# SYMMETRIC COMPOSITION OF ANISOTROPIC OPERATORS FOR GLOBAL SUBSEASONAL-TO-SEASONAL CLIMATE FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Accurate global Subseasonal-to-Seasonal (S2S) climate forecasting is critical for disaster preparedness and resource management, yet it remains challenging due to chaotic atmospheric dynamics. Despite advances in geometry-aware representations, existing methods do not specify how the zonal and meridional interactions that govern real atmospheric dynamics should be explicitly coordinated and composed, leaving an important architectural gap for S2S forecasting. In this paper, we propose **AnisoCast**, which explicitly models anisotropic global atmospheric dynamics for accurate S2S climate forecasting. It couples: (1) an Anisotropic Embedding strategy that tokenizes the global grid into latitudinal rings, preserving the integrity of zonal periodic structures; and (2) stacked Aniso Blocks that arrange latent Zonal and Meridional Operators in a weight-shared palindromic composition inspired by symmetric operator splitting. Extensive experiments on the ERA5 reanalysis dataset demonstrate that AnisoCast establishes a new state-of-the-art, significantly outperforming existing methods in both forecasting accuracy and computational efficiency. Code is available at https://anonymous.4open.science/r/AnisoCast.

## 1 INTRODUCTION

Global subseasonal-to-seasonal (S2S) climate forecasting aims to predict atmospheric conditions across the Earth, from several weeks (subseasonal) to months (seasonal) ahead. This is vital for decision-making in various sectors such as energy scheduling, agriculture, and disaster management (Vitart et al., 2017; White et al., 2017; Liu et al., 2025a; Wu et al., 2026). However, while data-driven models have advanced medium-range forecasting up to approximately two weeks (Bi et al., 2023; Lam et al., 2023; Price et al., 2025), the performance often degrades as the lead time extends into the subseasonal range (Nathaniel et al., 2024; Lyu et al., 2026). Forecasting at the S2S scale is even more challenging as much of the predictive information carried by the initial atmospheric state has already faded, whereas slower Earth system processes (such as ocean variability) generally require longer timescales to exert a strong influence on atmospheric evolution (Mariotti et al., 2018; Vitart & Robertson, 2018; Liu et al., 2025a; Gao et al., 2026).

To improve S2S forecasting performance, recent attempts introduce inductive biases based on spherical geometry to better capture global atmospheric dynamics (Chen et al., 2024; Zheng et al., 2025). For example, CirT adapts to the spherical grid structure via geometry-inspired circular patching (Liu et al., 2025a). EIMP uses equivariant message passing on spherical meshes (Liu et al., 2025b). Spectral approaches have also been explored to integrate teleconnection modeling with spherical harmonic representations (Lyu et al., 2026) and incorporate Fourier-based embeddings for enhancing wave pattern representations (Li et al., 2025).

Geometry-aware representations alone do not explicitly coordinate zonal and meridional interactions, leaving a key architectural gap in modeling anisotropic atmospheric dynamics. Global atmospheric fields exhibit anisotropic dynamics (Figure 1), in that their dynamical behavior (such as Rossby-wave propagation and the transport of heat and moisture) differs between the zonal (i.e., east–west) and meridional (i.e., north–south) directions (Rossby et al., 1939; Hoskins & Karoly, 1981). The two directions also differ geometrically. While longitude wraps around the globe and is periodic (e.g., $0^\circ$ and $360^\circ$ refer to the same meridian, so do $-180^\circ$ and $180^\circ$), latitude is finite (Peixoto & Oort,



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.

1991; Vallis, 2017). However, geometric design techniques such as 2D patching (Li et al., 2025), equivariant and invariant modeling (Liu et al., 2025b), and spherical harmonic aggregation (Lyu et al., 2026) do not specify how zonal and meridional interactions should be separately processed and composed. Encoding geometric structure therefore leaves an architectural gap between representing directional information and explicitly organizing its processing through complementary directional operations. This motivates an explicit design of complementary zonal and meridional operations and their composition within the forecasting backbone.

To alleviate these problems, we propose **AnisoCast**, which is designed to capture anisotropic atmospheric dynamics through a symmetric composition of zonal and meridional operators for effective global S2S climate forecasting. AnisoCast first employs an **Anisotropic Embedding** strategy that tokenizes the grid into latitudinal rings, preserving the integrity of zonal periodic structures while structurally distinguishing them from meridional representations. The **Aniso Blocks** then use **Zonal Operators** to capture the global spectral wave dynamics and **Meridional Operators** for the local spatial transport, allowing explicit modeling of their alternating interaction. These operators induce a geometrically orthogonal representation, thereby aligning the model with the anisotropic nature of global atmospheric dynamics. Each Aniso Block also introduces structural symmetry to mitigate the accumulation of discretization errors over the long S2S forecasting horizon. Finally, an MLP-based **Decoder** predicts all target variables, and the model is trained end-to-end.

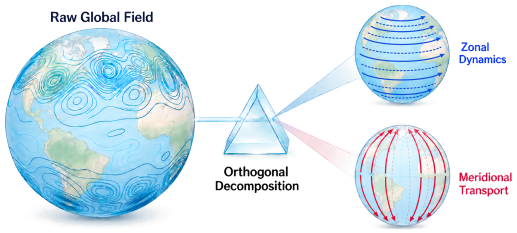


Figure 1: The anisotropic nature of global atmospheric dynamics. Global atmospheric circulation is governed by the coupling of two orthogonal physical processes: (i) Zonal dynamics, characterized by east–west wave propagation, and (ii) meridional transport, driven by the north-south exchange of heat and momentum.

Our main contributions are as follows:

- We identify that prevalent S2S forecasting models use isotropic modeling and highlight the importance of explicitly distinguishing zonal and meridional interactions in S2S forecasting architectures.
- We propose AnisoCast, which integrates Anisotropic Embedding with a symmetric composition of Zonal and Meridional Operators, that is theoretically and structurally aligned with long-term atmospheric dynamics at the S2S timescale.
- Extensive experiments on the Earth Reanalysis 5 (ERA5) dataset demonstrate that AnisoCast establishes a new state-of-the-art in S2S climate forecasting, outperforming existing methods in both accuracy and computational efficiency.

## 2 Preliminaries

### 2.1 Problem Definition

We consider global bi-weekly forecasting of $K$ meteorological variables defined on a latitude–longitude grid. Let $\Omega = [-90°, 90°] \times [-180°, 180°]$ be the spherical coordinate domain, discretized into $H$ latitude bands and $W$ longitude points. We represent the spatial coordinates by $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{H\times W\times 2}$ with $\boldsymbol{\mathcal{P}}_{h,w,:} = (\varphi_h, \lambda_w) \in \Omega$, where $\varphi_h$ and $\lambda_w$ are the latitude and longitude, respectively. At day $t$, the global atmospheric state is a tensor $\boldsymbol{X}_t \in \mathbb{R}^{K\times H\times W}$. In S2S benchmarking and S2S direct forecasting (Mouatadid et al., 2023; Nathaniel et al., 2024; Lyu et al., 2026; Liu et al., 2025a;b; Chen et al., 2024), we want to learn a predictor $\mathcal{F}_\Theta$ that, given an initial condition $(\boldsymbol{\mathcal{P}}, \boldsymbol{X}_{t_0})$, directly outputs the two verification targets for weeks 3-4 and weeks 5-6:

$$(\hat{\boldsymbol{Y}}_{15:28},\ \hat{\boldsymbol{Y}}_{29:42}) = \mathcal{F}_\Theta(\boldsymbol{\mathcal{P}}, \boldsymbol{X}_{t_0}), \tag{1}$$

where $\hat{\boldsymbol{Y}}_{15:28} \in \mathbb{R}^{K\times H\times W}$ and $\hat{\boldsymbol{Y}}_{29:42} \in \mathbb{R}^{K\times H\times W}$ denote the predicted averages over days 15-28 (weeks 3-4) and days 29-42 (weeks 5-6). A review on weather forecasting and S2S climate forecasting can be found in Appendix A.

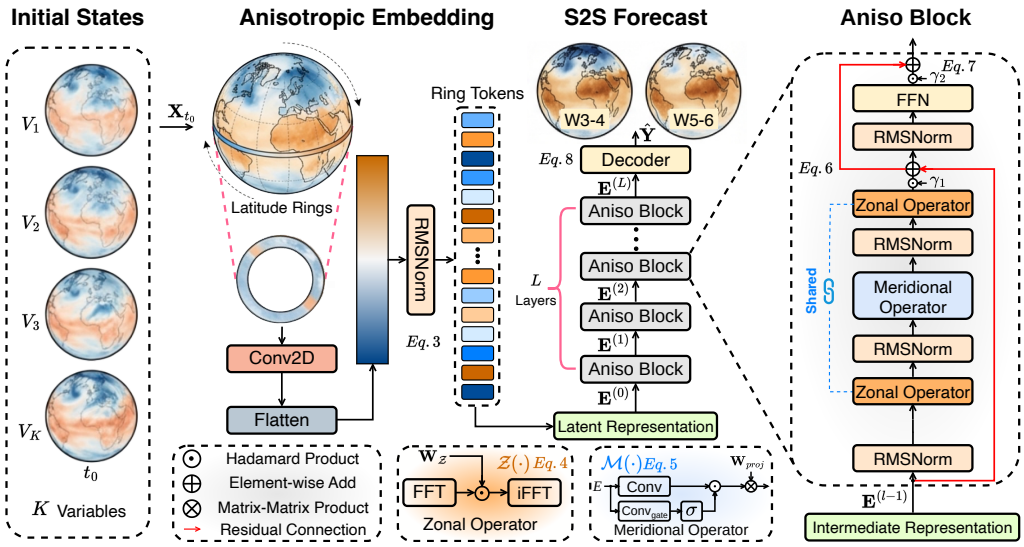


Figure 2: Main architecture of the proposed AnisoCast. The Anisotropic Embedding projects atmospheric initial states into latitudinal ring tokens by compressing the longitudinal dimension into feature channels. These tokens are then processed through a backbone of several stacked Aniso Blocks. Specifically, each block sequentially applies two weight-shared Zonal Operators and a Meridional Operator to separately process zonal- and meridional-oriented latent interactions. Finally, a Decoder maps the evolved features to bi-weekly forecast fields.

### 2.2 Operator Splitting for Non-commutative Dynamics

Atmospheric evolution can be abstracted as the dynamical system $\frac{\partial \mathbf{u}}{\partial t} \equiv \mathcal{F}(\mathbf{u}) = \mathcal{A}(\mathbf{u}) + \mathcal{B}(\mathbf{u})$, where $\mathbf{u}(t)$ is the global state vector, $\mathcal{A}$ is the zonal wave dynamics component, and $\mathcal{B}$ is the meridional transport component (Rossby et al., 1939; Hoskins & Karoly, 1981; Peixoto & Oort, 1991). The coupled flow, governed jointly by $\mathcal{A}$ and $\mathcal{B}$, is written as $\mathcal{F}(\mathbf{u}) = (\mathcal{A} + \mathcal{B})\mathbf{u}$.

Consider its spatially discretized linear analogue $\frac{\partial \mathbf{u}}{\partial t} = (\mathbf{A} + \mathbf{B})\mathbf{u}$, where $\mathbf{A}$ (resp. $\mathbf{B}$) is the matrix representation of the discretized operator for $\mathcal{A}$ (resp. $\mathcal{B}$) (McLachlan & Quispel, 2002; Hairer et al., 2006). The coupled flow over an interval of length $\tau$ can then be written as $\mathbf{u}(t+\tau) = \exp(\tau(\mathbf{A}+\mathbf{B}))\,\mathbf{u}(t)$, where $\exp(\cdot)$ denotes the matrix exponential. In general, $\mathbf{A}$ and $\mathbf{B}$ do not commute (i.e., $\mathbf{AB} - \mathbf{BA} \neq \mathbf{0}$) (Lanser & Verwer, 1999; Dimov et al., 2004), and so the flow cannot be factorized into separate evolutions.

Operator splitting (Trotter, 1959; Strang, 1968; McLachlan & Quispel, 2002) approximates this coupled flow by alternating two simpler subflows. Each subflow exploits its own structure while retaining the interaction through composition. Different splitting orders correspond to different discretizations of the underlying dynamics. The two representative methods are: (i) **Lie-Trotter splitting** (Trotter, 1959), which is a first-order approximation of the coupled flow and applies the two subflows sequentially as: $\mathcal{S}_{\text{Lie}}(\tau) = \exp(\tau\mathbf{B})\exp(\tau\mathbf{A})$. Its one-step local error is $\mathcal{O}(\tau^2)$ (Blanes & Casas, 2025). (ii) **Strang splitting** (Strang, 1968), which is a symmetric second-order decomposition that enforces time-reversal symmetry via:

$$\mathcal{S}_{\text{Strang}}(\tau) = \exp\left(\tfrac{\tau}{2}\mathbf{A}\right)\exp(\tau\mathbf{B})\exp\left(\tfrac{\tau}{2}\mathbf{A}\right). \tag{2}$$

This yields an $\mathcal{O}(\tau^3)$ one-step local error for a sufficiently small $\tau$ (McLachlan & Quispel, 2002).

In long-horizon forecasting, local approximation errors can accumulate under repeated composition. When modeling non-commutative systems, symmetric decomposition is often preferred over Lie-Trotter splitting as it typically reduces systematic bias and slows error accumulation (Hairer et al., 2006; Blanes & Casas, 2025).

# 3 METHODOLOGY

Figure 2 illustrates the architecture of the proposed AnisoCast, which is designed to align neural operations with the anisotropic nature of atmospheric dynamics. It comprises three modules: (i) **Anisotropic Embedding**, which projects each latitude ring into a latent token by encoding longitudinal information within feature channels; (ii) A stack of **Aniso Blocks**, which use a symmetric latent composition inspired by Strang splitting. It employs a **Zonal Operator** to capture the global spectral wave dynamics, and a **Meridional Operator** to model local spatial transport; (iii) A lightweight **Decoder**, which projects the encoded latent representations back to the physical space to generate the forecasts. The entire architecture is trained end-to-end by minimizing the prediction error over the S2S timescale. Appendix B shows a side-by-side comparison of AnisoCast and conventional isotropic atmospheric modeling.

## 3.1 ANISOTROPIC EMBEDDING

To tokenize global gridded fields, recent data-driven S2S models adopt various embedding strategies such as spherical harmonic embedding (Lyu et al., 2026), and Fourier embedding (Li et al., 2025). Inspired by the circular patching of CirT (Liu et al., 2025a), we adopt Anisotropic Embedding to encode longitude within token features while retaining latitude as the sequence dimension. Instead of extracting local patches, we project the entire latitudinal ring at each latitude index $h$ into a single high-dimensional latent token. Formally, given an input state $\boldsymbol{X}_{t_0} \in \mathbb{R}^{K\times H\times W}$ at time $t_0$, we apply 2D convolution $\mathrm{Conv}_{1\times W}$ with kernel size $(1, W)$ over the latitude and longitude dimensions. This compresses the longitude dimension to one spatial position, while its $C$ output filters project the $K$ input channels to $C$ latent channels. It is then followed by flattening $\mathrm{Flat}(\cdot)$, which reshapes the spatial dimension to a sequence of length $H$, and Root Mean Square Normalization (RMSNorm) (Zhang & Sennrich, 2019) $\mathrm{RMSN}(\cdot)$, which stabilizes the latent features without recentering:

$$\boldsymbol{E}^{(0)} = \mathrm{RMSN}\left(\mathrm{Flat}\left(\mathrm{Conv}_{1\times W}(\boldsymbol{X}_{t_0})\right)\right) \in \mathbb{R}^{H\times C}. \tag{3}$$

Here, dimension $H$ corresponds to the meridional domain, while the feature dimension $C$ encodes longitudinal information. We call each latitude-indexed row in $\boldsymbol{E}^{(0)}$ a *ring token*. Unlike local patching that fragments the grid (Pathak et al., 2022; Li et al., 2025), this aggregation constructs each ring token from the complete longitudinal context. Each ring token aggregates the complete longitude span, including both sides of the longitude-coordinate boundary, providing atmospheric context of the full latitudinal ring.

## 3.2 ANISO BLOCK

The ring tokens are processed through a backbone of $L$ stacked Aniso Blocks through direction-aware operations that separately model zonal- and meridional- interactions. This provides a novel way to capture the distinct atmospheric behavior along the two directions. Specifically, we define two operators acting in this latent space: Zonal Operator and Meridional Operator.

Since Anisotropic Embedding has encoded the longitudinal information into $C$ latent channels per ring token, the Zonal Operator performs spectral mixing across these channels, enabling amplitude and phase modulation of latent spectral components to model dependencies across the full longitude span. Let $\boldsymbol{E}^{(l-1)} \in \mathbb{R}^{H\times C}$ be the input to the $l$-th block. We define the Zonal Operator as:

$$\mathcal{Z}(\boldsymbol{E}^{(l-1)}) = \mathrm{iFFT}(\boldsymbol{W}_{\mathcal{Z}} \odot \mathrm{FFT}(\boldsymbol{E}^{(l-1)})), \tag{4}$$

where $\mathrm{FFT}(\cdot)$ and $\mathrm{iFFT}(\cdot)$ denote the one-dimensional Fast Fourier Transform and inverse Fast Fourier Transform applied along the latent channel dimension $C$ of each ring token, respectively, and $\odot$ is the element-wise Hadamard product. The parameter $\boldsymbol{W}_{\mathcal{Z}} \in \mathbb{C}^{C}$ is a learnable complex-valued weight initialized near identity. This latent spectral mixing enables each output channel to depend on the ring token's input channels in $\mathcal{O}(C \log C)$ time.

On the other hand, the Meridional Operator acts along the latitude dimension $H$, modeling local spatial transport as gated exchanges of latent information among neighboring latitude rings:

$$\mathcal{M}(\boldsymbol{E}^{(l-1)}) = (\mathrm{Conv}(\boldsymbol{E}^{(l-1)}) \odot \sigma(\mathrm{Conv}_{\mathrm{gate}}(\boldsymbol{E}^{(l-1)})))\boldsymbol{W}_{\mathrm{proj}}, \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid gate, $\mathrm{Conv}(\cdot)$ and $\mathrm{Conv}_{\mathrm{gate}}(\cdot)$ are two 1D convolution layers with kernel size $k$ sliding over the meridional sequence dimension $H$, and $\boldsymbol{W}_{\mathrm{proj}}$ is a learnable projection matrix.

The gated structure adaptively controls information flow along the meridional direction. Together, the two operators avoid pairwise attention across latitude rings, so their mixing costs scale linearly with the number of rings $H$.

Rather than treating the block as a physical time integrator on the primitive atmospheric variables, we adopt the symmetric composition underlying Strang splitting (Strang, 1968; Hairer et al., 2006) in the latent space. Specifically, the block updates the representation as:

$$\boldsymbol{E}' = \boldsymbol{E}^{(l-1)} + \boldsymbol{\gamma}_1 \odot (\mathcal{Z} \circ \mathrm{RMSN} \circ \mathcal{M} \circ \mathrm{RMSN} \circ \mathcal{Z}) \left(\mathrm{RMSN}(\boldsymbol{E}^{(l-1)})\right), \tag{6}$$

followed by a Feedforward Network ($\mathrm{FFN}(\cdot)$) with residual connections:

$$\boldsymbol{E}^{(l)} = \boldsymbol{E}' + \boldsymbol{\gamma}_2 \odot \mathrm{FFN}(\mathrm{RMSN}(\boldsymbol{E}')). \tag{7}$$

Here, $\circ$ denotes operator composition, and $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^C$ are learnable LayerScale (Touvron et al., 2021) parameters that modulate the residual branches and stabilize training. In Eq. (6), the Zonal Operator is applied twice with shared weights, yielding a symmetric $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ composition.[1] This also helps to mitigate the ordering bias, as the zonal wave propagation and meridional transport processes are non-commutative (Section 2.2).

To motivate the symmetric ordering, we apply the classical Strang-splitting result from Section 2.2 to a linear latent reference system. Let $\boldsymbol{z}_0 = \mathrm{vec}(\mathrm{RMSN}(\boldsymbol{E}^{(l-1)})) \in \mathbb{R}^{HC}$, where $\mathrm{vec}(\cdot)$ stacks the entries into a vector, and let $\mathbf{A}_{\mathrm{lat}}, \mathbf{B}_{\mathrm{lat}} \in \mathbb{R}^{HC \times HC}$ be fixed matrices defining its two components. For a composition step size $\tau > 0$, Eq. (2) gives $\mathcal{S}^{\mathrm{lat}}(\tau) = \exp\left(\frac{\tau}{2}\mathbf{A}_{\mathrm{lat}}\right) \exp(\tau \mathbf{B}_{\mathrm{lat}}) \exp\left(\frac{\tau}{2}\mathbf{A}_{\mathrm{lat}}\right)$. The classical Strang-splitting analysis (Strang, 1968; Hairer et al., 2006) establishes an $\mathcal{O}(\tau^3)$ local error bound for this linear composition relative to $\exp(\tau(\mathbf{A}_{\mathrm{lat}} + \mathbf{B}_{\mathrm{lat}}))$ as $\tau \to 0$. This provides motivation for the $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ ordering in Eq. (6).

While Layer Normalization (LayerNorm) (Ba et al., 2016) has been widely adopted in S2S architectures (Liu et al., 2025a;b; Lyu et al., 2026), its mean-centering discards each ring token's channel mean, which may encode background-state information relevant to S2S forecasting. We therefore adopt RMSNorm, whose normalization core rescales features without recentering. Appendix C provides a qualitative analysis showing that the LayerNorm core removes the zero-frequency component of the latent channel spectrum, whereas the RMSNorm core retains all spectral coefficients up to a common positive scaling factor, thereby avoiding this loss of channel-mean information. Ablation results in Section 4.4 further show that this choice improves forecasting accuracy.

### 3.3 Decoder

After passing through $L$ Aniso Blocks, the latent representation $\boldsymbol{E}^{(L)} \in \mathbb{R}^{H \times C}$ contains the encoded representation extracted from the initial atmospheric state $\boldsymbol{X}_{t_0}$. The Decoder projects this representation back to the original physical grid to generate the final forecasts. We employ a linear projection layer to reconstruct the longitudinal dimension $W$ from the $C$ feature channels:

$$(\hat{\boldsymbol{Y}}_{15:28},\ \hat{\boldsymbol{Y}}_{29:42}) = \mathrm{Reshape}\left(\mathrm{Linear}(\boldsymbol{E}^{(L)})\right), \tag{8}$$

where the linear layer projects the feature dimension from $C$ to $2 \times K \times W$. Finally, we obtain two tensors, $\hat{\boldsymbol{Y}}_{15:28}$ and $\hat{\boldsymbol{Y}}_{29:42}$, each with dimension $\mathbb{R}^{K \times H \times W}$, representing the two target bi-weekly mean fields corresponding to weeks 3–4 and weeks 5–6, respectively.

**Training Objective.** The model is trained in an end-to-end manner, minimizing the average squared error of the predicted bi-weekly means over the two forecast windows as in Eq. (1):

$$\mathcal{L} = \frac{1}{2KHW}\left(\left\|\hat{\boldsymbol{Y}}_{15:28} - \boldsymbol{Y}_{15:28}\right\|_{\mathcal{L}_2}^2 + \left\|\hat{\boldsymbol{Y}}_{29:42} - \boldsymbol{Y}_{29:42}\right\|_{\mathcal{L}_2}^2\right), \tag{9}$$

where $\boldsymbol{Y}_{15:28}$ and $\boldsymbol{Y}_{29:42}$ are the ground-truth averaged fields for weeks 3–4 and weeks 5–6, respectively, and $\|\cdot\|_{\mathcal{L}_2}^2$ is the latitude-weighted Mean Squared Error (Nguyen et al., 2023; 2024; Li et al., 2025) that accounts for the Earth's spherical geometry by weighting the error at latitude $h$ with $w_h = \cos(\varphi_h)$.

[1] We also evaluate the variant $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$ with shared outer Meridional Operators in Appendix G.1.

### 3.4 Computational Complexity

We analyze the computational complexity of AnisoCast in terms of the latitude resolution $H$, longitude resolution $W$, number of variables $K$, latent dimension $C$, number of blocks $L$, and meridional kernel size $k$. The full complexity of AnisoCast is $\mathcal{O}(HKWC) + \mathcal{O}(LH(C^2 + C\log C + Ck)) + \mathcal{O}(HCKW)$, corresponding to Anisotropic Embedding, the stacked Aniso Block, and the Decoder, respectively. Thus, the full model still scales linearly with the longitudinal resolution $W$, since the embedding must read the original longitude-resolved inputs and the decoder must reconstruct longitude-resolved forecasts. CirT (Liu et al., 2025a) also uses $H$ ring tokens, but its global attention incurs an $\mathcal{O}(H^2C)$ token-mixing cost per block. ClimaX (Nguyen et al., 2023) applies global attention over $N$ spatial patch tokens, with an $\mathcal{O}(N^2C)$ token-mixing cost per block, where $N \propto HW$ for a fixed patch size. AnisoCast avoids these quadratic attention terms through latent spectral mixing and local meridional convolutions, giving a backbone cost that scales linearly with $H$. Detailed analysis are provided in Appendix D.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset.** Following the established benchmarking protocols (Liu et al., 2025a;b; Lyu et al., 2026), we evaluate AnisoCast on the ERA5 reanalysis dataset (Hersbach et al., 2020). We regrid the data to a $1.5^\circ$ resolution, resulting in a $121 \times 240$ latitude–longitude grid. Following (Liu et al., 2025a;b), we use 63 meteorological variables as inputs at initialization and as prediction targets averaged over each future 14-day window, including six pressure-level variables (geopotential $z$, specific humidity $q$, temperature $t$, u-wind $u$, v-wind $v$, and vertical velocity $w$) at ten pressure levels (10, 50, 100, 200, 300, 500, 700, 850, 925, and 1000 hPa), and three surface variables (2-meter temperature $t2m$, 10-meter u-wind $u10$, and 10-meter v-wind $v10$). Following Liu et al. (2025a;b), we split the dataset chronologically, using the 1979–2016 data (38 years) for training, the 2017 data for validation, and the 2018 data for testing. Details can be found in Appendix E.1.

**Baselines.** We compare AnisoCast with two categories of methods: (1) **Operational NWP systems**: CMA (Wu et al., 2019), NCEP (Saha et al., 2014), and ECMWF (Molteni et al., 1996); (2) **Data-driven models**: PanguWeather (Pangu) (Bi et al., 2023), FourCastNetV2 (FCN2) (Pathak et al., 2022), ClimaX (Nguyen et al., 2023), ClimODE (Verma et al., 2024), CirT (Liu et al., 2025a), EIMP (Liu et al., 2025b), TelePiT (Lyu et al., 2026), TianQuan-S2S (Li et al., 2025), and OmniCast (Nguyen et al., 2025). We align predictions to the same bi-weekly targets and verification grid. More details are provided in Appendix E.2.

**Metrics.** Following existing evaluation standards (Rasp et al., 2020; Nathaniel et al., 2024), we employ the Latitude-weighted Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC) to quantify performance (Nathaniel et al., 2024), defined as:

$$\text{RMSE} = \sqrt{\frac{1}{HW}\sum_{h,w}\alpha(h)\,(\hat{\boldsymbol{X}}_{h,w} - \boldsymbol{X}_{h,w})^2}, \qquad \text{ACC} = \frac{\sum_{h,w}\alpha(h)\,\hat{\boldsymbol{X}}'_{h,w}\,\boldsymbol{X}'_{h,w}}{\sqrt{\sum_{h,w}\alpha(h)\,(\hat{\boldsymbol{X}}'_{h,w})^2}\,\sqrt{\sum_{h,w}\alpha(h)\,(\boldsymbol{X}'_{h,w})^2}}, \quad (10)$$

where $\alpha(h) = \cos(\phi_h) \big/ \left(\frac{1}{H}\sum_{h'}\cos(\phi_{h'})\right)$ is the latitude weighting factor, $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are the ground truth and prediction for a given bi-weekly verification window. The anomalies are computed as $\boldsymbol{X}' = \boldsymbol{X} - \boldsymbol{C}$ and $\hat{\boldsymbol{X}}' = \hat{\boldsymbol{X}} - \boldsymbol{C}$, where $\boldsymbol{C}$ is the observational climatology.

**Implementation Details.** For PanguWeather (Bi et al., 2023) and FourCastNetV2 (Pathak et al., 2022), we utilize the official pre-trained weights provided by the ECMWF project through API (see Appendix E.2 for details). For the other data-driven models, we train them from scratch using their officially recommended configurations on the same GPU server. Training is conducted on 8 NVIDIA H20 80G GPUs. Inference is performed on a single NVIDIA H100 80G GPU. We use $L = 7$ blocks for the proposed method. More implementation details are in Appendix E.3.

### 4.2 Overall Performance

Table 1 reports the quantitative comparison with operational NWP systems and data-driven baselines. AnisoCast often achieves the best performance, with particularly stable gains on large-scale upper-air variables ($z500$, $z850$, $t500$, and $t850$) and competitive performance on near-surface variables ($t2m$,

Table 1: Forecasting results for 14-day mean targets at weeks 3–4 and weeks 5–6. The best result is in **bold**, and the second-best is underlined. Results for the operational NWP systems are obtained from (Lyu et al., 2026), whereas we reproduce other data-driven models and report their results. Pressure-level variables $z500$ (gpm), $z850$ (gpm) and $t500$ (K), $t850$ (K) are at 500 and 850 hPa, while single-level variables $t2m$ (K), $u10$ (m/s), and $v10$ (m/s) are measured at 2m and 10m above the surface, respectively. N/A indicates the baseline does not predict the variables (Liu et al., 2025b).

| | Model | RMSE (↓) | | | | | | | ACC (↑) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pressure Level | | | | Single Level | | | Pressure Level | | | | Single Level | | |
| | | $z500$ | $z850$ | $t500$ | $t850$ | $t2m$ | $u10$ | $v10$ | $z500$ | $z850$ | $t500$ | $t850$ | $t2m$ | $u10$ | $v10$ |
| Weeks 3–4 (Days 15–28) | CMA | 69.833 | 47.056 | 2.598 | 2.918 | 4.221 | N/A | N/A | 0.968 | 0.919 | 0.975 | 0.974 | 0.960 | N/A | N/A |
| | NCEP | 68.441 | 44.787 | 2.225 | 2.742 | 4.028 | N/A | N/A | 0.969 | 0.921 | 0.979 | 0.977 | 0.966 | N/A | N/A |
| | ECMWF | 63.674 | 42.349 | 2.067 | 2.391 | **3.232** | N/A | N/A | 0.972 | 0.929 | 0.982 | 0.983 | 0.971 | N/A | N/A |
| | Pangu | 66.224 | 37.959 | 2.271 | 2.569 | N/A | 2.431 | 1.984 | 0.963 | 0.926 | 0.966 | 0.963 | N/A | 0.812 | 0.686 |
| | FCN2 | 62.755 | 41.020 | 2.093 | 2.390 | N/A | 2.328 | 1.896 | 0.947 | 0.896 | 0.966 | 0.957 | N/A | 0.830 | 0.712 |
| | ClimaX | 59.064 | 38.420 | 2.142 | 3.242 | 5.342 | 2.360 | 2.025 | 0.978 | 0.945 | 0.981 | 0.971 | 0.956 | 0.807 | 0.649 |
| | ClimODE | 68.767 | 39.326 | 2.518 | 3.147 | 3.782 | 2.326 | 1.793 | 0.970 | 0.946 | 0.972 | 0.970 | 0.975 | 0.827 | 0.738 |
| | CirT | 55.505 | 34.345 | 1.931 | 2.265 | 5.478 | 2.253 | 1.873 | 0.980 | 0.955 | 0.984 | 0.983 | 0.958 | 0.835 | 0.716 |
| | EIMP | 66.498 | 41.820 | 2.312 | 2.800 | 4.785 | 3.683 | 2.137 | 0.981 | 0.956 | 0.983 | 0.984 | 0.974 | 0.672 | 0.654 |
| | TelePiT | 59.772 | 35.199 | 2.035 | 2.448 | 4.002 | 2.122 | 1.746 | 0.977 | 0.953 | 0.982 | 0.980 | 0.974 | 0.854 | 0.741 |
| | TianQuan-S2S | 53.212 | 34.662 | 1.898 | 2.232 | 3.286 | 2.022 | 1.885 | 0.973 | 0.955 | 0.982 | 0.975 | 0.966 | 0.843 | 0.768 |
| | OmniCast | 54.423 | 35.561 | 1.992 | 2.341 | 3.366 | 2.213 | 1.903 | 0.972 | 0.955 | 0.980 | 0.974 | 0.965 | 0.843 | 0.766 |
| | **AnisoCast** | **51.956** | **33.453** | **1.824** | **2.145** | 3.255 | **1.994** | **1.621** | **0.982** | **0.958** | **0.985** | **0.985** | **0.982** | **0.872** | **0.783** |
| Weeks 5–6 (Days 29–42) | CMA | 71.400 | 48.583 | 2.639 | 2.933 | 5.521 | N/A | N/A | 0.966 | 0.912 | 0.974 | 0.974 | 0.950 | N/A | N/A |
| | NCEP | 71.332 | 46.834 | 2.343 | 2.833 | 4.406 | N/A | N/A | 0.967 | 0.913 | 0.977 | 0.976 | 0.952 | N/A | N/A |
| | ECMWF | 68.116 | 45.330 | 2.177 | 2.511 | 4.080 | N/A | N/A | 0.968 | 0.916 | 0.979 | 0.981 | 0.970 | N/A | N/A |
| | Pangu | 76.939 | 47.041 | 2.829 | 2.998 | N/A | 2.679 | 2.104 | 0.956 | 0.911 | 0.958 | 0.957 | N/A | 0.783 | 0.655 |
| | FCN2 | 66.531 | 43.469 | 2.250 | 2.567 | N/A | 2.479 | 1.980 | 0.943 | 0.889 | 0.963 | 0.953 | N/A | 0.812 | 0.691 |
| | ClimaX | 58.712 | 38.271 | 2.141 | 3.300 | 5.354 | 2.394 | 2.031 | 0.978 | 0.945 | 0.982 | 0.971 | 0.956 | 0.800 | 0.643 |
| | ClimODE | 68.218 | 39.287 | 2.554 | 3.213 | 4.094 | 2.336 | 1.844 | 0.971 | 0.946 | 0.972 | 0.969 | 0.971 | 0.829 | 0.717 |
| | CirT | 55.214 | 33.940 | 1.920 | 2.265 | 5.652 | 2.251 | 1.826 | 0.980 | 0.956 | 0.984 | 0.983 | 0.954 | 0.835 | 0.726 |
| | EIMP | 66.856 | 41.926 | 2.329 | 2.858 | 5.018 | 2.555 | 2.155 | 0.981 | 0.957 | 0.983 | 0.983 | 0.972 | 0.792 | 0.651 |
| | TelePiT | 58.770 | 35.164 | 1.953 | 2.451 | 4.056 | 2.120 | 1.750 | 0.978 | 0.953 | 0.982 | 0.981 | 0.975 | 0.854 | 0.741 |
| | TianQuan-S2S | 54.623 | 35.551 | 1.991 | 2.368 | 3.446 | 2.121 | 1.908 | 0.974 | 0.955 | 0.983 | 0.978 | 0.970 | 0.845 | 0.769 |
| | OmniCast | 55.522 | 36.015 | 1.991 | 2.340 | 3.347 | 2.326 | 1.912 | 0.973 | 0.956 | 0.980 | 0.973 | 0.964 | 0.841 | 0.765 |
| | **AnisoCast** | **52.268** | **33.437** | **1.830** | **2.156** | **3.231** | **1.993** | **1.606** | **0.982** | **0.958** | **0.985** | **0.985** | **0.982** | **0.873** | **0.786** |

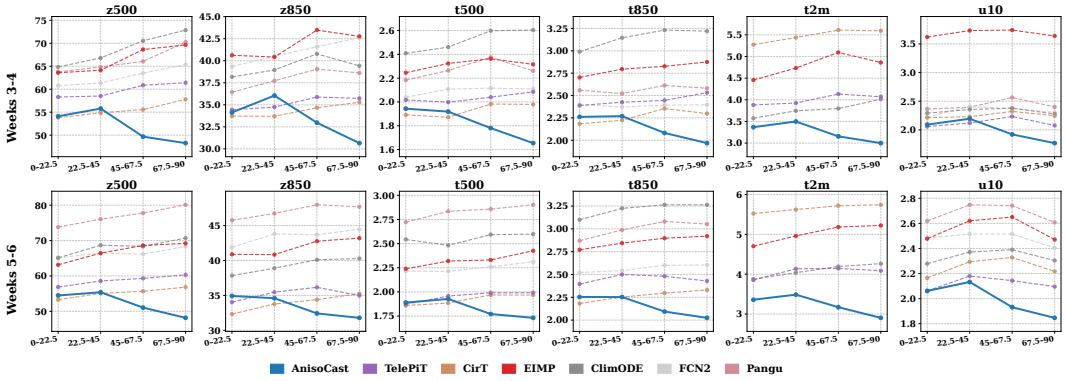


Figure 3: RMSE variation w.r.t. different latitude bands, comparing AnisoCast with baseline models.

$u10$, and $v10$). This suggests that the anisotropic ring-token representation and symmetric latent zonal–meridional composition provide an effective inductive bias for direct S2S forecasting.

Additional results are reported in the appendices. Appendix F.1 show that RMSE and ACC have small standard deviations across five seeds, indicating robustness to random initialization across training runs. Appendix F.2 extends testing from 2018 to 2018–2019. AnisoCast achieves the lowest RMSE on all seven reported variables and the highest ACC on six. Appendix F.3 shows that AnisoCast remains competitive across months, particularly on $t500$ and $t850$, indicating that its gains are not confined to a single season. Appendix F.4 reports higher temporal correlations for $z500$ and $z850$ in both windows and better extreme-event detection scores for $u10$, $v10$, and $t2m$, extending the evaluation beyond spatial RMSE and ACC.

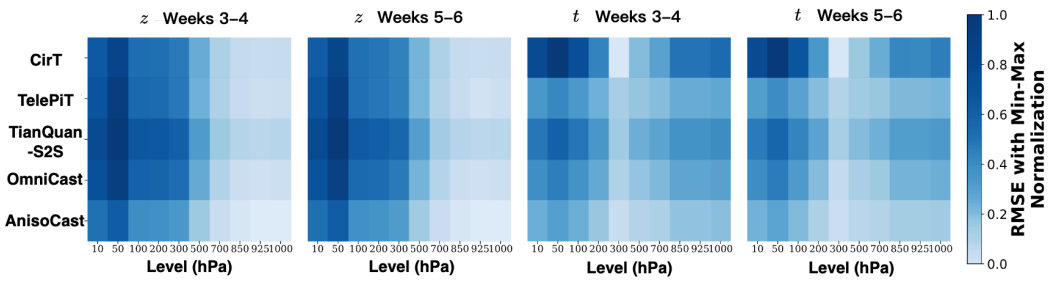


Figure 4: RMSE comparison on variable $z$ and $t$ across different pressure levels.

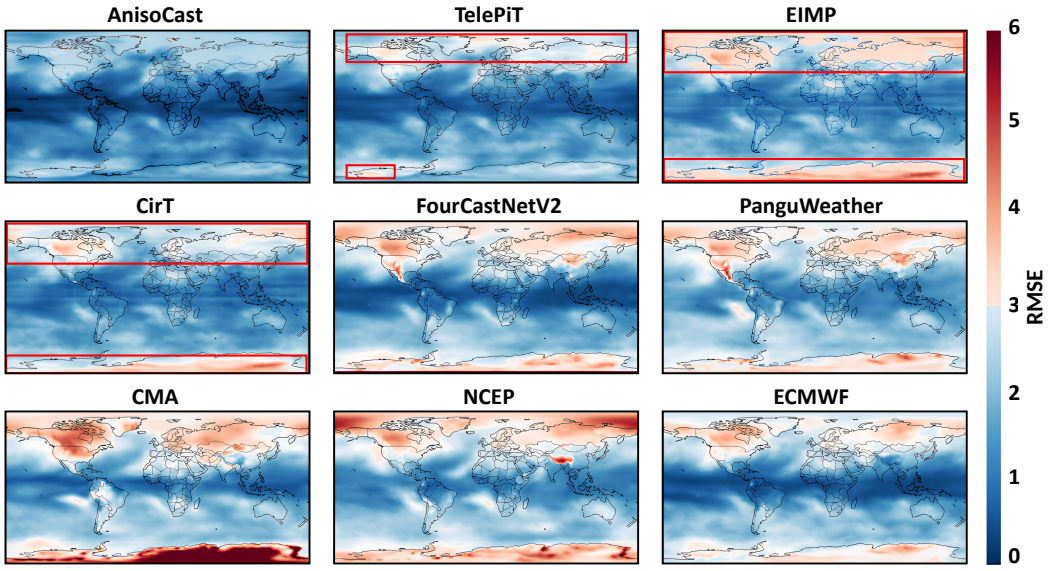


Figure 5: Global RMSE distribution of $t850$ with lead times weeks 3-4 in the testing set.

### 4.3 Additional Model Analysis

For model analysis, we select $z$ and $t$ at multiple pressure levels, including $z500$ and $t850$, together with near-surface temperature ($t2m$), following prior S2S studies (Liu et al., 2025a;b). These variables provide complementary information on atmospheric conditions: $z500$ characterizes synoptic-scale circulation and $t850$ reflects broad temperature variations (Rasp et al., 2020), while $t2m$ represents near-surface thermal conditions.

**Quantitative Latitudinal Analysis.** Figure 3 reports RMSE across four latitudinal bands from the equator ($0°$) to the pole ($90°$). AnisoCast shows smaller error growth toward high latitudes than competing baselines under the same ERA5-based verification protocol, suggesting that the anisotropic ring embedding reduces sensitivity to latitude-dependent grid convergence and projection distortion. Appendix F.5 reports lower $z500$ and $t2m$ RMSE under three grid resolutions ($5.625°$, $1.5°$, and $0.25°$). Appendix F.6 also reports lower RMSE for $z500$, $z850$, $u10$, and $v10$ after excluding regions beyond $60°$ latitude in either hemisphere. These results indicate that the gains of our proposed model are not confined to one resolution or to high-latitude regions.

**Performance across Pressure Levels.** Following (Liu et al., 2025a; Lyu et al., 2026), we visualize the RMSE for $z$ and $t$ across 10 pressure levels in Figure 4. AnisoCast maintains lower errors across most evaluated levels, including the upper atmosphere (50–200 hPa). This consistency reflects the joint encoding of variables across pressure levels, which allows each prediction to use cross-level context alongside complementary zonal and meridional interactions. Appendix H.1 reports lower spectral residuals (Lyu et al., 2026) across most evaluated levels, indicating closer agreement between predicted and target power spectra. Appendix H.2 further shows lower $z$ and $t$ RMSE than operational NWP systems (Wu et al., 2019; Saha et al., 2014; Molteni et al., 1996) across most pressure levels, supporting the consistency of these gains across the vertical atmospheric column.

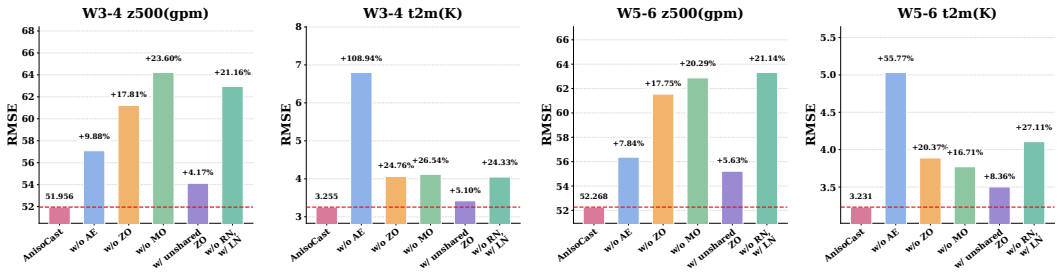


Figure 6: Ablation study quantifying the contribution of each AnisoCast component.

**Global Visualization.** Figure 5 visualizes the spatial distribution of RMSE for $t850$. Compared with the baselines, AnisoCast shows lower spatial errors, particularly in the high-latitude regions highlighted by the red boxes. Retaining complete longitudinal context within each ring and exchanging information between neighboring latitudes enables joint modeling of broad zonal structures and local meridional variations, consistent with these spatial improvements. Additional results in Appendix H.3 report slower error growth for separately trained daily-output models over days 15–42, extending the evaluation beyond bi-weekly mean prediction. Additional results in Appendix H.4 show similar spatial improvements across additional variables and both forecast windows.

### 4.4 Ablation Study

We conduct ablation studies to evaluate the contribution of each component of AnisoCast. Figure 6 compares AnisoCast with variants in which several key design choices are removed or replaced. Here, $z500$ and $t2m$ are selected to represent upper-air circulation and near-surface temperature, respectively. The detailed experimental settings are in Appendix E.4. We evaluate the following variants: (1) **w/o AE**, which replaces Anisotropic Embedding with the order-agnostic, shape-matched ring embedding; (2) **w/o ZO**, which replaces $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ with $\mathcal{M} \circ \mathcal{M}$; (3) **w/o MO**, which replaces $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ with $\mathcal{Z} \circ \mathcal{Z}$; (4) **unshared ZO**, which uses separate weights for the two Zonal Operators; and (5) **LayerNorm**, which replaces RMSNorm with LayerNorm. As can be seen, all five variants increase RMSE across the evaluated variables, supporting the contributions of the proposed embedding, complementary zonal and meridional mixing, weight sharing, and RMSNorm. Appendix G.1 reports lower RMSE for $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ than for $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$, supporting the adopted configuration. Appendix G.2 also shows that direct prediction outperforms autoregressive rollout in both windows, supporting direct optimization of the bi-weekly targets.

### 4.5 Efficiency Comparison

Figure 7 highlights the efficiency of AnisoCast. The reported metrics are measured with batch size 1 and hidden dimension 256 for a fair comparison. AnisoCast achieves the lowest FLOPs among the compared models and delivers inference speedups of over 2× and 5× relative to CirT (Liu et al., 2025a) and ClimaX (Nguyen et al., 2023), respectively. These gains reflect its attention-free design, which replaces quadratic attention with Zonal and Meridional Operators that avoid quadratic token-mixing costs.

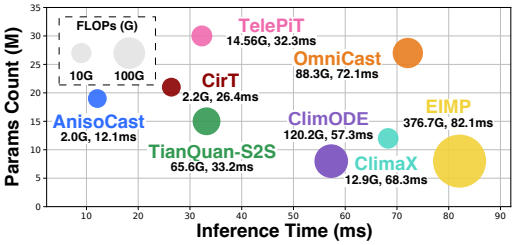


Figure 7: Efficiency analysis. The horizontal axis shows inference latency, the vertical axis shows parameter count, and bubble size depicts FLOPs.

## 5 Conclusion and Future Work

This paper explores directional structure as an inductive bias for modeling anisotropic atmospheric dynamics in S2S forecasting. We propose AnisoCast, which preserves complete latitude-ring context through anisotropic embedding and models zonal and meridional interactions with distinct operators. Their weight-shared symmetric composition couples global spectral mixing with local cross-latitude exchange while avoiding quadratic attention. Experiments on the ERA5 dataset show that AnisoCast achieves state-of-the-art forecasting accuracy and improved computational efficiency compared with NWP systems and data-driven models. We discuss its limitations in Appendix I. Future work will explore generative frameworks to quantify forecast uncertainty at chaotic S2S lead times.

## AI USE STATEMENT

The authors used AI-assisted tools for grammar checking and language refinement.

## REFERENCES


Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 2023.

Sergio Blanes and Fernando Casas. *A concise introduction to geometric numerical integration*. CRC press, 2025.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv:2304.02948*, 2023a.

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*, 2023b.

Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu, Bo Lu, Deliang Chen, Shang-Ping Xie, Libo Wu, Qingchen Chao, Chensen Lin, et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 2024.

I Dimov, István Faragó, Ágnes Havasi, and Zahari Zlatev. Operator splitting and commutativity analysis in the danish eulerian model. *Mathematics and Computers in Simulation*, 2004.

Yuchen Fang, Hao Miao, Yuxuan Liang, Liwei Deng, Yue Cui, Ximu Zeng, Yuyang Xia, Yan Zhao, Torben Bach Pedersen, Christian S. Jensen, Xiaofang Zhou, and Kai Zheng. Unraveling spatio-temporal foundation models via the pipeline lens: A comprehensive review. *TKDE*, 2026.

Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, et al. Oneforecast: A universal framework for global and regional weather forecasting. In *ICML*, 2025.

Yuan Gao, Hao Wu, Fan Xu, Yanfei Xiang, Ruijian Gou, Ruiqi Shu, Qingsong Wen, Xian Wu, Kun Wang, and Xiaomeng Huang. Neuralom: Neural ocean model for subseasonal-to-seasonal simulation. In *AAAI*, 2026.

Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 2006.

Shubhi Harbola and Volker Coors. One dimensional convolutional neural network architectures for wind prediction. *Energy Conversion and Management*, 2019.

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*, 2020.

Brian J Hoskins and David J Karoly. The steady linear response of a spherical atmosphere to thermal and orographic forcing. *Journal of Atmospheric Sciences*, 1981.

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

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

Debby Lanser and Jan G Verwer. Analysis of operator splitting for advection–diffusion–reaction problems from air pollution modelling. *Journal of computational and applied mathematics*, 1999.

Guowen Li, Xintong Liu, Shilei Cao, Haoyuan Liang, Mengxuan Chen, Lixian Zhang, Jinxiao Zhang, Jiuke Wang, Meng Jin, Juepeng Zheng, et al. Tianquan-climate: A subseasonal-to-seasonal global weather model via incorporate climatology state. *arXiv:2504.09940*, 2025.

Yang Liu, Zinan Zheng, Jiashun Cheng, Fugee Tsung, Deli Zhao, Yu Rong, and Jia Li. Cirt: Global subseasonal-to-seasonal forecasting with geometry-inspired transformer. In *ICLR*, 2025a.

Yang Liu, Zinan Zheng, Yu Rong, Deli Zhao, Hong Cheng, and Jia Li. Equivariant and invariant message passing for global subseasonal-to-seasonal forecasting. In *SIGKDD*, 2025b.

Tengfei Lyu, Weijia Zhang, and Hao Liu. Physics-informed teleconnection-aware transformer for global subseasonal-to-seasonal forecasting. In *SIGKDD*, 2026.

Annarita Mariotti, Paolo M Ruti, and Michel Rixen. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *Npj Climate and Atmospheric Science*, 2018.

Robert I McLachlan and G Reinout W Quispel. Splitting methods. *Acta Numerica*, 2002.

Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 1996.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Judah Cohen, Miruna Oprescu, Ernest Fraenkel, and Lester Mackey. Adaptive bias correction for improved subseasonal forecasting. *Nature Communications*, 2023.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. In *NeurIPS*, 2024.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *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 *NeurIPS*, 2024.

Tung Nguyen, Tuan Pham, Troy Arcomano, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Omnicast: A masked latent diffusion model for weather forecasting across time scales. In *NeurIPS*, 2025.

Peisong Niu, Ziqing Ma, Tian Zhou, Weiqi Chen, Lefei Shen, Rong Jin, and Liang Sun. Utilizing strategic pre-training to reduce overfitting: Baguan - a pre-trained weather forecasting model. In *SIGKDD*, 2025.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv:2202.11214*, 2022.

José Pinto Peixoto and Abraham H Oort. Physics of climate. 1991.

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

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

CG Rossby, HC Willett, Messrs Holmboe, J Namias, L Page, R Allen, et al. Relation between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the permanent centers of action atmosphere and the displacements of the permanent centers of action. 1939.

Suranjana Saha, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, et al. The ncep climate forecast system version 2. *Journal of Climate*, 2014.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015.

Ruiqi Shu, Ruijian Gou, Yanfei Xiang, and Xiaomeng Huang. Ocean-e2e: Hybrid physics-based and data-driven global forecasting of marine heatwaves with end-to-end neural assimilation. In *SIGKDD*, 2026.

Gilbert Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 1968.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021.

Hale F Trotter. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society*, 1959.

Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017.

Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. In *ICLR*, 2024.

F Vitart and AW Robertson. The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. npj climate atmos. sci., 1, 3, 2018.

F. Vitart, C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, et al. The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 2017.

Chunzai Wang, Clara Deser, Jin-Yi Yu, Pedro DiNezio, and Amy Clement. El niño and southern oscillation (enso): a review. *Coral reefs of the eastern tropical Pacific: Persistence and loss in a dynamic environment*, 2016.

Yanyun Wang, Dehui Du, Haibo Hu, Zi Liang, and Yuanhao Liu. Tsfool: Crafting highly-imperceptible adversarial time series through multi-objective attack. In *ECAI*, 2024.

Christopher J White, Henrik Carlsen, Andrew W Robertson, Richard JT Klein, Jeffrey K Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J Ray, Virginia Murray, et al. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications*, 2017.

Hao Wu, Fan Xu, Yuxu Lu, Penghao Zhao, Fan Zhang, Hao Jia, Yuxuan Liang, Ruijian Gou, Qingsong Wen, Xian Wu, Xiaomeng Huang, and Yuan Gao. Pnp-corrector: A universal correction framework for coupled spatiotemporal forecasting. In *ICML*, 2026.

Tongwen Wu, Yixiong Lu, Yongjie Fang, Xiaoge Xin, Laurent Li, Weiping Li, Weihua Jie, Jie Zhang, Yiming Liu, Li Zhang, et al. The beijing climate center climate system model (bcc-csm): The main progress from cmip5 to cmip6. *Geoscientific Model Development*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. arXiv:1910.07467, 2019.

Zinan Zheng, Yang Liu, and Jia Li. Mesh interpolation graph network for dynamic and spatially irregular global weather forecasting. *arXiv:2509.20911*, 2025.

Ziyu Zhou, Gengyu Lyu, Yiming Huang, Zihao Wang, Ziyu Jia, and Zhen Yang. Sdformer: Transformer with spectral filter and dynamic attention for multivariate time series long-term forecasting. In *IJCAI*, 2024.

# Appendix (Supplementary Material)

## TABLE OF CONTENTS

**A Related Works** **14**
A.1 Data-driven Weather Forecasting . . . . . 14
A.2 Data-driven S2S Forecasting . . . . . 14

**B Isotropic and Anisotropic Modeling Paradigms** **14**

**C Channel-Spectral Analysis of LayerNorm and RMSNorm** **15**

**D Theoretical Model Complexity** **15**

**E Experimental Details** **16**
E.1 Dataset . . . . . 16
E.2 Baselines . . . . . 16
E.3 Implementation Details . . . . . 18
E.4 Ablation Settings . . . . . 18

**F Additional Results on Robustness and Generalization** **18**
F.1 Multi-seed Robustness . . . . . 18
F.2 Extended Evaluation Period for Interannual Robustness . . . . . 19
F.3 Seasonal Robustness Analysis . . . . . 19
F.4 Evaluation under Additional Criteria . . . . . 19
F.5 Multi-resolution Evaluation . . . . . 21
F.6 Evaluation Excluding High-latitude Regions . . . . . 21

**G Additional Results on Architectural and Forecasting Choices** **22**
G.1 Full Ablation Results . . . . . 22
G.2 Direct vs. Autoregressive Setting . . . . . 23
G.3 Performance w.r.t. the number of Aniso Blocks . . . . . 23

**H Complementary Forecast Diagnostics** **23**
H.1 Spectral Energy Distribution Analysis . . . . . 23
H.2 Vertical Profile Analysis . . . . . 24
H.3 Daily Visualization and Temporal Stability . . . . . 25
H.4 Additional Global Visualizations . . . . . 25

**I Limitations** **25**

# A Related Works

## A.1 Data-driven Weather Forecasting

Spatio-temporal data, together with time series data, are prevalent across various domains of daily life, among which weather data are closely related to human activities (Fang et al., 2026; Wang et al., 2024; Zhou et al., 2024). Recently, deep learning has emerged as a computationally efficient alternative to traditional Numerical Weather Prediction (NWP) systems (Vitart et al., 2017; White et al., 2017). While early works focused on regional tasks (Shi et al., 2015; Harbola & Coors, 2019), the release of ERA5 (Hersbach et al., 2020) has shifted focus to global forecasting. Foundational benchmarks (Rasp et al., 2020) paved the way for models that rival operational NWP systems (e.g., IFS) in the medium range. Prominent architectures include Adaptive Fourier Neural Operators (FourCastNet (Pathak et al., 2022)), Graph Neural Networks (GraphCast (Lam et al., 2023), AIFS (Lang et al., 2024)), and Transformers (PanguWeather (Bi et al., 2023), FengWu (Chen et al., 2023a), FuXi (Chen et al., 2023b)). Recent research also explores probabilistic forecasting via diffusion models (Price et al., 2025). However, the performance of these medium-range models often degrades at subseasonal scales due to the lack of physical constraints and error accumulation.

## A.2 Data-driven S2S Forecasting

Subseasonal-to-Seasonal (S2S) forecasting (2–6 weeks) remains a grand challenge due to chaotic atmospheric dynamics. While traditional dynamical models (e.g., ECMWF (Molteni et al., 1996)) face computational bottlenecks, recent global data-driven models have shown promise. ClimaX (Nguyen et al., 2023) adapts a vision foundation model for climate tasks via fine-tuning. FuXi-S2S (Chen et al., 2024) employs a cascade architecture to forecast daily averaged statistics. CirT (Liu et al., 2025a) introduces circular patching strategies to better handle spherical boundary conditions. EIMP (Liu et al., 2025b) leverages graph neural networks with rotation-equivariant message passing to model vector fields. TelePiT (Lyu et al., 2026) combines spherical harmonics with physics-informed neural ODEs to capture multi-scale teleconnections, while TianQuan-S2S (Li et al., 2025) further enhances predictability by incorporating climatology priors and noise injection. Although these approaches introduce different geometric, spectral, or physical inductive biases, they do not explicitly formulate zonal- and meridional-oriented latent transformations as a symmetric operator composition. In contrast, AnisoCast organizes these complementary transformations through a weight-shared $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ structure, providing an efficient direction-specific backbone for global S2S forecasting.

# B Isotropic and Anisotropic Modeling Paradigms

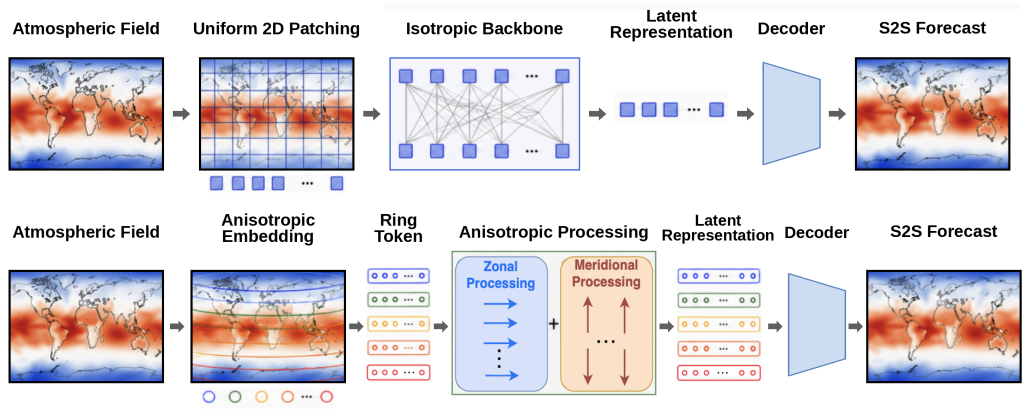


Figure 8: Paradigm-level comparison between isotropic modeling and the proposed anisotropic modeling for S2S forecasting.

Figure 8 holds the atmospheric input and S2S forecast target fixed and illustrates the difference in spatial modeling paradigms. The upper workflow summarizes the conventional isotropic modeling paradigm, in which latitude–longitude fields are processed without explicitly decomposing zonal- and meridional-oriented latent transformations. It is a schematic abstraction and does not imply

that all existing methods employ identical uniform 2D patches. Recent geometry-aware methods introduce circular patching, spherical message passing, or spherical-harmonic representations (Liu et al., 2025a;b; Lyu et al., 2026). These designs account for important properties of global geometry, but they do not explicitly organize zonal- and meridional-oriented latent transformations. AnisoCast changes this spatial inductive bias rather than the forecasting objective. Its Anisotropic Embedding maps each complete latitude ring to one token, retaining the full longitudinal span within a single representation. The subsequent processing assigns distinct roles to zonal-oriented interactions within the learned ring representation and meridional-oriented interactions across the ordered latitude sequence. This comparison concerns architectural inductive biases and does not imply a unique decomposition of the atmospheric governing equations.

## C Channel-Spectral Analysis of LayerNorm and RMSNorm

We analyze how normalization affects the channel spectrum of a latent ring token $\boldsymbol{z} \in \mathbb{R}^C$. Let $\mu = C^{-1}\sum_{n=0}^{C-1} z_n$ and $\widehat{\boldsymbol{z}} = \mathcal{F}(\boldsymbol{z})$ denote its channel mean and unnormalized channel-wise DFT, respectively. Ignoring affine parameters, the normalization cores are $\mathrm{LN}_0(\boldsymbol{z}) = (\boldsymbol{z} - \mu\mathbf{1})/s_{\mathrm{LN}}$ and $\mathrm{RMS}_0(\boldsymbol{z}) = \boldsymbol{z}/s_{\mathrm{RMS}}$, where $s_{\mathrm{LN}} = \sqrt{C^{-1}\sum_{n=0}^{C-1}(z_n-\mu)^2}$ and $s_{\mathrm{RMS}} = \sqrt{C^{-1}\sum_{n=0}^{C-1} z_n^2}$ are assumed to be nonzero. Since $\widehat{z}_0 = C\mu$, subtracting the channel mean gives $\mathcal{F}(\mathrm{LN}_0(\boldsymbol{z}))_0 = 0$, while $\mathcal{F}(\mathrm{LN}_0(\boldsymbol{z}))_k = \widehat{z}_k/s_{\mathrm{LN}}$ for $k \neq 0$. LayerNorm therefore removes the channel-mean component, which may contain information useful for forecasting. In contrast, RMSNorm satisfies $\mathcal{F}(\mathrm{RMS}_0(\boldsymbol{z}))_k = \widehat{z}_k/s_{\mathrm{RMS}}$ for all $k = 0, \ldots, C-1$, retaining the entire channel spectrum, including its zero-frequency component, up to a common positive scaling factor. Both cores preserve the phases and relative magnitudes of nonzero-frequency coefficients whenever these coefficients are nonzero; their key difference is the removal or retention of the channel-mean component. These properties apply to the normalization cores, as learned channel-wise gains and biases can further alter the spectrum. They concern latent-channel coefficients and do not establish preservation of all atmospheric information or physical wave phases. The forecasting effect of the normalization choice is evaluated in Section 4.4.

## D Theoretical Model Complexity

We provide the detailed complexity derivation of AnisoCast. Let $H$ and $W$ denote the latitude and longitude resolutions, $K$ the number of meteorological variables, $C$ the latent dimension, $L$ the number of Aniso Blocks, and $k$ the meridional convolution kernel size. Standard Transformer-based models typically tokenize the latitude–longitude grid into $T \propto HW$ tokens. Therefore, global self-attention has complexity $\mathcal{O}(T^2C) = \mathcal{O}(H^2W^2C)$, which scales quadratically with the two-dimensional token count. In contrast, AnisoCast first applies Anisotropic Embedding to compress each latitudinal ring into one latent token. This embedding reads the original grid $\boldsymbol{X}_{t_0} \in \mathbb{R}^{K\times H\times W}$ and maps the $K$ input variables over $W$ longitudes into $C$ latent channels for each latitude, giving complexity $\mathcal{O}(HKWC)$. After this step, the latent representation has shape $\mathbb{R}^{H\times C}$, so the Aniso Block operates on only $H$ latitudinal ring tokens rather than $HW$ two-dimensional tokens. For each Aniso Block, the main costs come from channel-wise projections and FFN layers with complexity $\mathcal{O}(HC^2)$, latent spectral mixing in the Zonal Operator with complexity $\mathcal{O}(HC\log C)$, and meridional convolution with complexity $\mathcal{O}(HCk)$. Therefore, one Aniso Block costs $\mathcal{O}(H(C^2 + C\log C + Ck))$, and $L$ stacked blocks cost $\mathcal{O}(LH(C^2 + C\log C + Ck))$. Importantly, this backbone complexity is independent of $W$ because the longitude dimension has already been compressed into latent channels by the anisotropic embedding. Finally, the Decoder maps the latent representation back to the physical grid. Since each latitude token is projected from $C$ latent channels to longitude-resolved outputs over $K$ variables and $W$ longitudes, the decoder costs $\mathcal{O}(HCKW)$, with the constant factor for the two bi-weekly targets omitted in big-$\mathcal{O}$ notation. Combining the embedding, backbone, and decoder terms, the full complexity of AnisoCast is $\mathcal{O}(HKWC) + \mathcal{O}(LH(C^2 + C\log C + Ck)) + \mathcal{O}(HCKW)$. Both AnisoCast and CirT (Liu et al., 2025a) compress the longitude dimension into ring-token features. AnisoCast further avoids the $\mathcal{O}(LH^2C)$ attention term through its spectral and convolutional backbone.

# E EXPERIMENTAL DETAILS

## E.1 DATASET

We evaluate AnisoCast on the ERA5 reanalysis dataset (Hersbach et al., 2020), which provides a comprehensive record of global atmospheric conditions. To construct the training and evaluation data for our S2S forecasting task, we downsample the original $0.25^\circ$ ERA5 data to a coarser resolution of $1.5^\circ \times 1.5^\circ$ by sub-sampling grid points that align with the target grid. Specifically, we extract coordinates $(\varphi, \lambda)$ on the latitude–longitude domain $\Omega = [-90^\circ, 90^\circ] \times [-180^\circ, 180^\circ)$, where $\varphi \in \{-90^\circ, -88.5^\circ, \ldots, 90^\circ\}$ and $\lambda \in \{-180^\circ, -178.5^\circ, \ldots, 178.5^\circ\}$, resulting in a spatial resolution of $121 \times 240$. This regridding strategy ensures that our input data structure is consistent with the outputs of baseline models such as PanguWeather (Bi et al., 2023) and FourCastNetV2 (Pathak et al., 2022) when evaluated at this resolution. We select a set of 63 variables critical for capturing atmospheric dynamics, covering the time period from 1979 to 2018 with a temporal resolution of 1 day (daily averages). A detailed summary of the dataset configuration is provided in Table 2.

Table 2: Detailed configuration of the ERA5 dataset used in this work.

| Category | Variable Name | Symbol | Levels | Resolution | Lat-Lon Range | Time |
|---|---|---|---|---|---|---|
| **Pressure Level** | Geopotential | $z$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | Specific Humidity | $q$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | Temperature | $t$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | U Component of Wind | $u$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | V Component of Wind | $v$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | Vertical Velocity | $w$ | 10 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| **Single Level** | 2 Metre Temperature | $t2m$ | 1 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | 10 Metre U Wind | $u10$ | 1 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |
| | 10 Metre V Wind | $v10$ | 1 | $1.5^\circ$ | $-90^\circ$S $180^\circ$W $\sim 90^\circ$N $180^\circ$E | 1979 $\sim$ 2018 |

## E.2 BASELINES

To rigorously evaluate the performance of AnisoCast, we benchmark it against a comprehensive set of baseline models:

**Operational Forecasting Systems.** These established numerical weather prediction (NWP) systems represent the current gold standard in operational meteorological forecasting.

- **CMA:** The China Meteorological Administration implements the Beijing Climate Center's fully-coupled BCC-CSM2-HR model (Wu et al., 2019), providing operational forecasts with a 60-day lead time.
- **NCEP:** The National Centers for Environmental Prediction deploys the Climate Forecast System version 2 (CFSv2) (Saha et al., 2014) to generate daily control forecasts extending to a 45-day prediction window.
- **ECMWF:** The European Centre for Medium-Range Weather Forecasts employs the Integrated Forecasting System (IFS) (Molteni et al., 1996).

For NWP baselines, we follow the variable availability of the standardized S2S evaluation archives used in prior work (Lyu et al., 2026). CMA, NCEP, and ECMWF do not provide $u10$/$v10$ forecasts in this setting; this is a data-availability constraint of the archived products.

**Data-Driven Models.** We compare against both pre-trained foundation models and models trained from scratch.

- **PanguWeather (Bi et al., 2023):** A 3D Earth-specific Transformer that tokenizes pressure-level and surface data separately, trained using a hierarchical temporal aggregation strategy. We use the API https://github.com/ecmwf-lab/ai-models to perform inference.

- **FourCastNetV2 (FCN2) (Pathak et al., 2022):** An iterative model based on Vision Transformers that incorporates Adaptive Fourier Neural Operators (AFNO) for efficient spatial token mixing. We use the API https://github.com/ecmwf-lab/ai-models to perform inference.
- **ClimaX (Nguyen et al., 2023):** A foundation model for weather and climate using a Vision Transformer backbone with variable-specific tokenization and aggregation.
- **ClimODE (Verma et al., 2024):** A physics-informed approach that models atmospheric dynamics using Neural Ordinary Differential Equations (ODEs) to capture continuous-time evolution. We download the official code from https://github.com/Aalto-QuML/ClimODE.
- **CirT (Liu et al., 2025a):** A Transformer-based model designed for S2S forecasting that introduces circular patching and frequency-domain mixing to address spherical boundary conditions. We download the official code from https://github.com/compasszzn/CirT.
- **EIMP (Liu et al., 2025b):** A graph-based model utilizing equivariant and invariant message passing to strictly enforce rotational symmetries on spherical grids. We download the official code from https://zenodo.org/records/15516468.
- **TelePiT (Lyu et al., 2026):** A physics-informed Transformer that integrates spherical harmonic embeddings with neural ODEs to explicitly capture multi-scale teleconnections. We download the official code from https://github.com/usail-hkust/TelePiT.
- **TianQuan-S2S (Li et al., 2025):** We use its official repository[2] and aggregate daily-mean forecasts over the target days to obtain weeks 3–4 and weeks 5–6 means.
- **OmniCast (Nguyen et al., 2025):** We use its official repository [3], spatially align its outputs to the benchmark grid, and compute the same latitude-weighted metrics on the shared verification targets.

Since PanguWeather (Bi et al., 2023) and FourCastNetV2 (Pathak et al., 2022) are pretrained medium-range autoregressive models, we roll them out to the required target days and average the generated fields into the same bi-weekly verification windows. Therefore, they serve as autoregressive reference baselines under our standardized S2S evaluation protocol, rather than as models trained with the same direct bi-weekly objective as AnisoCast. For controlled comparison, all trainable learned baselines in the main tables are evaluated under our unified preprocessing, inference, and metric-computation protocol. For data-driven models, we do not mix originally reported scores from prior papers into the main benchmark because public repositories differ in preprocessing details, checkpoint availability, or inference pipelines; such reported numbers are therefore treated as non-directly comparable to the unified results.

Moreover, our reproduced CirT scores differ from Table 1 of the original paper (Liu et al., 2025a). The released repository [4] leaves key evaluation choices unspecified: `test_step` reports a single normalized, tensor-level RMSE without de-normalization or variable-wise RMSE/ACC aggregation, and the available ACC function is not called. The default configuration also reverses the reported validation/test years, while the climatology script does not restrict input years. Together with unavailable published checkpoints and frozen preprocessing artifacts, these omissions prevent exact reconstruction of the published evaluation. We therefore evaluate CirT using the same pipeline as the other baselines: the fixed 1979–2016/2017/2018 train/validation/test split, training-only normalization and climatology, and common physical-unit RMSE and ACC aggregation. Our released code [5] records these settings and verifies frozen inputs and result tables with SHA-256 manifests. This makes the comparison in Table 1 consistent across models while clarifying why its CirT scores may differ from the originally reported values.

Following CirT (Liu et al., 2025a), we compute latitude-weighted ACC on the 14-day mean fields for weeks 3–4 and weeks 5–6. For each evaluated variable, anomalies are obtained by subtracting the same observational climatology from predictions and targets. In our evaluation, this climatology is estimated from the 1979–2016 training data and is shared across all models, ensuring a consistent reference for the comparison. Their absolute values should be interpreted within this setting. Comparisons with WeatherBench (Rasp et al., 2020) or ECMWF forecasting evaluations (Molteni et al.,

[2] `https://github.com/zhangminglang42/TianQuan`
[3] `https://github.com/tung-nd/omnicast`
[4] `https://github.com/compasszzn/CirT`
[5] `https://anonymous.4open.science/r/AnisoCast`

1996) require accounting for differences in forecast targets, temporal aggregation, spatial domains, seasons, variables, and climatology definitions.

### E.3 Implementation Details

The AnisoCast model is implemented using PyTorch 2.3.1 and trained on 8 NVIDIA GPUs via Distributed Data Parallel (DDP) with mixed precision (FP16), while FFT operations within the Zonal Operator are forced to FP32 to ensure numerical stability. We utilize an anisotropic embedding kernel size of $(1, 240)$ and a meridional spatial kernel size of $k = 7$. The model is trained for 20 epochs with a global batch size of 256 using the AdamW optimizer (weight decay $10^{-5}$, $\epsilon = 10^{-8}$) and gradient clipping at 1.0. The learning rate follows a cosine annealing schedule, decaying from a peak of $10^{-3}$ to $10^{-4}$. To determine the optimal architecture, we conducted a grid search on the validation set over the following hyperparameter spaces: embedding dimension $C \in \{128, 192, 256, 384, 512\}$, Aniso Blocks $L \in \{4, 5, 6, 7, 8, 9\}$, decoder depth $\in \{1, 2, 3\}$, learning rate $\in \{5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}\}$, MLP ratio $\in \{2.0, 4.0, 6.0\}$, and dropout rates $\in \{0.0, 0.1, 0.2\}$. The final selected configuration is $C = 256$, $L = 7$, decoder depth 1, learning rate $10^{-3}$, MLP ratio 4.0, and dropout 0.1, resulting in a parameter count of approximately 16.58M.

### E.4 Ablation Settings

The following variants are studied in the ablation studies in Section 4.4:

1. **Variant without Anisotropic Embedding (w/o AE).** We replace Anisotropic Embedding with an order-agnostic ring embedding constructed specifically as a shape-matched control. Formally, the replacement embedding is defined as
$$\widetilde{\boldsymbol{E}}^{(0)} = \mathrm{RMSN}\left(\mathrm{Flat}\left(\mathrm{Linear}\left(\mathrm{AvgPool}_{1\times W}(\boldsymbol{X}_{t_0})\right)\right)\right) \in \mathbb{R}^{H\times C}.$$
$\mathrm{AvgPool}_{1\times W}$ averages each latitude ring over longitude, and $\mathrm{Linear}$ maps the resulting $K$-dimensional vector at each latitude to $C$ channels. This control evaluates the effect of removing longitude-dependent input information, rather than isolating the contribution of the embedding geometry.
2. **Variant without Zonal Operator (w/o ZO).** We remove the spectral mechanism and replace the $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ symmetric structure with stacked Meridional Operators ($\mathcal{M} \circ \mathcal{M}$).
3. **Variant without Meridional Operator (w/o MO).** We replace the spatial operator with a double Zonal structure ($\mathcal{Z} \circ \mathcal{Z}$).
4. **Variant with unshared-weight Zonal Operator (w/ unshared ZO).** We define two independent Zonal Operators, $\mathcal{Z}_1$ and $\mathcal{Z}_2$, resulting in an asymmetric structure $\mathcal{Z}_1 \circ \mathcal{M} \circ \mathcal{Z}_2$.
5. **Variant with LayerNorm (w/ LN).** We replace all RMSNorm layers with standard Layer Normalization to evaluate the impact of mean-centering on long-term stability.
6. **Variant with alternative symmetric ordering.** We replace $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ with $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$, where the two outer Meridional Operators share weights, while keeping all other settings unchanged.

## F Additional Results on Robustness and Generalization

### F.1 Multi-seed Robustness

To evaluate sensitivity to random initialization, we train AnisoCast with five random seeds and report the mean and standard deviation of RMSE and ACC in Table 3.

Table 3: Multi-seed robustness of AnisoCast over five random seeds. We report mean±std for RMSE and ACC on weeks 3–4 and weeks 5–6. Lower RMSE and higher ACC indicate better performance.

| **Period** | $z500$ **RMSE** | $z850$ **RMSE** | $u10$ **RMSE** | $v10$ **RMSE** | $z500$ **ACC** | $z850$ **ACC** | $u10$ **ACC** | $v10$ **ACC** |
|---|---|---|---|---|---|---|---|---|
| w3–4 | 51.993±0.233 | 33.755±0.412 | 1.995±0.016 | 1.719±0.143 | 0.981±0.002 | 0.958±0.002 | 0.871±0.001 | 0.783±0.002 |
| w5–6 | 52.518±0.275 | 33.400±0.382 | 1.995±0.090 | 1.671±0.114 | 0.981±0.002 | 0.958±0.002 | 0.872±0.002 | 0.786±0.002 |

Across both forecast windows, AnisoCast's mean RMSE remains below the reported baseline values on all four evaluated variables, while ACC standard deviations are at most 0.002. These small ACC

variations indicate limited sensitivity to random initialization, supporting robustness across training runs. The larger relative variation in $v10$ RMSE nevertheless indicates greater seed sensitivity for near-surface wind errors.

### F.2 Extended Evaluation Period for Interannual Robustness

To demonstrate that AnisoCast's performance improvements are robust to interannual climate variability and not an artifact of the specific 2018 test period, we extend our evaluation to encompass both 2018 and 2019. S2S atmospheric dynamics often exhibit significant year-to-year shifts driven by broader Earth system signals (e.g., ENSO phases (Wang et al., 2016)). Expanding the test horizon allows us to examine whether the model has successfully learned generalized physical laws rather than overfitting to a single year's weather regimes.

Table 4: Extended-period evaluation on the 2018–2019 held-out years. Scores are averaged over the weeks 3–4 and weeks 5–6 forecast windows. The best result across all models is in **bold**, and the best baseline result is underlined. Lower RMSE and higher ACC indicate better performance. ACC is dimensionless. The variable groups follow Table 1.

| | Model | RMSE (↓) | | | | | | | ACC (↑) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pressure Level | | | | Single Level | | | Pressure Level | | | | Single Level | | |
| | | $z500$ | $z850$ | $t500$ | $t850$ | $t2m$ | $u10$ | $v10$ | $z500$ | $z850$ | $t500$ | $t850$ | $t2m$ | $u10$ | $v10$ |
| 2018–2019 | ClimaX | 67.287 | 40.132 | 2.293 | 2.692 | 5.298 | 3.268 | 2.665 | 0.960 | 0.948 | 0.970 | 0.968 | 0.955 | 0.806 | 0.641 |
| | ClimODE | 70.326 | 42.260 | 2.665 | 3.166 | 5.122 | 4.076 | 2.892 | 0.957 | 0.947 | 0.968 | 0.965 | 0.958 | 0.828 | 0.735 |
| | CirT | 65.211 | 38.544 | 2.302 | 2.987 | 4.696 | 2.451 | 1.989 | 0.968 | 0.948 | 0.974 | 0.982 | 0.973 | 0.831 | 0.709 |
| | EIMP | 66.724 | 39.138 | 2.465 | 2.489 | 4.592 | 2.512 | 2.332 | 0.974 | 0.950 | 0.975 | **0.983** | 0.972 | 0.798 | 0.732 |
| | TelePiT | 64.016 | 38.246 | 2.320 | 2.677 | 4.311 | 2.392 | 2.264 | 0.972 | 0.949 | 0.976 | 0.978 | 0.962 | 0.853 | 0.739 |
| | TianQuan-S2S | 65.981 | 41.022 | 2.412 | 2.774 | 4.599 | 2.522 | 2.032 | 0.975 | 0.953 | 0.975 | 0.973 | 0.963 | 0.837 | 0.764 |
| | OmniCast | 66.403 | 39.293 | 2.121 | 2.629 | 4.482 | 2.496 | 2.091 | 0.977 | 0.952 | 0.979 | 0.973 | 0.963 | 0.839 | 0.763 |
| | **AnisoCast** | **62.381** | **36.489** | **1.936** | **2.442** | **4.298** | **2.233** | **1.899** | **0.978** | **0.954** | **0.982** | 0.982 | **0.977** | **0.869** | **0.783** |

As shown in Table 4, AnisoCast consistently achieves the lowest RMSE across all reported variables under this extended evaluation period. Quantitatively, it reduces the RMSE relative to the strongest respective baseline by approximately 2.6% for upper-air circulation ($z500$), 8.7% for mid-troposphere temperature ($t500$), and 6.6% and 4.5% for near-surface winds ($u10$ and $v10$). AnisoCast also achieves the highest ACC on six of the seven reported variables, while EIMP performs slightly better on $t850$. These results carry an important implication: the geometric and physical inductive biases introduced by AnisoCast's anisotropic modeling yield robust forecasting skill that persists through different climatic backgrounds. It verifies that our performance gains are not transient phenomena tied to a specific temporal window. We maintain 2018 as the primary benchmark in the main text purely to ensure strictly controlled comparisons with existing direct S2S literature.

### F.3 Seasonal Robustness Analysis

S2S forecasting skill often varies significantly by season due to changing atmospheric dynamics. To evaluate robustness, we break down the testing set performance by month. Figure 9 to Figure 12 illustrate the monthly RMSE for key variables ($z500, z850, t500, t850$). AnisoCast remains competitive across months, with particularly low errors for $t500$ and $t850$, although rankings vary by variable and month. Whether in Northern Hemisphere winter (where baroclinic instability is high) or summer, AnisoCast maintains the lowest error margins. This suggests that the model has successfully learned generalized representations of seasonal cycle dynamics rather than overfitting to specific regimes.

### F.4 Evaluation under Additional Criteria

In addition to RMSE and spatial ACC reported in the main text, we further evaluate AnisoCast using Temporal Correlation Coefficient (TCC) and extreme-event metrics. Following FuXi-S2S (Chen et al., 2024), TCC measures the temporal phase consistency of anomaly forecasts. For each variable and lead window, let $a_{n,i}$ and $\hat{a}_{n,i}$ denote the observed and predicted anomalies at evaluation sample

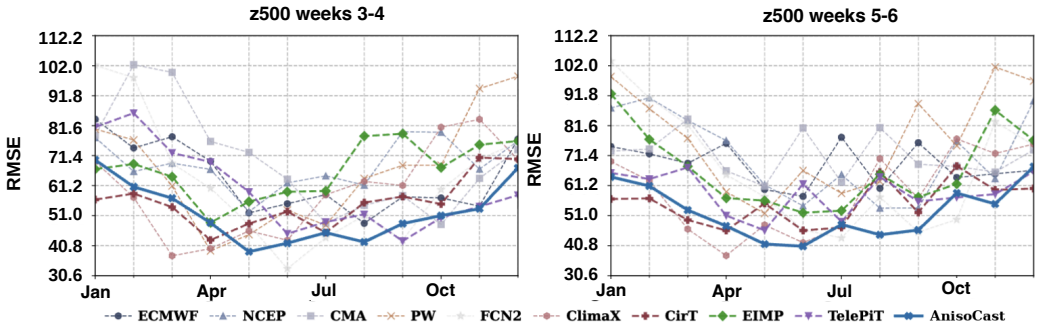


Figure 9: Monthly RMSE breakdown for $z500$.

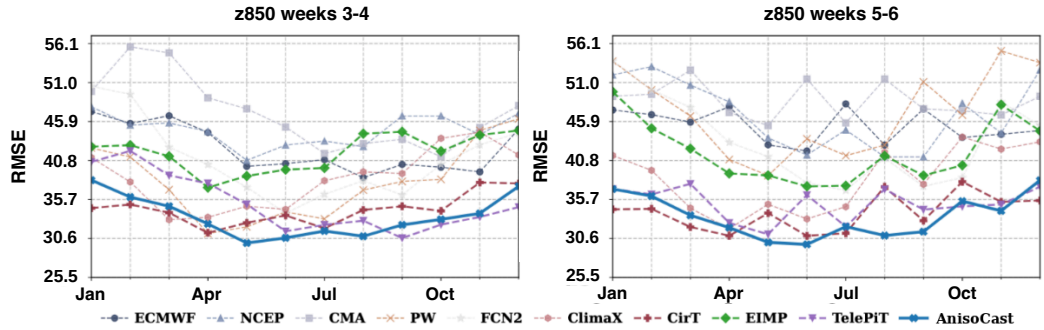


Figure 10: Monthly RMSE breakdown for $z850$.

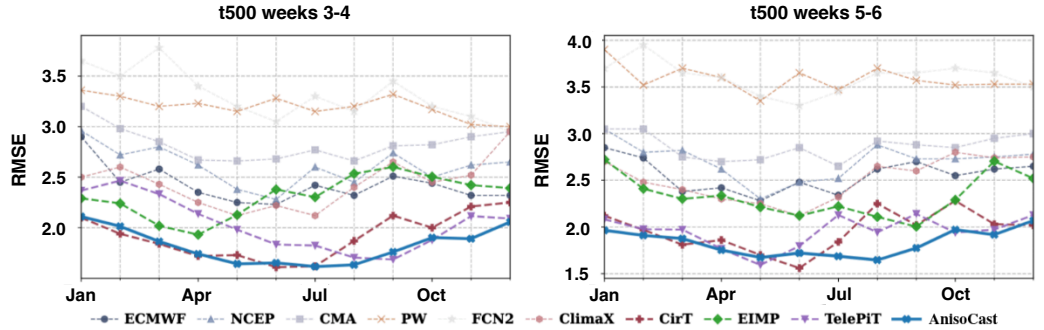


Figure 11: Monthly RMSE breakdown for $t500$.

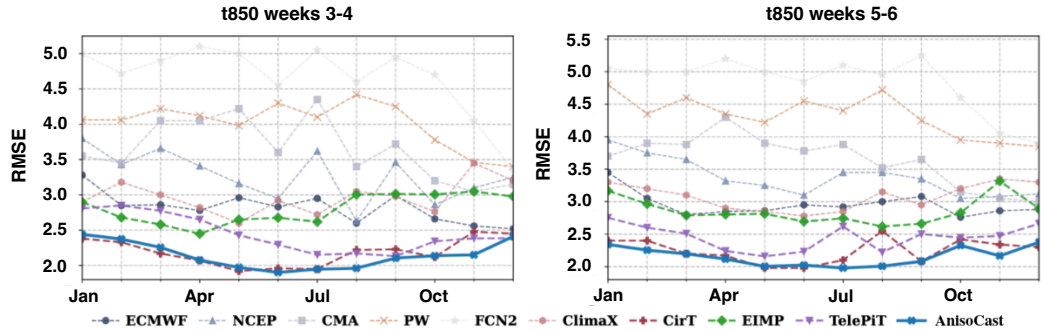


Figure 12: Monthly RMSE breakdown for $t850$.

$n$ and grid point $i$, and let $\alpha_i$ be the latitude weight. We compute

$$\rho_i = \frac{\sum_n (\hat{a}_{n,i} - \bar{\hat{a}}_i)(a_{n,i} - \bar{a}_i)}{\sqrt{\sum_n (\hat{a}_{n,i} - \bar{\hat{a}}_i)^2} \sqrt{\sum_n (a_{n,i} - \bar{a}_i)^2}}, \qquad \mathrm{TCC} = \frac{\sum_i \alpha_i \rho_i}{\sum_i \alpha_i}.$$

We report TCC on representative atmospheric variables and total precipitation (tp),an additional key surface variable for climate forecasting (Chen et al., 2023a;b; 2024; Niu et al., 2025).

Table 5: Evaluation under additional criteria. We report TCC for $z500$, $z850$, $t2m$, $v10$, and total precipitation (tp). CSI and SEDI are reported for extreme-event evaluation on $u10$, $v10$, and $t2m$. All metrics are dimensionless, and higher values indicate better performance.

| Model | TCC W3–4 | | | | | TCC W5–6 | | | | | CSI | | | SEDI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $z500$ | $z850$ | $t2m$ | $v10$ | tp | $z500$ | $z850$ | $t2m$ | $v10$ | tp | $u10$ | $v10$ | $t2m$ | $u10$ | $v10$ | $t2m$ |
| CirT | 0.420 | 0.419 | 0.365 | 0.301 | 0.261 | 0.351 | 0.350 | 0.370 | 0.303 | 0.169 | 0.09 | 0.12 | 0.22 | 0.29 | 0.33 | 0.40 |
| EIMP | 0.412 | 0.411 | 0.368 | 0.292 | 0.251 | 0.334 | 0.331 | 0.360 | **0.313** | 0.177 | 0.10 | 0.12 | 0.22 | 0.31 | 0.34 | 0.39 |
| TelePiT | 0.428 | 0.424 | 0.370 | 0.289 | 0.276 | 0.362 | 0.358 | **0.372** | 0.312 | 0.182 | 0.10 | 0.11 | 0.22 | 0.32 | 0.32 | 0.38 |
| AnisoCast | **0.442** | **0.440** | **0.372** | **0.313** | **0.282** | **0.373** | **0.370** | 0.368 | 0.309 | **0.196** | **0.12** | **0.13** | **0.24** | **0.34** | **0.35** | **0.42** |

For extreme-event evaluation, we report Critical Success Index (CSI) (Shu et al., 2026) and Symmetric Extremal Dependence Index (SEDI) (Gao et al., 2025). Extreme events are defined using the same grid-wise threshold $k_i$ for all models. In our implementation, $k_i$ is computed as the 90th percentile of the training-period climatology at each grid point and for each variable. We compute

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \quad \mathrm{SEDI} = \frac{\log(F) - \log(H) - \log(1-F) + \log(1-H)}{\log(F) + \log(H) + \log(1-F) + \log(1-H)},$$

where $H = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ is the hit rate and $F = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$ is the false alarm rate. Higher TCC, CSI, and SEDI indicate better performance.

As shown in Table 5, AnisoCast achieves the best TCC on $z500$, $z850$, and total precipitation in both forecast windows, while remaining competitive on near-surface variables. It also achieves the best CSI and SEDI on $u10$, $v10$, and $t2m$. These gains suggest that its learned representations capture temporal variations and extreme atmospheric states across different variables, supporting broader applicability beyond the primary evaluation variables and metrics.

### F.5 Multi-resolution Evaluation

To examine whether the performance gain of AnisoCast is sensitive to the spatial resolution, we further evaluate different models under three grid settings: $5.625°$ ($32 \times 64$), $1.5°$ ($121 \times 240$), and $0.25°$ ($721 \times 1440$). The results are summarized in Table 6.

Across all resolutions, AnisoCast consistently achieves the best performance on both $z500$ and $t2m$, indicating that its advantage does not rely on a particular grid configuration. We also observe that increasing the spatial resolution does not uniformly improve S2S forecasting skill, which is consistent with the long-range nature of the task: at subseasonal horizons, predictability is often governed more by large-scale circulation patterns than by fine-scale spatial details. These results support our use of the $1.5°$ setting as a practical and effective trade-off between computational cost and forecasting accuracy. This analysis is based on ERA5 verification and should not be interpreted as independent polar validation against radiosonde or satellite observations.

### F.6 Evaluation Excluding High-latitude Regions

To verify that the overall improvement of AnisoCast is not mainly driven by polar-region performance, we additionally evaluate all models over the non-polar latitude band $|\phi| \leq 60°$, excluding high-latitude regions above $60°$. The results are visualized in Figure 13. Each bar reports the region-aggregated RMSE for one variable, averaged over the selected latitude band, all longitudes, and the evaluation samples. AnisoCast achieves the lowest RMSE across all four reported variables, including $z500$, $z850$, $u10$, and $v10$. Specifically, AnisoCast obtains $z500$ RMSE of $43.781$, $z850$ RMSE of $28.326$,

Table 6: Multi-resolution evaluation on Weeks 3–4. We compare different models under $5.625°$, $1.5°$, and $0.25°$ spatial resolutions. The best results under each resolution are highlighted in bold. Lower RMSE and higher ACC indicate better performance.

| **Model** | **Resolution** | $z500$ **RMSE** | $t2m$ **RMSE** | $z500$ **ACC** | $t2m$ **ACC** |
|---|---|---|---|---|---|
| CirT | $5.625°$ | 56.103 | 5.510 | 0.978 | 0.957 |
| EIMP | $5.625°$ | 66.326 | 4.783 | 0.980 | 0.974 |
| TelePiT | $5.625°$ | 59.780 | 4.121 | 0.976 | 0.973 |
| AnisoCast | $5.625°$ | **52.015** | **3.320** | **0.981** | **0.983** |
| CirT | $1.5°$ | 55.505 | 5.478 | 0.980 | 0.958 |
| EIMP | $1.5°$ | 66.498 | 4.785 | 0.981 | 0.974 |
| TelePiT | $1.5°$ | 59.772 | 4.002 | 0.977 | 0.974 |
| AnisoCast | $1.5°$ | **51.956** | **3.255** | **0.982** | **0.982** |
| CirT | $0.25°$ | 56.124 | 5.682 | 0.977 | 0.956 |
| EIMP | $0.25°$ | 67.589 | 4.851 | 0.980 | 0.972 |
| TelePiT | $0.25°$ | 58.761 | 3.988 | 0.977 | 0.975 |
| AnisoCast | $0.25°$ | **51.262** | **3.188** | **0.983** | **0.982** |

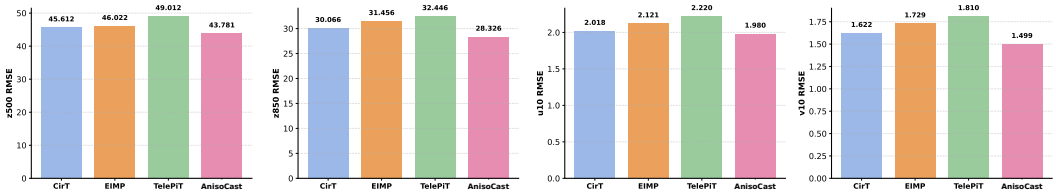


Figure 13: Evaluation over the non-polar latitude band $|\phi| \leq 60°$, excluding high-latitude regions.

$u10$ RMSE of $1.980$, and $v10$ RMSE of $1.499$, outperforming CirT, EIMP, and TelePiT. These results indicate that the advantage of AnisoCast does not rely solely on the high-latitude or polar regions, but remains evident under the standard S2S evaluation protocol after those regions are excluded.

# G Additional Results on Architectural and Forecasting Choices

## G.1 Full Ablation Results

Table 7: Full ablation study results of AnisoCast on weeks 3–4 and weeks 5–6.

| | **Model** | **RMSE** (↓) | | | | | | | **ACC** (↑) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Pressure Level** | | | | **Single Level** | | | **Pressure Level** | | | | **Single Level** | | |
| | | $z500(gpm)$ | $z850(gpm)$ | $t500(K)$ | $t850(K)$ | $t2m(K)$ | $u10(m/s)$ | $v10(m/s)$ | $z500(gpm)$ | $z850(gpm)$ | $t500(K)$ | $t850(K)$ | $t2m(K)$ | $u10(m/s)$ | $v10(m/s)$ |
| Weeks 3-4 | **AnisoCast** | **51.956** | **33.453** | **1.824** | **2.145** | **3.255** | **1.994** | **1.621** | **0.982** | **0.958** | **0.985** | **0.985** | **0.982** | **0.872** | **0.783** |
| | w/o AE | 57.088 | 38.211 | 2.481 | 2.630 | 6.801 | 3.012 | 2.531 | 0.953 | 0.952 | 0.948 | 0.920 | 0.957 | 0.833 | 0.732 |
| | w/o ZO | 61.207 | 39.118 | 2.467 | 2.742 | 4.061 | 3.053 | 3.441 | 0.970 | 0.939 | 0.953 | 0.934 | 0.961 | 0.818 | 0.721 |
| | w/o MO | 64.218 | 42.507 | 2.334 | 3.176 | 4.119 | 3.284 | 3.121 | 0.955 | 0.926 | 0.962 | 0.908 | 0.958 | 0.844 | 0.764 |
| | w/ unshared ZO | 54.124 | 35.021 | 1.940 | 2.217 | 3.421 | 2.128 | 1.822 | 0.968 | 0.956 | 0.977 | 0.980 | 0.976 | 0.860 | 0.776 |
| | w/o RN, w/ LN | 62.948 | 40.917 | 2.412 | 2.957 | 4.047 | 3.067 | 3.302 | 0.968 | 0.932 | 0.959 | 0.924 | 0.966 | 0.856 | 0.777 |
| | $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$ | 52.366 | 33.692 | 1.902 | 2.165 | 3.292 | 2.002 | 1.790 | 0.982 | 0.958 | 0.984 | 0.983 | 0.982 | 0.870 | 0.782 |
| Weeks 5-6 | **AnisoCast** | **52.268** | **33.437** | **1.830** | **2.156** | **3.231** | **1.993** | **1.606** | **0.982** | **0.958** | **0.985** | **0.985** | **0.982** | **0.873** | **0.786** |
| | w/o AE | 56.368 | 36.271 | 2.662 | 2.920 | 5.033 | 3.541 | 1.918 | 0.961 | 0.939 | 0.955 | 0.945 | 0.960 | 0.836 | 0.726 |
| | w/o ZO | 61.543 | 40.643 | 2.721 | 2.783 | 3.889 | 3.318 | 2.952 | 0.974 | 0.936 | 0.951 | 0.953 | 0.963 | 0.821 | 0.728 |
| | w/o MO | 62.872 | 44.019 | 2.407 | 2.792 | 3.771 | 3.544 | 2.386 | 0.961 | 0.949 | 0.960 | 0.938 | 0.971 | 0.846 | 0.759 |
| | w/ unshared ZO | 55.213 | 36.221 | 1.960 | 2.322 | 3.501 | 2.012 | 1.708 | 0.970 | 0.950 | 0.978 | 0.980 | 0.980 | 0.862 | 0.766 |
| | w/o RN, w/ LN | 63.318 | 42.674 | 2.581 | 2.979 | 4.107 | 3.702 | 2.561 | 0.970 | 0.952 | 0.967 | 0.936 | 0.973 | 0.857 | 0.772 |
| | $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$ | 54.447 | 34.690 | 1.843 | 2.166 | 3.341 | 2.008 | 1.705 | 0.981 | 0.957 | 0.983 | 0.983 | 0.982 | 0.872 | 0.785 |

Table 7 provides the full ablation results corresponding to the main-text ablation study. We report both RMSE and ACC across the two verification windows and all evaluated variables. The variants isolate the effects of ordered ring-wise embedding, latent zonal mixing, meridional sequence modeling, shared symmetric composition, and RMSNorm. Overall, each component contributes to the final performance, with consistent performance degradation observed when replacing the ordered anisotropic embedding with the matched order-agnostic ring embedding. Additionally, we test an alternative symmetric ordering by replacing $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ with $\mathcal{M} \circ \mathcal{Z} \circ \mathcal{M}$ in each Aniso Block. As

shown in Table 7, the adopted $\mathcal{Z} \circ \mathcal{M} \circ \mathcal{Z}$ ordering achieves lower RMSE for all seven variables over both weeks 3–4 and weeks 5–6, while attaining equal or higher ACC in every case. These results empirically support our choice of this ordering under the evaluated setting.

### G.2 Direct vs. Autoregressive Setting

Table 8: Comparison between autoregressive and direct S2S formulations. The direct formulation predicts the bi-weekly mean targets directly from the initial state, while the autoregressive variant recursively rolls out future states before computing the target averages.

| Paradigm | Period | z500 RMSE | z850 RMSE | t2m RMSE | v10 RMSE | z500 ACC | z850 ACC | t2m ACC | v10 ACC |
|---|---|---|---|---|---|---|---|---|---|
| AR | W3–4 | 56.122 | 40.112 | 5.881 | 2.421 | 0.952 | 0.951 | 0.954 | 0.730 |
| Direct | W3–4 | **51.956** | **33.453** | **3.255** | **1.621** | **0.982** | **0.958** | **0.982** | **0.783** |
| AR | W5–6 | 57.468 | 37.011 | 5.035 | 2.012 | 0.960 | 0.938 | 0.960 | 0.725 |
| Direct | W5–6 | **52.268** | **33.437** | **3.231** | **1.606** | **0.982** | **0.958** | **0.982** | **0.786** |

Recent direct S2S forecasting models commonly predict the target bi-weekly averages directly from the initial atmospheric state Liu et al. (2025a;b); Lyu et al. (2026), rather than generating daily forecasts autoregressively and then averaging them Pathak et al. (2022). To verify the effect of this formulation, we compare two variants of AnisoCast: an autoregressive variant (AR), which recursively rolls out future states before computing the bi-weekly targets, and our default direct variant (Direct), which predicts the weeks 3–4 and weeks 5–6 averages in a single forward pass. The results are reported in Table 8. Direct prediction consistently outperforms the autoregressive variant across both forecast windows and all evaluated variables. This suggests that directly optimizing the evaluation-aligned bi-weekly targets is more suitable for the S2S setting, where long autoregressive rollouts can accumulate errors and degrade the final averaged forecasts.

### G.3 Performance w.r.t. the number of Aniso Blocks

We evaluate the impact of network depth by varying the number of Aniso Blocks $L \in \{4, 5, 6, 7, 8, 9\}$. As can be seen from Figure 14, forecasting performance consistently improves as depth increases, peaking at $L = 7$ where RMSE reaches a minimum for key variables. Increasing the depth further yields diminishing returns or slight degradation. Thus, we select $L = 7$ to balance accuracy and computational efficiency.

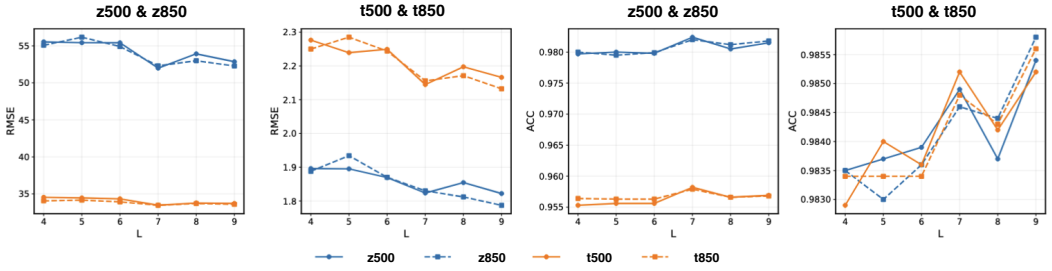


Figure 14: RMSE and ACC variations across different numbers of Aniso Blocks ($L$).

## H Complementary Forecast Diagnostics

### H.1 Spectral Energy Distribution Analysis

We further assess spectral energy distribution accuracy using SpecRes (Lyu et al., 2026) across all pressure levels for the prognostic variables $z$, $t$, $u$, and $v$ under weeks 3–4 and weeks 5–6 forecasts. SpecRes quantifies the root of the expected squared residual in the spectral domain, i.e., $\text{SpecRes} = \sqrt{\mathbb{E}_k\left[(\hat{S}'(k) - S'(k))^2\right]}$, where $\hat{S}'(k)$ and $S'(k)$ denote the normalized power spectra of prediction and target, respectively. Therefore, lower SpecRes values indicate a more accurate preservation of spectral energy distribution. As shown in Figure 15, AnisoCast consistently exhibits

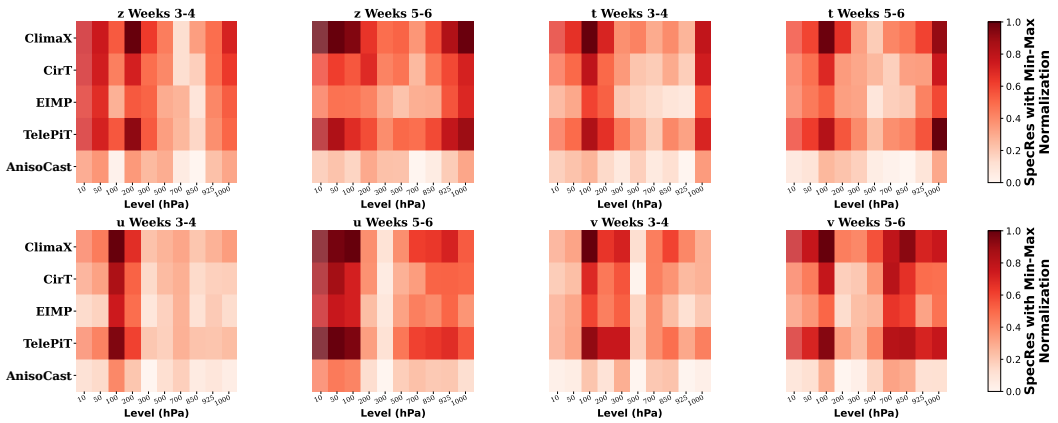


Figure 15: SpecRes comparison across pressure levels.

substantially lower SpecRes values than all competing models over nearly all vertical levels, indicating a more faithful preservation of spectral energy distribution at long forecast lead times. This indicates closer agreement with the target power spectra, complementing the grid-space RMSE evaluation.

### H.2 Vertical Profile Analysis

We further analyze the model's performance across the vertical atmospheric column to ensure physical consistency at different altitudes. Figure 16 compares the RMSE of $u, v$ against data-driven baselines, while Figure 17 benchmarks $z, t$ against operational NWP systems. AnisoCast achieves the lowest normalized RMSE (indicated by the lightest colors) across almost all 10 pressure levels, across ten pressure levels from 10 to 1000 hPa. This vertical consistency is particularly notable in comparison to models like OmniCast (Nguyen et al., 2025) and CirT (Liu et al., 2025a), which exhibit degradation in the upper atmosphere. The results demonstrate that AnisoCast's anisotropic embedding and operator design effectively capture the distinct dynamics of different atmospheric layers without requiring level-specific fine-tuning.

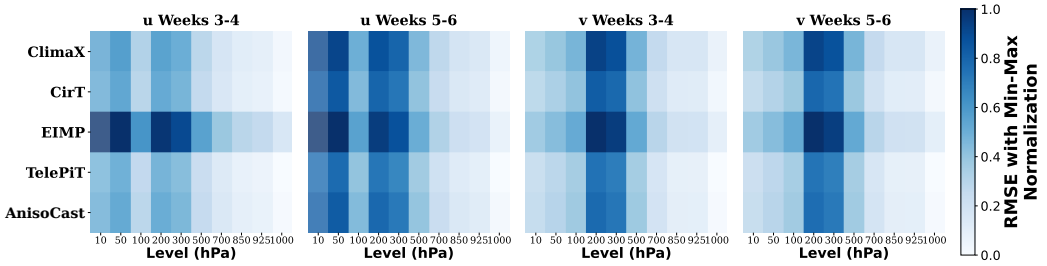


Figure 16: RMSE comparison on variables $u, v$ across different pressure levels.

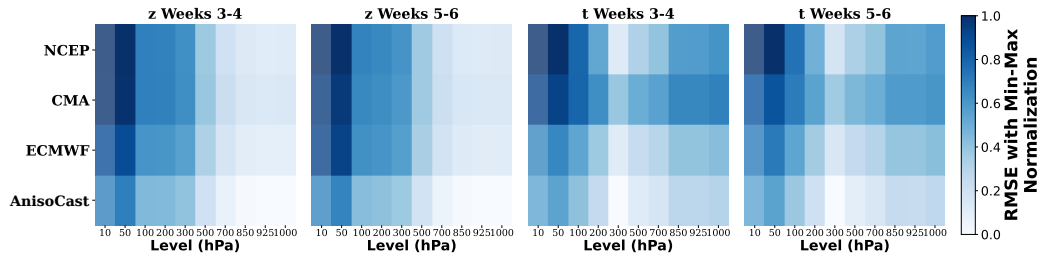


Figure 17: RMSE comparison on variables $z$ and $t$ across different pressure levels.

### H.3 Daily Visualization and Temporal Stability

To probe fine-grained temporal stability, we trained AnisoCast and other baselines to output the full daily trajectory. Figure 18 and Figure 19 visualize the day-by-day evolution of RMSE and ACC for weeks 3–4 (days 15–28) and weeks 5–6 (days 29–42), respectively. In the weeks 3–4 window (Figure 18), AnisoCast consistently maintains superior performance across all variables. While baselines suffer from rapid degradation over time (manifested as steep error slopes), AnisoCast exhibits a remarkably flat error growth rate. This advantage becomes even more pronounced in the extended weeks 5–6 window (Figure 19). For highly chaotic surface variables like $v10$, the performance gap between AnisoCast and the nearest competitor widens significantly as the lead time increases. These results support the temporal performance of the daily-output variant.

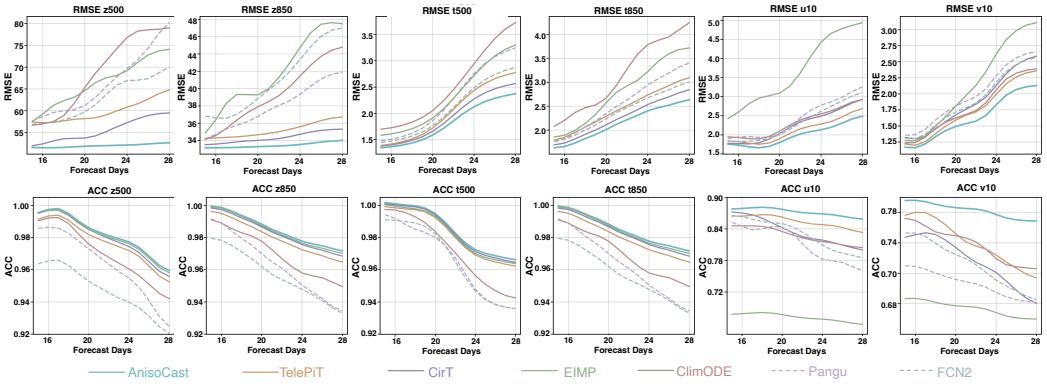


Figure 18: Daily RMSE and ACC variations across critical variables during weeks 3–4 (Days 15–28).

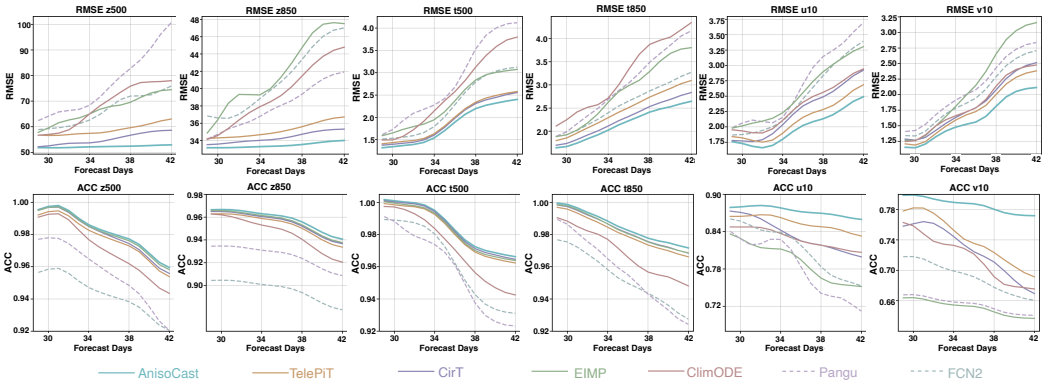


Figure 19: Daily RMSE and ACC variations across critical variables during weeks 5–6 (Days 29–42).

### H.4 Additional Global Visualizations

We provide expanded global error maps to visually assess spatial performance distributions across different variables and lead times. Figure 20 through Figure 26 display the RMSE distributions for $t850, t500, z500$, and $z850$ during weeks 3–4 and weeks 5–6. Consistent with the findings in the main text, AnisoCast exhibits deep blue patterns globally, indicating low error. The maps show pronounced error differences in several high-latitude regions, with AnisoCast exhibiting lower errors in the highlighted areas across both forecast windows. These comparisons assess spatial forecast accuracy rather than directly identifying the physical cause of the improvements.

## I Limitations

While AnisoCast achieves strong performance under the ERA5-based direct bi-weekly S2S benchmark, it currently relies on a deterministic regression objective and therefore does not explicitly

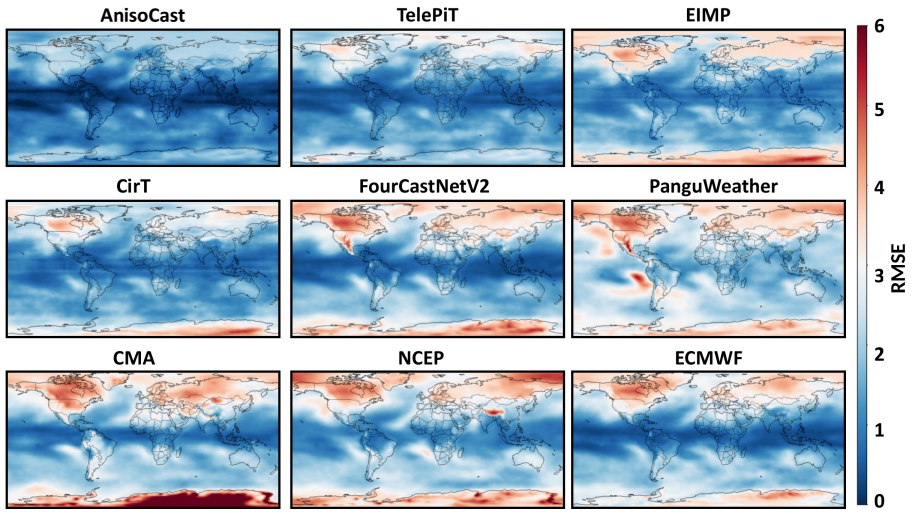


Figure 20: Global RMSE distribution for $t850$ at weeks 5–6 lead time.

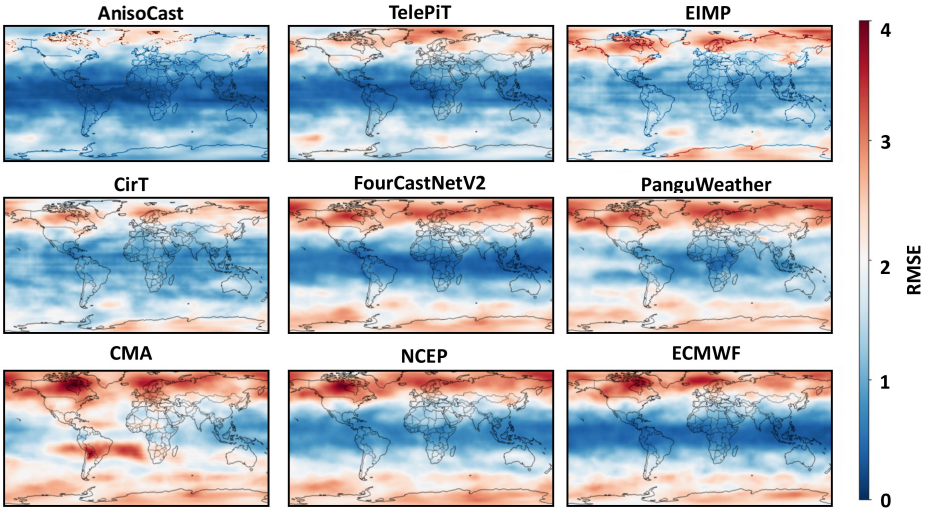


Figure 21: Global RMSE distribution for $t500$ at weeks 3–4 lead time.

quantify forecast uncertainty or provide probabilistic skill scores. Our evaluation focuses on grid-field accuracy metrics and selected deterministic diagnostics; it does not yet include a full set of process-oriented S2S diagnostics, such as MJO propagation skill and teleconnection fidelity. Future work will extend AnisoCast toward probabilistic or ensemble forecasting, process-oriented S2S verification, and higher-resolution operational evaluation.

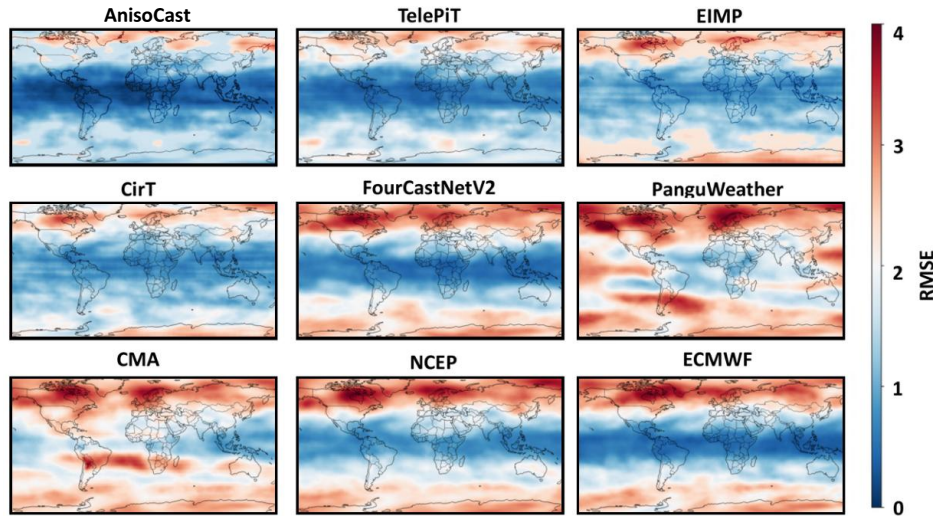


Figure 22: Global RMSE distribution for $t500$ at weeks 5–6 lead time.

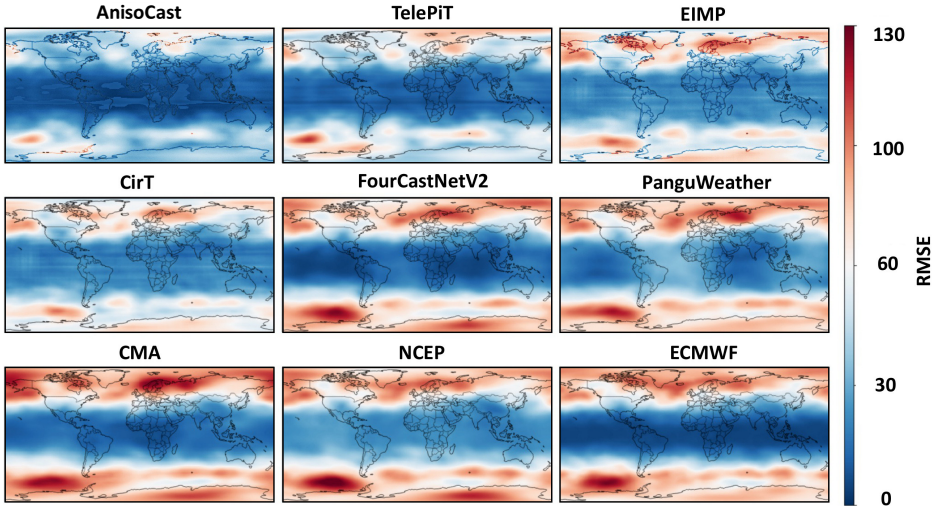


Figure 23: Global RMSE distribution for $z500$ at weeks 3–4 lead time.

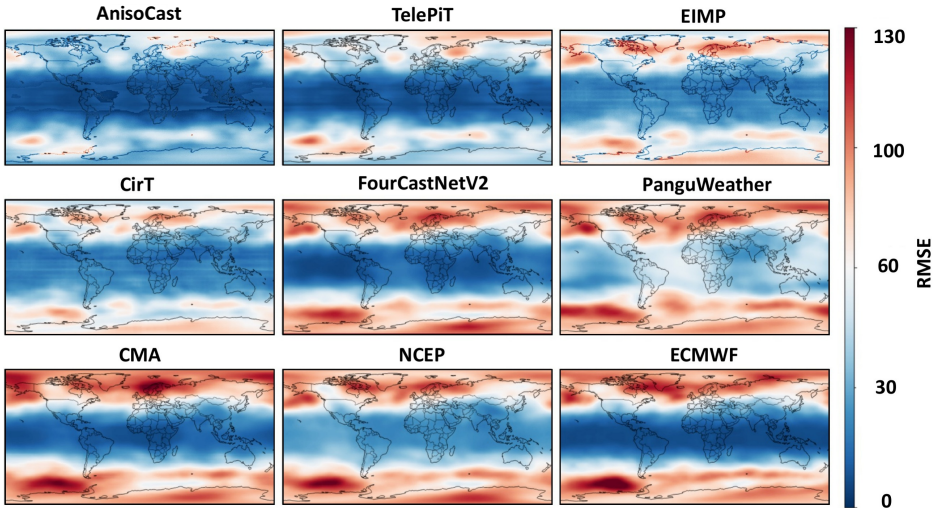


Figure 24: Global RMSE distribution for $z500$ at weeks 5–6 lead time.

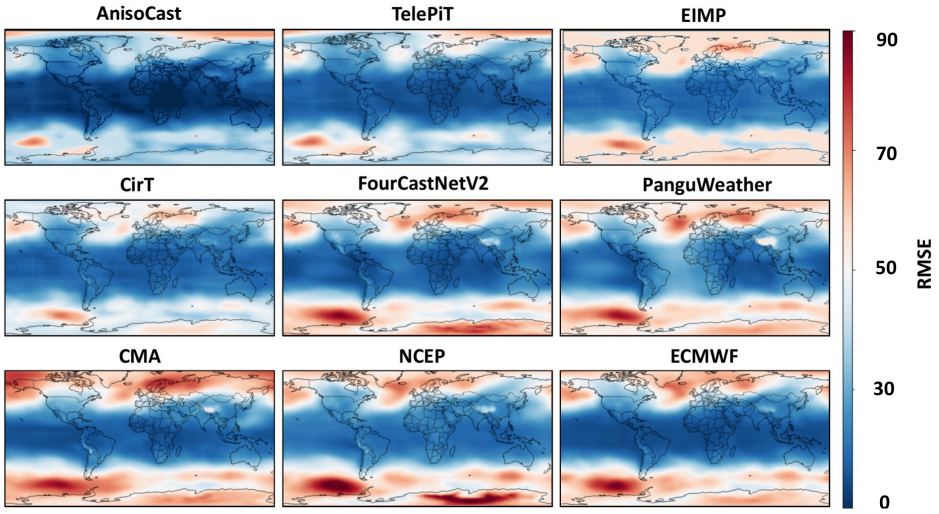


Figure 25: Global RMSE distribution for $z850$ at weeks 3–4 lead time.

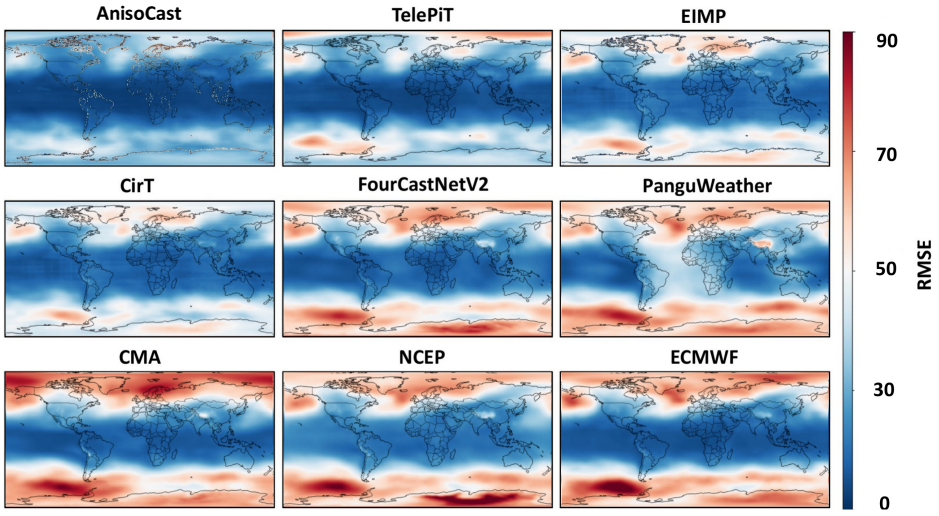


Figure 26: Global RMSE distribution for $z850$ at weeks 5–6 lead time.