ARGO: A REGIONAL-GLOBAL JOINT METHOD FOR REGION-FLEXIBLE WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Regional weather forecasting is crucial for many domains such as transportation, energy and disaster prevention. Existing data-driven methods typically perform global forecasting directly or build weather forecasting systems tailored to specific regions. However these paradigms present two key limitations: (1) generalization across regions with different geographical locations and spatial extents, and (2) filtering relevant global information to guide regional prediction. To address these challenges, we propose ARGO, a regional–global joint method for region-flexible weather forecasting. To support regions with varying locations and spatial extents, ARGO combines Geo-aware Position Encoding (GaPE), random regional sampling, and an auxiliary alignment loss to learn transferable forecasting patterns and compatible geographical representations. To filter global information, a Global Token Router (GTR) selects relevant nearby and remote tokens while retaining broad global context, followed by latent-space cross-attention for regional–global fusion. Experiments at 0.25° resolution show that ARGO outperforms state-of-the-art baselines in regional forecasting and generalizes across different regions without retraining from scratch. The code is available at: https://anonymous.4open.science/r/ARGO-8B6E.

1 INTRODUCTION

Weather is closely intertwined with human daily life (Tian et al., 2025; Allen et al., 2025). Various applications, including energy management (Xu et al., 2024; 2026) and transportation planning (Wolfson & Clark, 2006), focus on **regional weather forecasting** (RWF), which provides high-resolution, location-specific predictions of atmospheric evolution.

Traditional weather forecasting relies on numerical weather prediction (NWP). It solves atmospheric governing equations under initial and boundary conditions and external forcings related with target regions but requires labor-intensive expert design and calibration of physical parameterizations (Bauer et al., 2015; Skamarock & Klemp, 2008; Roberts et al., 2018). Recent data-driven models have advanced weather forecasting substantially (Bi et al., 2023; Lam et al., 2023; Nguyen et al., 2024; Tian et al., 2026), but mostly focus on **global weather forecasting** (GWF), leaving regional weather forecasting comparatively underexplored. A straightforward approach from global to regional forecasting is to first perform global forecasting and then crop the target region from the global prediction, as illustrated in Figure 1(a). However, the cropped forecast is limited by the coarse resolution of the GWF. Increasing the global resolution will incur substantial computational and memory costs over the entire globe. For example, increasing the resolution from $0.25°$ to $0.1°$ results in approximately $6.25\times$ more grid points. This motivates us to develop

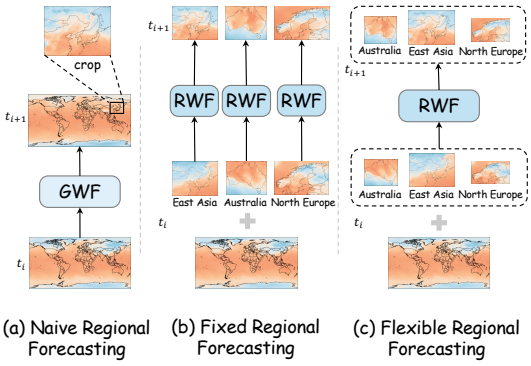


Figure 1: Different settings for regional weather forecasting. (a) Cropping from GWF. (b) Fixed regional-global joint forecasting. (c) Regional-global joint forecasting for arbitrary region.

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.

fine-grained regional weather forecasting guided by coarse global fields, which presents two important limitations:

Cross-Region Generalization. Generalizing across target regions requires a single model to produce accurate forecasts for regions with different geographical locations and spatial extents without being trained from scratch for each region. Existing regional-global forecasting methods are typically trained on a predefined target region, as illustrated in Figure 1(b) (Gao et al., 2025; Chen et al., 2026). Such training exposes the model to a restricted range of geographical conditions and regional–global relationships. Applying such methods to a new region often requires region-specific training or adaptation, making it costly to serve diverse regional forecasting needs. Therefore, this motivates us to develop a region-global joint forecasting model with cross-region generalization, as illustrated in Figure 1(c). It requires both learning fine-grained atmospheric evolving patterns across diverse regions and consistently representing their geographical relationships with the coarse global field.

Relevant Global Information Filtering. Regional atmospheric evolution depends not only on local conditions but also on large-scale circulation and remote atmospheric forcing, making global background information essential for accurate regional forecasting (Pathak et al., 2024). Existing approaches incorporate all available global information through input concatenation or representation fusion (Nguyen et al., 2025; Gao et al., 2025; Chen et al., 2026). Although these operations provide access to global context, they do not explicitly distinguish relevant information from redundant or weakly related signals prior to fusion. Consequently, passing excessive global information to the regional model may hinder its ability to identify useful predictive guidance. Conversely, restricting global predictions to a fixed neighborhood around the target region may exclude relevant remote processes. These limitations motivate filtering global information based on the current regional state while preserving the broader atmospheric context necessary for accurate prediction.

In this paper, we propose **ARGO**, **A R**egional–**G**lobal joint forecasting method for region-flexible weather f**O**recasting. To address the first challenge, ARGO combines Geo-aware Position Encoding (GaPE), random regional sampling, and an auxiliary alignment loss. We first pretrain a forecasting backbone on coarse global fields and provide global information for the regional model. GaPE represents regional and global token positions using their geographical coordinates and respects longitude periodicity, allowing their positional relationships to remain meaningful when the target region changes. During regional training, we randomly sample target regions so that the same model learns from different geographical conditions and their associated global backgrounds. The alignment loss further regularizes the regional and global encoder–head mappings toward a common zero-increment reference, encouraging compatible representations for their interaction. Together, these designs support the transfer of learned forecasting capabilities across target regions.

We introduce a **Global Token Router (GTR)** followed by cross-attention in the latent space. GTR partitions the regional grid into subregions and uses their pooled features to select remote global tokens according to the current weather state. It additionally retains nearby tokens based on spherical distance and supplements them with area-weighted background tokens, preserving local details, dynamic remote signals, and coarse global context. The resulting compact candidate set is processed by GaPE-equipped cross-attention. Although regional tokens within each subregion share the same candidates, they independently determine how to attend to them, preserving fine-grained regional–global interactions. This design avoids dense interaction with every global token while maintaining access to relevant global information. In summary, our contributions can be summarized as follows:

- We propose ARGO, **A R**egional–**G**lobal joint method for region-flexible weather f**O**recasting. It includes Geo-aware Position Encoding with Random Region Sampling and Auxiliary Alignment Loss, supporting cross-region generalization without training from scratch.
- To filter global information according to geographical proximity and the current weather state, we introduce Global Token Router. Followed by latent-space cross-attention, it can integrate the selected coarse global information into regional forecasts.
- All experiments are conducted at a spatial resolution of 0.25° for fine-grained regional forecasting and 1.5° for coarse-resolution global forecasting. The results demonstrate improved regional forecasting accuracy over state-of-the-art baselines and effective generalization to a target region excluded from regional training.

2 PRELIMINARIES

2.1 PROBLEM STATEMENT

Regional weather forecasting aims to predict future atmospheric states over a target region using both regional observations and the corresponding global background. Let $R_t \in \mathbb{R}^{c\times h\times w}$ and $G_t \in \mathbb{R}^{C\times H\times W}$ denote the regional and global states at time t, respectively, where c and C are the numbers of regional and global input channels. Here, $h \times w$ and $H \times W$ are the regional and global spatial grid sizes. Given a base forecast interval $\delta \in \{6\mathrm{h}, 12\mathrm{h}, 24\mathrm{h}\}$, we define a global forecasting model $g(\cdot;\delta)$ and a regional forecasting model $f(\cdot,\cdot;\delta)$. Rather than predicting the future states directly, they predict the corresponding weather increments:

$$g(G_t;\delta) = G_{t+\delta} - G_t, \qquad f(R_t, G_t;\delta) = R_{t+\delta} - R_t. \tag{1}$$

2.2 RELATED WORK

Global Weather Forecasting. Numerical weather prediction (NWP) forecasts atmospheric evolution by assimilating observations and numerically solving discretized governing equations. Operational systems such as the ECMWF Integrated Forecasting System (IFS) provide skillful medium-range forecasts, but require substantial computational resources at high spatial resolutions (Bauer et al., 2015; Roberts et al., 2018). With the availability of ERA5 (Hersbach et al., 2020) and WeatherBench2 (Rasp et al., 2024), data-driven forecasting has progressed rapidly. Representative methods include the neural-operator-based FourCastNet (Pathak et al., 2022), the Transformer-based Pangu-Weather (Bi et al., 2023), the graph-based GraphCast (Lam et al., 2023), and the cascaded FuXi system (Chen et al., 2024). These models achieve competitive global forecasts at substantially lower inference cost. However, most operate at a fixed global resolution, making it inefficient to obtain finer predictions only for selected regions; directly cropping global forecasts also cannot recover unresolved local structures.

Data-driven Regional Forecasting and Geographical Modeling. Regional forecasting resolves finer local dynamics and typically relies on large-scale background fields or boundary conditions from a global model (Skamarock & Klemp, 2008). Recent data-driven approaches either downscale coarse fields, as in CorrDiff (Mardani et al., 2025), or learn autoregressive regional dynamics with global guidance, as in StormCast (Pathak et al., 2024) and YingLong (Xu et al., 2025). STCast (Chen et al., 2026) introduces Spatial-Aligned Attention (SAA), which initializes a learnable regional–global distribution using geographical distance and refines it through cross-attention. Although SAA improves regional–global coupling beyond static boundaries, its learned spatial alignment is tied to a predefined regional domain and grid, making it difficult to transfer directly to unseen regions or spatial resolutions without retraining.

3 METHODOLOGY

3.1 MODEL OVERVIEW

To better incorporate global background information into high-resolution regional forecasts and enable generalization across target regions, we propose ARGO, a regional–global joint forecasting framework for arbitrary regional weather forecasting. As shown in Figure 2, ARGO comprises three main components:

- **Regional-Global Joint Weather Forecasting.** To provide an evolving global background for regional prediction, we pretrain a coarse-resolution global branch to predict weather increments at multiple forecast intervals. Its patch tokens pass through alternating global and window-based Transformer blocks, capturing large-scale atmospheric dependencies and local spatial structures. A high-resolution regional branch uses the same forecasting architecture and is trained for regional prediction. GaPE represents token locations through geographical coordinates, while random regional sampling exposes the regional branch to different locations and spatial extents.
- **Effective Global Information Selection.** Before each regional forecasting block, the Global Token Router (GTR) forms a compact candidate set for each regional subregion. It retains nearby global tokens, selects weather-relevant remote tokens according to the current features, and adds

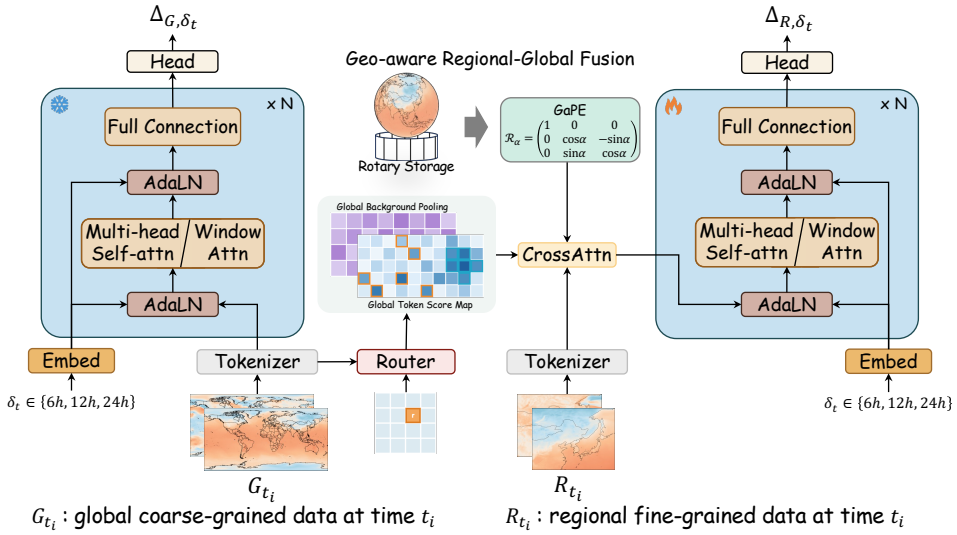


Figure 2: Overview of ARGO's three components. Regional–Global Joint Weather Forecasting uses a coarse global branch to guide high-resolution regional prediction, with GaPE encoding geographical token positions. Effective Global Information Selection combines GTR and selective cross-attention to fuse relevant global tokens into regional features. The Training Objective combines prediction and zero-increment reconstruction losses across global pre-training, regional training, and two-step fine-tuning.

pooled tokens that summarize the global background. Selective geo-aware cross-attention then uses regional tokens as queries and the selected global tokens as keys and values. GaPE encodes their geographical positions, while router scores bias the attention toward relevant candidates. Each regional token independently attends to the shared candidate set, preserving fine-grained interactions without processing the entire global grid.

- **Training Objective.** To encourage compatible global and regional token representations, we use a common objective that combines forecast-increment prediction with zero-increment reconstruction. Training then proceeds in three stages. We first pretrain the global branch on coarse-resolution data; next, we freeze it and train the regional branch and fusion modules for single-step prediction under the chosen regional sampling setting. Finally, we unfreeze the global branch and jointly fine-tune both branches over two autoregressive steps, helping them adapt to errors in their own forecasts.

3.2 Regional-Global Joint Weather Forecasting

3.2.1 Global Weather Forecasting Pre-training

The global and regional branches share the same forecasting backbone, as shown in Figure 2. Each weather field is divided into non-overlapping $P \times P$ patches and embedded as tokens. Alternating standard ViT and window-based Transformer blocks model large-scale dependencies and local structures, with each block conditioned on the forecast interval $\delta \in \{6\text{h}, 12\text{h}, 24\text{h}\}$. A linear head predicts the weather increment. We first pretrain the backbone on coarse global fields and use its weights to initialize regional forecasting. Tokenization and block details are given in Appendix B.1.

3.2.2 Regional Weather Forecasting

Geo-aware Position Encoding (GaPE). To encode the geographical relationships between regional and global tokens, we propose GaPE. For an angle θ, we define

$$p_\theta = (1, \cos\theta, \sin\theta)^\top, \qquad p_\alpha^\top p_\beta = 1 + \cos(\alpha - \beta). \tag{2}$$

Thus, similarity is periodic in the angular difference.

Equivalently, we use the rotation matrix

$$\mathcal{R}_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}, \tag{3}$$

since $p_{\alpha+\Delta\alpha} = \mathcal{R}_{\Delta\alpha} p_\alpha$, rotated queries and keys satisfy

$$(\mathcal{R}_\alpha q)^\top (\mathcal{R}_\beta k) = q^\top \mathcal{R}_{\beta-\alpha} k, \tag{4}$$

Hence attention depends on relative angle. With $d_h = 3\kappa_1$, we apply the rotation to each three-dimensional group:

$$\tilde{q} = \mathrm{diag}(\underbrace{\mathcal{R}_\alpha, \ldots, \mathcal{R}_\alpha}_{\kappa_1 \text{ blocks}}) q. \tag{5}$$

Figure 4 shows that GaPE preserves similarity across the longitude boundary, unlike standard RoPE (Su et al., 2024).

For longitude λ and latitude ϕ, we split an even head dimension equally, giving $d_h = 6\kappa$ and κ three-dimensional groups per coordinate. GaPE applies the corresponding rotation to each group:

$$\mathcal{G}(\lambda, \phi) = \mathrm{diag}(\underbrace{\mathcal{R}_\phi, \ldots, \mathcal{R}_\phi}_{\kappa \text{ blocks}}, \underbrace{\mathcal{R}_\lambda, \ldots, \mathcal{R}_\lambda}_{\kappa \text{ blocks}}), \qquad \tilde{q}_i = \mathcal{G}(\lambda_i, \phi_i) q_i, \qquad \tilde{k}_j = \mathcal{G}(\lambda_j, \phi_j) k_j. \tag{6}$$

Regional queries and global keys thus interact according to their relative geographical positions.

Random Regional Sampling. Because GaPE derives positional relationships from geographical coordinates, the regional model can process windows at different locations. To exploit this flexibility and learn weather patterns across regions, we consider three regional training settings. In full-shot, the regional training data always come from a fixed target region. In few-shot, we randomly sample crops across the global 0.25° grid, with heights of 64–150 latitude grid points and widths of 80–200 longitude grid points; these crops may overlap the target region. Zero-shot uses the same sampling range but rejects crops intersecting the target region or its 20-grid-point buffer. All three settings are validated and tested on the same fixed target region.

3.3 Effective Global Information Selection

3.3.1 Global Token Router (GTR)

To capture global information efficiently, we introduce a global token router (GTR). At each forecasting block, GTR builds a compact global candidate set for each regional subregion. Let $x_{R,i}, x_{G,j} \in \mathbb{R}^d$ denote regional and global token features, omitting the layer index. We divide the regional grid into 4×4 subregions $\{\mathcal{I}_r\}$ and mean-pool their valid features. A lightweight router scores each global token by content relevance:

$$\bar{x}_{R,r} = \frac{1}{|\mathcal{I}_r|} \sum_{i \in \mathcal{I}_r} x_{R,i}, \quad u_{rj} = \frac{\big(W_q^s \,\mathrm{LN}(\bar{x}_{R,r})\big)^\top \big(W_k^s \,\mathrm{LN}(x_{G,j})\big)}{\sqrt{d_s}}, \quad s_{rj} = 2\tanh\Big(\frac{u_{rj}}{2}\Big). \tag{7}$$

Here $d_s = 32$ is the projection dimension; the bounded score s_{rj} selects candidates and biases cross-attention.

Nearby and remote tokens. The normalized mean of a subregion's three-dimensional unit position vectors defines its spherical center. Its $K_n = 64$ nearest global tokens by great-circle distance form $\mathcal{N}_r$; from the remainder, we select the $K_f = 128$ highest-scoring tokens:

$$\mathcal{F}_r = \mathrm{TopK}_{j \notin \mathcal{N}_r}(s_{rj}, K_f). \tag{8}$$

Here, *remote* means outside $\mathcal{N}_r$; selection depends on current features and is repeated at each block.

Global background. To retain broad context, we divide the global token grid into 4×8 cells $\{\mathcal{C}_m\}$ and pool each with latitude-based area weights:

$$b_m = \sum_{j \in \mathcal{C}_m} w_{mj} x_{G,j}, \qquad w_{mj} = \frac{\max(\cos\phi_j, \epsilon)}{\sum_{j' \in \mathcal{C}_m} \max(\cos\phi_{j'}, \epsilon)}, \tag{9}$$

Here, ϕ_j is latitude and $\epsilon = 10^{-6}$. Each pooled token uses its normalized, area-weighted spherical centroid as its position. The 32 background tokens form $\mathcal{B}$, giving $\mathcal{S}_r = \mathcal{N}_r \cup \mathcal{F}_r \cup \mathcal{B}$ with 224 candidates. Tokens in one subregion share this set but attend independently. Background scores use equation 7 with b_m in place of $x_{G,j}$.

Training the router. From scratch, we use all global tokens plus background for five epochs, then reduce the remote budget through 512 and 256 (five epochs each) to 128. Sparse training reserves eight remote slots for random unselected candidates; inference uses only top-scoring candidates. Later fine-tuning starts with the final budget. The additive attention bias provides gradients to selected scores, but not to discrete Top-k indices.

3.3.2 Selective Geo-aware Cross-Attention

To fuse the GTR candidates with regional features, we apply a cross-attention before each regional block. Regional features serve as queries, while the candidates constructed in Section 3.3.1 provide keys and values. GaPE encodes their geographical relationships, and the router scores provide a subregion-specific relevance bias.

Let $X_R^{(n)} \in \mathbb{R}^{N_R \times d}$ and $X_G^{(n)} \in \mathbb{R}^{N_G \times d}$ denote the regional and global features at the input of the n-th block. For attention head a, let $q_{i,a}^{(n)}$ denote the query of regional token $i \in \mathcal{I}_r$, and let $k_{j,a}^{(n)}$ and $v_{j,a}^{(n)}$ denote the key and value of candidate token $j \in \mathcal{S}_r$. GaPE is applied according to their geographical coordinates:

$$\tilde{q}_{i,a}^{(n)} = \mathcal{G}(\lambda_i, \phi_i) q_{i,a}^{(n)}, \qquad \tilde{k}_{j,a}^{(n)} = \mathcal{G}(\lambda_j, \phi_j) k_{j,a}^{(n)}. \tag{10}$$

The resulting selective cross-attention is

$$E_{i,a}^{(n)} = \sum_{j \in \mathcal{S}_r} \mathrm{Softmax}_{j \in \mathcal{S}_r} \left(\frac{(\tilde{q}_{i,a}^{(n)})^\top \tilde{k}_{j,a}^{(n)}}{\sqrt{d_h}} + s_{rj}^{(n)} \right) v_{j,a}^{(n)}, \qquad i \in \mathcal{I}_r, \tag{11}$$

where d_h is the attention-head dimension. The candidate set and score bias are shared across heads, while each regional token retains its own query-dependent attention distribution. Concatenating the head outputs, applying an output projection $W_O^{(n)}$, and restoring the regional token order gives the residual update

$$\tilde{X}_{R,r}^{(n)} = X_{R,r}^{(n)} + \left[\mathrm{Concat}_a \big(E_{i,a}^{(n)} \big) \right]_{i \in \mathcal{I}_r} W_O^{(n)}. \tag{12}$$

The enhanced regional features are then fed into the subsequent forecasting block. With $M = 16$ subregions and $K = |\mathcal{S}_r| = 224$ candidates per subregion, router scoring costs $O(MN_G d_s)$ and fine attention costs $O(N_R K d)$, compared with $O(N_R N_G d)$ for dense attention, excluding feature projections and pooling.

3.4 Training Objective

Regional–global fusion requires token representations from the two branches to be compatible. To help keep them in a shared latent space, we initialize the regional branch from the global backbone and use the same forecasting and zero-increment reconstruction objective for both. For branch $b \in \{G, R\}$,

$$\mathcal{L}^b = (1 - \eta_{\mathrm{rec}}) \mathcal{L}_{\mathrm{pred}}^b + \eta_{\mathrm{rec}} \mathcal{L}_{\mathrm{rec}}^b, \tag{13}$$

where $\mathcal{L}_{\mathrm{pred}}^b$ measures forecast-increment error and $\mathcal{L}_{\mathrm{rec}}^b$ encourages a zero increment when the forecasting blocks are bypassed. The coefficient η_{rec} balances the two terms; details are given in Appendix B.2. We apply this objective across three training stages:

Stage 1: Global Pre-training. We first train the global backbone on coarse-resolution data for single-step weather-increment prediction at the base forecast intervals (Section 3.2.1), optimizing $\mathcal{L}^G$. Its weights initialize the regional branch in Stage 2.

Stage 2: Regional Training. Then, we freeze the pretrained global branch and train the regional branch and fusion modules for single-step high-resolution prediction, optimizing $\mathcal{L}^R$. The global branch supplies background features for regional fusion at each forecasting block. Regional inputs follow the full-shot, few-shot, or zero-shot sampling scheme in Section 3.2.2.

Table 1: Overall prediction performance comparison. The best and second-best results are highlighted in **bold** and underline, respectively.

Methods		Pangu		Stormer		OneForecast		STCast		ARGO	
Variable	Lead Time	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
T2m	1-day	1.45	0.88	1.30	0.91	1.73	0.84	1.32	0.91	**1.02**	**0.94**
	4-day	3.34	0.41	3.30	0.43	2.53	0.70	2.70	0.63	**1.77**	**0.83**
	7-day	4.30	0.20	4.04	0.22	3.31	0.51	3.35	0.44	**2.59**	**0.63**
	10-day	4.75	0.15	4.34	0.14	4.12	0.29	3.87	0.29	**3.21**	**0.42**
U10	1-day	1.60	0.86	1.43	0.89	1.35	0.90	1.33	0.91	**1.03**	**0.94**
	4-day	3.48	0.27	3.48	0.27	2.26	0.72	3.11	0.47	**2.07**	**0.76**
	7-day	4.00	0.08	3.79	0.08	3.25	0.44	3.74	0.27	**3.00**	**0.48**
	10-day	4.16	0.06	3.88	0.03	3.93	0.23	4.06	0.16	**3.43**	**0.27**
V10	1-day	1.61	0.86	1.46	0.89	1.39	0.90	1.33	0.91	**1.05**	**0.94**
	4-day	3.65	0.25	3.64	0.25	2.34	0.72	3.14	0.49	**2.11**	**0.76**
	7-day	4.19	0.05	3.96	0.03	3.39	0.43	3.82	0.27	**3.10**	**0.47**
	10-day	4.41	0.02	3.95	0.01	4.10	0.20	4.25	0.13	**3.54**	**0.26**
MSLP	1-day	227.39	0.91	207.83	0.93	104.27	0.98	154.19	0.96	**73.18**	**0.99**
	4-day	624.87	0.31	620.01	0.32	265.01	0.88	428.26	0.68	**230.54**	**0.91**
	7-day	768.57	0.03	701.66	0.06	476.93	0.65	572.91	0.43	**442.22**	**0.66**
	10-day	805.07	0.00	714.08	0.00	657.73	0.38	663.30	0.27	**564.75**	**0.42**
Z500	1-day	213.55	0.94	194.80	0.95	75.80	0.99	140.33	0.98	**60.66**	**1.00**
	4-day	658.60	0.45	677.99	0.44	257.16	0.92	456.51	0.74	**236.76**	**0.93**
	7-day	798.38	0.22	800.20	0.23	496.50	0.71	599.86	0.56	**462.51**	**0.73**
	10-day	849.66	0.17	827.11	0.17	715.83	0.44	719.86	0.38	**622.50**	**0.48**
T850	1-day	1.38	0.92	1.22	0.94	1.07	0.95	1.05	0.96	**0.77**	**0.98**
	4-day	4.03	0.39	4.01	0.39	1.94	0.85	2.98	0.66	**1.71**	**0.88**
	7-day	5.26	0.14	4.87	0.15	3.16	0.63	3.84	0.44	**2.84**	**0.67**
	10-day	5.79	0.10	5.13	0.08	4.30	0.35	4.50	0.27	**3.71**	**0.41**
Aggregate	1-day	49.25	0.87	45.37	0.89	20.01	0.93	33.04	0.91	**16.01**	**0.95**
	4-day	147.60	0.35	150.37	0.34	59.23	0.80	100.62	0.61	**52.77**	**0.82**
	7-day	180.56	0.15	177.96	0.15	108.70	0.59	132.28	0.43	**99.70**	**0.62**
	10-day	192.87	0.09	185.41	0.09	156.24	0.36	157.41	0.29	**135.76**	**0.40**

Stage 3: Two-Step Fine-tuning. Finally, we unfreeze the global branch and unroll both branches for two autoregressive steps, feeding their first predictions into the second step. We jointly update both branches and the fusion modules by optimizing their objectives over both steps, exposing the model to errors in its own forecasts.

The full training procedure is given in Algorithm 1.

4 Experiments

4.1 Experiment Settings

Datasets. We use ERA5 data from WeatherBench2 at 6-hour intervals, with global and regional resolutions of 1.5° (121×240) and 0.25° (721×1440), respectively. Both inputs contain five atmospheric variables (Z, Q, T, U, and V) at 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa) and four surface variables (T2m, U10, V10, and MSLP), with seven auxiliary channels included only in the regional input (land–sea mask, surface geopotential, latitude, longitude, local time of day, year progress, and top-of-atmosphere incident solar radiation). We use 2010–2018 for training, 2019 for validation, and 2020 for testing, and report results on T2m, U10, V10, MSLP, Z500, T850 and Aggregate (average over all values).

Sampling Settings. We compare three regional training settings. ARGO (full-shot) uses a fixed East Asian domain (25.25°–59.75°N, 102.75°–149.75°E), whereas ARGO few-shot samples regions randomly without excluding East Asia. ARGO zero-shot excludes East Asia from regional training and is evaluated there without target-region fine-tuning; the coarse global background still covers the entire globe. At evaluation, all three variants receive the same high-resolution East Asian initial state and coarse-resolution global initial state.

Baselines. We evaluate ARGO against baselines from two categories: 1) Graph Neural Network-based Methods: OneForecast (Gao et al., 2025). 2) Transformer-based Methods: Pangu-Weather (Bi et al., 2023), Stormer (Nguyen et al., 2024), and STCast (Chen et al., 2026). Since Pangu-Weather

Table 2: Cross-region generalization performance comparison of aggregated RMSE and ACC.

Model	6-hour		1-day		3-day		5-day		7-day	
	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
OneForecast	12.32	0.96	20.01	0.93	44.80	**0.86**	74.81	0.74	108.70	0.59
STCast	11.21	**0.97**	33.04	0.91	84.23	0.69	112.09	0.54	132.28	0.43
ARGO few-shot	**9.58**	**0.97**	**17.01**	**0.94**	**40.16**	**0.86**	**69.84**	**0.75**	**102.18**	**0.61**
ARGO zero-shot	9.89	**0.97**	17.60	**0.94**	41.78	**0.86**	71.53	0.74	102.70	0.60

and Stormer do not include a regional–global interaction module, we train them directly on the high-resolution regional data.

Evaluation Metrics. We evaluate de-normalized predictions at each lead time separately using latitude-weighted root mean squared error (RMSE) and anomaly correlation coefficient (ACC). Lower RMSE and higher ACC indicate better performance; detailed definitions are provided in Appendix D.

4.2 Performance Comparison

Overall Performance. Table 1 compares ARGO with baseline models across six meteorological variables at lead times of 1, 4, 7, and 10 days. ARGO achieves the lowest RMSE across all evaluated variables and lead times, while matching or exceeding the ACC of all baselines. At the 4-day lead time, ARGO reduces aggregate RMSE from 59.23 for the strongest baseline, OneForecast, to 52.77 (a 10.9% reduction), while improving aggregate ACC from 0.80 to 0.82.

Generalization Ability. Table 2 evaluates ARGO on East Asia under few-shot and zero-shot regional training. Although East Asia is excluded from zero-shot regional training, its aggregate RMSE is at most 4.1% higher than few-shot across the five lead times, with ACC differing by at most 0.01. At 7 days, zero-shot achieves an RMSE of 102.70 and an ACC of 0.60, compared with 108.70 and 0.59 for the strongest baseline, OneForecast. These results show that ARGO retains forecasting accuracy on a held-out region without target-region fine-tuning. Detailed few-shot and zero-shot results for each variable are provided in Table 5.

From Tables 1 and 2, we make three observations: 1) Pangu-Weather and Stormer, trained only on regional inputs, fall further behind at longer lead times, underscoring the value of global context for regional prediction. 2) OneForecast outperforms STCast in aggregate at lead times of one day or longer, but both trail ARGO; access to global information alone is therefore insufficient to match ARGO's regional–global fusion. 3) ARGO few-shot and zero-shot remain close on East Asia even when zero-shot excludes the target region from regional training, indicating that its forecasting ability extends beyond a fixed domain. Together, these results support effective regional–global fusion and transferable geographical modeling.

4.3 Ablation Studies

Effect of GTR. To assess the effect of global token routing, we replace GTR with dense cross-attention over all global tokens (w/o GTR). As shown in Table 3, ARGO few-shot matches or improves both RMSE and ACC for all four variables at 72 hours. In particular, MSLP ACC increases from 0.93 to 0.95, while T2m RMSE decreases from 1.63 to 1.61. These results suggest that the selected global candidates retain useful information for regional prediction.

Effect of GaPE. To examine GaPE-based regional–global fusion, we replace it with Spatial-Aligned Attention (w SAA). Table 3 shows higher RMSE for all four variables under this replacement; T2m RMSE rises from 1.61 to 1.69 and MSLP RMSE from 174.70 to 182.75. ACC also decreases for three variables, while Z500 remains unchanged at the reported precision. These results favor fine-grained regional–global interactions over the shared global summaries used by SAA in this setting.

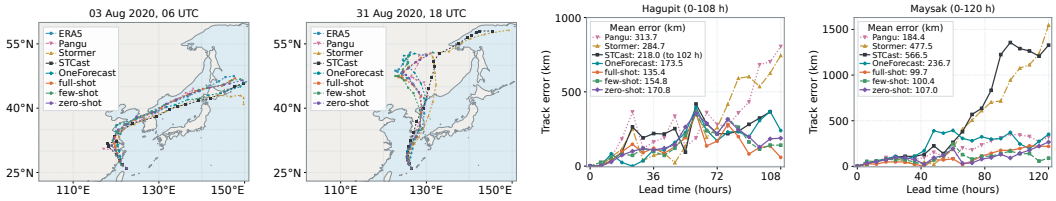
Effect of two-step fine-tuning. To evaluate this training stage, we retain only single-step regional adaptation (w/o two-step fine-tuning). As shown in Table 3, removing two-step fine-tuning increases RMSE for all four variables at 72 hours; Z500 rises from 169.18 to 175.12 and MSLP from 174.70 to 178.20. T2m and U10 ACC each decrease by 0.01, while MSLP and Z500 ACC remain unchanged. These results suggest that training on the model's own predictions improves its multi-step forecasts.

Table 3: Effect of GTR, GaPE and two-step fine-tuning.

Model	T2m (72 h)		U10 (72 h)		MSLP (72 h)		Z500 (72 h)	
	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑	RMSE↓	ACC↑
w/o GTR	1.63	0.85	**1.81**	0.81	174.89	0.93	169.45	**0.96**
w SAA	1.69	0.84	1.88	0.80	182.75	0.94	175.07	**0.96**
w/o two-step fine-tuning	1.64	0.85	1.83	0.81	178.20	**0.95**	175.12	**0.96**
ARGO few-shot	**1.61**	**0.86**	**1.81**	**0.82**	**174.70**	**0.95**	**169.18**	**0.96**

4.4 Case Studies

Extreme weather events such as typhoons threaten human lives and cause substantial socioeconomic disruption (Young & Hsiang, 2024; Newman & Noy, 2023). We examine two representative typhoons in 2020, Hagupit (Shu et al., 2023) and Maysak (Yang et al., 2025), to assess ARGO's ability to capture the evolution of regional weather systems. Figure 3 compares predicted trajectories with ERA5-derived reference tracks. All models use fixed 6-hour autoregression in these cases.



(a) Hagupit (2020.08) (b) Maysak (2020.08) (c) Hagupit track error (d) Maysak track error

Figure 3: Typhoon tracks and errors over East Asia. (a,b) Hagupit (108 h) and Maysak (120 h) trajectories at 6-hour intervals. (c,d) Corresponding track errors relative to ERA5; legends report mean errors over each extracted track. STCast's Hagupit track reaches the eastern boundary at 102 h.

As shown in Figure 3(a,b), the ARGO variants follow both storms' northward motion and subsequent turns, while several baselines drift away at later lead times. The error curves in Figure 3(c,d) show broadly comparable short-range performance but clearer separation as the forecast extends. For Hagupit, OneForecast remains competitive, yet ARGO full-shot reduces the mean 6–108-hour track error from 173.5 to 135.4 km; few-shot and zero-shot achieve 154.8 and 170.8 km, respectively. For Maysak, all three ARGO variants achieve mean 6–120-hour errors of 99.7–107.0 km, compared with 184.4 km for the next-best Pangu-Weather. These cases illustrate that ARGO retains useful track-forecasting skill across distinct storm paths, including when East Asia is excluded from regional training.

5 Conclusion

In this paper, we present ARGO, a regional–global joint forecasting framework designed for high-resolution weather prediction over arbitrary target regions. ARGO combines a coarse-resolution global forecasting backbone with geo-aware regional–global fusion. Geo-aware Position Encoding represents geographical relationships directly from longitude and latitude, allowing positional representations to accommodate changes in the target region. The Global Token Router combines nearby tokens, dynamically selected remote tokens, and pooled global background tokens to reduce the cost of cross-attention while preserving fine-grained interactions with selected global locations. Experiments on ERA5 data show competitive performance retained in the few-shot setting. Future work will investigate robustness across more diverse regions and spatial resolutions, as well as mechanisms for better capturing localized extreme weather events.

AI use statement

In this work, we have not used generative AI tools for "generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingre-

dients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results". We take responsibility for the final content of this work, including text, claims or artifacts.

REFERENCES

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nat.*, 641(8065):1172–1179, 2025. doi: 10.1038/S41586-025-08897-0.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015. doi: 10.1038/nature14956.

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

Hao Chen, Tao Han, Jie Zhang, Song Guo, and Lei Bai. Stcast: Adaptive boundary alignment for global and regional weather forecasting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2026.

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*, 7:6, 2024. doi: 10.1038/s41612-023-00512-1.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 18658–18697. PMLR, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

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

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6:124, 2025. doi: 10.1038/s43247-025-02042-5.

Rebecca Newman and Ilan Noy. The global costs of extreme weather that are attributable to climate change. *Nature Communications*, 14:6103, 2023. doi: 10.1038/s41467-023-41888-1.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2196.

Tung Nguyen, Harkanwar Singh, Nilay Naharas, Lucas Bandarkar, and Aditya Grover. IndiaWeatherBench: A dataset and benchmark for data-driven regional weather forecasting over india. *arXiv preprint arXiv:2509.00653*, 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 preprint arXiv:2202.11214*, 2022.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, and Michael Pritchard. Kilometer-scale convection allowing model emulation using generative diffusion modeling. *arXiv preprint arXiv:2408.10958*, 2024.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Christopher D Roberts, Retish Senan, Franco Molteni, Souhail Boussetta, Michael Mayer, and Sarah PE Keeley. Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Geoscientific model development*, 11(9):3681–3712, 2018.

Aiqing Shu, Feifei Shen, Lipeng Jiang, Tao Zhang, and Dongmei Xu. Assimilation of clear-sky FY-4A AGRI radiances within the WRFDA system for the prediction of a landfalling typhoon Hagupit (2020). *Atmospheric Research*, 283:106556, 2023. doi: 10.1016/j.atmosres.2022.106556.

William C. Skamarock and Joseph B. Klemp. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227(7):3465–3485, 2008. doi: 10.1016/j.jcp.2007.01.037.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

J. Tian, Y. Liang, R. Xu, P. Chen, C. Guo, A. Zhou, L. Pan, Z. Rao, and B. Yang. Air quality prediction with physics-guided dual neural ODEs in open systems. In *International Conference on Learning Representations*, 2025.

Jindong Tian, Yifei Ding, Ronghui Xu, Hao Miao, Chenjuan Guo, and Bin Yang. Arrow: An adaptive rollout and routing method for global weather forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 54823–54847, 2026.

Marilyn M. Wolfson and David A. Clark. Advanced aviation weather forecasts. *Lincoln Laboratory Journal*, 16(1):31–58, 2006.

Pengbo Xu, Xiaogu Zheng, Tianyan Gao, Yu Wang, Junping Yin, Juan Zhang, Xuanze Zhang, San Luo, Zhonglei Wang, Zhimin Zhang, et al. An artificial intelligence-based limited area model for forecasting of surface meteorological variables. *Communications Earth & Environment*, 6:372, 2025. doi: 10.1038/s43247-025-02347-5.

Ronghui Xu, Tongxin Wu, Guozhen Zhang, Yihan Li, Chenjuan Guo, Bin Yang, and Yong Li. Uniwind: Toward unified day-ahead wind power forecasting via physics-informed state routing. *CoRR*, abs/2607.01670, 2026.

Shiwei Xu, Yongjun Wang, Xinglei Xu, Guang Shi, Yingya Zheng, He Huang, and Chengqiu Hong. A multi-step wind power group forecasting seq2seq architecture with spatial–temporal feature fusion and numerical weather prediction correction. *Energy*, pp. 130352, 2024. doi: 10.1016/j.energy.2024.130352.

Sinil Yang, Hyo-Jun Bae, Mark Bourassa, Chaehyeon Chelsea Nam, Steven Cocke, D. W. Shin, Benjamin W. Barr, Hyodae Seo, Dong-Hyun Cha, Min-Ho Kwon, Daehyun Kim, Moon-Soo Park, Kwang-Young Jeong, and Baek-Min Kim. Sea spray effects on typhoon prediction in the Yellow and East China Seas: case studies using a coupled atmosphere-ocean-wave model for Lingling (2019) and Maysak (2020). *Environmental Research Letters*, 20(5):054028, 2025. doi: 10.1088/1748-9326/adc616.

Rachel Young and Solomon Hsiang. Mortality caused by tropical cyclones in the United States. *Nature*, 635:121–128, 2024. doi: 10.1038/s41586-024-07945-5.

APPENDIX

A NOTATION

Table 4: Summary of the main notations used in the paper.

Symbol	Description
δ	Base forecast interval.
c, C	Numbers of regional and global input channels, respectively.
h, w and H, W	Regional and global spatial grid sizes, respectively.
R_t, G_t	Regional and global weather states at time t, in $\mathbb{R}^{c\times h\times w}$ and $\mathbb{R}^{C\times H\times W}$.
f, g	Regional and global forecasting models, respectively.
P, d, d_s, d_h	Patch size, token dimension, router dimension, and attention-head dimension.
N, n, $\Phi^{(n)}$	Number of forecasting blocks, block index, and the mapping of the n-th block.
λ_i, ϕ_i	Longitude and latitude of token i.
$\mathcal{R}_\theta$	Rotation matrix of the one-dimensional GaPE at angle θ.
$\mathcal{G}(\lambda,\phi)$	Two-dimensional GaPE transformation at coordinate (λ,ϕ).
$X_R^{(n)}$, $X_G^{(n)}$	Regional and global token representations at forecasting block n.
$x_{R,i}$, $x_{G,j}$	Individual regional and global token features.
$\mathcal{I}_r$, $\bar{x}_{R,r}$	Token set and pooled feature representation of regional subregion r.
s_{rj}	Routing score between regional subregion r and global token j.
K_n, K_f	Numbers of retained nearby and dynamically selected remote tokens.
$\mathcal{N}_r$, $\mathcal{F}_r$, $\mathcal{B}$	Nearby, dynamically selected remote, and pooled global background token sets.
$\mathcal{S}_r$	Candidate set used by subregion r for cross-attention, $\mathcal{S}_r=\mathcal{N}_r\cup\mathcal{F}_r\cup\mathcal{B}$.
N_R, N_G, M, K	Numbers of regional tokens, global tokens, subregions, and candidates per subregion.

B BACKBONE AND REGIONAL TRAINING DETAILS

B.1 GLOBAL FORECASTING BACKBONE

B.1.1 WEATHER TOKENIZATION AND LEAD-TIME CONDITIONING

Given a global field G_t or a regional field R_t, we partition it into non-overlapping $P\times P$ patches. Each patch is flattened and linearly projected into a d-dimensional token, producing spatial token grids for the two domains. This maps fields with different spatial extents into the representation used by the forecasting blocks while preserving their spatial arrangement. The forecast interval $\delta\in\{6\text{h},12\text{h},24\text{h}\}$ is encoded as an embedding e_δ and injected into each block, allowing the same backbone to predict increments over different lead times.

B.1.2 FORECASTING BLOCKS AND GLOBAL PRE-TRAINING

The resulting weather tokens and lead-time embedding are then processed by alternating standard ViT and window-based Transformer blocks, which model large-scale atmospheric dependencies and finer local structures, respectively. If $X^{(n)}$ is the input to block n, its update is

$$X^{(n+1)}=\Phi^{(n)}(X^{(n)},e_\delta), \tag{14}$$

where $\Phi^{(n)}$ is either a standard ViT or a window-based Transformer block under lead-time conditioning. After N blocks, a linear head maps the latent representation to a meteorological increment. The backbone is first trained on coarse global fields; its learned weights initialize the regional branch.

B.2 REGIONAL FINE-TUNING AND LOSSES

During single-step regional training, the global branch supplies large-scale features while remaining frozen. At each layer, GTR constructs a candidate set for every regional subregion, and selective geo-aware cross-attention injects the selected global information before the regional forecasting

block. In the subsequent two-step stage, both branches are unfrozen and their first predictions are fed back as the next inputs, exposing them to their own forecast errors.

Both global and regional training use the auxiliary zero-increment reconstruction loss. For branch $b \in \{G, R\}$, let $\hat{\Delta Y}^b$ and ΔY^b be the predicted and target weather increments, and let $\hat{\Delta Y}^{b,(0)}$ be the increment produced by its encoder and prediction head without forecasting blocks. The terms in Eq. equation 13 are

$$\mathcal{L}^b_{\text{pred}} = \ell_b(\hat{\Delta Y}^b, \Delta Y^b), \tag{15}$$

$$\mathcal{L}^b_{\text{rec}} = \ell_b(\hat{\Delta Y}^{b,(0)}, 0), \tag{16}$$

where ℓ_b is the latitude-weighted forecasting error on the corresponding grid. During two-step fine-tuning, the objectives for both branches are accumulated over both autoregressive steps.

C Baseline Details

C.1 Stormer

Stormer (Nguyen et al., 2024) is a global weather prediction model based on the Transformer architecture (derived from DiT). Its training process consists of two phases: initial pre-training via single-step prediction to establish short-term forecasting capabilities, followed by critical multi-step fine-tuning that focuses on 4-step and 8-step prediction performance (including ablation studies validating its significance). During inference, Stormer adopts an innovative composite prediction strategy, generating weather forecasts by averaging outputs across multiple time intervals, thereby substantially improving forecast stability and accuracy. For example, the results of three 24-hour, six 12-hour, and twelve 6-hour predictions are ensembled to produce the final 72h forecast.

C.2 Pangu-Weather

Pangu-Weather (Bi et al., 2023) is a model specialized in global high-resolution weather prediction. Built upon a 3D Earth-Specific Transformer architecture—tailored to process the three-dimensional spatial characteristics of Earth's meteorological data (longitude, latitude, altitude)—its core training strategy involves concurrently training three independent models. During inference, the system employs a greedy strategy for any target forecast horizon: it iteratively selects and utilizes the model with the largest available timestep (24h, 12h or 6h) to efficiently achieve long-term forecasting.

C.3 OneForecast

OneForecast (Gao et al., 2025) is a graph neural network framework designed to support both global and regional weather forecasting. It constructs a multi-scale graph based on multi-grid theory and increases the node density within the target region to represent fine-grained local structures while retaining the surrounding global context. Its adaptive messaging mechanism employs dynamic gating to integrate node and edge features during information propagation. For high-resolution regional forecasting, OneForecast further introduces a neural nested grid that couples the coarse global grid with a denser regional grid, reducing information loss near the regional boundary.

C.4 STCast

STCast (Chen et al., 2026) is a regional–global forecasting framework that combines Spatial-Aligned Attention (SAA) with a Temporal Mixture-of-Experts (TMoE). SAA initializes a learnable regional–global distribution using geographical distance and incorporates it into linear cross-attention, allowing the spatial alignment to be refined during training. TMoE introduces month-specific Gaussian priors to route atmospheric features to specialized experts, enabling the model to capture seasonal variations. Together, these modules integrate global spatial information with high-resolution regional dynamics while adapting the forecast to different temporal patterns.

D METRICS

D.1 ROOT MEAN SQUARED ERROR (RMSE)

For each forecast lead time, we calculate the latitude-weighted RMSE over all evaluation samples and variables as

$$\text{RMSE} = \frac{1}{TV}\sum_{t=1}^{T}\sum_{v=1}^{V}\sqrt{\frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}L(i)\left(\hat{Y}_{t,v,i,j}-Y_{t,v,i,j}\right)^2}, \tag{17}$$

where T is the number of evaluation samples, V is the number of evaluated variables, and $h \times w$ is the regional grid size. $L(i)$ denotes the normalized cosine-latitude weight at row i, while $\hat{Y}_{t,v,i,j}$ and $Y_{t,v,i,j}$ denote the prediction and ground truth, respectively.

D.2 ANOMALY CORRELATION COEFFICIENT (ACC)

Let $\hat{Y}'_{t,v,i,j} = \hat{Y}_{t,v,i,j} - \mu_{t,v,i,j}$ and $Y'_{t,v,i,j} = Y_{t,v,i,j} - \mu_{t,v,i,j}$ denote the predicted and ground-truth anomalies relative to climatology $\mu_{t,v,i,j}$. We compute the latitude-weighted spatial ACC separately for each sample and variable and then average the results:

$$\text{ACC} = \frac{1}{TV}\sum_{t=1}^{T}\sum_{v=1}^{V}\frac{\sum_{i,j}L(i)\hat{Y}'_{t,v,i,j}Y'_{t,v,i,j}}{\sqrt{\left(\sum_{i,j}L(i){\hat{Y}'_{t,v,i,j}}^2\right)\left(\sum_{i,j}L(i){Y'_{t,v,i,j}}^2\right)}}. \tag{18}$$

Higher ACC indicates better spatial agreement between the predicted and observed anomaly patterns.

E EXPERIMENTS DETAILS

E.1 SOFTWARE AND HARDWARE ENVIRONMENT

All experiments are implemented using PyTorch 2.1.0 and PyTorch Lightning 2.2.1 on 8 NVIDIA GeForce RTX 4090 GPUs (24 GB each). We use full-precision training and the Distributed Data Parallel (DDP) strategy. The software environment includes CUDA 11.8 and Python 3.11, running on Ubuntu 22.04.

E.2 IMPLEMENTATION DETAILS

Stage 1: Global Pre-training. In ARGO, the patch size is set to $P = 4$. The backbone comprises 16 Transformer blocks with a hidden dimension of 512, eight attention heads, an MLP ratio of 2, and a window size of 8×8. The forecasting interval satisfies $\delta \in \{6\text{h}, 12\text{h}, 24\text{h}\}$. The global model optimizes $\mathcal{L}^G$ in Eq. equation 13 with a batch size of 16 and an initial learning rate of 5×10^{-4}, which follows a cosine decay schedule after a five-epoch warm-up. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 10^{-5}, together with early stopping with a patience of 10 epochs.

Stage 2: Regional Training. We initialize both forecasting branches from the Stage 1 checkpoint, freeze the global branch, and optimize the regional branch and fusion modules on the 0.25° regional data using $\mathcal{L}^R$. Regional crops follow the full-shot, few-shot, or zero-shot sampling settings described in Section 3.2.2. The attention-head dimension of selective geo-aware cross-attention is set to $d_h = 66$. GTR divides the regional grid into $M = 16$ subregions arranged in a 4×4 layout and retains $K_n = 64$ nearby, $K_f = 128$ dynamically selected remote, and $|\mathcal{B}| = 32$ pooled background global tokens for each subregion, giving $K = 224$ candidates in total. After a five-epoch warm-up using all global tokens, the remote-token budget is progressively reduced to 512, 256, and 128. We use an effective batch size of 48, with learning rates of 5×10^{-4} for router training and 10^{-5} for final fine-tuning.

Stage 3: Two-Step Fine-tuning. Starting from the best checkpoint of Stage 2, we unroll both forecasting branches for two steps and feed their first predictions back as the inputs to the second step. The time-varying auxiliary fields are updated at each forecast time. We unfreeze the global branch and jointly optimize both branches and the fusion modules using the global and regional objectives at both steps. We use an effective batch size of 48 and a learning rate of 10^{-5} for up to 50 epochs, with a five-epoch warm-up and early stopping with a patience of 10 epochs.

Inference. Following Stormer (Nguyen et al., 2024), we independently generate predictions using each applicable base interval in $[6\mathrm{h}, 12\mathrm{h}, 24\mathrm{h}]$ and average the resulting forecasts. For example, a $72\mathrm{h}$ forecast averages the predictions obtained from twelve $6\mathrm{h}$ steps, six $12\mathrm{h}$ steps, and three $24\mathrm{h}$ steps.

E.3 Three-stage Training Algorithm

Algorithm 1 summarizes the training procedure of ARGO. We first learn transferable atmospheric dynamics from global data, adapt the pretrained backbone to single-step regional forecasting, and finally improve autoregressive robustness through two-step fine-tuning.

Algorithm 1 Three-stage training procedure of ARGO

Input: Global dataset $\mathcal{D}_G$, regional dataset $\mathcal{D}_R$, forecast intervals $\{6\mathrm{h}, 12\mathrm{h}, 24\mathrm{h}\}$
Output: Global parameters θ_G, regional parameters θ_R, and router parameters θ_S

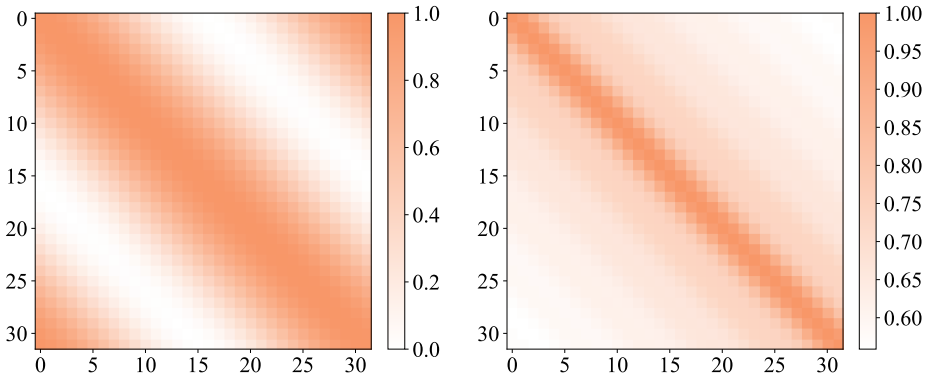
Stage I: Global Forecasting Pre-training
Initialize the global forecasting model g_{θ_G}
for each global pre-training iteration **do**
  Sample $(G_t, G_{t+\delta}) \sim \mathcal{D}_G$ and a forecast interval δ
  $\hat{G}_{t+\delta} \leftarrow G_t + g_{\theta_G}(G_t; \delta)$
  Update θ_G using the one-step global forecasting objective
end for

Stage II: Regional Single-step Training
Initialize the regional model θ_R and router θ_S; freeze θ_G
for each regional single-step iteration **do**
  Sample $(R_t, G_t, R_{t+\delta}) \sim \mathcal{D}_R$ and a forecast interval δ
  $\{\mathcal{S}_r\} \leftarrow \mathrm{GTR}_{\theta_S}(R_t, G_t)$
  Fuse the selected global tokens into regional features using GaPE cross-attention
  $\hat{R}_{t+\delta} \leftarrow R_t + f_{\theta_R, \theta_S}(R_t, G_t; \delta)$
  Update θ_R and θ_S using $\mathcal{L}(\hat{R}_{t+\delta}, R_{t+\delta})$
end for

Stage III: Regional Two-step Autoregressive Fine-tuning
Initialize θ_R and θ_S from the best Stage-II checkpoint and unfreeze θ_G
for each regional two-step iteration **do**
  Sample a forecast interval δ and paired regional and global trajectories at t, $t+\delta$, and $t+2\delta$ from $\mathcal{D}_R$
  $(\hat{R}^{(0)}, \hat{G}^{(0)}) \leftarrow (R_t, G_t)$ and $\mathcal{L}_{\mathrm{2step}} \leftarrow 0$
  for $s = 1, 2$ **do**
    Update the time-varying auxiliary fields to time $t + (s-1)\delta$
    $\{\mathcal{S}_r^{(s)}\} \leftarrow \mathrm{GTR}_{\theta_S}(\hat{R}^{(s-1)}, \hat{G}^{(s-1)})$
    $\hat{R}^{(s)} \leftarrow \hat{R}^{(s-1)} + f_{\theta_R, \theta_S}(\hat{R}^{(s-1)}, \hat{G}^{(s-1)}; \delta)$
    $\hat{G}^{(s)} \leftarrow \hat{G}^{(s-1)} + g_{\theta_G}(\hat{G}^{(s-1)}; \delta)$
    Compute $\mathcal{L}_s^R$ and $\mathcal{L}_s^G$ from Eq. equation 13 using targets at $t + s\delta$
    $\mathcal{L}_{\mathrm{2step}} \leftarrow \mathcal{L}_{\mathrm{2step}} + \mathcal{L}_s^R + \mathcal{L}_s^G$
  end for
  Update θ_G, θ_R, and θ_S using $\mathcal{L}_{\mathrm{2step}}$
end for
return $\theta_G, \theta_R, \theta_S$

F Visualization of Positional Similarity

Figure 4 visualizes the one-dimensional positional similarity induced by GaPE and RoPE. GaPE preserves similarity across the longitude wrap-around boundary, whereas RoPE treats the two ends of the coordinate range as distant.



(a) Geo-aware positional encoding (b) Rotary positional encoding

Figure 4: Similarity matrices of GaPE and RoPE in the one-dimensional setting.

G Per-variable Cross-region Generalization

Table 5 expands the aggregate comparison in Table 2 into per-variable RMSE and ACC. Across the six variables, full-shot generally achieves the lowest RMSE, while few-shot and zero-shot results remain close to one another.

Table 5: Cross-region generalization performance of ARGO under full-shot, few-shot, and zero-shot settings.

Variable		**T2m**			**U10**			**V10**		
Metric	Lead Time	Full-shot	Few-shot	Zero-shot	Full-shot	Few-shot	Zero-shot	Full-shot	Few-shot	Zero-shot
RMSE↓	1-day	**1.02**	1.12	1.18	**1.03**	1.18	1.24	**1.05**	1.21	1.27
	4-day	**1.77**	1.86	1.93	**2.07**	2.15	2.22	**2.11**	2.20	2.24
	7-day	**2.59**	2.66	2.72	**3.00**	3.06	3.14	**3.10**	3.15	3.21
	10-day	**3.21**	3.30	3.39	**3.43**	3.52	3.61	**3.54**	3.64	3.70
ACC↑	1-day	**0.94**	0.93	0.92	**0.94**	0.92	0.91	**0.94**	0.92	0.91
	4-day	**0.83**	0.81	0.80	**0.76**	0.74	0.72	**0.76**	0.74	0.73
	7-day	**0.63**	0.61	0.61	**0.48**	0.46	0.45	**0.47**	0.45	0.45
	10-day	**0.42**	0.39	0.39	**0.27**	0.25	0.24	**0.26**	0.23	0.23
Variable		**MSLP**			**Z500**			**T850**		
Metric	Lead Time	Full-shot	Few-shot	Zero-shot	Full-shot	Few-shot	Zero-shot	Full-shot	Few-shot	Zero-shot
RMSE↓	1-day	**73.18**	79.56	85.83	**60.66**	63.78	65.58	**0.77**	0.87	0.92
	4-day	**230.54**	235.14	249.41	**236.76**	240.59	248.63	**1.71**	1.78	1.82
	7-day	**442.22**	446.03	448.21	**462.51**	466.30	469.70	**2.84**	2.89	2.94
	10-day	**564.75**	571.90	570.87	**622.50**	629.11	637.29	**3.71**	3.77	3.87
ACC↑	1-day	**0.99**	**0.99**	**0.99**	**1.00**	0.99	0.99	**0.98**	0.97	0.96
	4-day	**0.91**	**0.91**	0.90	**0.93**	**0.93**	0.92	**0.88**	0.87	0.87
	7-day	**0.66**	**0.66**	**0.66**	**0.73**	**0.73**	**0.73**	**0.67**	0.66	0.66
	10-day	0.42	0.41	**0.43**	0.48	0.49	**0.50**	**0.41**	0.40	0.40