ROTARY POSITION EMBEDDINGS ON THE SPHERE FOR GLOBAL WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern transformers rely on positional encodings in the attention mechanism to account for the positions of input tokens. Rotary Position Embeddings (RoPE) are among the most popular approaches to date for this. For transformers applied to sequences defined on the sphere, for example in machine learning-based weather forecasting, extending RoPE to S^2 is non-trivial. We therefore develop S2RoPE, a spherical rotary position encoding for queries and keys defined on the sphere. We propose two variants: channel-wise modulation by normalized complex spherical harmonics (S2RoPE-SH) and rotations within complete harmonic degree spaces (S2RoPE-Wigner). The first approach modulates both amplitude and phase, whereas the second one preserves feature norms and encodes the relative position of two tokens through a rotation between their local frames in one spherical harmonics band. We apply both forms to two global medium-range forecasting models, the WeatherGenerator and Stormer. We find improvements over baseline methods in standard metrics, particularly at short and medium lead times. In WeatherGenerator, we further find improvements in the preservation of fine-scale detail and less drift in the balance between wind and geopotential during extended rollouts.

1 INTRODUCTION

Numerical global weather models have long had a native answer to the question of how position should be represented on the sphere. Since the introduction of the spectral transform method (Orszag, 1970; Eliasen et al., 1970; Bourke, 1972), leading forecasting systems such as ECMWF's Integrated Forecasting System have represented the atmosphere as expansions in spherical harmonics. This choice gives a resolution that is uniform over the globe and independent of where the coordinate poles are placed (Durran, 2010). Data-driven models have begun to adopt the same structure, for instance in spherical neural operators (Bonev et al., 2023) and location encoders (Rußwurm et al., 2024).

The representation of spherical geometry in transformers is so far, however, imperfect: the positional encoding of queries and keys usually does not respect the spherical geometry. Current weather forecasting models that employ transformers use learned embeddings (Nguyen et al., 2023), Fourier features of latitude and longitude (Bodnar et al., 2025), learned positional biases (Bi et al., 2023), or rotary position embeddings (RoPE) borrowed from language and vision (Su et al., 2024; Heo et al., 2024), which treat latitude and longitude as flat, independent axes. The same distance in longitude, however, spans very different distances at the equator and near the poles. An existing approach (Zhuang & Duraisamy, 2025) adapts RoPE to geographic coordinates through latitude- and longitude-specific mappings, while retaining a coordinate-wise rotary structure.

We close this gap with S2RoPE, which employs spherical harmonics for rotary positional encodings that closely mimic the 1-dimensional counterpart. Classical RoPE can be read in two equivalent ways: as multiplying features by phase factors, or as rotating pairs of features. Replacing its one-dimensional Fourier modes with spherical harmonics turns these two readings into two constructions. S2RoPE-SH multiplies query and key features by spherical harmonic values, so that attention reflects the angular distance between tokens, similar to how this is achieved in classical RoPE. S2RoPE-Wigner rotates groups of features according to each token's position, so that attention depends on the relative position of two tokens while preserving feature norms. Using several

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.

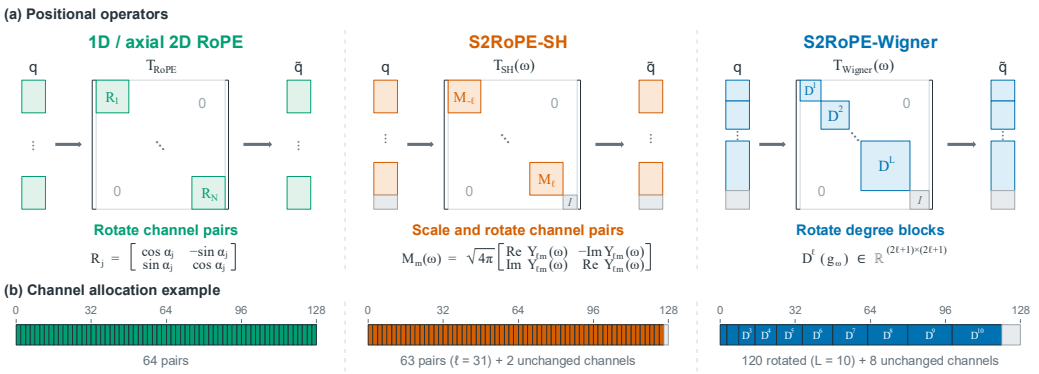


Figure 1: **S2RoPE positional operators and channel allocation.** (a) 1D and axial 2D RoPE rotate channel pairs; S2RoPE-SH scales and rotates pairs; S2RoPE-Wigner rotates complete harmonic degree blocks. Operators act on queries and keys; values are unchanged. Gray denotes identity blocks or unchanged channels. (b) Example channel allocation for a head with $d = 128$.

harmonic degrees gives a multi-scale representation, much like RoPE's range of frequencies. Neither construction is fully invariant to rotations of the sphere. For weather this is a natural choice: Earth's rotation singles out the polar axis, which S2RoPE retains.

We evaluate both constructions in two forecasting architectures, WeatherGenerator by Lessig et al. (2026) and Stormer (Nguyen et al., 2024), on global forecasts up to 14 days, and find improvements in RMSE and ACC. S2RoPE also better preserves fine-scale spectral structure, and reduces drift from atmospheric balances during long rollout. A controlled ablation further shows that using multiple harmonic degrees outperforms repeating low-degree representations at the same channel budget.

Our contributions are: (i) **S2RoPE**, two rotary position encodings on the sphere based on spherical harmonics, derived from the two views of classical RoPE; (ii) **an evaluation in two weather forecasting architectures**, with analyses of spectral fidelity and physical consistency; and (iii) **a controlled study** of how the choice of harmonic degrees affects model performance at a fixed channel budget.

2 Method

We study positional encoding in transformer models when tokens are associated with locations on the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$. Our approach provides an extension of RoPE (Su et al., 2024), which is one of the most popular approaches for positional encodings in 1-dimensional sequences, for example in LLMs. Building on classical spherical harmonic theory, we develop two constructions: harmonic modulation (S2RoPE-SH) and orthogonal rotations of complete harmonic degree spaces (S2RoPE-Wigner). We first review classical and axial RoPE, and subsequently present our two constructions and their relationship.

For one attention head, let $q_i, k_i \in \mathbb{R}^d$ be the d-dimensional query and key associated with token i, and let $\omega_i \in \mathbb{S}^2$ denote the token's position on the sphere. All positional encodings considered here act only on queries and keys through position-dependent maps $q_i \mapsto \widetilde{q}_i$ and $k_i \mapsto \widetilde{k}_i$; values are left unchanged, so position influences the output only through the attention weights. We write $A_{ij} = \widetilde{q}_i^\top \widetilde{k}_j$ for the dot product between the transformed queries and keys. The attention logit is $A_{ij}/\sqrt{d}$.

2.1 Classical rotary position embeddings

Classical RoPE (Su et al., 2024) augments the queries and keys with unit-modulus complex phase factors

$$\widetilde{q}_{i,t} = e^{\mathrm{i} n_i \xi_t} q_{i,t}, \qquad \widetilde{k}_{i,t} = e^{\mathrm{i} n_i \xi_t} k_{i,t}, \qquad \xi_t = b^{-2t/d}, \quad t = 0, \ldots, d/2 - 1. \tag{1}$$

where the d real channels have been identified with $d/2$ complex channels $q_{i,t}, k_{i,t} \in \mathbb{C}$. Here n_i is the position index of token i, and $b > 1$ is the frequency base. The query-key product now becomes

$$A_{ij} = \mathrm{Re}\left[\sum_t q_{i,t} k^*_{j,t} e^{\mathrm{i}(n_i - n_j)\xi_t}\right], \tag{2}$$

where $(\cdot)^*$ is complex conjugation. The positional factor $e^{\mathrm{i}(n_i-n_j)\xi_t}$ thus depends on relative displacement, while the unit-modulus multipliers $e^{\mathrm{i}n_i\xi_t}$ and $e^{\mathrm{i}n_j\xi_t}$ preserve query and key norms.

2.2 Axial rotary embeddings in two dimensions

Axial 2D RoPE (Heo et al., 2024) provides an extension of classical RoPE to two dimensions. It assigns half the complex channels to each coordinate. For the sphere $\mathbb{S}^2$ parametrized with a latitude λ and longitude φ coordinate system (in radians), its real operator is

$$T_{2\mathrm{D}}(\lambda, \varphi) = \left[\bigoplus_{t=0}^{d/4-1} R(\lambda\nu_t)\right] \oplus \left[\bigoplus_{t=0}^{d/4-1} R(\varphi\nu_t)\right], \qquad \nu_t = b^{-4t/d}, \tag{3}$$

where d is divisible by four and $R(\alpha)$ is the planar rotation corresponding to multiplication by $e^{\mathrm{i}\alpha}$, up to a fixed reordering of feature channels. Here, $\bigoplus$ denotes block-diagonal composition, with each block acting on a separate group of feature channels. The positional factor now depends on latitude and longitude differences. However, these are chart coordinates: a longitude increment has different physical lengths at different latitudes, and non-integer frequencies do not identify longitudes separated by 2π. A positional encoding adapted to the sphere should respect its symmetries and avoid the above degeneracies.

2.3 From RoPE to spherical harmonics

An alternative approach to generalize RoPE to the sphere is to use the analogues of the complex exponential phase factors on $\mathbb{S}^2$. These are given by zonal functions built from Legendre polynomials. For $\omega, \gamma \in \mathbb{S}^2$ being two points on the sphere, the fundamental zonal function of degree ℓ is $P_\ell(\omega \cdot \gamma)$, where P_ℓ is the Legendre polynomial of degree ℓ. Summing over degrees gives the band-limited zonal kernel of degree up to L,

$$k_L(\omega \cdot \gamma) = \sum_{\ell=0}^{L} P_\ell(\omega \cdot \gamma). \tag{4}$$

The function k_L is axially symmetric (with either ω or γ fixed) and depends only on the cosine of the angle α (or spherical distance) between ω and γ, i.e. $\cos(\alpha) = \omega \cdot \gamma$. This is the analogue of $e^{\mathrm{i}(n_i-n_j)\xi_t}$ in one-dimensional RoPE, which depends on the signed displacement $n_i - n_j$ between the two positions.

A key for the definition of 1-dimensional RoPE is the multiplicative decomposition $e^{\mathrm{i}(n_i-n_j)\xi_t} = e^{\mathrm{i}n_i\xi_t}\left(e^{\mathrm{i}n_j\xi_t}\right)^*$. For the zonal function k_L, an analogous decomposition exists through the spherical harmonics addition theorem (the existence of such a decomposition can be traced back to Lie group-theoretic arguments):

$$P_\ell(\omega \cdot \gamma) = \frac{4\pi}{2\ell+1} \sum_{m=-\ell}^{\ell} Y_{\ell m}(\omega)\, Y^*_{\ell m}(\gamma), \tag{5}$$

where the $Y_{\ell m}$ are the complex spherical harmonics. Using latitude $\lambda \in [-\pi/2, \pi/2]$, colatitude $\theta = \pi/2 - \lambda \in [0, \pi]$, and longitude $\varphi \in [0, 2\pi)$, these can be written as

$$Y_{\ell m}(\theta, \varphi) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}} P_\ell^m(\cos\theta) e^{\mathrm{i}m\varphi}, \qquad -\ell \le m \le \ell. \tag{6}$$

where P_ℓ^m denotes the associated Legendre function including the Condon–Shortley phase, extended to negative orders via $P_\ell^{-m} = (-1)^m \frac{(\ell-m)!}{(\ell+m)!} P_\ell^m$.

Classical RoPE can be viewed either as multiplication of complex channels by phase factors or as rotation of pairs of real channels. These two equivalent views motivate two distinct spherical constructions. S2RoPE-SH uses spherical harmonic values as channel-wise multipliers, whereas S2RoPE-Wigner uses their rotation representations to transform blocks of channels associated with harmonic subspaces that are closed under rotations. Figure 1 illustrates both constructions alongside classical RoPE and shows their channel allocation.

2.4 S2RoPE-SH: CHANNEL-WISE HARMONIC MODULATION

S2RoPE-SH extends the complex-multiplication view of RoPE by using spherical harmonic values in place of complex phase factors. The addition theorem in Equation (5) motivates this choice: products of harmonic values evaluated at two positions, summed over all orders of a degree, yield a function of their angular distance alone.

For a head with d real channels, we choose a harmonic degree ℓ satisfying $2(2\ell+1) \le d$. We assign one complex query channel and one complex key channel to each order $m = -\ell, \dots, \ell$, using $2(2\ell+1)$ real channels in each vector. For a token at position $\omega_i \in \mathbb{S}^2$ and with $a_m(\omega) = \sqrt{4\pi}\, Y_{\ell m}(\omega)$, we define

$$\widetilde{q}_{i,m} = a_m(\omega_i)\, q_{i,m}, \qquad \widetilde{k}_{i,m} = a_m(\omega_i)\, k_{i,m}. \tag{7}$$

Unlike the unit-modulus phase factors in Equation (1), these multipliers have position-dependent magnitudes and therefore modulate both amplitude and phase. In real coordinates, each complex multiplication scales and rotates a pair of feature channels without mixing different pairs, as illustrated in Figure 1(a).

Normalization. Because harmonic modulation changes amplitudes as well as phases, we normalize its mean squared gain across the complete degree. Setting $\gamma = \omega$ in the addition theorem (5) and using $P_\ell(1) = 1$ gives $\sum_{m=-\ell}^{\ell} |Y_{\ell m}(\omega)|^2 = (2\ell+1)/(4\pi)$ for every $\omega \in \mathbb{S}^2$. Multiplying by $\sqrt{4\pi}$ therefore makes the mean squared gain across the $2\ell+1$ channels equal to one:

$$\frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |a_m(\omega)|^2 = 1, \qquad \omega \in \mathbb{S}^2. \tag{8}$$

Since the individual gains $|a_m(\omega)|$ vary with m and ω, this normalization does not in general preserve query and key norms. Nonetheless, it ensures that the mean squared gain across the modulated channels equals one at every position.

Effect on attention. The contribution of the modulated channels to the query–key dot product is

$$A_{ij}^{\mathrm{SH}} = \mathrm{Re}\left[\sum_{m=-\ell}^{\ell} q_{i,m} k_{j,m}^{*}\, a_m(\omega_i) a_m(\omega_j)^{*} \right]. \tag{9}$$

Each summand combines a query–key content product $q_{i,m}k_{j,m}^{*}$ with a positional factor $a_m(\omega_i)a_m(\omega_j)^{*}$. By the addition theorem, the unweighted positional factors satisfy

$$\sum_{m=-\ell}^{\ell} a_m(\omega_i) a_m(\omega_j)^{*} = (2\ell+1)\, P_\ell(\omega_i^{\top}\omega_j). \tag{10}$$

Consequently, if the content products are equal across orders, $q_{i,m}k_{j,m}^{*} = c$ for all m, then $A_{ij}^{\mathrm{SH}} = (2\ell+1)\,\mathrm{Re}[c]\, P_\ell(\omega_i^{\top}\omega_j)$ depends on position only through the angular separation. For general learned features, order-dependent content weights prevent the positional dependence from reducing to angular separation alone. The positional factors are, however, invariant under joint longitude shifts of the two positions, since each depends on longitude only through $e^{\mathrm{i}m(\varphi_i - \varphi_j)}$.

Outside the modulated block, the remaining $d - 2(2\ell+1)$ real channels are left unchanged and contribute their ordinary query–key inner product. The default choice of ℓ, the real-valued implementation, and the full head operator are detailed in Appendix A.1.

2.5 S2RoPE-Wigner: rotations of complete degree spaces

We now develop a complementary construction that retains RoPE's orthogonal rotation structure. Unlike S2RoPE-SH, it preserves query and key norms and, as in classical RoPE, its positional operators combine into a single relative transformation in attention. It associates each token position with a spatial rotation through a chosen local frame, then uses representations of that rotation on complete spherical harmonic degree spaces (the spaces spanned by the spherical harmonics with fixed ℓ) to transform learned query and key blocks.

From a position to a rotation. We use the standard spherical-coordinate frame (Arfken, 2013) to associate each token position with a three-dimensional rotation. Its two tangent axes e_θ and e_φ point in the directions of increasing colatitude and longitude, while its radial axis points outward from the sphere's center through the token position.

For a position with colatitude θ and longitude φ, the rotation mapping the reference Cartesian frame to this local frame is

$$g_\omega = R_z(\varphi)R_y(\theta), \qquad g_\omega e_x = e_\theta, \quad g_\omega e_y = e_\varphi, \quad g_\omega e_z = \omega. \tag{11}$$

Here $R_z(\varphi)$ and $R_y(\theta)$ denote right-handed rotations about the z- and y-axes, respectively. The frame is aligned with Earth's rotation axis, the physically privileged direction discussed in Section 1. At the poles, where longitude is undefined, we use the longitude assigned by the input grid.

From a rotation to feature transformations. We seek matrices that represent the spatial rotation g_ω on feature blocks of different dimensions. These matrices should be orthogonal to preserve query and key norms, and respect rotation composition so that query–key interactions can be expressed through relative frame rotations.

To connect this construction to classical RoPE, consider $f(x) = a\cos(\xi x) + b\sin(\xi x)$ at a fixed frequency ξ. Translating this function by s gives $f(x-s)$, whose coefficients in the same basis are $R(\xi s)(a,b)^\top$. Thus, the rotation used in RoPE can be interpreted as the coefficient transformation induced by translating a sinusoidal function. On the sphere, spatial rotations similarly transform the coefficients of spherical harmonic expansions.

The corresponding coefficient transformations are described by the Wigner-D matrices (Edmonds, 1996), which we use in their real form. The space $\mathcal{H}_\ell$ of degree-ℓ spherical harmonics has dimension $2\ell+1$ and is invariant under rotations (Freeden et al., 1998). A spatial rotation acting on a function $f \in \mathcal{H}_\ell$ can thus be expressed by a square matrix of the same dimension. These are the Wigner-D matrices. Let $\boldsymbol{y}_\ell(u) \in \mathbb{R}^{2\ell+1}$ be the vector of spherical harmonics $y_{\ell m}$ evaluated at an arbitrary $u \in \mathbb{S}^2$. A spherical harmonics coefficient vector c then defines the function $f_c(u) = \boldsymbol{y}_\ell(u)^\top c$. We specify the matrix $D^\ell(R)$ through

$$f_c(R^{-1}u) = \boldsymbol{y}_\ell(u)^\top D^\ell(R)c, \qquad R \in \mathrm{SO}(3), \tag{12}$$

for every coefficient vector c and every $u \in \mathbb{S}^2$. Here $f_c(R^{-1}u)$ is the rotated function, which is defined by evaluating the original f at the point $R^{-1}u$ with the inverse rotation R^{-1}. The right-hand side expresses this rotated function in the original basis, with updated coefficients $D^\ell(R)c$.

Spatial rotations are unitary transformations and preserve the L^2 inner product on S^2 and respect composition. For the orthonormal spherical harmonics basis, their coefficient transformations therefore satisfy

$$D^\ell(R)^\top D^\ell(R) = I_{2\ell+1}, \qquad D^\ell(R_1R_2) = D^\ell(R_1)D^\ell(R_2) \tag{13}$$

since the Wigner-D matrices D^ℓ are representations of $\mathrm{SO}(3)$ on the spaces $\mathcal{H}_\ell$. In particular, $D^\ell(R)^\top = D^\ell(R)^{-1} = D^\ell(R^{-1})$. We next apply the $D^\ell(R)$ to learned query and key blocks.

Relative rotations for attention. For each degree ℓ, let $q_i^{(\ell)}, k_i^{(\ell)} \in \mathbb{R}^{2\ell+1}$ denote the corresponding subvectors of token i's query and key, each containing $2\ell+1$ real channels. We refer to these subvectors as degree-ℓ feature blocks. We apply the matrix associated with the token's position to both blocks:

$$\widetilde{q}_i^{(\ell)} = D^\ell(g_{\omega_i})\, q_i^{(\ell)}, \qquad \widetilde{k}_i^{(\ell)} = D^\ell(g_{\omega_i})\, k_i^{(\ell)}. \tag{14}$$

Each matrix acts jointly on the channels within its degree block and preserves the L^2 norm of the channel block, as illustrated in Figure 1(a). Orthogonality and composition then give the contribution of this block to the query–key dot product as

$$\big(\widetilde{q}_i^{(\ell)}\big)^\top \widetilde{k}_j^{(\ell)} = \big(q_i^{(\ell)}\big)^\top D^\ell(g_{\omega_i})^\top D^\ell(g_{\omega_j})\, k_j^{(\ell)} = \big(q_i^{(\ell)}\big)^\top D^\ell\big(g_{\omega_i}^{-1} g_{\omega_j}\big) k_j^{(\ell)}. \tag{15}$$

Just as the phase factors in classical RoPE combine into a factor determined by the index difference, the two Wigner matrices combine into a representation of the relative rotation between the selected local frames. This relative rotation, $g_{\omega_i}^{-1} g_{\omega_j} = R_y(-\theta_i)\, R_z(\varphi_j - \varphi_i)\, R_y(\theta_j)$, maps e_z to the position of token j expressed in token i's local frame, and thus encodes the spherical distance between the two tokens. It depends on longitude only through $\varphi_j - \varphi_i$, so attention is exactly invariant under common longitude shifts; because the frames are tied to the polar axis, it is not invariant under arbitrary rotations of the sphere.

Our default construction combines one block for each degree $\ell = 1, \ldots, L$ into a block-diagonal positional operator, leaving any remaining channels unchanged, analogous to the approach used in LLMs, e.g. Black et al. (2022). The choice of L, the channel allocation, and the full head operator are detailed in Appendix A.2.

3 Experiments

3.1 Experimental setup

Architectures. We evaluate S2RoPE in two transformer-based weather forecasting models to test whether it provides consistent performance across different spatial representations and rollout designs. In our experiments, WeatherGenerator (Lessig et al., 2026) encodes inputs from a Gaussian grid into a latent representation organized over HEALPix cells, advances the forecast in latent space, and decodes to physical fields. S2RoPE is applied in its global aggregation and latent self-attention blocks, using the HEALPix cell centre associated with each spatial token as ω. Stormer (Nguyen et al., 2024) uses a ViT backbone with weather-specific patch embeddings on a regular latitude–longitude grid and autoregressively updates the physical fields. S2RoPE acts in its self-attention blocks, with ω denoting the patch centre. Further architectural details for the two models and how S2RoPE is applied are detailed in Appendix A.

Baselines. We compare S2RoPE-SH and S2RoPE-Wigner against absolute positional encoding (No RoPE), axial 2D RoPE, and GeoRoPE. No RoPE retains each backbone's absolute positional encodings without applying rotary transformations to queries and keys. 2D RoPE is the axial construction of Equation (3), rotating half of the channel pairs by latitude and half by longitude (Heo et al., 2024). GeoRoPE, introduced in LaDCast (Zhuang & Duraisamy, 2025), uses separate rotary channel groups for time and geographic coordinates. For WeatherGenerator, we adapt its spatial encoding to HEALPix cell centres while retaining its coordinate scaling, frequency base, and channel allocation. The temporal block is left unchanged because the spatial tokens share the same forecast time.

Data. All experiments use ERA5 (Hersbach et al., 2020), ECMWF's global reanalysis, which provides hourly estimates of the atmospheric state from 1940 onward at 0.25° ($\sim$25 km) spatial resolution. We use ERA5 on two grids that differ in both geometry and resolution. *Lat–lon ERA5* is a regular equiangular grid at 1.40625° (128×256 points, 6-hourly) following the WeatherBench 2 configuration (Rasp et al., 2024). It is used by Stormer (Nguyen et al., 2024) since its image-like layout suits ViTs. However, it oversamples the poles where fields are strongly distorted. *O96 ERA5* uses ECMWF's octahedral reduced Gaussian grid (Hortal & Simmons, 1990) with 96 latitude lines per hemisphere ($\sim 1^\circ$, $\sim$100 km, 40,320 points, 6-hourly). Its near-uniform spacing avoids polar oversampling. Because of this, it is used by the WeatherGenerator.

Metrics. Following standard practice (Rasp et al., 2024), we evaluate both Stormer and WeatherGenerator using the root mean square error (RMSE) and anomaly correlation coefficient (ACC). Both scores measure gridpoint-wise accuracy. RMSE punishes errors in the mean and variance of the prediction, as well as feature locations in space and time. ACC focuses only on the latter aspect by removing the long-term mean fields and then computing the spatial pattern correlation; it is insensitive to mean biases and misspecified variances. For the WeatherGenerator, we further evaluate

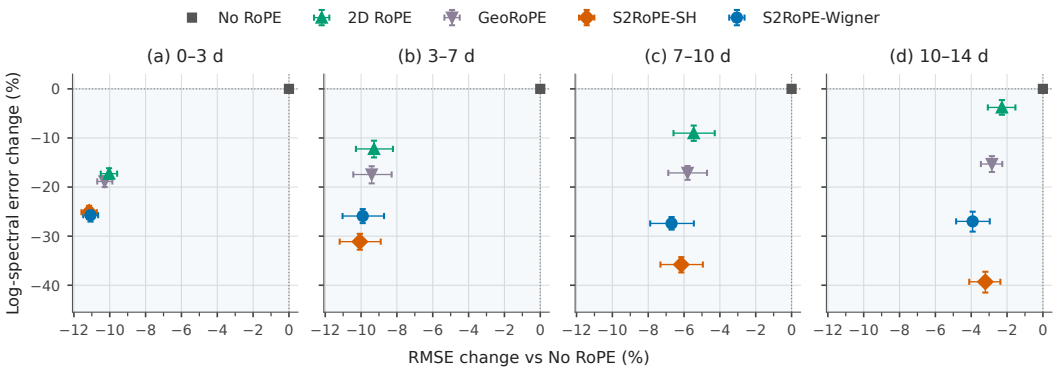


Figure 2: **Forecast accuracy and spectral fidelity in WeatherGenerator.** Four panels summarize forecast leads in $(0, 3]$, $(3, 7]$, $(7, 10]$, and $(10, 14]$ days. Both metrics report percentage changes relative to No RoPE. Lower-left positions indicate lower RMSE and log-spectral error. Whiskers show marginal 95% paired circular-block bootstrap intervals.

Table 1: WeatherGenerator forecasts from 6 h to 14 days for different positional embedding methods. ΔR denotes the percentage change in RMSE relative to No RoPE; negative values indicate lower error. Bold and underlined entries indicate the **best** and second-best results in each column.

	6 h		1 day		3 days		5 days		7 days		10 days		14 days	
Method	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑
No RoPE	+0.00	0.9738	+0.00	0.9583	+0.00	0.8997	+0.00	0.7808	+0.00	0.6100	+0.00	0.3606	+0.00	0.1779
2D RoPE	-10.00	0.9779	-9.79	0.9640	-10.43	0.9153	-9.62	0.8149	-7.34	0.6587	-4.11	0.4106	-1.00	0.2028
GeoRoPE	-10.11	0.9780	-9.92	0.9641	-10.82	0.9159	-9.70	0.8150	-7.41	0.6582	-4.59	0.4121	-1.57	0.2032
S2RoPE-SH	**-11.22**	**0.9784**	-10.83	0.9648	**-11.55**	**0.9173**	**-10.32**	**0.8176**	-8.01	**0.6618**	-4.75	0.4111	-2.14	**0.2077**
S2RoPE-Wigner	-11.11	0.9784	**-10.88**	**0.9649**	-11.30	0.9168	-10.14	0.8164	**-8.14**	0.6607	**-5.50**	**0.4139**	**-2.56**	0.2004

the spatial structure of the forecasts using spherical harmonics power spectra. Individual forecast and target spectra are log-transformed and then compared in terms of mean squared errors. The resulting score is referred to as log-spectral error throughout this manuscript.

For headline comparisons, we use nine variables (Z500, T850, Q850, U850, V850, T2M, U10, V10, and MSLP) with equal weights: at each lead time, we average the per-variable RMSE changes relative to No RoPE and the ACC values separately, while spectral changes are computed from the ratio of the variable-averaged log-spectral errors to the corresponding baseline average.

Training and evaluation. Both models are trained with a latitude-weighted error, averaged equally over forecast steps. WeatherGenerator uses MSE and is trained in two stages: 96 epochs of pre-training with a three-step rollout, followed by 32 epochs of fine-tuning with a nine-step rollout, during which only the forecasting engine is updated and the remaining modules are frozen. Stormer additionally applies pressure-level weighting to the training loss and is trained in three stages of increasing rollout length: 100 epochs single-step, 20 epochs at four steps, and 20 epochs at eight steps. Step size is fixed within a trajectory during training, but at inference Stormer combines mixed-step rollouts for an ensemble prediction. Both models are evaluated on 14-day rollouts over one year of test data, across nine variables. All variants retain the original absolute positional embeddings and share the same training and evaluation recipe within each backbone, except that WeatherGenerator's No RoPE reference retains its default latent noise in both training stages, whereas its RoPE variants use zero latent noise (Appendix C).

3.2 Forecast Performance

Tables 1 and 2 report forecast accuracy for WeatherGenerator and Stormer. In WeatherGenerator, both S2RoPE variants achieve lower RMSE than 2D RoPE and GeoRoPE at every reported lead, and an S2RoPE variant attains the highest ACC in every column. At day 3, S2RoPE-SH and S2RoPE-Wigner reduce RMSE relative to No RoPE by 11.55% and 11.30%, respectively. In Stormer, both

Table 2: Stormer forecasts from 1 to 14 days for different positional embedding methods. Bold and underlined entries indicate the **best** and second-best results in each column.

	1 day		3 days		5 days		7 days		10 days		14 days	
Method	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑	ΔR% ↓	ACC ↑
No RoPE	+0.00	0.9707	+0.00	0.9170	+0.00	0.8121	+0.00	0.6602	+0.00	0.4288	+0.00	0.1900
2D RoPE	-1.02	0.9711	-0.69	0.9178	-0.43	0.8131	-0.22	0.6606	-0.18	0.4301	-0.33	**0.1957**
S2RoPE-SH	-2.90	**0.9721**	-1.97	0.9199	-1.21	0.8167	-1.04	**0.6673**	-0.92	0.4330	+0.52	0.1627
S2RoPE-Wigner	**-2.91**	0.9720	**-2.18**	**0.9199**	**-1.78**	**0.8178**	**-1.23**	0.6671	**-1.10**	**0.4354**	**-0.46**	0.1878

Table 3: WeatherGenerator performance across overall, early, and late forecast windows for different variants of S2RoPE with No RoPE as baseline. Bold and underlined entries indicate the **best** and second-best results in each column.

	Overall 6 h–14 d			Early 6 h–3 d	Late 10–14 d
Method	Δ R% ↓	ACC ↑	ΔPSD% ↓	Δ R% ↓	Δ R% ↓
2D RoPE	-6.63	0.6246	-8.62	-10.04	-2.29
S2RoPE-SH	-7.50	**0.6271**	**-34.92**	**-11.15**	-3.21
Repeated D^1	-7.17	0.6252	-14.05	-10.17	-3.37
Repeated $D^1 \oplus D^2$	-6.93	0.6240	-20.26	-10.22	-3.11
Projected S2RoPE-Wigner	-7.12	0.6266	-4.90	-9.57	-3.50
S2RoPE-Wigner	**-7.76**	0.6269	-26.69	-11.07	**-3.92**

variants outperform 2D RoPE in RMSE and ACC through day 10, with S2RoPE-Wigner achieving the lowest RMSE at all reported leads. Results at day 14 are mixed: S2RoPE-Wigner retains a 0.46% RMSE reduction relative to No RoPE, but 2D RoPE has the highest ACC, and SH performs worse than No RoPE on both metrics. Figure 2 complements these accuracy results with spectral evaluation in WeatherGenerator. Both S2RoPE variants have lower log-spectral error than 2D RoPE and GeoRoPE across all four forecast windows, indicating better fidelity to the target power spectra. Beyond day 3, SH achieves lower log-spectral error than Wigner, whereas Wigner achieves lower RMSE. Smoother forecasts with less high-frequency detail and therefore larger log-spectral error suffer less from RMSE's double penalty and thus achieve in general better results in this metric. S2RoPE-SH and S2RoPE-Wigner thus allow one to trade-off forecast sharpness with accuracy.

3.3 Ablation study

We examine the effects of harmonic degree composition and zonal projection in S2RoPE-Wigner within WeatherGenerator. Table 3 compares the default configuration using degrees 1–10 with three variants: repeated degree-1 blocks, repeated degree-1/2 blocks, and a projected version of the default configuration. The degree-composition variants retain the same allocation of 120 transformed real channels; the projected variant retains only the $m = 0$ input component of each degree block before rotation (Appendix B.3). Percentage changes are relative to the No-RoPE reference. The default configuration achieves lower overall, early-window, and late-window RMSE, as well as lower log-spectral error, than both repeated low-degree variants. Its overall RMSE reduction is 7.76%, compared with 7.17% and 6.93%, respectively. The projected variant has higher overall RMSE and log-spectral error than the default configuration, although projection also reduces the effective content rank. Across all listed methods, S2RoPE-SH achieves the highest overall ACC and the lowest log-spectral error and early-window RMSE, while S2RoPE-Wigner achieves the lowest overall and late-window RMSE.

3.4 Physical consistency

Figure 3 shows a diagnostic of the physical consistency between variables and spatial locations in our WeatherGenerator experiments: in the extra-tropical atmosphere, part of the horizontal wind field $\mathbf{v}$ can be approximated by the geostrophic wind $\mathbf{v}_g$ which depends only on the spatial derivatives of the geopotential field ϕ. In a physically consistent model, the ratio between ageostrophic wind speed $|\mathbf{v} - \mathbf{v}_g|$ and geostrophic wind speed $|\mathbf{v}_g|$ should remain in agreement with the target throughout the rollout (Bonavita, 2024). To monitor even slowly accruing deviations from the target ratio, we have

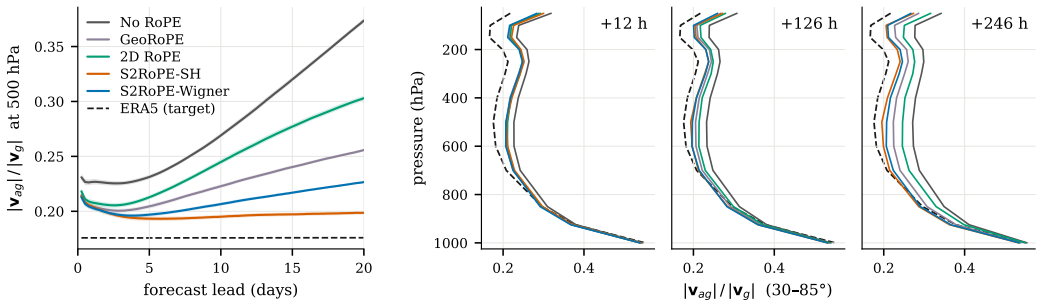


Figure 3: **Physical consistency measured by the ratio between ageostrophic and geostrophic wind.** Left panel shows the ratio between average ageostrophic and geostrophic wind speed in the 500 hPa level for ERA5 and various WeatherGenerator simulations as a function of forecast lead time. Right panels show the same ratio across all pressure levels for three selected lead times.

extended the rollout to 20 days for Figure 3. Without RoPE, we observe a continuous drift away from the target ratio, indicating loss of physical consistency between the wind and geopotential fields. Both 2D RoPE and GeoRoPE improve upon the baseline but still contain a substantial drift. Both versions of our S2RoPE remain significantly more stable, with S2RoPE-SH in particular showing near zero evidence of drift throughout the entire 20 day roll-out.

4 Related work

Rotary encodings beyond sequences. RoPE (Su et al., 2024) has been extended to image coordinates (Heo et al., 2024; van de Geijn et al., 2025) and learned rotations for Euclidean coordinates (Ostmeier et al., 2024; Schenck et al., 2025), with broader formulations based on group representations (Lim et al., 2023) and group actions (Zhang et al., 2026). For spherical and geographic settings, SpheRoPE (Hirschorn et al., 2026) adapts rotary frequencies and coordinates for panorama generation, while StereoRoPE in STRATA (Hu et al., 2026) uses local stereographic coordinates. SoPE (Yip et al., 2026) uses spherical coordinates and coordinate-specific frequencies for 3D point clouds. In weather forecasting, GeoRoPE in LaDCast (Zhuang & Duraisamy, 2025) assigns separate rotary channel groups to time, latitude, and longitude. S2RoPE instead constructs positional operators from spherical harmonics and their rotation representations.

Harmonic and representation-based attention. AllScAIP's Legendre Angular Encoding (Qu et al., 2026) modulates repeated query and key channels with real spherical harmonics of edge directions, using the addition theorem. S2RoPE-SH instead evaluates complex harmonics at token positions and assigns independent content channels to individual orders. GTA (Miyato et al., 2024) transforms queries, keys, values, and outputs using representations of camera geometry, including Wigner-D blocks. Its formulation supports general degrees, with degrees one and two used in experiments. S2RoPE-Wigner instead obtains rotations from the standard spherical-coordinate frame at each token position and transforms only queries and keys. Additional discussion of spherical location encoders and architectures is provided in Appendix D.

5 Conclusion and Future Work

We present S2RoPE to adapt rotary positional encoding to the spherical geometry of global weather forecasting. Its two constructions modulate query and key features with spherical harmonic values or rotate feature blocks with Wigner-D matrices. Experiments show that our S2RoPE improves forecast accuracy in two architectures, with additional improvements in spectral fidelity and evidence of improved physical consistency in WeatherGenerator. Future work could explore degree allocations tailored to different layers and forecasting tasks, and evaluate whether these benefits extend to higher-resolution forecasts and additional model architectures.

AI USE STATEMENT

We used generative AI tools only for language polishing and to assist in writing code for line plots of precomputed evaluation metrics. The plotting code was verified by at least two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

G.B. Arfken. *Mathematical Methods for Physicists*. Academic Press, 2013. ISBN 9781483277820. URL https://books.google.de/books?id=X9U3BQAAQBAJ.

Yaniv Benny and Lior Wolf. Sphereuformer: A u-shaped transformer for spherical 360 perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 940–950. IEEE, 2025.

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

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022.

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

Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024.

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

William Bourke. An efficient, one-level, primitive-equation spectral model. *Monthly Weather Review*, 100(9):683 – 689, 1972. doi: 10.1175/1520-0493(1972)100<0683:AEOPSM>2.3.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/100/9/1520-0493_1972_100_0683_aeopsm_2_3_co_2.xml.

Oscar Carlsson, Jan E Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Heal-swin: A vision transformer on the sphere. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6067–6077. IEEE, 2024.

Dale R. Durran. *Numerical Methods for Fluid Dynamics: With Applications to Geophysics*, volume 32 of *Texts in Applied Mathematics*. Springer, New York, 2 edition, 2010. doi: 10.1007/978-1-4419-6412-0.

A. R. Edmonds. *Angular Momentum in Quantum Mechanics*. Princeton University Press, Princeton, NJ, 2 edition, 1996. Fourth printing, with corrections.

Erik Eliasen, Bennert Machenhauer, and Erik Rasmussen. On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields. Technical report, 01 1970.

W Freeden, T Gervens, and M Schreiner. *Constructive Approximation on the Sphere: With Applications to Geomathematics*. Oxford University Press, 01 1998. ISBN 9780198536826. doi: 10.1093/oso/9780198536826.001.0001. URL https://doi.org/10.1093/oso/9780198536826.001.0001.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision (ECCV)*, pp. 289–305. Springer, 2024.

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 (QJRMS)*, 146(730):1999–2049, 2020.

Or Hirschorn, Aaron Olender, Eli Alshan, Ianir Ideses, Lior Fritz, and Sagie Benaim. Spherope: Zero-shot optimization-free 360 panorama generation with spherical rope. *arXiv preprint arXiv:2606.32033*, 2026.

Mariano Hortal and Adrian J. Simmons. Use of reduced Gaussian grids in spectral models. ECMWF Technical Memorandum 168, European Centre for Medium-Range Weather Forecasts, Reading, UK, June 1990. URL https://www.ecmwf.int/en/elibrary/74898-use-reduced-gaussian-grids-spectral-models.

Zeyuan Hu, Akshay Subramaniam, Noel Keen, Tao Ge, Jaideep Pathak, Mohammad Shoaib Abbas, Suman Ravuri, Karthik Kashinath, Naser Mahfouz, Peter Caldwell, et al. Scaling storm-resolving atmospheric ai simulation to the entire planet. *arXiv preprint arXiv:2606.31248*, 2026.

Lessig et al. *WeatherGenerator: The Machine Learning Earth System Model*, 2026. URL https://github.com/ecmwf/WeatherGenerator. Accessed: 2026-07-20.

Guowen Li, Xintong Liu, Yang Liu, Mengxuan Chen, Shilei Cao, Xuehe Wang, Juepeng Zheng, Jinxiao Zhang, Haoyuan Liang, Lixian Zhang, et al. Tianquan-s2s: A subseasonal-to-seasonal global weather model via incorporate climatology state. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 106063–106088, 2026.

Derek Lim, Hannah Lawrence, Ningyuan Teresa Huang, and Erik Henning Thiede. Positional encodings as group representations: A unified framework. In *ICML 2023 Workshop on Topology, Algebra, and Geometry in Machine Learning*, 2023.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 8172–8208, 2024.

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

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 68740–68771, 2024.

Steven A. Orszag. Transform method for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. *Journal of Atmospheric Sciences*, 27(6):890 – 895, 1970. doi: 10.1175/1520-0469(1970)027<0890:TMFTCO>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/27/6/1520-0469_1970_027_0890_tmftco_2_0_co_2.xml.

Sophie Ostmeier, Brian Axelrod, Maya Varma, Michael E Moseley, Akshay Chaudhari, and Curtis Langlotz. Liere: Lie rotational positional encodings. *arXiv preprint arXiv:2406.10322*, 2024.

Eric Qu, Brandon M Wood, Aditi S Krishnapriyan, and Zachary W Ulissi. A recipe for scalable attention-based mlips: unlocking long-range accuracy with all-to-all node attention. *arXiv preprint arXiv:2603.06567*, 2026.

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

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 1746–1759, 2024.

Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Avinava Dubey, Ayzaan Wahid, et al. Learning the ropes: Better 2d and 3d position encodings with string. *arXiv preprint arXiv:2502.02562*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Chase van de Geijn, Timo Lüddecke, Polina Turishcheva, and Alexander Ecker. A circular argument: Does rope need to be equivariant for vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 20388–20424, 2025.

Koonting Yip, Qiyan Zhao, Wenhao Yu, Liangyu Yuen, Mingkai Li, Xiaofeng Zhang, Jianmin Ji, Yanyong Zhang, Qing Jiang, and Ka-Veng Yuen. Sope: Spherical coordinate-based positional embedding for enhancing spatial perception of 3d lvlms. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33714–33726, 2026.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Qin Zhen, Rina Hughes, Kangping Xu, Yang Yuan, Quanquan Gu, and Andrew Yao. Group representational position encoding. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 56496–56519, 2026.

Yilin Zhuang and Karthik Duraisamy. Ladcast: A latent diffusion model for medium-range ensemble weather forecasting. *arXiv preprint arXiv:2506.09193*, 2025.

A IMPLEMENTATION DETAILS

A.1 S2ROPE-SH IMPLEMENTATION

For head dimension d, we choose a harmonic degree ℓ satisfying $r_{\mathrm{SH}} = 2(2\ell + 1) \leq d$. The modulated channels are grouped into $2\ell + 1$ real channel pairs, each representing one complex channel indexed by $m = -\ell, \ldots, \ell$. Multiplication by a complex scalar a is implemented as

$$M(a) = \begin{bmatrix} \operatorname{Re} a & -\operatorname{Im} a \\ \operatorname{Im} a & \operatorname{Re} a \end{bmatrix}. \tag{16}$$

Using the multipliers defined in Equation (7), the full head operator is

$$T_{\mathrm{SH}}(\omega) = \left[\bigoplus_{m=-\ell}^{\ell} M\big(a_m(\omega)\big) \right] \oplus I_{d-r_{\mathrm{SH}}}. \tag{17}$$

Each block acts independently on one channel pair, while the identity block leaves the remaining channels unchanged. At each token, this operator is applied to both queries and keys; values are unchanged.

Our default configuration uses the largest complete degree that fits within the head dimension. For WeatherGenerator, $d = 128$ gives $\ell = 31$, with 126 modulated real channels and two unchanged channels. For Stormer, $d = 64$ gives $\ell = 15$, with 62 modulated real channels and two unchanged channels.

A.2 S2ROPE-WIGNER IMPLEMENTATION

For an attention head with d real channels, our default construction uses one feature block for each degree $\ell = 1, \ldots, L$. We choose the largest integer L satisfying $L(L+2) \leq d$. Each degree-ℓ block contains $2\ell + 1$ real channels. The full positional operator is

$$\begin{aligned} T_{\mathrm{Wigner}}(\omega) &= \left[\bigoplus_{\ell=1}^{L} D^{\ell}(g_\omega) \right] \oplus I_{d-r_{\mathrm{Wigner}}}, \\ r_{\mathrm{Wigner}} &= \sum_{\ell=1}^{L} (2\ell + 1) = L(L+2). \end{aligned} \tag{18}$$

Here r_{Wigner} is the number of transformed real channels, and the identity block leaves the remaining $d - r_{\mathrm{Wigner}}$ channels unchanged. All degree blocks represent the same spatial rotation g_ω and act on separate channel groups. At each token, the operator is applied to both queries and keys; values are unchanged.

For WeatherGenerator, $d = 128$ gives $L = 10$, with $3 + 5 + \cdots + 21 = 120$ transformed real channels and eight unchanged channels, as illustrated in Figure 1(b). For Stormer, $d = 64$ gives $L = 7$, with $3 + 5 + \cdots + 15 = 63$ transformed real channels and one unchanged channel.

A.3 BACKBONE ARCHITECTURES AND TRAINING

The two models we evaluate are representative of the two dominant families of transformer-based AIWP architectures, which is why we do not evaluate further models separately. WeatherGenerator belongs to the Perceiver-based family, together with Aurora(Bodnar et al., 2025): both employ Perceiver modules that compress the state into a resolution-independent latent set and perform the bulk of their computation there. WeatherGenerator is the more demanding of the two cases, since its rollout is performed entirely in latent space. Stormer belongs to the ViT-based family, together with Pangu-Weather (Bi et al., 2023): both apply attention over patch tokens that inherit the geometry of the underlying grid, differing mainly in how the atmospheric state is embedded beforehand. Covering one model from each family lets us test whether our variants of RoPE transfer across both regimes. We use WeatherGenerator and Stormer in their original form.

WeatherGenerator is a multi-modular, Perceiver-based architecture organized around *streams*, allowing multiple data sources to be embedded and ingested jointly; we use ERA5 alone. The stream

embedder learns per-grid-point interactions across variables, and the local-to-global assimilator aggregates grid points into HEALPix level-5 cells (12,288 cells), restricting attention to points within a cell to yield one latent token per cell; both modules use sinusoidal positional embeddings. The global engine then attends over neighbouring cell tokens to model inter-cellular interactions, and the forecasting engine autoregressively propagates the latent state via self-attention. The decoder maps latents back to physical space through cross-attention, with each grid point querying its HEALPix cell's latent state. Table 4 lists the hyperparameters used to train WG models.

Stormer is built on a simple ViT backbone with minimal weather-specific modifications. First, each variable is patchified independently, mapping a field of size $H \times W$ to $(H/P) \times (W/P)$ tokens of dimension D ($H \times W$ the grid size, P the patch size), with sinusoidal positional embeddings added to the tokens; stacking the V variables yields $(H/P)(W/P)V$ tokens in total, so sequence length grows linearly with the number of variables. To keep this tractable, a cross-attention layer aggregates across the variable dimension, capturing inter-variable relationships while collapsing the V tokens at each spatial location into one. Second, Stormer conditions on lead time: it takes the forecast interval δt as an additional input and predicts the increment $X_{t+\delta t} - X_t$ rather than the state itself, with each transformer block using δt to scale and shift its normalization parameters via adaptive layer normalization in place of standard LayerNorm. Table 5 lists the hyperparameters used to train Stormer models.

Table 4: Model configuration of the WG model.

Model Component	Configuration	Value
Embedding	Embed_unembed_mode	block
	HEALPix level	5
	Dropout rate	0.1
Global Engine	Attention blocks	4
	Attention heads	16
	Embedding dimension	2048
	Dropout rate	0.1
Decoder	Attention blocks	2
	Attention heads	4
	Embedding dimension	256
	Dropout rate	0.1
Forecasting Engine	Attention blocks	16
	Attention heads	16
	Dropout rate	0.1

Table 5: Model configuration of the Stormer model.

Model Component	Configuration	Value
Weather Embedding	Patch size	4
	Embedding dimension	1024
	Attention heads	16
Time Embedder	Linear embedding dimension	1024
ViT	Attention blocks	24
	Attention heads	16
	Embedding dimension	1024

A.4 Integration into the backbones

Token geometry. We evaluate positional operators at HEALPix cell centers in WeatherGenerator and latitude–longitude patch centers in Stormer. Tokens sharing a center share its operator; nonspatial register tokens use the identity transformation.

Channel layouts. Our axial 2D RoPE baseline pairs real channel t with channel $t + d/2$, for $0 \le t < d/2$. Latitude acts on pairs with $t < d/4$, and longitude acts on pairs with $d/4 \le t < d/2$, giving Equation (3) after a fixed channel permutation. S2RoPE-SH instead uses adjacent real/imaginary channel pairs. S2RoPE-Wigner groups channels into complete degree blocks. The

degree choices and transformed-channel counts for both backbones are specified in Appendices A.1 and A.2.

Operator placement. In WeatherGenerator, the positional operators are applied in the global query-aggregation blocks and the self-attention blocks of the global assimilation and forecasting engines. Local assimilation and decoder attention retain their original positional treatment. In Stormer, the positional operators are applied in every backbone Transformer block.

The operators act on projected queries and keys after any enabled Q/K normalization and before the backbone's fused attention operation. S2RoPE-SH uses no additional post-modulation normalization. Values are unchanged.

A.5 Harmonic conventions and numerical construction

Harmonic conventions. S2RoPE-SH uses the complex spherical harmonics defined in Section 2, with multipliers $a_m(\omega) = \sqrt{4\pi}Y_{\ell m}(\omega)$ for the selected degree ℓ. For S2RoPE-Wigner, the real orthonormal basis is ordered by $m = -\ell, \dots, \ell$ and defined by

$$y_{\ell 0} = Y_{\ell 0}, \qquad y_{\ell m} = \sqrt{2}\,\mathrm{Re}\,Y_{\ell m}, \qquad y_{\ell,-m} = \sqrt{2}\,\mathrm{Im}\,Y_{\ell m}, \quad m > 0. \tag{19}$$

With the coefficient-transformation convention in Equation (12), $D^\ell(R_z(\varphi))$ acts by $R(m\varphi)$ on each coefficient pair ordered as $(+m, -m)$ and leaves the $m = 0$ component unchanged.

Numerical construction. For S2RoPE-Wigner, we precompute

$$D^\ell(g_\omega) = D^\ell(R_z(\varphi))D^\ell(R_y(\theta)). \tag{20}$$

The first factor is constructed analytically from the planar rotations described above. For each degree and each distinct colatitude θ, we obtain the second factor by fitting the basis-value relation implied by Equation (12):

$$\boldsymbol{y}_\ell(R_y(\theta)^{-1}u)^\top = \boldsymbol{y}_\ell(u)^\top D^\ell(R_y(\theta)). \tag{21}$$

The least-squares fit uses 768 fixed sample directions: HEALPix centers with $N_{\mathrm{side}} = 8$ for WeatherGenerator and Fibonacci-sphere points for Stormer. If the fitted matrix has singular value decomposition $U\Sigma V^\top$, we replace it by $UV^\top$ to enforce orthogonality up to numerical precision.

The matrices are constructed in float64. The resulting geometry tensors are cached, shared across attention heads, and used at model precision. Consequently, norm preservation holds up to numerical precision in execution.

B Experimental details

B.1 Spectral evaluation

Spherical harmonic power spectra. Here, spherical harmonics decompose forecast and target physical fields for spectral evaluation, whereas S2RoPE-SH uses their values at token positions to modulate query and key features within attention. We compute spherical harmonic transforms (SHTs) of the global WeatherGenerator forecasts and their corresponding targets on the O96 grid. Let i, t, and v index initialization, forecast lead, and variable, respectively, and let $c_{itv,\ell m}$ denote the complex spherical harmonic coefficient of degree ℓ and order m. The power at each degree is

$$P_{itv,\ell} = |c_{itv,\ell 0}|^2 + 2\sum_{m=1}^{\ell} |c_{itv,\ell m}|^2. \tag{22}$$

The factor of two accounts for the equal power in the negative-order coefficients of a real-valued field, since the implementation stores only nonnegative orders. We retain degrees $\ell = 0, \dots, 95$, including the degree-zero component, and apply the same transform to forecasts and targets.

Log-spectral error. For method a, we compare its forecast power $P^{(a)}_{itv,\ell}$ with the target power $P^{\star}_{itv,\ell}$:

$$E^{(a)}_{itv} = \frac{1}{|\mathcal{L}^{(a)}_{itv}|} \sum_{\ell \in \mathcal{L}^{(a)}_{itv}} \left[\log_{10} P^{(a)}_{itv,\ell} - \log_{10} P^{\star}_{itv,\ell}\right]^2, \tag{23}$$

where $\mathcal{L}^{(a)}_{itv}$ contains degrees with finite, strictly positive forecast and target power. Valid degrees receive equal weight. The error is computed separately for each initialization before averaging; spectra are not averaged across initializations before computing the score. This metric measures discrepancies in power across spatial scales and complements RMSE and ACC because it does not retain spatial phase information.

Aggregation and relative changes. For a set of forecast leads $\mathcal{T}$, we average the errors equally over N initializations, $V = 9$ variables, and the selected leads:

$$\overline{E}^{(a)}_{\mathcal{T}} = \frac{1}{NV|\mathcal{T}|} \sum_{i=1}^{N} \sum_{v=1}^{V} \sum_{t \in \mathcal{T}} E^{(a)}_{itv}. \tag{24}$$

A singleton $\mathcal{T}$ gives the mean log-spectral error at one lead time. Relative spectral error changes against reference method b are reported as

$$\Delta^{(a|b)}_{\mathrm{PSD}}(\mathcal{T}) = 100 \left(\frac{\overline{E}^{(a)}_{\mathcal{T}}}{\overline{E}^{(b)}_{\mathcal{T}}} - 1 \right). \tag{25}$$

The reference is No RoPE unless otherwise specified. Negative values indicate lower spectral error. Thus, Δ_{PSD} denotes a relative change in mean log-spectral error, rather than in power itself. We compute a ratio of aggregated errors, rather than averaging per-variable or per-lead percentage changes.

B.2 Geostrophic balance as a diagnostic of physical consistency

On large spatio-temporal scales, the flow in the extra-tropical free troposphere is in approximate balance between the Coriolis force and the pressure gradient force. Using pressure as the vertical coordinate, this balance reads

$$f \mathbf{e}_z \times \mathbf{v}_g = -\nabla_p \Phi \,, \tag{26}$$

where f is the latitude dependent Coriolis parameter, $\mathbf{e}_z$ denotes the vertical unit vector and $\nabla_p \Phi$ is the horizontal gradient of geopotential on a pressure level. The equilibrium wind that fulfills this equation is referred to as the geostrophic wind

$$\mathbf{v}_g = -\frac{1}{f} \frac{\partial \Phi}{\partial y} \cdot \mathbf{e}_x + \frac{1}{f} \frac{\partial \Phi}{\partial x} \cdot \mathbf{e}_y \,, \tag{27}$$

whereas the residual $\mathbf{v}_{ag} = \mathbf{v} - \mathbf{v}_g$ is called the ageostrophic wind.

A data driven weather model does not explicitly compute forces or accelerations and may therefore introduce independent errors as well as artifacts with unrealistic spatial structure into the geopotential and wind field. Bonavita (2024) suggested to monitor such behavior in terms of the ratio

$$\frac{|\mathbf{v}_{ag}|}{|\mathbf{v}_g|} \,, \tag{28}$$

where $|\cdot|$ denotes the absolute values, i.e., the wind speeds. For our experiment, we compute the average values of $|\mathbf{v}_{ag}|$ and $|\mathbf{v}_g|$ over the whole evaluation period and then take their ratio. Any natural variability in the wind components averages out and the result should ideally remain stable and close to the target ratio (calculated analogously from ERA5).

B.3 Projected Wigner control

In addition to the repeated-degree controls, we consider a control that projects each input feature block onto its $m = 0$ component before applying the Wigner rotation. In the real spherical harmonic

basis used above, let $e_0 \in \mathbb{R}^{2\ell+1}$ be the standard unit vector with components $(e_0)_m = \delta_{m0}$, $m = -\ell, \ldots, \ell$, where δ is the Kronecker delta. We use $y_{\ell 0} = Y_{\ell 0}$. The projected operator is

$$B^\ell(\omega) = D^\ell(g_\omega) e_0 e_0^\top. \tag{29}$$

The implementation keeps the $m' = 0$ column of $D^\ell(g_\omega)$ and sets all other columns to zero before applying the matrix to queries and keys. Equivalently, it retains the $m = 0$ input component and maps it along $D^\ell(g_\omega)e_0$.

Under our harmonic and rotation conventions,

$$D^\ell(g_\omega) e_0 = \sqrt{\frac{4\pi}{2\ell+1}}\, \boldsymbol{y}_\ell(\omega). \tag{30}$$

The addition theorem therefore gives

$$\begin{aligned}\left(B^\ell(\omega_i) q_i^{(\ell)}\right)^\top \left(B^\ell(\omega_j) k_j^{(\ell)}\right) = (e_0^\top q_i^{(\ell)})(e_0^\top k_j^{(\ell)}) \\ \times P_\ell(\omega_i^\top \omega_j).\end{aligned} \tag{31}$$

At each position, $B^\ell(\omega)$ has rank one. This concerns the feature transformation, not the rank of the Legendre kernel matrix across positions. The control retains the block width but reduces the effective content dimension and does not preserve arbitrary feature norms. Its performance difference from the full construction therefore cannot be attributed solely to directional dependence.

This control also differs from S2RoPE-SH: it retains one input component per degree block, whereas S2RoPE-SH multiplies independent complex channels by harmonic values.

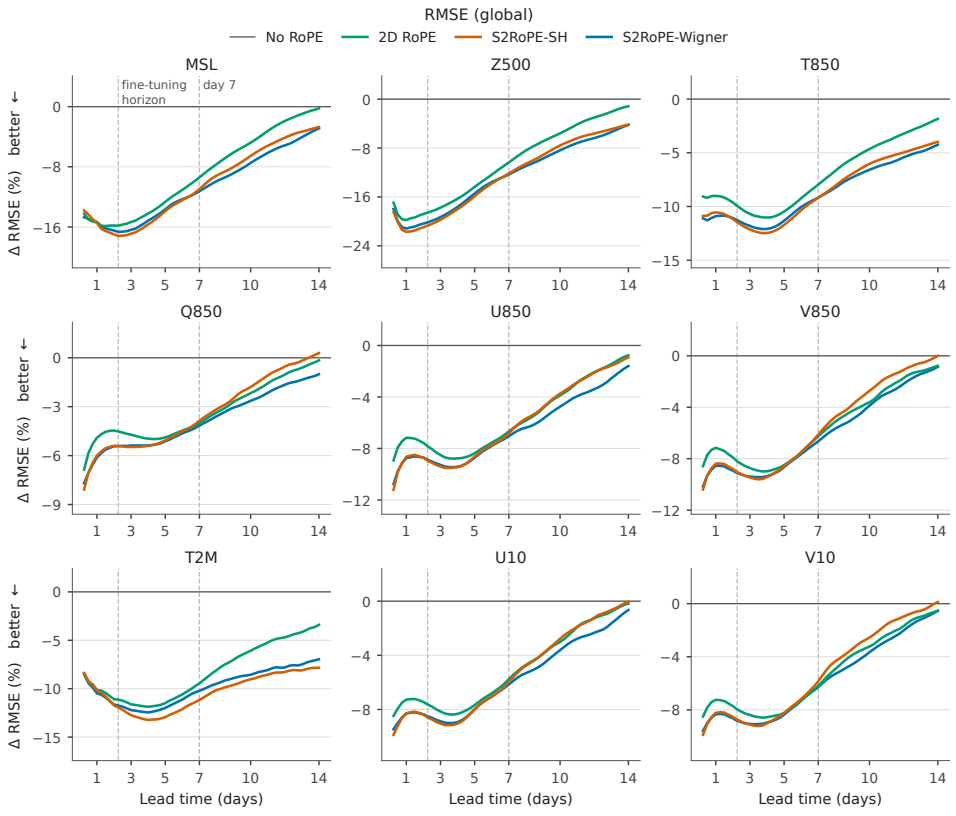


Figure 4: WeatherGenerator RMSE changes across nine variables and forecast lead times. Each panel shows the percentage change in RMSE relative to No RoPE for one variable; negative values indicate lower error. Dashed vertical lines mark the 9-step (54-hour) fine-tuning horizon and day 7.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 FORECAST PERFORMANCE BY VARIABLE

Figure 4 shows how RMSE changes vary across variables and forecast lead times in WeatherGenerator. Both S2RoPE variants generally achieve lower RMSE than 2D RoPE, with clearer gains for mean sea-level pressure (MSL), 500-hPa geopotential (Z500), and temperature at 850 hPa (T850) and 2 m (T2M). Improvements relative to No RoPE generally diminish at longer lead times, particularly for specific humidity at 850 hPa (Q850) and wind components at 850 hPa (U850, V850) and 10 m (U10, V10). Neither S2RoPE variant is uniformly better: S2RoPE-SH achieves lower RMSE for 2-m temperature (T2M), whereas S2RoPE-Wigner generally retains larger improvements for humidity and winds toward the end of the forecast.

C.2 EFFECT OF GLOBAL POSITIONAL EMBEDDINGS

We examine how retaining or removing WeatherGenerator's original global positional embeddings affects each RoPE method (Table 6 and Figure 5). All percentage changes use the same default No RoPE reference. S2RoPE-SH changes little across the two settings, retaining similar RMSE, ACC, and log-spectral error. For S2RoPE-Wigner, removing global PE increases RMSE across the overall, early, and late windows and lowers ACC, but reduces log-spectral error. For 2D RoPE, removal reduces overall and late-window RMSE and log-spectral error, while increasing early-window RMSE. These results show that the effect of global positional embeddings depends on both the rotary construction and the evaluation metric, rather than providing a uniform benefit.

C.3 EFFECT OF LATENT TRAINING NOISE

Feature-space noise has been used in weather forecasting to mitigate over-smoothing and retain forecast variability (Li et al., 2026). We examine whether WeatherGenerator exhibits a similar dependence on latent training noise when equipped with RoPE. Figure 6 compares three No RoPE settings: noise of 10^{-4} in both training stages, noise only during pre-training, and no noise in either stage. All RoPE variants use zero latent noise. Global positional embeddings are retained throughout.

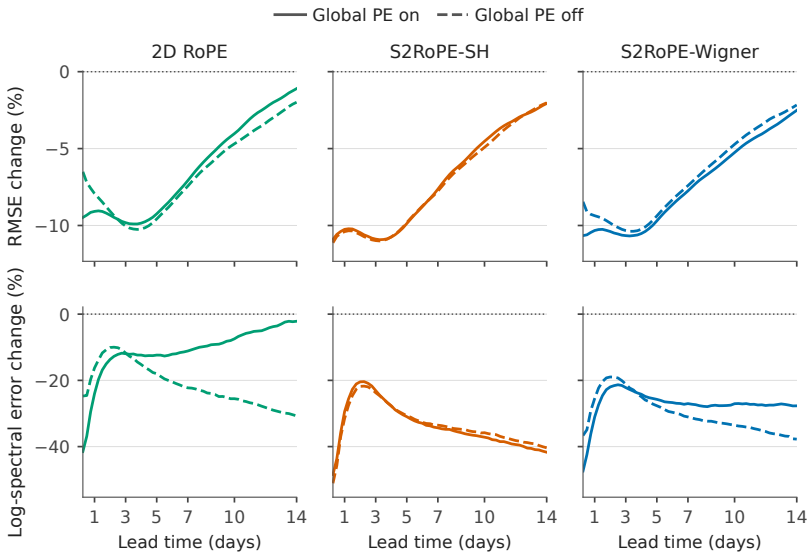


Figure 5: Effect of global positional embeddings (PE) on WeatherGenerator forecasts for each RoPE method. Solid and dashed lines denote global PE on and off, respectively. Both RMSE and log-spectral error are reported as percentage changes relative to the default No RoPE reference, represented by the zero line. Negative values indicate lower error.

Table 6: Global positional embeddings (PE) ablation in WeatherGenerator across overall, early, and late forecast windows. Bold and underlined entries indicate the **best** and second-best results in each column.

	Overall 6 h–14 d			Early 6 h–3 d	Late 10–14 d
Method	Δ R% ↓	ACC ↑	ΔPSD% ↓	Δ R% ↓	Δ R% ↓
No RoPE (default)	+0.00	0.5917	+0.00	+0.00	+0.00
2D RoPE / PE on	-6.63	0.6246	-8.62	-10.04	-2.29
2D RoPE / PE off	-6.99	0.6253	-23.21	-9.22	-3.32
S2RoPE-SH / PE on	-7.50	**0.6271**	**-34.92**	-11.15	-3.21
S2RoPE-SH / PE off	-7.61	0.6267	-34.23	**-11.29**	-3.32
S2RoPE-Wigner / PE on	**-7.76**	0.6269	-26.69	-11.07	**-3.92**
S2RoPE-Wigner / PE off	-7.22	0.6252	-31.56	-10.26	-3.35

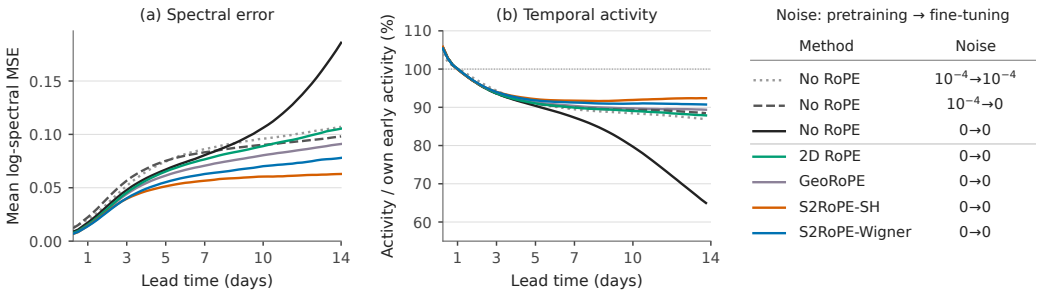


Figure 6: **Effect of latent training noise in WeatherGenerator.** Legend entries specify noise levels during pre-training and fine-tuning. (a) Mean log-spectral MSE. (b) Normalized temporal activity derived from FROCT, relative to each model's own early-lead activity.

Diagnostics. We assess both log-spectral error and temporal activity. The former measures agreement between forecast and target power spectra. We derive temporal activity from the forecast rate of change over time (FROCT), measured here as the mean absolute forecast change between consecutive six-hour leads. We first average FROCT over initializations, then normalize each model–variable curve by its own mean over the first eight transitions (6–12 h through 48–54 h). The normalized curves are averaged equally across variables and multiplied by 100. This measures the retention of temporal variation relative to each model's own early forecasts.

Results. Without latent noise, No RoPE exhibits a pronounced decline in temporal activity and a sharp increase in spectral error at longer lead times. Noise during pre-training substantially reduces these effects, even when omitted during fine-tuning. Under the same zero-noise setting, all RoPE variants show a smaller decline in activity than No RoPE. S2RoPE-SH and S2RoPE-Wigner additionally achieve lower late-lead spectral error than all three No RoPE configurations. Thus, in these experiments, S2RoPE retains more of its early-lead temporal activity while achieving better spectral fidelity, without the latent noise used by the default reference. We retain that default No RoPE configuration in the main comparison rather than using the zero-noise configuration that deteriorates at longer leads.

D Additional related work

Spherical representations and architectures. Spherical-harmonic location encoders (Rußwurm et al., 2024) and Sphere2Vec (Mai et al., 2023) map individual geographic locations to feature vectors. HEAL-SWIN (Carlsson et al., 2024), spherical Fourier neural operators (Bonev et al., 2023), and SphereUFormer (Benny & Wolf, 2025), which combines spherical local attention with relative position biases for panoramic perception, incorporate spherical structure into the architecture itself. Our focus is complementary: spherical geometry enters through position-dependent transformations of queries and keys, rather than through a redesign of the forecasting backbone.