $\mathcal{L}_{\mathrm{spec}}$, $\mathcal{L}_{\mathrm{bridge}}$, and $\mathcal{L}_{\mathrm{coup}}$
4: Update moving references using Equation 30 and form the objective in Equation 10
5: Apply one AdamW step with mixed-precision scaling and the scheduled learning rate
6: **return** $\Theta, \{m_q\}$

The moving references are initialized from the first detached loss values and carried across optimization steps. The recurrent state is initialized at the start of each forecast sequence.

D Discussions

D.1 Cross-Scale Transfer and Grid Interactions

STBridge connects three forms of information exchange through the forecast grid. Adaptive Multi-Grid Bridging transfers features across resolutions, Dynamic Spatial Graph Routing aggregates information between cells at the forecast resolution, and Grid-Wise Multi-Element Coupling exchanges information between variables within each cell. Spatio-Temporal Learnable Neighborhood supplies the edge weights and lead-dependent states used by routing. Together, these operations link coarse circulation context, local spatial relations, and meteorological variables in the forecasting computation described in Section 4.

D.2 Compound Objective and Coupling

The compound objective in Equation 10 combines field risk with the geometric factor $G = [\prod_{q\in\mathcal{Q}}(1 + \lambda_q\widetilde{\mathcal{L}}_q)]^{1/4}$. Its auxiliary-loss derivative is $\partial\mathcal{L}/\partial\widetilde{\mathcal{L}}_q = \mathcal{L}_{\mathrm{lat}}G\lambda_q/[4(1 + \lambda_q\widetilde{\mathcal{L}}_q)]$, so field risk scales the structural gradients during joint optimization. The coupling risk approaches

zero when attention concentrates on the target's own projected token; cross-variable contributions are assessed through the coupling ablation.

E Additional Results

The following analyses extend the main results with input perturbations, regional errors, and prediction–truth statistics.

E.1 Sensitivity to Input Perturbations

Figure 8 evaluates the sensitivity of WS10, Z500, Q700, and T850 forecasts to Gaussian perturbations of the input grid features. Each heatmap reports RMSE after per-lead min–max normalization, with lighter cells indicating lower normalized error. WS10 and Z500 show higher normalized errors at larger perturbation scales across most leads, while Q700 and T850 exhibit stronger variation between forecast times. The panels compare noise responses separately within each lead.

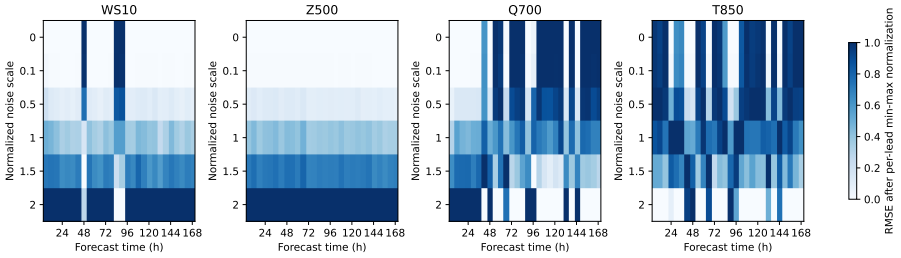

Figure 8: Forecast RMSE after per-lead min–max normalization under Gaussian input perturbations.

E.2 Regional Forecasting Performance

To locate the global forecast gains, Figure 9 examines the three variable–latitude bands with the largest average RMSE margins among bands led by STBridge. The selected panels cover MSLP at 60°–90°N, Q700 at 30°S–0°, and Z500 at 0°–30°N; columns identify bands and rows show RMSE and signed bias. STBridge has the lowest time-mean RMSE and smallest absolute bias in each selected band, outperforming the four baselines in these regional comparisons.

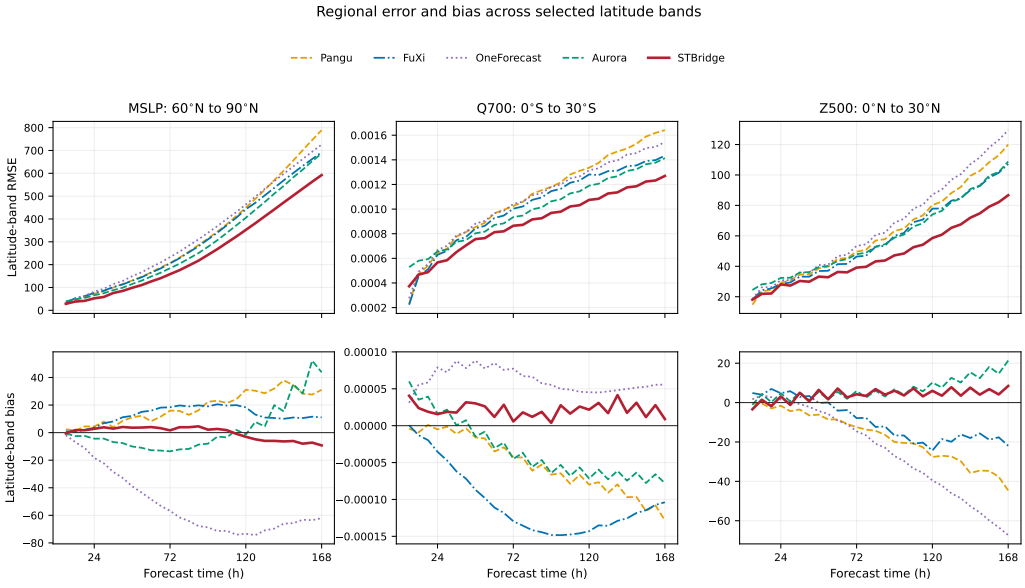

Figure 9: RMSE and bias across selected latitude bands.

E.3 Prediction–Truth Agreement

Beyond mean errors and regional averages, Figure 10 compares 120h predictions with ERA5 for T2M, T850, MSLP, and Z500. Rows identify the variables and columns the five systems; each

panel reports R^2, MAE, and signed bias. STBridge obtains the highest R^2 and lowest MAE in all four variables: $(0.9940, 1.001)$ for T2M, $(0.9884, 1.121)$ for T850, $(0.9570, 186.45)$ for MSLP, and $(0.9927, 182.72)$ for Z500. Its density remains concentrated around the identity line across low-, central-, and high-value regions, extending the aggregate RMSE advantage across the forecast distribution. For MSLP, STBridge also reaches the smallest absolute signed bias at 120h, 4.85 compared with 5.52 for the closest baseline.

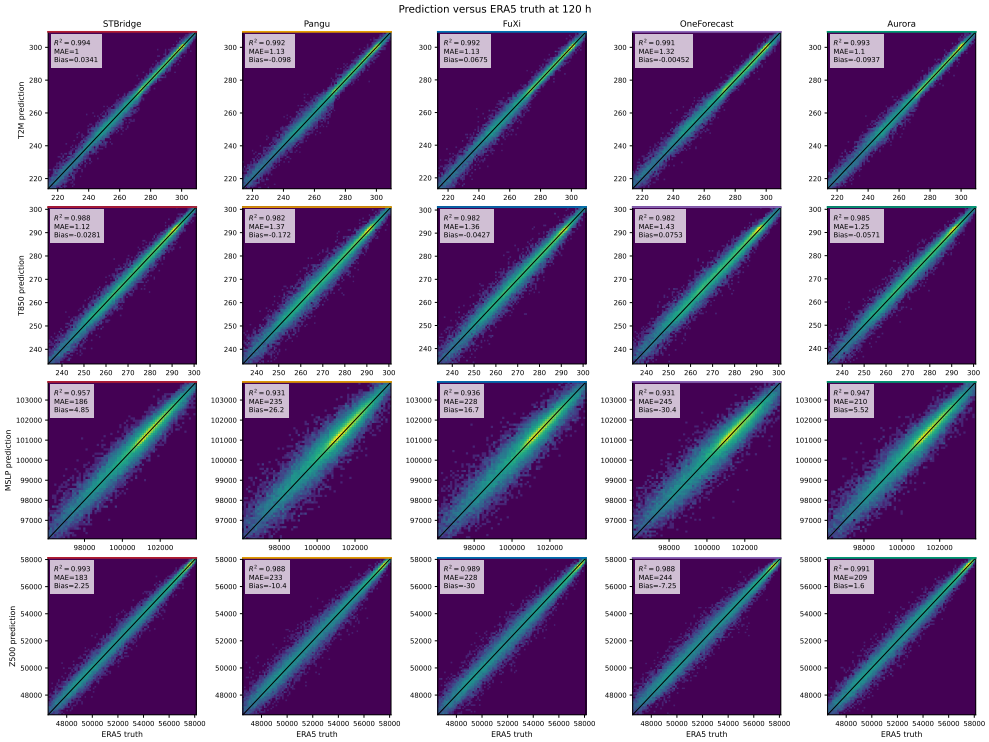

Figure 10: Prediction–truth density at 120 hours across four variables.

E.4 Full-Horizon Forecast Distribution

Extending this density comparison to the full forecast horizon, Figure 11 aggregates T2M, MSLP, T850, Q700, and Z500 over all six-hourly forecast leads through 168 hours. Every system uses the same initializations and spatial samples. Each field is standardized with statistics from the complete 2022 ERA5 verification collection, giving the five variables equal weight in the pooled density; panels report macro R^2, MAE, and mean absolute standardized bias. Under the shared identity line and sampling rule, STBridge achieves the highest macro R^2, lowest macro MAE, and smallest mean absolute standardized bias.

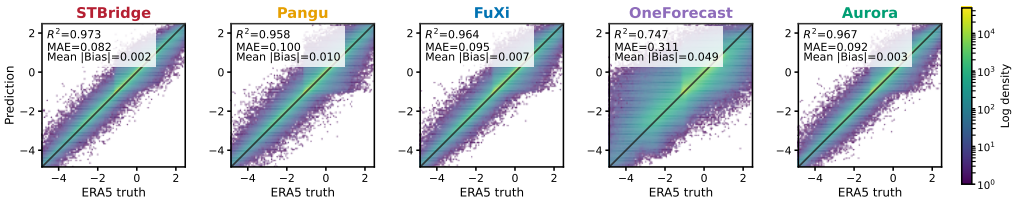

Figure 11: Pooled prediction–truth density across variables and forecast times.

F Visualization

The visualizations connect internal mechanism activity with forecast spectra and matched meteorological fields.

Fixed-point selection. For the fixed-location trajectories in Figure 7, the registered selection first requires STBridge to have the smallest time-mean MAE and RMSE among the displayed systems. Within that eligible set, each point is ranked by its normalized margin over the strongest competitor. The selected WS10, Q700, and Z500 locations are then held fixed across all forecast leads and models.

F.1 Mechanism Activity

Figure 12 visualizes the activity of Adaptive Multi-Grid Bridging, Spatio-Temporal Learnable Neighborhood, Dynamic Spatial Graph Routing, and Multi-Element Coupling across six specialists and 28 forecast times. Neighborhood levels 1, 2, and 3 correspond to grid distances 1, 2, and 4. The multi-grid and graph routes vary across both target and horizon, and the coupling panel separates the element-specific computation allocated to each atmospheric field. The heatmaps complement the component removals in Table 2 and the neighborhood-depth analysis in Table 3, demonstrating that the specialists execute distinct grid computations within one architecture.

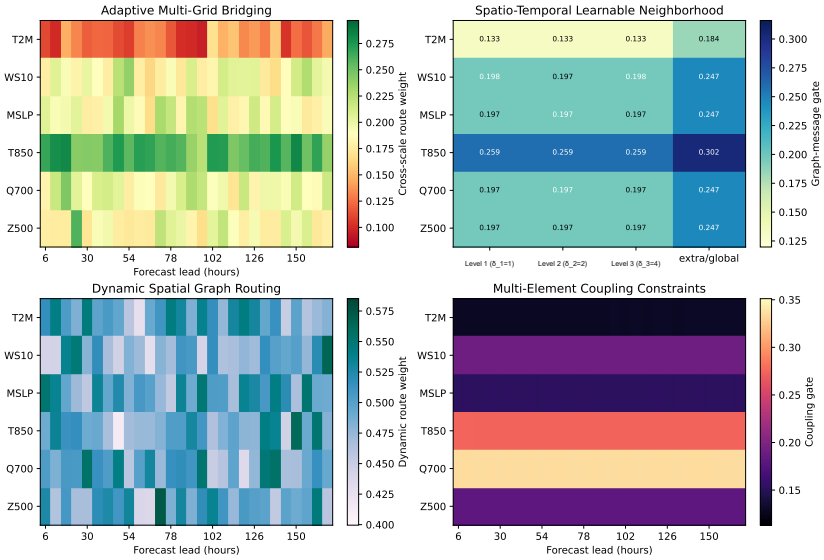

Figure 12: Activity of the four STBridge mechanisms across variables and forecast times.

F.2 Spatial Spectra

Turning from module activity to forecast-field structure, Figure 13 evaluates zonal spectral structure over the six variables at 24, 72, and 168 hours and retains the two panels with the smallest integrated log-spectrum error for STBridge. At 24h, STBridge records errors of 0.0601 for T2M and 0.1173 for Z500; the strongest baseline values are 0.0699 and 0.1194. In these selected cases, the near-surface temperature and geopotential spectra closely match those of ERA5.

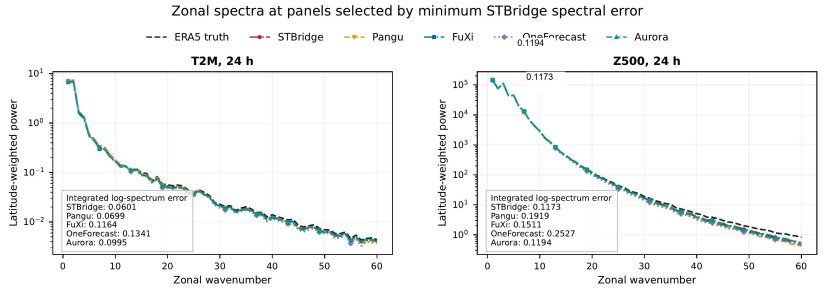

Figure 13: Zonal power spectra for selected T2M and Z500 forecasts.

F.3 Matched Spatial Forecasts

To relate the distributional results to individual weather fields, Figures 14 and 15 compare the same 3 November 2022 12:00 UTC event at 24, 72, and 168 hours. Their rows show forecast times and

columns show ERA5, STBridge, FuXi, and Aurora. All panels share one color range within each variable, and every forecast panel reports independently recomputed latitude-weighted signed bias. STBridge obtains the smallest absolute bias in all six variable-and-time comparisons. The matched fields retain spatial wind-speed patterns and large-scale geopotential structure from one to seven days in this event.

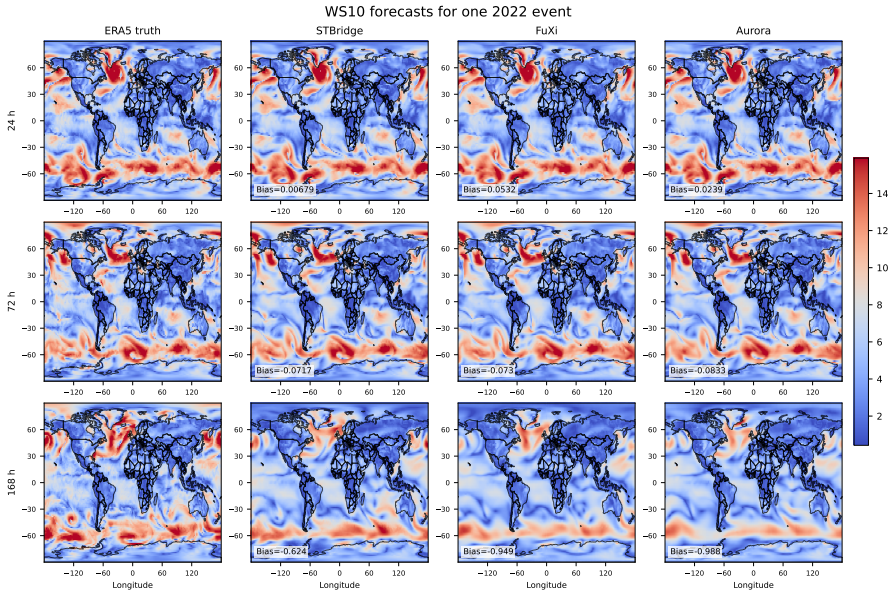

Figure 14: WS10 forecast fields and bias for a fixed 2022 event.

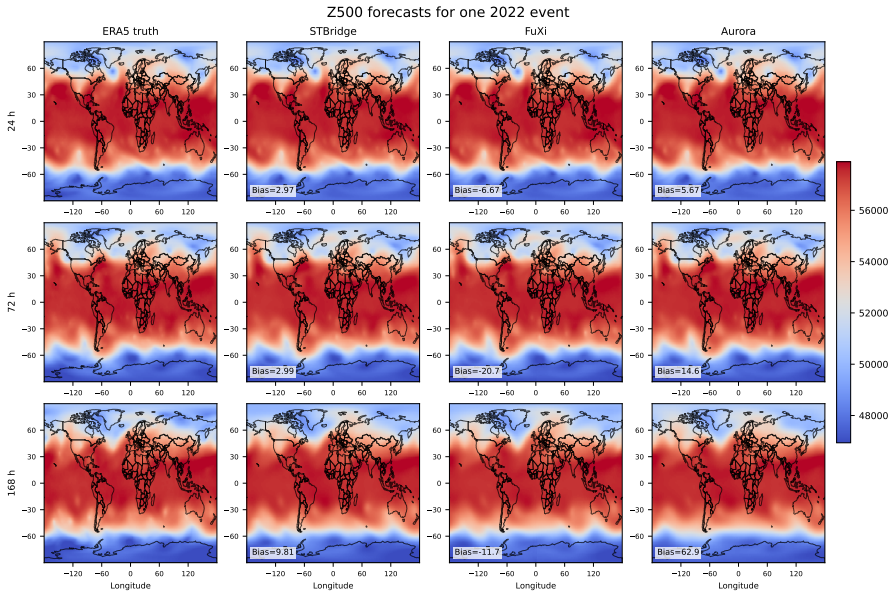

Figure 15: Z500 forecast fields and bias for the same event.