https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Enhancing Data-Driven Weather Forecasting via Gated Relative Position Encoding and Spatial-Aware Feed-Forward Network

Leyi Wang[1], Duo Zhang[1,2], Jerry Zhijian Yang[1,2,3], Baoxiang Pan[4], Dazhi Xi[5], Xiaoyu Huang[6]

[1]Hubei Center for Applied Mathematics, Wuhan University, Wuhan, 430072, China
[2]Institute for Math and AI, Wuhan University, Wuhan, 430072, China
[3]School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China
[4]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100029, China
[5]Department of Earth and Planetary Sciences, The University of Hong Kong, Hong Kong SAR, 999077, China
[6]Dongfang Electric Co., Ltd., Chengdu, 611731, China

Correspondence to: Xiaoyu Huang (huangxy4721@dongfang.com), Jerry Zhijian Yang (zjyang.math@whu.edu.cn)

Abstract. Data-driven weather models have emerged to address the immense computational costs of traditional numerical weather prediction by generating highly accurate, global forecasts in seconds. While Transformer-based architectures have achieved higher accuracy than numerical weather predictions, their existing position encodings typically embed limited spatial and temporal context, failing to fully account for the time variability, directionality, and location-dependency inherent in atmospheric motions. To resolve this, we introduce a novel model, Neighborhood Attention Transformer for atmospheric prediction (AtmoNAT). We propose two unique architectural components: a Gated Relative Position Encoding (GRPE) and a Spatial-Aware Feed-Forward Network (SAFN). The GRPE maintains independent positional biases based on absolute coordinates to secure location-dependency with a negligible increase in model size, while effectively capturing the directionality and temporal variations of the atmosphere. Simultaneously, the SAFN incorporates parallel input and gating branches, alongside a global positional bias, to explicitly simulate non-local interactions between atmospheric variables and integrate terrain effects. Evaluated on the WeatherBench 2 data at a 1.5° spatial resolution, AtmoNAT's deterministic forecasts demonstrate lower prediction errors on key variables up to a 72-hour lead time when compared to other coarse-resolution ensemble forecasts. Furthermore, AtmoNAT achieves state-of-the-art forecasting performance over global land areas, highlighting the profound potential of GRPE and SAFN in advancing next-generation weather forecasting.

1 Introduction

Weather forecasting is essential for saving lives and supporting emergency management during extreme events like hurricanes, while also mitigating economic losses in sectors like agriculture and transport. Numerical weather prediction works by treating the prediction as an initial value problem, pinning down current conditions through data assimilation and then integrating nonlinear partial differential equations across a global grid of enormous points (Bauer et al., 2015).

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Numerical weather prediction is considered to have a high degree of accuracy up to approximately 7 days (Haiden et al., 2025). However, its wide range accessibility is limited by immense computational costs.

To address the immense computational costs, data-driven weather models have emerged, training deep neural network to learn atmospheric patterns directly from decades of historical reanalysis data (Bi et al., 2023; Bodnar et al., 2025; Chen et al., 2023; Chen et al., 2025; Couairon et al., 2024; Du et al., 2025; Esteves et al., 2023; Keisler, 2022; Lam et al., 2023; Nguyen et al., 2024; Niu et al., 2025). These models can yield global forecasts in seconds, orders of magnitude faster than numerical weather prediction, while demonstrating comparable statistical accuracy in medium-range forecasts (Rasp et al., 2024).

Over the past decade, the Transformer (Vaswani et al., 2017) has become the dominant architecture, enabling models to learn long-range dependencies. The performance also scales predictably with model size (Kaplan et al., 2020). Most data-driven weather models utilize Transformers and have achieved state-of-the-art accuracy (Bi et al., 2023; Bodnar et al., 2025; Chen et al., 2023; Chen et al., 2025; Couairon et al., 2024; Du et al., 2025; Nguyen et al., 2024; Niu et al., 2025). Because self-attention processes all tokens in parallel, Transformers must rely on position encodings to retain spatial information. Models like FuXi, Fengwu, and Stormer (Chen et al., 2023; Chen et al., 2025; Nguyen et al., 2024) employ the vanilla position encodings of the Swin Transformer (Liu et al., 2021) and Vision Transformer (Dosovitskiy et al., 2020). Aurora (Bodnar et al., 2025) applies additive encodings for patch position of latitude and longitude, patch area, and absolute time. Similarly, Baguan (Niu et al., 2025) incorporates lead time and position encodings into its tokens.

However, these approaches largely adopt standard computer vision solutions of position bias or embed only limited spatial and temporal context, failing to fully account for the time variability, directionality, and location-dependency of atmospheric motions. Alternatively, Pangu-Weather (Bi et al., 2023) utilizes a 3D Transformer block with the Earth-Specific Positional Bias (ESPB), which adds a positional bias to each token based on its absolute coordinates to accommodate the atmosphere's spherical geometry and vertical stratification. Yet, To et al. (2024) find that the ESPB can be replaced with a relative positional bias without any degradation on validation loss, reducing the model size by nearly 40%. ESPB is an inefficient method for characterizing the location-dependency of atmospheric motions and does not contribute significantly to Pangu-Weather's performance. In summary, there remains a critical need for efficient position encodings tailored for atmospheric motions to further enhance performance of data-driven weather models.

In this work, we introduce **N**eighborhood **A**ttention **T**ransformer (Hassani et al., 2023) for **Atmo**spheric prediction (AtmoNAT). Neighborhood attention bridges the gap between the powerful modeling capabilities of Transformers and the efficient, localized characteristics of convolutional neural networks. It significantly reduces the computational complexity of standard Transformer blocks and has already been successfully applied to data-driven weather forecasting (e.g., Du et al., 2025; Pathak et al., 2026). We propose a Gated Relative Position Encoding (GRPE) for neighborhood attention. The design

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

of GRPE explicitly accounts for the time variability, directionality, and location-dependency of atmospheric motions. GRPE can evolve with time to adapt to the dynamical features of atmosphere. GRPE is a linear combination of learnable prototypes, which gives sufficient flexibility to align with the direction of local atmospheric motions. GRPE also maintains independent positional biases based on absolute coordinates, similar to ESPB, but with a negligible increase in model size. Not like ESPB, integrating GRPE into AtmoNAT yields significantly higher accuracy on WeatherBench 2 (Rasp et al., 2024) compared to models without it.

A gating mechanism is a structural component that uses a learned "gate" to control information flow, determining how much information is passed through, filtered out, or updated. Initially adopted in LSTMs and GRUs (Hochreiter et al., 1997; Chung et al., 2014), gating is now widely applied across various neural networks (e.g., Munir et al., 2025; S. Yang et al., 2024; Zamir et al., 2022). Shazeer (2020) first introduced gating mechanisms into Transformers, enhancing their expressive power and establishing them as standard components in many open-source large language models (Grattafiori et al., 2024; Qiu et al., 2025; A. Yang et al., 2024). Variable interactions are fundamental in atmospheric dynamics. For instance, wind speed control the advections of water vapor. Gating mechanisms can directly simulate these interactions in feature space through control of information flow, thereby offering a promising avenue to boost performance of data-driven weather models. To our knowledge, gating mechanisms have not yet been explicitly applied in data-driven weather models.

We introduce a Spatial-Aware Feed-Forward Network (SAFN) to AtmoNAT. SAFN comprises two branches that facilitate feature interactions through a gating mechanism. Furthermore, we propose a global positional bias to explicitly incorporate terrain effects in gating mechanism. Our results demonstrate that SAFN significantly improves AtmoNAT's weather forecasting performance.

The remainder of this paper is organized as follows. Section 2 details the data, training procedures, and the architectural design of AtmoNAT, specifically focusing on GRPE and SAFN. Section 3 presents an ablation study and compares AtmoNAT's performance with other data-driven weather models. Section 4 provides a visualization of the GRPE and SAFN components to further understand the function of these two structures. Finally, Section 5 concludes the study and offers perspectives for future research.

2 Methods

2.1 Data and training

AtmoNAT is trained and validated on ERA5 reanalysis data (Hersbach et al., 2020). AtmoNAT predicts the evolution of both upper-air and surface fields. The upper-air components comprise five variables across 13 pressure levels: geopotential (Z), specific humidity (Q), temperature (T), and zonal and meridional wind speeds (U, V). Surface fields include 10-meter

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

wind speeds (U10m, V10m), 2-meter temperature (T2m), and mean sea-level pressure (MSLP). Additionally, static geographic inputs, are incorporated to provide spatial context, including altitude, latitude, longitude, the cosine of the solar zenith angle, and a land-sea mask. We utilize ERA5 data at a 1.5° spatial resolution with a 12-hour temporal resolution (00:00 and 12:00 UTC). The dataset is chronologically partitioned into training (1979–2018; 29,200 samples), validation (2019; 730 samples), and testing (2020; 730 samples) sets.

Implemented in PyTorch (Paszke et al., 2019), AtmoNAT optimizes a latitude- and pressure-level-weighted loss function during the pretraining stage, defined in Eq. (1):

$$loss = \frac{1}{C_A \times H \times W \times L}\sum_{c=1}^{C_A}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{L} w_c a_i b_k |\hat{X}_{c,i,j,k}^{t+1} - X_{c,i,j,k}^{t+1}| + \frac{1}{C_S \times H \times W}\sum_{c=C_A+1}^{C_S}\sum_{i=1}^{H}\sum_{j=1}^{W} w_c a_i |\hat{X}_{c,i,j}^{t+1} - X_{c,i,j}^{t+1}|, \quad (1)$$

where C_A, C_S, H, W, and L denote the number of upper-air variables, surface variables, and grid points in the latitudinal, longitudinal, and vertical directions, respectively. The subscripts c, i, j, and k represent their corresponding indices. The terms $\hat{X}_{c,i,j,k}^{t+1}$ and $\hat{X}_{c,i,j}^{t+1}$ denote the upper-air and surface variables predicted by AtmoNAT at a 12-hour lead time $(t+1)$ based on current states (t), while $X_{c,i,j,k}^{t+1}$ and $X_{c,i,j}^{t+1}$ are the corresponding ERA5 ground truth values. The variables a_i and b_k represent the latitude and pressure-level weights, respectively, both of which decrease as latitude and vertical level increase. To account for the varying physical importance of different variables, we apply variable-specific weights (w_c) following the approach of Pangu-Weather (Bi et al., 2023).

Training utilizes the AdamW optimizer (Loshchilov et al., 2017), beginning with a 5,000-step linear warmup followed by a cosine annealing schedule with a peak learning rate of 2×10^{-4}. We employ gradient accumulation to achieve an effective batch size of 8. The pretraining phase spans 280,000 iterations, followed by 20,000 iterations of multi-step autoregressive fine-tuning. Consistent with other models, AtmoNAT generates long-lead-time predictions autoregressively. During multi-step autoregressive fine-tuning (Couairon et al., 2024; Keisler, 2022; Lam et al., 2022; Nguyen et al., 2023; Niu et al., 2025), the model is rolled out K times, and the loss function in Eq. (1) is averaged across all K steps. During fine-tuning, the rollout steps are progressively increased: K = 2 for the first 8,000 steps, K = 3 for the next 8,000, and K = 4 for the final 4,000 steps. Due to GPU memory constraints, fine-tuning is capped at K = 4 (representing a 48-hour lead time).

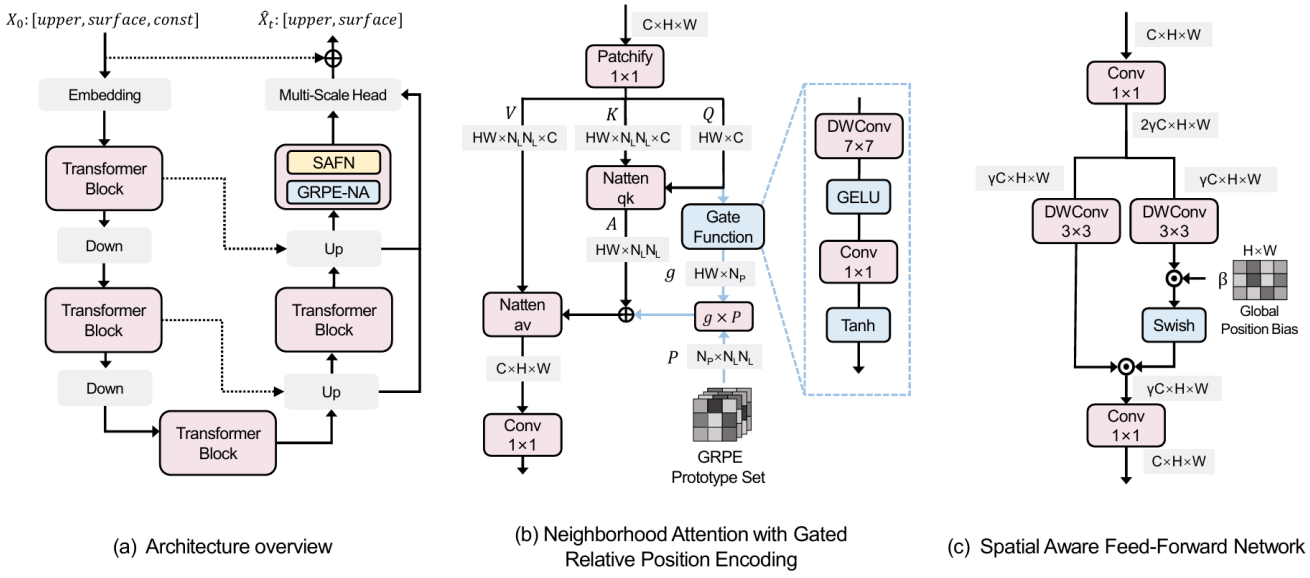
2.2 Model architecture

2.2.1 Backbone network

Inspired by feature pyramid network (Lin et al., 2017), The architecture of AtmoNAT (Figure 1a) employs a backbone comprising five Transformer blocks. The input data X_0 consists of the current time-step upper-air and surface variables

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

outlined in Section 2.1, alongside static geographic inputs. The output data $\hat{X}_t$ represents the predicted upper-air and surface variables 12 hours later. In the encoder pathway, the spatial resolution of the feature maps is halved at each downsampling block ("Down" in Figure 1a) while the feature dimension doubles. The decoder pathway performs the inverse operations. Spatial downsampling is achieved using patch merging (Liu et al., 2021), whereas upsampling and feature alignment across varying resolutions ("Up" in Figure 1a) are accomplished via frequency fusion (Figure S1; Chen et al., 2024). A multi-scale-head block (Figure S1) fuses features from three distinct resolutions.



(a) Architecture overview (b) Neighborhood Attention with Gated Relative Position Encoding (c) Spatial Aware Feed-Forward Network

Figure 1: Schematic diagram of model architecture (a), Neighborhood Attention with Gated Relative Position Encoding (b), and Spatial Aware Feed-Forward Network (c). The dimensions of matrixes at each node are shown.

Following an initial embedding layer, extracted features pass through the five Transformer blocks. Each block contains several submodules, each of which consists of two functional components: Neighborhood Attention with GRPE (GRPE-NA) and SAFN. For illustration, the fifth Transformer block in Figure 1a displays a single submodule of GRPE-NA and SAFN. Similar to other models (Nguyen et al., 2024; Niu et al., 2025), AtmoNAT employs adaptive layer normalization to modulate the inputs and outputs of both the GRPE-NA and SAFN within each submodule. The required scale and shift parameters are regressed from the embeddings of the lead time and forecast initialization time. The detailed version of AtmoNAT's structure with time-embedding input, the structure of embedding block, and the structure of adaptive layer normalization are illustrated in Figure S2.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

2.2.2 Neighborhood Attention with Gated Relative Position Encoding (GRPE-NA)

The GRPE-NA is detailed in Figure 1b. Given an input feature map $x \in \mathbb{R}^{C\times H\times W}$ (where C, H, and W denote the channel, height, and width dimensions, respectively), a 1×1 convolution generates the query (Q), key (K), and value (V) tensors, all preserving the spatial dimensions of x.

Following Hassani et al. (2023), neighborhood attention is computed by first defining a query vector $\boldsymbol{q}_{x,y} \in \mathbb{R}^{C}$ at spatial location (x, y) in Q. The corresponding key and value vectors within a local $N_L \times N_L$ neighborhood of K and V centered at (x, y) are denoted as $\boldsymbol{k}_{n_L\times n_L} \in \mathbb{R}^{C\times N_L\times N_L}$ and $\boldsymbol{v}_{N_L\times N_L} \in \mathbb{R}^{C\times N_L\times N_L}$. For any point (i, j) within this neighborhood, the scaled dot-product $prod(i, j)$ is calculated as:

$$prod(i,j) = \frac{1}{\sqrt{c}}\boldsymbol{q}_{x,y} \cdot \boldsymbol{k}_{N_L\times N_L}(i,j), \tag{2}$$

The attention score $a_{N_L\times N_L} \in \mathbb{R}^{N_L\times N_L}$ at (i, j), is then derived via a SoftMax operation over the neighborhood:

$$a_{N_L\times N_L}(i,j) = \frac{\exp\big(prod(i,j)\big)}{\sum_{a=1}^{N_L}\sum_{b=1}^{N_L}\exp\big(prod(a,b)\big)}, \tag{3}$$

Finally, the neighborhood attention's output vector at (x, y), $\boldsymbol{o}_{x,y}$, is the weighted sum of the value vectors $\boldsymbol{v}_{N_L\times N_L}$:

$$\boldsymbol{o}_{x,y} = \sum_{i=1}^{N_L}\sum_{j=1}^{N_L} a_{N_L\times N_L}(i,j) * \boldsymbol{v}_{N_L\times N_L}(i,j), \tag{4}$$

In Figure 1b, Eq. (2) and Eq. (3) correspond to the operations in the "Natten-qk" module, while Eq. (4) characterizes the "Natten-av" module. Applying Eq. (2) and Eq. (3) across all $H \times W$ spatial locations yields the final attention map $A \in \mathbb{R}^{H\times W\times N_L\times N_L}$. A is reshaped into a two-dimensional tensor $A \in \mathbb{R}^{HW\times N_L N_L}$ for convenience.

To compute the GRPE (E), we define N_P learnable prototypes, collectively denoted as $P \in \mathbb{R}^{N_P\times N_L N_L}$. The query tensor Q is processed through a gating function, comprising a 7×7 depth wise separable convolution, a GeLU activation (Hendrycks and Gimpel, 2016), a 1×1 convolution, and a Tanh activation:

$$g = Tanh\Big(Conv2d\Big(GeLU\big(DWConv(Q)\big)\Big)\Big), \tag{5}$$

Equation (5) yields the partition variable $g \in \mathbb{R}^{HW\times N_P}$. The GRPE is then calculated as:

$$E = g \times P, \tag{6}$$

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

The resulting $E \in \mathbb{R}^{HW \times N_L N_L}$ is added directly the attention map A. The calculations of GRPE through Eq. (5) and Eq. (6) are illustrated by the blue arrows in Figure 1b. Consequently, the scaled dot-product in Eq. (2) is updated to incorporate this spatial encoding:

$$prod(i,j) = \frac{1}{\sqrt{c}} \boldsymbol{q}_{x,y} \cdot \boldsymbol{k}_{N_L * N_L}(i,j) + E_{x,y}(i,j), \tag{7}$$

This updated product in Eq. (7) generates the final neighborhood attention output, which subsequently passes through a 1×1 convolution to produce the input features for the SAFN.

Atmospheric motions are inherently directional; for instance, substance transport follows wind direction, which is driven by pressure gradients. To account for this directionality, the GRPE calculation must aggregate information across a large neighborhood. Therefore, the 7×7 depth-wise separable convolution in Eq. (5) analyzes features in query tensor Q over an extended receptive field to determine the partition variable g.

Let the GRPE and partition variable g at coordinate (x, y) be $E_{x,y} \in \mathbb{R}^{N_L N_L}$ and $\boldsymbol{g}_{x,y} \in \mathbb{R}^{N_P}$, respectively, where $\boldsymbol{g}_{x,y}(k)$ is the k-th partition variable. If $P_k \in \mathbb{R}^{N_L N_L}$ is the k-th prototype in P, Eq. (6) yields:

$$E_{x,y} = \sum_{k=1}^{N_P} \boldsymbol{g}_{x,y}(k) * P_k, \tag{8}$$

In Eq. (8), GRPE is represented as a linear combination of learnable prototypes P_k, weighted by $\boldsymbol{g}_{x,y}$, which generalizes across all spatial coordinates. In summary, g provides the weighting coefficients for the prototypes, independently determining the GRPE at each location to capture the location-dependency of atmospheric motions. Furthermore, GRPE's reliance on Q introduces essential temporal variations, accounting for the dynamic nature of the atmosphere. The directionality, location-dependency, and time variations of GRPE are further discussed in section 4.

2.2.3 Spatial Aware Feed-Forward Network (SAFN)

The SAFN architecture, shown in Figure 1c, consists of parallel input and gating branches. For an input feature map $x \in \mathbb{R}^{C \times H \times W}$, a 1×1 convolution projects the features into a $2\gamma C$-dimensional space. This projection is split evenly along the channel dimension and processed by two parallel 3×3 depth-wise separable convolutions. This yields the input features $I \in \mathbb{R}^{\gamma C \times H \times W}$ and the gating features $G \in \mathbb{R}^{\gamma C \times H \times W}$.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

The gating branch spatially modulates the information flowing from the input branch. To ensure precise spatial control, the gating features G are multiplied element-wise by a global positional bias $\beta \in \mathbb{R}^{H\times W}$ for each channel. The product passes through a Swish activation function (Ramachandran et al., 2017) to produce the gating coefficients. The final output O is calculated as:

$$O = I \odot Swish(G \odot \beta), \tag{9}$$

where $\odot$ denotes element-wise multiplication.

SAFN's gating mechanism facilitates direct interaction between the input features I and gating features G, thereby mimicking variable interactions inherent in atmospheric motions. Atmospheric interactions are fundamentally non-local. For example, neighboring wind speeds are also strongly correlated with local substance transport. Inspired by this, the parallel 3×3 depth-wise separable convolutions extract crucial spatial context to model these non-local dependencies. Terrain is expected to exert permanent and location-dependent influence on atmospheric motions, which inspires us the inclusion of global position bias β in gating coefficient calculation. The ablation studies in section 3.2 evaluate the impact of SAFN on AtmoNAT's performance.

3 Results

3.1 Comparisons with other data-driven weather models

In this section, we compare the root mean square error (RMSE) and anomaly correlation coefficient (ACC) of AtmoNAT against several state-of-the-art data-driven weather models, including Pangu-Weather (Bi et al., 2023), NeuralGCM (Kochkov et al., 2024), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), Aurora (Bodnar et al., 2025), FengWu (Chen et al., 2025), Baguan (Niu et al., 2025), ArchesWeather (Couairon et al., 2024), and Stormer (Nguyen et al., 2024). For NeuralGCM, Stormer, and ArchesWeather, we evaluate the mean of its ensemble predictions. Ensemble prediction can substantially improve their performances. Regarding training resolutions, AtmoNAT and ArchesWeather are trained at 1.5°, NeuralGCM and Stormer at 1.4°, and all other models at 0.25°. Both the RMSE and ACC metrics incorporate area weighting, consistent with Rasp et al. (2024). Furthermore, RMSEs are normalized against the High-Resolution Ensemble Forecast (HRES) system from the ECMWF.

As detailed in Section 2.2.1, we define the combination of GRPE-NA and SAFN as a single feature extraction submodule. The five Transformer blocks in AtmoNAT (Figure 1a) contain 10, 8, 4, 8, and 10 of these submodules, respectively. Due to GPU memory constraints, AtmoNAT's embedding dimension is restricted to 128, resulting in a total of 43.4 million parameters. The model was trained on two NVIDIA RTX 4090 GPUs for 28 days and is utilized for all subsequent performance comparisons.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Figure 2 illustrates the relative RMSE across different models. For a fair evaluation, we restrict this comparison to models trained at coarse resolutions: AtmoNAT and the ensemble forecasts of ArchesWeather, Stormer and NeuralGCM. AtmoNAT significantly outperforms the other models on 850 hPa zonal (U850) and meridional (V850) winds up to a 72-hour lead time. It also demonstrates distinctly lower RMSEs for 10-meter winds (U10m, V10m) and 850 hPa temperature (T850) up to an 84-hour lead time. Furthermore, its lower prediction error for 700 hPa specific humidity (Q700) is sustained through a 60-hour lead time. For 500 hPa geopotential (Z500) and 2-meter temperature (T2m), AtmoNAT surpasses Stormer, and is only outperformed by NeuralGCM on Z500 at lead times exceeding 48 hours. Notably, AtmoNAT remains nearly the most accurate model for MSLP up to a 120-hour lead time. In summary, single AtmoNAT's deterministic forecasts achieve superior performance across most key variables up to 72 hours when compared with the ensemble forecasts of other coarse-resolution models.

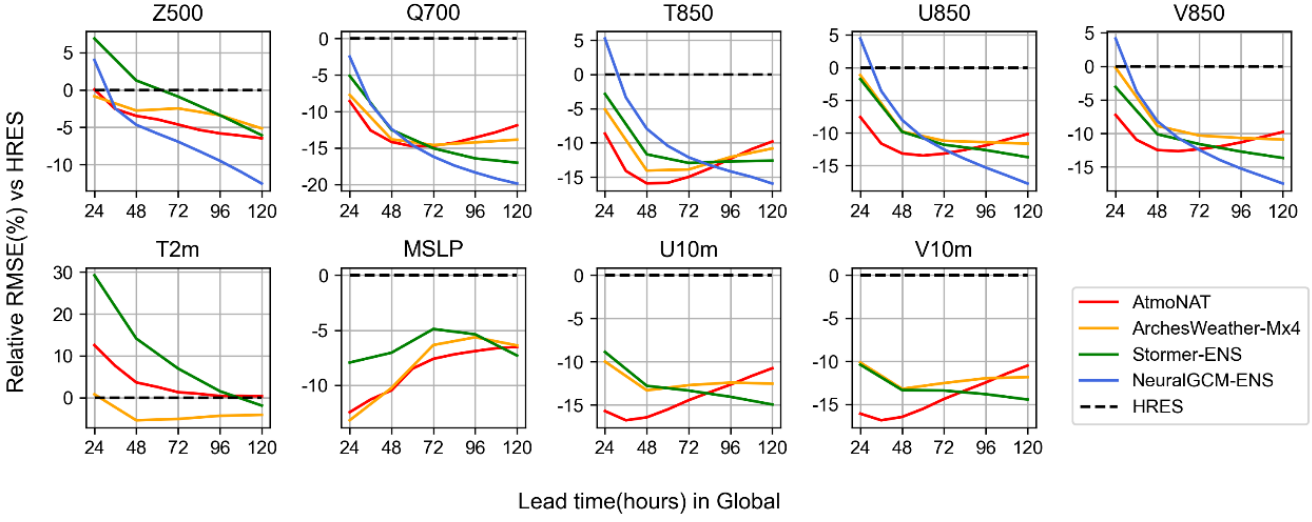


Figure 2. Relative RMSE of data-driven weather models trained on coarse resolution on key variables. Relative RMSE is calculated based on HRES performances.

The relative RMSEs for specific geographic regions (Figures S4 – S9) align with the global findings in Figure 2. Additionally, Figure S3 compares AtmoNAT's RMSE against all evaluated data-driven weather models, including those trained at 0.25° resolution. While AtmoNAT performs significantly better than Pangu-Weather, it trails behind FuXi and GraphCast. On certain variables, AtmoNAT's performance degrades beyond a 72-hour lead time; this is expected, as resource limitations restricted multi-step fine-tuning to a maximum 48-hour lead time.

Figure 3 presents the ACC of the coarse-resolution models for key variables up to a 72-hour lead time. These ACC trends are highly consistent with the RMSE results in Figure 2. AtmoNAT achieves the highest ACC for MSLP, T850, U850, V850,

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

U10m, and V10m. Notably, its ACC for wind velocities (U850, V850, U10m, V10m) is significantly higher than that of the competing models. It ranks second best for Z500 and outperforms Stormer on T2m, while ACC values for Q700 are nearly identical across all models. Extended ACC results up to a 120-hour lead time (Figure S10) reveal a performance degradation at day 5, again attributable to the 48-hour fine-tuning limit imposed by computational constraints.

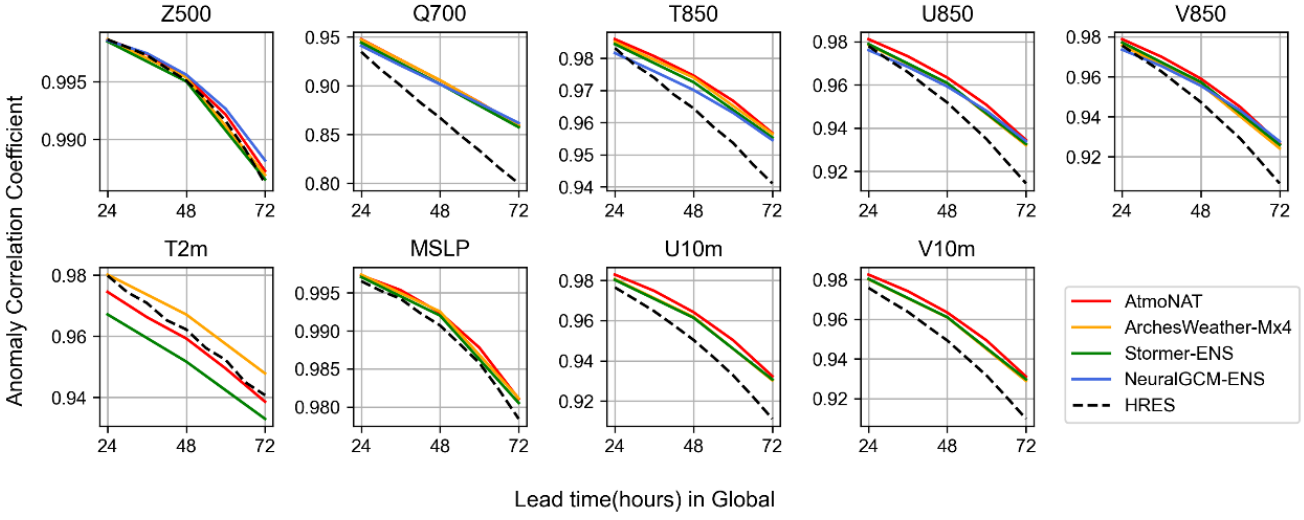


Figure 3. ACC of different data-driven weather models trained on coarse resolution on key variables till 72-hour lead time.

Figure 4 depicts the relative RMSE evaluated specifically over global land areas, incorporating state-of-the-art models trained at 0.25° resolution. Notably, AtmoNAT significantly outperforms all other models, including FuXi and GraphCast, on U10m, V10m, and Q700 up to a 120-hour lead time. It also achieves the lowest RMSE for U850 and V850 between 48 and 108 hours. For Z500, T850, and MSLP, AtmoNAT's performance remains comparable to other coarse-resolution models and distinctly superior to Pangu-Weather. Regional land-area RMSE analyses (Figures S11 – S16) are generally similar with Figure 4. However, over the North Pacific land areas, AtmoNAT yields significantly lower RMSEs across all key variables except T850 and Q700.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

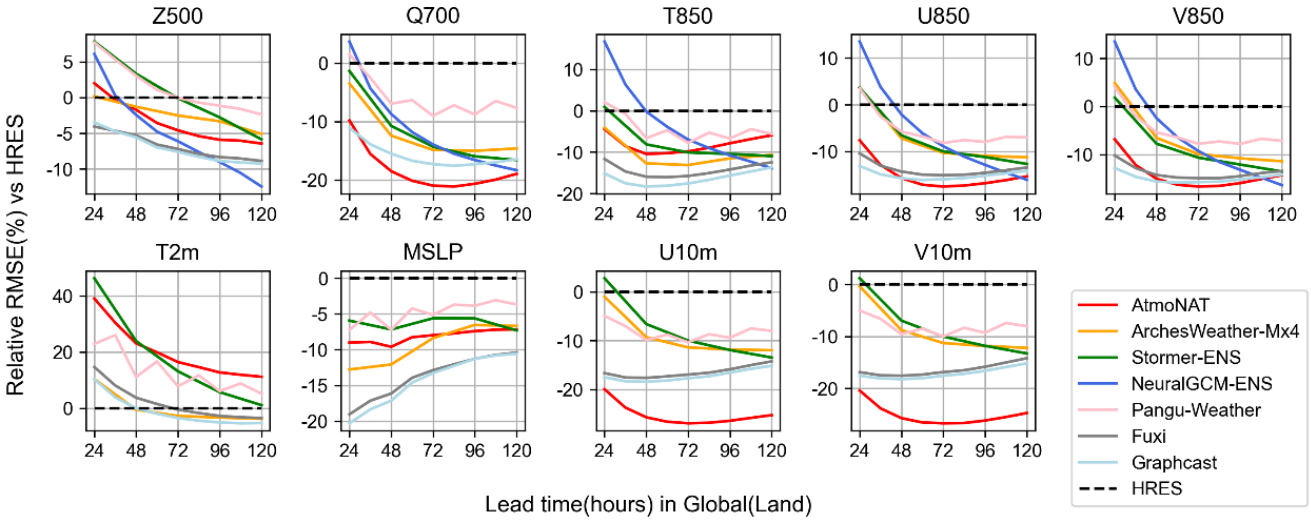


Figure 4. Relative RMSE of data-driven weather models in global land area on key variables. Relative RMSE is calculated based on HRES performances.

Figure 5 compares forecasts of key variables over Eurasia generated by AtmoNAT, HRES, and Pangu-Weather. AtmoNAT successfully reproduces the spatial distribution of these variables. Error field analysis indicates that AtmoNAT consistently maintains lower prediction errors across the middle-to-high latitudes of Eurasia. For example, while HRES and Pangu-Weather exhibit distinct Z500 errors over northern Eurasia, AtmoNAT mitigates these inaccuracies. Similarly, pronounced positive MSLP errors are visible for HRES and Pangu-Weather over northwestern Eurasia, whereas AtmoNAT achieves a significantly lower RMSE. For Q700 and wind velocities (U850, V850, U10m, V10m), AtmoNAT's error fields are uniformly near zero, corroborating the state-of-the-art performance demonstrated in Figure 4. Forecasts initialized at two alternative times (Figures S17 and S18) display consistent results.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

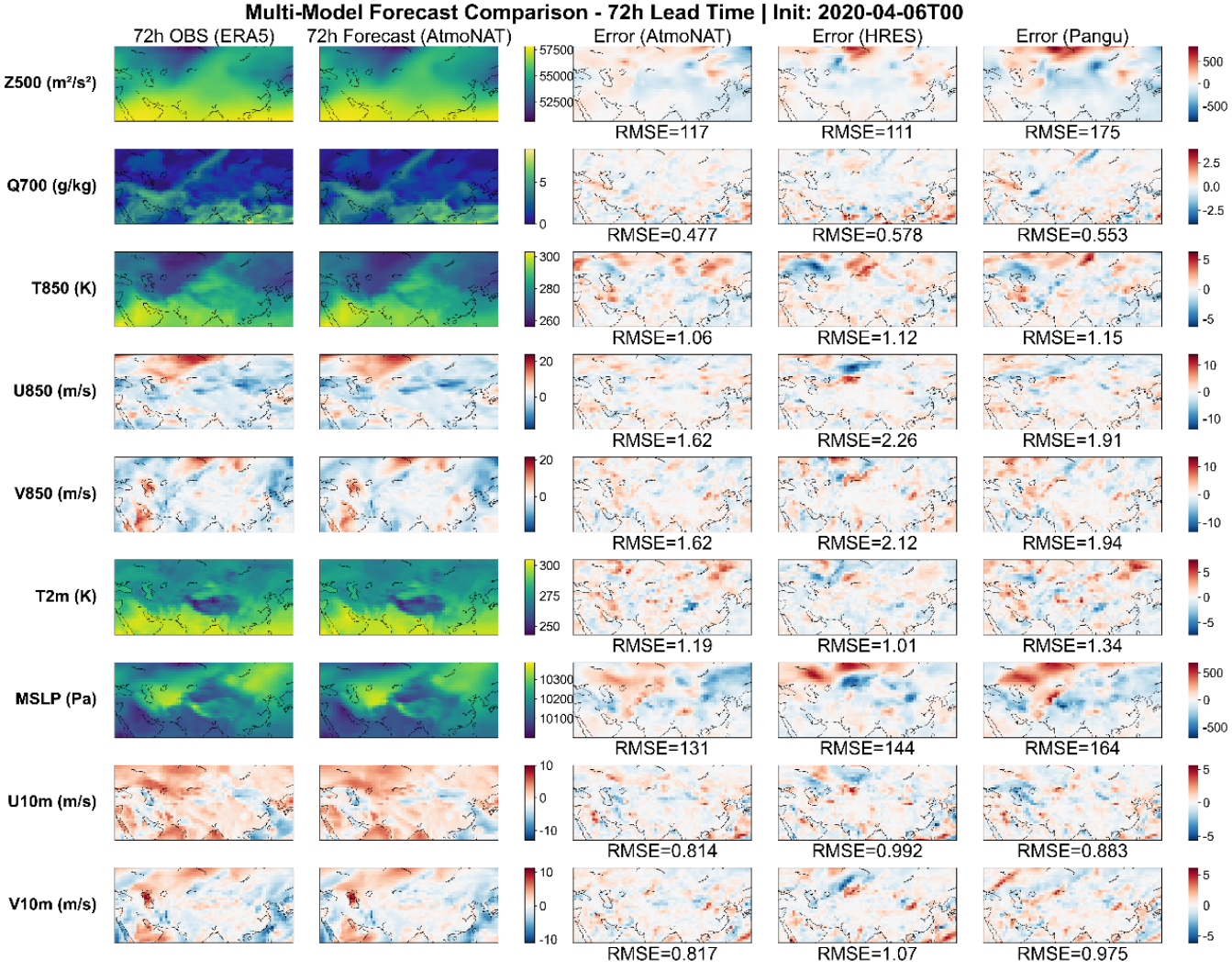


Figure 5. Forecast comparisons between AtmoNAT, HRES, and Pangu-Weather on Eurasia at 72-hour lead time initialized at 6 April 2020, 00:00:00 UTC. From left to right columns are ground truth data from ERA5, predicted fields from AtmoNAT, error fields from AtmoNAT, HRES, and Pangu-Weather. RMSE of each error field is located at the bottom of each subplot.

3.2 Ablation study

To evaluate the relative contributions of AtmoNAT's architectural components, we conducted a step-by-step ablation study, comparing the 24-hour lead time RMSEs of seven model variants (Table 1). We established a baseline model, denoted as "Swin Attention". The latent dimension of this model is 96. It only uses Swin Transformer blocks (Liu et al., 2021). Multi-Scale Head is removed. Multi-step autoregressive fine-tuning is not applied. The remaining six models were constructed by sequentially integrating our proposed modifications: (1) replacing Shifted Window Self-Attention of Swin Transformer with

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Neighborhood Attention, (2) adding the Multi-Scale Head, (3) incorporating GRPE into the Neighborhood Attention, (4) substituting the standard Transformer feed-forward network (FFN; Vaswani et al., 2017) with SAFN, (5) increasing the latent dimension to 128, and finally, (6) applying multi-step autoregressive fine-tuning.

Except for the fine-tuned variant, all models in this section were trained for 110,000 iterations with a batch size of 8. Table 1 also reports the parameter counts and floating-point operations (FLOPs) during inference. The results indicate that standard Neighborhood Attention performs similarly to Swin Attention. While the Multi-Scale Head reduces RMSE, its impact is less pronounced than that of GRPE and SAFN. Scaling the model capacity and applying autoregressive fine-tuning also yield substantial performance gains.

Table 1. RMSE, parameter counts, and FLOPs of AtmoNATs for ablation study.

Model	Parameters	FLOPs	Z500	T850	U850	T2M	V10M
Swin Attention	22.9M	268G	57.3	0.68	1.30	0.68	0.90
Neighborhood Attention	22.8M (-0.1)	267G (-1)	57.7	0.69	1.31	0.70	0.90
+ Multi-Scale Head	24.2M (+1.4)	307G (+40)	55.2	0.68	1.28	0.67	0.89
+ GRPE	24.9M (+0.7)	316G (+9)	52.0	0.66	1.24	0.66	0.85
+ SAFN	25.9M (+1.0)	320G (+4)	47.1	0.62	1.18	0.63	0.80
96 dims→128 dims	43.4M (+17.5)	590G (+270)	44.5	0.58	1.11	0.59	0.75
AtmoNAT (finetuned)	43.4M (0)	590G (0)	**42.9**	**0.57**	**1.10**	**0.56**	**0.73**

Note: Numbers in the parentheses are the parameter count or FLOPs change compared to the model in the former row. The lowest RMSEs are bolded. The performances of different models are validated at 24-hour lead time on testing data.

Figure 6 extends the analysis in Table 1 up to a 120-hour lead time. Consistent with Table 1, forecasting accuracy steadily improves with the progressive addition of new components, increased capacity, and fine-tuning. Neighborhood Attention and the Multi-Scale Head offer marginal improvements. The integration of GRPE yields a significant 5% to 10% RMSE reduction across most variables (excluding T2m and Q700) at all lead times. Similarly, replacing the standard FFN with SAFN provides a stable 5% to 10% RMSE reduction across all variables. GRPE and SAFN achieve these substantial error reductions while increasing model size and FLOPs by only 1% to 4% (Table 1). Although increasing the latent dimension from 96 to 128 enhances performance, it nearly doubles the parameter count and FLOPs without yielding obvious improvements at longer lead times. Finally, consistent with previous studies (Nguyen et al., 2024), multi-step autoregressive fine-tuning is vital for long-range forecasting stability, delivering an approximately 10% RMSE reduction across all key variables, particularly at 96- to 120-hour lead times.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

320

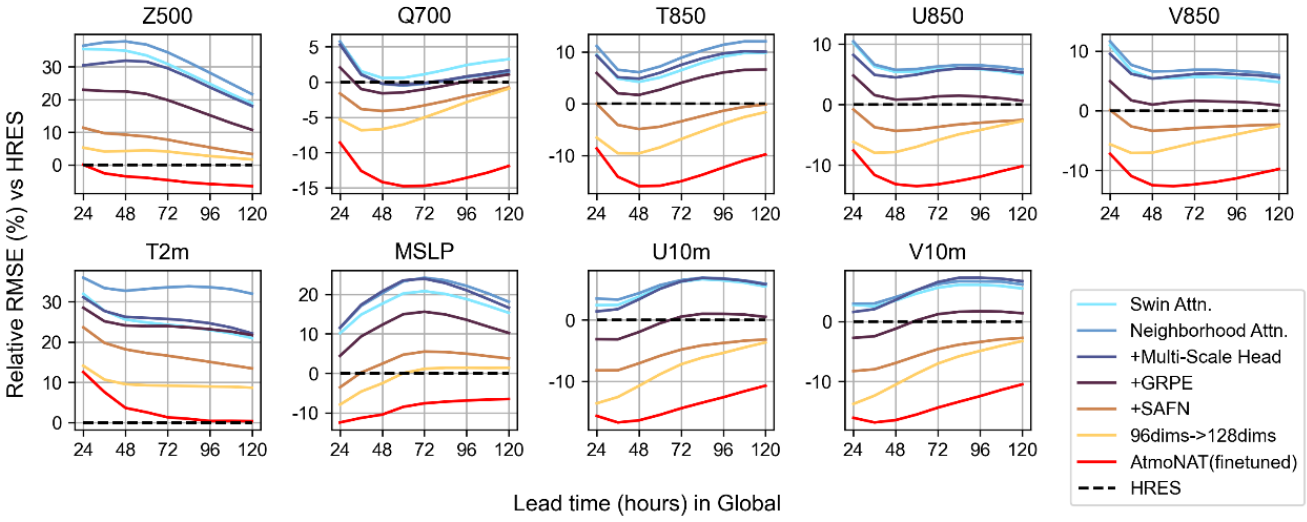


Figure 6. Relative RMSE of AtmoNATs for ablation study in Table 1 on key variables. Relative RMSE is calculated based on HRES performances.

325 Figures 7 and 8 present scorecards detailing the relative RMSE improvements following the respective additions of GRPE and SAFN. Incorporating GRPE consistently reduces the RMSE of U, V, MSLP, U10m, and V10m by 5% to 7.5% across all vertical levels up to day 5. The most substantial improvement occurs for Z, which sees an overall RMSE reduction exceeding 7.5%. Forecast accuracy also noticeably improves for T at pressure levels above 850 hPa, though gains are marginal for Q and T2m. Notably, both Q and Z exhibit large RMSE reductions at altitudes above 150 hPa. The addition of
330 SAFN produces a similar, but generally more pronounced, pattern of improvement (Figure 8). SAFN enhances the predictions of U, V, and T by more than 7.5% at levels above 400 hPa up to a 24-hour lead time, while driving RMSE reductions of over 10% for Z across all vertical levels up to day 3. Consistent with Figure 6, both GRPE and SAFN are critical to AtmoNAT's high predictive accuracy.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

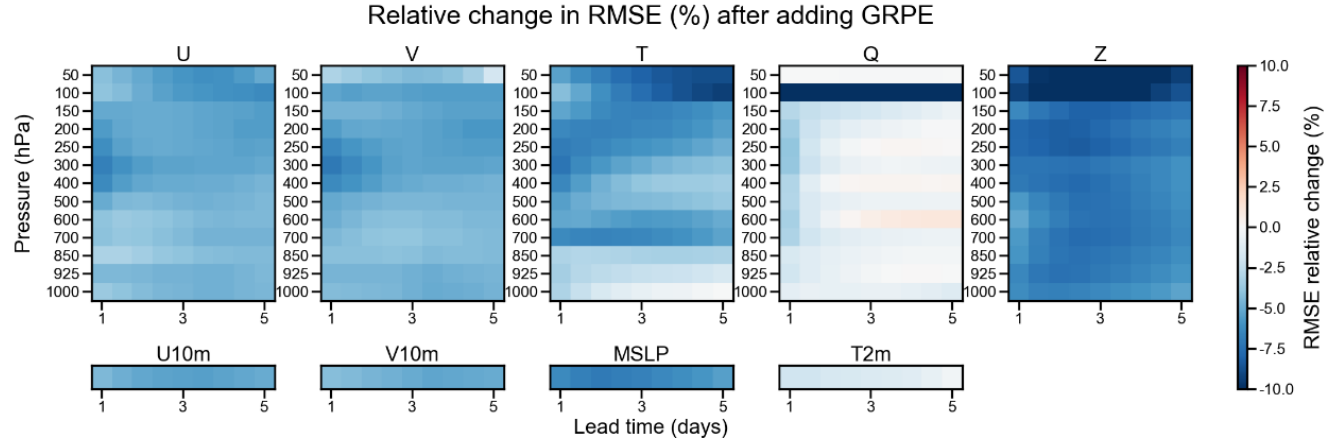


Figure 7. Scorecard of relative RMSE change after adding GRPE to Neighborhood Attention.

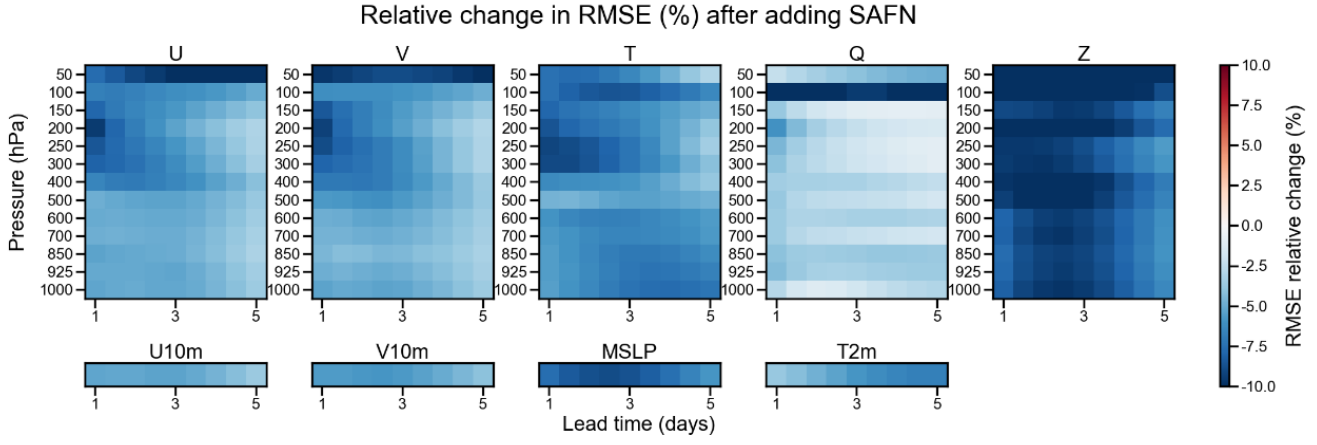


Figure 8. Scorecard of relative RMSE change after changing multi-layer perceptron into SAFN.

Table 2 isolates the impact of various position encodings on model RMSE. For these experiments, SAFN was replaced by a standard Transformer FFN, fine-tuning was disabled, and all models utilized a latent dimension of 96 and were trained for 1,825,000 iterations. We compared variants completely lacking position encodings against those utilizing standard relative position encodings (Liu et al., 2021), GRPE, and RoPE-Mixed (Heo et al., 2024)—a 2D adaptation of Rotary Positional Embeddings (Su et al., 2024) tailored for Vision Transformers. Because RoPE-Mixed is incompatible with Neighborhood Attention, it is evaluated exclusively within a Swin Transformer-based AtmoNAT backbone. Introducing any position encoding dramatically reduces AtmoNAT's RMSE compared to the baseline without encodings. Notably, GRPE achieves the lowest RMSE of all tested encodings.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Table 2. RMSE, parameter counts, and FLOPs of AtmoNATs with different position encodings.

Attention	PE	Parameters	FLOPs	Z500	T850	U850	T2m	V10m
NA	w/o	24.2M	307G	58.1	0.712	1.338	0.682	0.919
NA	Relative PE	24.2M	307G	55.2	0.682	1.283	0.673	0.885
Swin	RoPE-Mixed	24.2M	309G	53.1	0.672	1.251	0.667	0.854
NA	GRPE	24.9M	316G	**52.0**	**0.661**	**1.242**	**0.659**	**0.848**

Note: The lowest RMSEs are bolded. The performances of different models are validated at 24-hour lead time on testing data.

Table 3 evaluates AtmoNAT variants equipped with different feed-forward networks, maintaining the training protocols and latent dimensions used in Table 2. The baseline model employs a standard Transformer FFN. The second variant replaces SAFN with a standard gating mechanism, the Swish-Gated Linear Unit (SwiGLU; Shazeer, 2020). SwiGLU resembles SAFN but applies uniform linear projections across all spatial locations of the global feature map. The results demonstrate that SwiGLU outperforms the standard FFN, affirming the effectiveness of gating mechanisms in data-driven weather forecasting. However, SAFN achieves distinctly lower RMSEs than SwiGLU. We attribute this superior performance to SAFN's use of parallel depth-wise separable convolutions, which extract non-local spatial context. This extended receptive field explicitly enables non-local feature interactions, a critical requirement for accurately predicting atmospheric dynamics.

Table 3. RMSE, parameter counts, and FLOPs of AtmoNATs with different feed-forward networks.

Feed-forword network	Parameters	FLOPs	Z500	T850	U850	T2m	V10m
FFN	24.9M	316G	52.0	0.661	1.242	0.659	0.848
SwiGLU	24.9M	315G	51.05	0.646	1.220	0.645	0.841
SAFN	25.6M	320G	**47.1**	**0.624**	**1.176**	**0.634**	**0.799**

Note: The lowest RMSEs are bolded. The performances of different models are validated at 24-hour lead time on testing data.

Figure 9 illustrates the relative RMSE improvement over HRES for various data-driven weather models, aggregated across Z500, T850, Q700, and 850 hPa wind vectors following Rasp et al. (2024). To ensure fairness, this comparison is restricted to the first three forecast days, matching AtmoNAT's maximum 48-hour fine-tuning lead time. Baseline models' RMSEs are sourced from the WeatherBench 2 website, while training costs, model sizes, and FLOPs are drawn from published works. Training costs were normalized to NVIDIA V100 GPU days.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

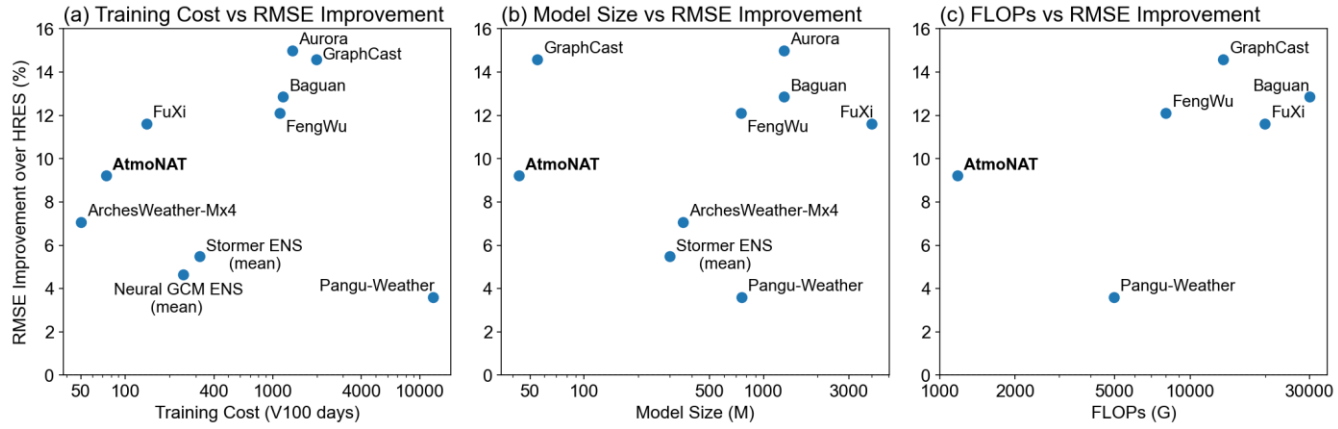


Figure 9. Scatter plot of RMSE improvement over HRES versus training cost (a), model size (b), and floating-point operations (FLOPs) per inference (c). Training costs are normalized to GPU days of NVIDIA V100.

As shown in Figure 9a, AtmoNAT, ArchesWeather, and FuXi are the most efficient models regarding RMSE improvement divided by training cost. When evaluating RMSE improvement divided by parameter count (Figure 9b), AtmoNAT ranks second only to GraphCast; notably, competing models contain 7.5 to 75 times more parameters. Furthermore, AtmoNAT stands out as the most computationally efficient model, demonstrating the highest RMSE improvement divided by FLOPs per inference (Figure 9c). Other models outperform AtmoNAT at the expense of requiring 5 to 30 times more FLOPs. Overall, AtmoNAT achieves the lowest RMSE among coarse-resolution models and Pangu-Weather, proving to be the most efficient model when taking training cost, model size, and computational cost all into account.

4 Discussion

To understand the efficacy of GRPE, Figure 10 visualizes the learned GRPE prototypes, the corresponding partition variables (g), and the resulting aggregated GRPEs at two distinct geographic locations and forecast initialization times. For this analysis, we extracted four representative prototypes and their partition variables from the second GRPE-NA module within the final Transformer block.

As shown, the prototypes capture distinct spatial patterns learned dynamically during training: Prototype 1 divides positive and negative areas along a diagonal, Prototypes 2 and 4 feature a horizontal division across the center, and Prototype 3 centers a large negative bias at its core. These prototypes explicitly encode the underlying directionality of atmospheric motions. The final GRPE is a linear combination of the prototypes, enabling flexible alignment of the GRPE to actual direction of motion at each spatial location.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Furthermore, the partition variable g independently determines the prototype weights at each spatial location, enabling the encoding to adapt to localized weather phenomena and successfully preserving the location-dependency of the atmosphere. Additionally, g exhibits clear temporal variations (Figure 10a), inducing dynamic changes in the resulting GRPE over South America (Figure 10b). This demonstrates GRPE's capacity to adjust to the temporal variability of atmospheric systems. Ultimately, GRPE's success in data-driven weather prediction stems from its comprehensive integration of temporal variability, spatial directionality, and location-dependency of atmospheric motions.

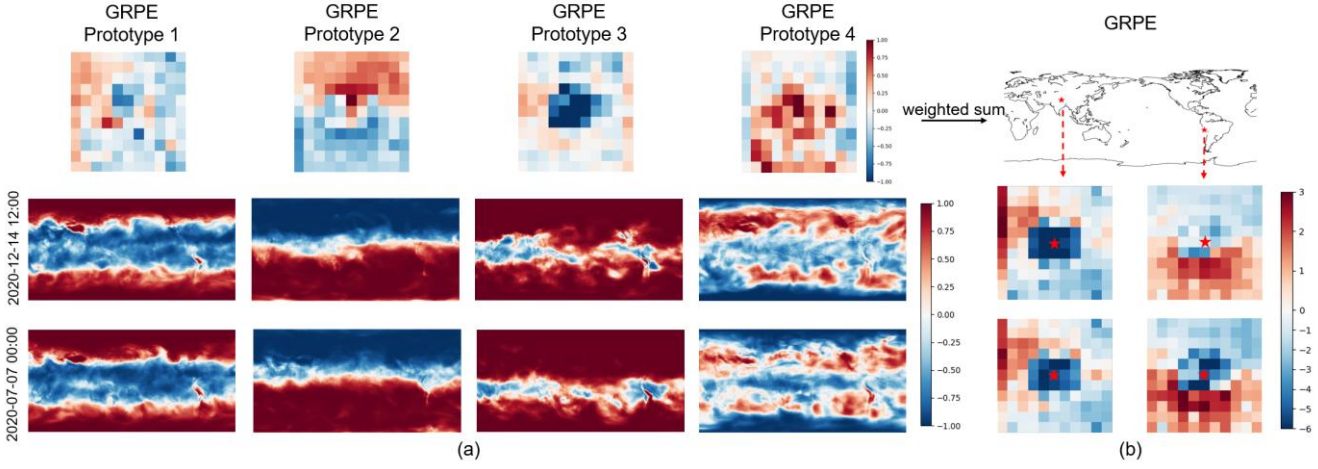


Figure 10. Subplot (a) presents GRPE prototypes and corresponding gating variables g at each spatial coordinate. GRPEs in subplot (b) are the weighted sum of prototypes with weighting coefficients defined by g. The spatial locations of GRPEs are also presented in the world map at upper-right region. g and GRPEs are shown at prediction starting time 7 July 2020, 00:00:00 UTC and 14 December 2020 12:00:00 UTC.

To understand the function of the global position bias β, Figure 11 illustrates β of the final four SAFNs within the last Transformer block. These biases strongly correlate with global terrain, exhibiting absolute values that are generally proportional to altitude. By accounting for the permanent influence of topography, the global positional bias provides the SAFN with an additional degree of freedom to generate more accurate gating coefficients [$G \odot \beta$ in Eq. (9)]. Consequently, this mechanism likely contributes to AtmoNAT's superior forecasting performance over global land areas.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

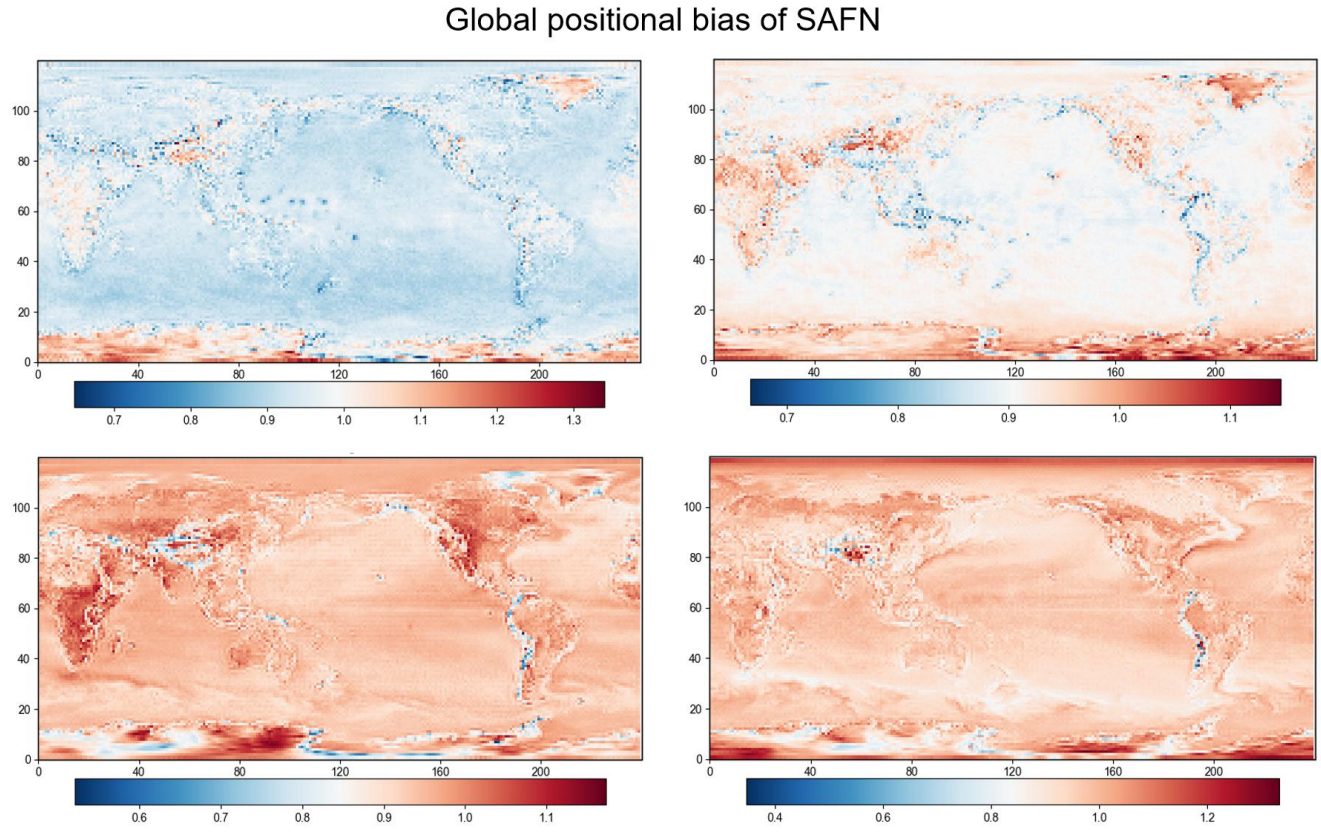


Figure 11. The visualization of global positional bias β of the last four SAFNs in the last Transformer block.

5 Conclusion

In this study, we propose AtmoNAT, a data-driven weather model based on Neighborhood Attention. The model consists two novel modules, namely GRPE and SAFN. GRPE can capture the time variability and directionality of atmospheric motions, and accomplishes independent encodings at every spatial location with omittable model size and FLOPs increase. Furthermore, we propose the Spatial-Aware Feed-Forward Network (SAFN), a gating mechanism enhanced by a global positional bias, to allow direct feature interactions under the influence of terrain's permanent effect.

AtmoNAT's deterministic forecasts achieve the preferable performance across most key variables at lead times up to 72 hours, outperforming Pangu-Weather (Bi et al., 2023) and the ensemble forecasts of other coarse-resolution models (Kochkov et al., 2024; Couairon et al., 2024; Nguyen et al., 2024). Over global land areas, it delivers advantageous accuracy in predicting U10m, V10m, U850, V850, and Q700, surpassing even high-resolution models such as FuXi (Chen et al., 2023) and GraphCast (Lam et al., 2023). Moreover, AtmoNAT is exceptionally resource-efficient, ranking as one of the most

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

efficient models in terms of RMSE improvement divided by training cost and FLOPs, and the second most efficient regarding RMSE improvement divided by model parameter count.

Ablation studies demonstrate that both GRPE and SAFN significantly enhance AtmoNAT's predictive accuracy, doing so at the cost of only negligible increases in parameter counts and FLOPs. When evaluated against alternative position encodings and feed-forward neural networks, the AtmoNAT architecture equipped with GRPE and SAFN consistently yields the lowest weather prediction errors.

Building on the proof-of-concept established in this study, future research should evaluate GRPE and SAFN within AtmoNAT architectures featuring larger latent dimensions and extended fine-tuning lead times. Because Neighborhood Attention requires significantly less computational power and memory than standard Transformer blocks, it has rapidly become the choice of architecture for recent data-driven weather models (e.g., Du et al., 2025; Pathak et al., 2026). Integrating GRPE and SAFN holds tremendous potential for developing earth-specific, highly efficient structures for the next generation of data-driven weather models.

Code and data availability

Source code of AtmoNAT can be accessed at https://doi.org/10.5281/zenodo.19369025 (Huang and Wang, 2026).
ERA5 data for training, validation and testing AtmoNAT can be accessed at data repository of WeatherBench 2: https://console.cloud.google.com/storage/browser/weatherbench2.

Author contributions

Conceptualization: XH; Formal analysis: LW; Funding acquisition: JZY; Methodology: XH; Project administration: JZY; Writing (original draft preparation): LW; Writing (review and editing): LW, DZ, JZY, BP, DX, XH.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The authors thank P. Zhang for his guidance on formulating the idea. The authors thank the insights from the anonymous reviewers to improve the quality of this work. During the preparation of this manuscript, the authors used Gemini 3 for the purposes of language polishing.

Financial support

This work is funded by Smart-Grid National Science and Technology Major Project (Grant No. 2025ZD0805500); Science and Technology Project of the North China Branch of State Grid Corporation of China (Power Meteorological Forecasting Technology Combining Physical Mechanisms and Data-Driven Approaches, Supporting Project for Subproject 3 of National Science and Technology Major Project, Grant No. 52992326000X).

References

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47-55, https://doi.org/10.1038/nature14956, 2015.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533-538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K, Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth system, Nature, 641, 1180-1187, https://doi.org/10.1038/s41586-025-09005-y, 2025.

Chen, K., Han, T., Ling, F., Gong, J., Bai, L., Wang, X., Luo, J.-J., Fei, B., Zhang, W., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Yang, X., and Ouyang, W.: The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time, Communications Earth & Environment, 6, 518, https://doi.org/10.1038/s43247-025-02502-y, 2025.

Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., and Huang, G.: Frequency-aware feature fusion for dense image prediction, IEEE Trans. Pattern Anal. Mach. Intell., 46, 10763–10780, https://doi.org/10.1109/TPAMI.2024.3449959, 2024.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H.: FuXi: A cascade machine learning forecasting system for 15-day global weather forecast, npj Clim. Atmos. Sci., 6, 190, https://doi.org/10.1038/s41612-023-00512-1, 2023.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv [preprint], arXiv:1412.3555, 11 December 2014.

Couairon, G., Lessig, C., Charantonis, A., and Monteleoni, C.: ArchesWeather: An efficient AI weather forecasting model at 1.5° resolution, arXiv [preprint], arXiv:2405.14527, 23 May 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv [preprint], arXiv:2010.11929, 22 October 2020.

Du, H., Kim, L., Creus-Costa, J., Michaels, J., Shetty, A., Hutchinson, T., Riedel, C., and Dean, J.: WeatherMesh-3: Fast and accurate operational global weather forecasting, arXiv [preprint], arXiv:2503.22235, 28 March 2025.

Esteves, C., Slotine, J. J., and Makadia, A.: Scaling spherical CNNs, in: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, Hawaii, USA, 23–29 July 2023, 9396–9411, 2023.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., …, and Ma, Z.: The Llama 3 herd of models, arXiv [preprint], arXiv:2407.21783, 31 July 2024.

Haiden, T., Janousek, M., Prates, F., Maier-Gerber, M., Li, C., and Chevallier, M.: Evaluation of ECMWF forecasts, ECMWF Tech. Memo. 931, European Centre for Medium-Range Weather Forecasts, Reading, UK, 2025.

Hassani, A., Walton, S., Li, J., Li, S., and Shi, H.: Neighborhood attention transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, Canada, 18–22 June 2023, 6185–6194, https://doi.org/10.1109/CVPR52729.2023.00599, 2023.

Hendrycks, D. and Gimpel, K.: Gaussian error linear units (GELUs), arXiv [preprint], arXiv:1606.08415, 27 June 2016.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Heo, B., Park, S., Han, D., and Yun, S.: Rotary position embedding for vision transformer, in: Proceedings of the European Conference on Computer Vision (ECCV 2024), Milan, Italy, 29 September–4 October 2024, 289–305, https://doi.org/10.1007/978-3-031-72684-2_17, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Huang, X. and Wang, L.: Neighborhood Attention Transformer for Atmospheric prediction (AtmoNAT), Zenodo [code], https://doi.org/10.5281/zenodo.19369025, 2026.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D.: Scaling laws for neural language models, arXiv [preprint], arXiv:2001.08361, 23 January 2020.

Keisler, R.: Forecasting global weather with graph neural networks, arXiv [preprint], arXiv:2202.07575, 15 February 2022.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural general circulation models for weather and climate, Nature, 632, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y, 2024.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S.: Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawaii, USA, 21–26 July 2017, 2117–2125, https://doi.org/10.1109/CVPR.2017.106, 2017.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, Canada, 11–17 October 2021, 10012–10022, 2021.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, arXiv [preprint], arXiv:1711.05101, 14 November 2017.

Munir, M., Rahman, M. M., and Marculescu, R.: AdaptViG: Adaptive Vision GNN with exponential decay gating, arXiv [preprint], arXiv:2511.09942, 15 November 2025.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Maulik, R., Kotamarthi, V., Foster, I., Madireddy, S., and Grover, A.: Scaling transformer neural networks for skillful and reliable medium-range weather forecasting, in: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada, 10–15 December 2024, 68740–68771, 2024.

Niu, P., Ma, Z., Zhou, T., Chen, W., Shen, L., Jin, R., and Sun, L.: Utilizing strategic pre-training to reduce overfitting: Baguan – a pre-trained weather forecasting model, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington D.C., USA, 3–7 August 2025, 2186–2197, https://doi.org/10.1145/3711896.3737178, 2025.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An imperative style, high-performance deep learning library, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 8–14 December 2019, 8026–8037, 2019.

Pathak, J., Abbas, M. S., Harrington, P., Hu, Z., Brenowitz, N., Ravuri, S., Carpentieri, A., Leinonen, J., Adams, C., Hennigh, O., Geneva, N., Durran, D., and Pritchard, M.: Learning accurate storm-scale evolution from observations, arXiv [preprint], arXiv:2601.17268, 24 January 2026.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., and Lin, J.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free, in: Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, USA, 2–7 December 2025.

https://doi.org/10.5194/egusphere-2026-1990
Preprint. Discussion started: 14 April 2026

Ramachandran, P., Zoph, B., and Le, Q. V.: Searching for activation functions, arXiv [preprint], arXiv:1710.05941, 16 October 2017.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A benchmark for the next generation of data-driven global weather models, J. Adv. Model. Earth Syst., 16, e2023MS004019, https://doi.org/10.1029/2023MS004019, 2024.

Shazeer, N.: GLU variants improve transformer, arXiv [preprint], arXiv:2002.05202, 12 February 2020.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding, Neurocomputing, 568, 127063, https://doi.org/10.1016/j.neucom.2023.127063, 2024.

To, D., Quinting, J., Hoshyaripour, G. A., Götz, M., Streit, A., and Debus, C.: Architectural insights into and training methodology optimization of Pangu-Weather, Geosci. Model Dev., 17, 8873–8884, https://doi.org/10.5194/gmd-17-8873-2024, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, in: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, USA, 4–9 December 2017, 5998–6008, 2017.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z.: Qwen2.5 technical report, arXiv [preprint], arXiv:2412.15115, 19 December 2024.

Yang, S., Kautz, J., and Hatamizadeh, A.: Gated delta networks: Improving Mamba2 with delta rule, arXiv [preprint], arXiv:2412.06464, 9 December 2024.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M. H.: Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, USA, 19–24 June 2022, 5718–5729, 2022.