https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Mamba-Stormer v1.0: A Bidirectional Vision Mamba Backbone for Accurate and Scalable Global Weather Forecasting

Jiayou Chao[1], Guanchao Tong[2, 3], Zhenhua Liu[1], Wuyin Lin[4], Minghua Zhang[5], Merry Ma[6], and Wei Zhu[1]

[1]Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11790, USA
[2]College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, 325060, China
[3]College of Science, Mathematics and Technology, Kean University, 1000 Morris Avenue, Union, 07083, USA
[4]Environmental and Climate Sciences Department, Brookhaven National Laboratory, Upton, NY, 11973-5000, USA
[5]School of Marine and Atmospheric Sciences, State University of New York at Stony Brook, Stony Brook, NY 11790, USA
[6]Center for Data Science, New York University, New York, NY 10011, USA

Correspondence: Wei Zhu (wei.zhu@stonybrook.edu)

Abstract. Accurate global weather forecasting at increasing spatial resolution is a central challenge in atmospheric science. Machine learning models have recently matched or exceeded numerical weather prediction systems on standard medium-range benchmarks, with transformer-based architectures at their core. However, the self-attention mechanism underlying these models scales quadratically in memory and compute with sequence length — a critical bottleneck as training and inference resolutions approach operational standards. We describe Mamba-Stormer, which replaces the transformer backbone in the Stormer architecture with a bidirectional Vision Mamba (BiMamba) backbone. Our key design contributions are: (1) bidirectional Mamba scanning applied in an alternating horizontal–vertical pattern across 14 layers to capture 2D atmospheric structure, and (2) the integration of bidirectional SSMs into Stormer's adaLN-Zero residual block by replacing the self-attention sub-block while retaining the feed-forward sub-block and zero-initialized lead-time conditioning. In a controlled local comparison on ERA5 WeatherBench 2 (240x121 grid, 69 atmospheric variables, multi-step finetuned), Mamba-Stormer outperforms the 24-layer Transformer baseline: +2.73% mean per-variable RMSE improvement at 6h, +2.34% at 72h, and +1.07% at 120h (all $p < 0.01$), with BiMamba winning on 67/69, 69/69, and 66/69 variables respectively. Simultaneously, its $O(L)$ backbone delivers a 1.10x computational throughput speedup at the WeatherBench 2 1.5° training resolution (121x240), growing to 2.59x at 512x512 as the $O(L^2)$ attention cost increasingly dominates. These results suggest that bidirectional Vision Mamba is a strong backbone for neural weather prediction — achieving better accuracy and growing computational efficiency, consistent with a more suitable spatial inductive bias for atmospheric modeling.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

1 Introduction

In recent years, machine-learning models have matched or exceeded numerical weather prediction (NWP) on standard medium-range forecasting benchmarks. Traditional NWP, which has undergone a "quiet revolution" over the past several decades (Bauer et al., 2015), is typically represented by operational systems such as the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS). Models such as Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FourCastNet (Pathak et al., 2022), Aurora (Bodnar et al., 2024), GenCast (Price et al., 2024), AIFS (Lang et al., 2024), and the original Stormer (Nguyen et al., 2023b) have demonstrated that data-driven approaches can match or exceed the ECMWF IFS on standard benchmarks (Rasp et al., 2024).

At the core of these high-performing models lies the vision transformer (ViT) (Dosovitskiy et al., 2021), adapted to the spherical geometry and physical structure of the atmosphere. The transformer's self-attention mechanism captures global spatial dependencies in a single layer — a property that is critically important for representing large-scale atmospheric dynamics such as Rossby waves, jet streams, and teleconnections. However, standard self-attention computes pairwise interactions among all L tokens, yielding $O(L^2)$ memory and compute complexity. For a 1° resolution global forecast grid (360×181 latitude-longitude points), a patch size of 4 already yields $L \approx 16{,}000$ tokens — thereby making the quadratic scaling a serious bottleneck for both training and inference at resolutions approaching the operational 0.25° standard.

Mamba (Gu and Dao, 2023) and its theoretical successor Mamba-2 (Dao and Gu, 2024) are recently proposed selective state space models (SSMs) that achieve $O(L)$ time and memory complexity through an input-dependent selective scan algorithm. Unlike prior SSMs (e.g., S4 (Gu et al., 2022)), Mamba uses content-dependent gating that allows the model to selectively remember or forget information along the sequence, providing an expressiveness more comparable to standard self-attention (Vaswani et al., 2017). Hardware-aware implementations with parallel scan further reduce wall-clock time, making Mamba competitive with highly optimized attention kernels (e.g., FlashAttention-1 (Dao et al., 2022) and FlashAttention-2 (Dao, 2023)). Since its introduction, Mamba has been successfully adapted for computer vision and structured sequence modeling. Vision Mamba (Vim) (Zhu et al., 2024) introduced bidirectional Mamba blocks that were reported to match or exceed DeiT on ImageNet classification, COCO detection, and ADE20K segmentation. The critical insight of Vim is that for spatial data — where there is no inherent directional ordering — processing sequences bidirectionally (both left-to-right and right-to-left) significantly improves representational capacity relative to the causal (unidirectional) baseline. VMamba (Liu et al., 2024a) further introduced a four-way cross-scan strategy that was reported to outperform Swin Transformer with linear complexity. To extend SSMs to non-Euclidean structures, Graph-Mamba (Wang et al., 2024) investigated permutation and sequence-mapping strategies for graph-structured data. In the weather and climate domain, SSMs have begun to appear in specialized applications: MetMamba (Qin et al., 2024) applied Mamba to regional weather forecasting, MambaDS (Liu et al., 2024b) to meteorological downscaling, IceMamba (Wang et al., 2025) to seasonal sea ice prediction, and VMRNN

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

(Tang et al., 2024) to spatiotemporal forecasting by integrating Vision Mamba and LSTM blocks. However, to our knowledge, no prior work has conducted a direct, controlled replacement of the full transformer backbone with a bidirectional Vision Mamba in a global weather forecasting system.

In this work, we present Mamba-Stormer with a BiMamba backbone. Our contributions are:

- **BiMamba backbone**: We introduce a 14-layer bidirectional Vision Mamba backbone with alternating horizontal and vertical scans, designed to capture 2D atmospheric spatial structure without the $O(L^2)$ cost of attention. In our controlled local comparison, Mamba-Stormer outperforms the 24-layer Transformer baseline while using approximately 12% fewer parameters (~405M vs. ~464M).
- **Accuracy comparison**: We conduct a systematic evaluation after multi-step finetuning. Mamba-Stormer achieves +2.73% mean per-variable RMSE improvement at 6h, +2.34% at 72h, and +1.07% at 120h (all $p < 0.01$), with BiMamba improving on all 69 variables at the 72h horizon.
- **Throughput analysis**: We quantify the throughput advantage, observing speedups of 1.10x at the WeatherBench 2 1.5° training resolution and 2.59x at 512x512 (NVIDIA H200 NVL), with the speedup growing monotonically with sequence length.
- **Block-level integration**: We integrate bidirectional SSMs into Stormer's adaLN-Zero residual block, replacing the self-attention sub-block with BiMamba while retaining the feed-forward sub-block and zero-initialization stability.

2 Background

Deep learning-based weather prediction has advanced rapidly in recent years. Early models laid the groundwork by using convolutional neural networks on cubed sphere projections to predict global atmospheric fields (Weyn et al., 2020), which led to the creation of the original WeatherBench benchmark (Rasp et al., 2020) to standardize data-driven forecasting performance. Subsequently, FourCastNet (Pathak et al., 2022) was among the first large-scale deep learning models to demonstrate competitive performance with the ECMWF IFS, using Adaptive Fourier Neural Operators (AFNO) to efficiently capture global spatial dependencies on 0.25° ERA5 data. Pangu-Weather (Bi et al., 2023), published in *Nature*, introduced a 3D Earth-specific Transformer with local attention windows that respect the vertical structure of the atmosphere, becoming the first ML model to clearly surpass IFS HRES. GraphCast (Lam et al., 2023), published in *Science*, built upon early graph-based atmospheric architectures (Keisler, 2022) by adopting graph neural networks on an icosahedral mesh to further advance forecast skill. FengWu (Chen et al., 2025) pushed skillful forecasts beyond 10.75 days through multi-modal cross-variable fusion, while FuXi (Chen et al., 2023) employed a cascade of specialized sub-models for different forecast horizons to mitigate error accumulation out to 15 days. NeuralGCM (Kochkov et al., 2024), published in *Nature*, took a fundamentally different hybrid approach, combining a differentiable dynamical core with neural parameterizations for competitive weather forecasts

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

alongside climate-scale stability. More recently, ViT-based approaches such as ClimaX (Nguyen et al., 2023a), Stormer (Nguyen et al., 2023b), and ArchesWeather (Couairon et al., 2024) have shown that simple, minimally modified Vision Transformers can achieve competitive forecasting at moderate resolution with substantially less compute. Foundation models such as Aurora (Bodnar et al., 2024) have extended this paradigm to 1.3B-parameter multi-task settings trained on heterogeneous reanalysis sources, while GenCast (Price et al., 2024) demonstrated that diffusion-based ensemble generation can outperform ECMWF ENS on the majority of skill metrics. ECMWF's own data-driven system, AIFS (Lang et al., 2024), has reached operational deployment based on a GNN–transformer architecture. WeatherBench 2 (Rasp et al., 2024) has become the standard evaluation framework, providing ERA5 data at multiple resolutions and standardized metrics for fair comparison across architectures.

The Vision Transformer (Dosovitskiy et al., 2021) (ViT) demonstrated that a pure transformer operating on sequences of image patches can match or exceed CNNs at sufficient scale, and its global receptive field makes it well-suited for capturing large-scale atmospheric dynamics such as Rossby waves and teleconnections. However, standard self-attention scales as $O(L^2)$ in both time and memory, which becomes prohibitive at high-resolution weather grids — a 0.25° global grid yields ~65,000 grid points, and even with patch size 4, the resulting ~16,000 tokens make full self-attention prohibitively expensive. The Swin Transformer (Liu et al., 2021) introduced hierarchical local window-based attention to achieve $O(L)$ complexity for a fixed window size, though at the cost of global receptive field within each layer. FlashAttention (Dao et al., 2022) and xFormers (Rabe and Staats, 2021) take a hardware-centric approach, achieving 2–4x wall-clock speedup without approximation — the latter is used by Stormer's transformer blocks, which is why Mamba's asymptotic advantage becomes most pronounced at longer sequence lengths.

S4 (Gu et al., 2022) introduced the structured state space model via a global convolution in $O(L \log L)$ time. Mamba (Gu and Dao, 2023) addressed S4's expressiveness limitation by making SSM parameters input-dependent ("selective"), achieving $O(L)$ time and memory through a hardware-aware parallel scan. In computer vision, Vision Mamba (Zhu et al., 2024) proposed bidirectional Mamba blocks that matched DeiT on ImageNet, COCO, and ADE20K — demonstrating that bidirectionality is the key ingredient for spatial sequence modeling with SSMs, since causal scanning imposes an unnatural directional bias on 2D patches. VMamba (Liu et al., 2024a) further showed with a four-way cross-scan that capturing multiple spatial directions yields advantages most pronounced at higher resolutions. In weather and climate applications, SSMs remain nascent: VMRNN (Tang et al., 2024) combined Vision Mamba and LSTM for efficient spatiotemporal forecasting; MetMamba (Qin et al., 2024) applied Mamba to regional weather forecasting with a 3D cross-scan strategy; MambaDS (Liu et al., 2024b) targeted meteorological downscaling; and IceMamba (Wang et al., 2025) addressed seasonal sea ice prediction. To our knowledge, no prior work has conducted a controlled architectural comparison of bidirectional Vision Mamba versus the transformer backbone in global weather forecasting, which is the contribution of the present work.

The adaptive layer normalization zero (adaLN-Zero) conditioning mechanism, introduced in Diffusion Transformers (DiT; (Peebles and Xie, 2023)), has emerged as a widely adopted method for incorporating conditioning signals

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

into transformer architectures. In DiT, a conditioning vector is projected to produce 6 modulation parameters per block — scale, shift, and gate for each of the two sub-blocks (attention and MLP) — where zero initialization ensures that each block initially behaves approximately as an identity mapping during early-stage optimization. Stormer adopted adaLN-Zero to condition its transformer blocks on the forecast lead time τ, enabling a single model to handle variable forecast horizons. In the present work, we retain this two-sub-block conditioning formulation while replacing the self-attention sub-block with a bidirectional Mamba layer, so the BiMamba block preserves an identical zero-initialized lead-time conditioning interface as the Transformer baseline.

3 Model Description

3.1 Stormer Architecture

The original Stormer (Nguyen et al., 2023b) is a vision transformer architecture adapted for global weather forecasting. Given a set of V atmospheric variables on a lat-lon grid of size $H \times W$, the model maps an input state $\mathbf{x} \in \mathbb{R}^{V \times H \times W}$ and a lead time τ (in hours) to a predicted difference $\Delta\hat{\mathbf{x}} \in \mathbb{R}^{V \times H \times W}$, from which the forecast is recovered as $\hat{\mathbf{x}}_\tau = \mathbf{x} + \Delta\hat{\mathbf{x}}$.

Weather Embedding. Each variable is independently processed by a separate patch-embedding module (a 2D convolutional projection) that maps $p \times p$ spatial patches to D-dimensional token vectors. Variable-specific embeddings are aggregated via a cross-attention layer using a learnable query before passing to the main backbone, yielding a sequence of $L = \lceil H'/p \rceil \times \lceil W'/p \rceil$ tokens of dimension D, where H' and W' are the padded grid dimensions. In our experiments, $p = 4$, $D = 1024$, and the grid is padded from 240x121 to 240x124 to make both dimensions divisible by p, giving $L = 60 \times 31 = 1{,}860$ tokens.

Timestep Embedding. The lead time τ is embedded via a linear projection $\mathbf{c} = W_t\tau \in \mathbb{R}^D$, where $W_t \in \mathbb{R}^{D \times 1}$. This conditioning vector $\mathbf{c}$ is shared across all backbone blocks.

Transformer Blocks (Baseline). The 24-layer baseline backbone applies multi-head self-attention followed by a position-wise feed-forward network (MLP), both conditioned on $\mathbf{c}$ via adaptive layer norm zero (adaLN-Zero; (Peebles and Xie, 2023)):

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{g}_{\text{msa}} \odot \text{Attention}(\text{LN}(\mathbf{x}; \boldsymbol{\alpha}_{\text{msa}}, \boldsymbol{\beta}_{\text{msa}})) \tag{1}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{g}_{\text{mlp}} \odot \text{MLP}\big(\text{LN}(\mathbf{x}; \boldsymbol{\alpha}_{\text{mlp}}, \boldsymbol{\beta}_{\text{mlp}})\big) \tag{2}$$

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

where $(\boldsymbol{\alpha}_{\text{msa}}, \boldsymbol{\beta}_{\text{msa}}, \mathbf{g}_{\text{msa}}, \boldsymbol{\alpha}_{\text{mlp}}, \boldsymbol{\beta}_{\text{mlp}}, \mathbf{g}_{\text{mlp}}) = \text{Linear}_{6D}(\text{SiLU}(\mathbf{c}))$ are 6 learned modulation vectors of dimension D, produced by a linear map $\text{Linear}_{6D} : \mathbb{R}^D \to \mathbb{R}^{6D}$, and $\text{LN}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = (1 + \boldsymbol{\alpha}) \odot \text{LayerNorm}(\mathbf{x}) + \boldsymbol{\beta}$ is the modulated layer norm. All modulation weights are initialized to zero, ensuring identity mapping at initialization (the "Zero" in adaLN-Zero).

The attention sub-block uses memory-efficient attention from xFormers (Rabe and Staats, 2021) with 16 heads, and the MLP uses a 4x expansion ratio with approximate GELU activation.

Output Head. A linear output module decodes the token sequence back to weather variable patches via a linear projection, conditioned on $\mathbf{c}$ by a 2-parameter adaLN modulation. The patches are reassembled into the output grid $\Delta\hat{\mathbf{x}}$.

3.2 BiMamba Block Design

We replace all transformer blocks with bidirectional Mamba blocks implementing Vision Mamba's bidirectional scanning (Zhu et al., 2024).

3.2.1 Bidirectional Mamba Processing

Each block processes the sequence $\mathbf{x} \in \mathbb{R}^{B \times L \times D}$ simultaneously in both causal directions. The following is a simplified presentation; the full formulation including gating, input projection, and output projection follows (Gu and Dao, 2023). The standard (causal) Mamba scan maintains a recurrent hidden state $\mathbf{h}_t \in \mathbb{R}^N$ per channel governed by the selective state-space equations:

$$\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t x_t, \qquad y_t = \mathbf{C}_t \mathbf{h}_t + D x_t \tag{3}$$

where the discretized matrices $\bar{\mathbf{A}}_t \in \mathbb{R}^{N \times N}$, $\bar{\mathbf{B}}_t \in \mathbb{R}^N$, the output projection $\mathbf{C}_t \in \mathbb{R}^N$, and the time-step Δ_t are all functions of the input x_t — the key "selective" mechanism. The bidirectional extension runs two such scans in parallel: one forward ($t = 1, \ldots, L$) and one backward ($t = L, \ldots, 1$). Their outputs $\mathbf{y}^{\text{fwd}}, \mathbf{y}^{\text{bwd}} \in \mathbb{R}^{B \times L \times D}$ are normalized and fused element-wise:

$$\mathbf{y} = \frac{1}{2}\left[\text{LayerNorm}(\mathbf{y}^{\text{fwd}}) + \text{LayerNorm}(\mathbf{y}^{\text{bwd}})\right] \tag{4}$$

We apply LayerNorm to each directional output before fusion; empirically, omitting this step leads to mismatched output magnitudes and numerical instability during training on weather data distributions.

We use Mamba with state dimension $d_{\text{state}} = 32$, convolution width $d_{\text{conv}} = 4$, and expansion factor $\text{expand} = 2$ (inner dimension $2D = 2048$).

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

3.2.2 Alternating 2D Scans

A 1D scan sequence is insufficient for 2D atmospheric grids, as it imposes an arbitrary spatial ordering. We employ alternating scan directions across layers to build up 2D spatial receptive fields:

- **Even layers** ($\ell = 0, 2, 4, \ldots$): Bidirectional horizontal scan in row-major order (each row processed left-to-right and right-to-left).
- **Odd layers** ($\ell = 1, 3, 5, \ldots$): Bidirectional vertical scan, implemented by first transposing the token grid to column-major order, scanning, then transposing back.

After every two layers, each token has integrated information from both its full spatial row and full spatial column, thereby providing structured global spatial coverage along both horizontal and vertical axes. Across 14 layers, corresponding to seven horizontal–vertical scan pairs, this alternating design progressively builds dense, multi-scale two-dimensional spatial representations. Vertical scans are implemented by first transposing the token grid into column-major order before applying the scan, and then applying the inverse permutation to restore the original layout. This design preserves contiguous memory access patterns for the optimized CUDA kernel while enabling efficient directional scanning.

3.2.3 Training Stabilization for BiMamba

SSM training on weather data requires several architectural and optimization considerations beyond those typically used in image classification tasks. First, we use a state dimension of $d_{\text{state}} = 32$ (versus 16 in the original Mamba paper) to increase hidden memory capacity for long-range atmospheric patterns such as Rossby waves and jet streams. Second, the weight initialization routine preserves the specialized internal Mamba parameters (time-step bias and log-diagonal state matrix) rather than overwriting them with generic random initialization, maintaining the SSM's intended temporal inductive bias at the start of training.

3.3 adaLN-Zero Conditioning for BiMamba

The BiMamba block preserves Stormer's two-sub-block adaLN-Zero residual skeleton. The attention sub-block is replaced by a bidirectional Mamba operator, while the position-wise MLP is retained:

$$(\boldsymbol{\alpha}_{\text{bm}}, \boldsymbol{\beta}_{\text{bm}}, \mathbf{g}_{\text{bm}}, \boldsymbol{\alpha}_{\text{mlp}}, \boldsymbol{\beta}_{\text{mlp}}, \mathbf{g}_{\text{mlp}}) = \text{Linear}_{6D}(\text{SiLU}(\mathbf{c})) \tag{5}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{g}_{\text{bm}} \odot \text{BiMamba}(\text{LN}(\mathbf{x}; \boldsymbol{\alpha}_{\text{bm}}, \boldsymbol{\beta}_{\text{bm}})) \tag{6}$$

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{g}_{\text{mlp}} \odot \text{MLP}\big(\text{LN}(\mathbf{x}; \boldsymbol{\alpha}_{\text{mlp}}, \boldsymbol{\beta}_{\text{mlp}})\big) \tag{7}$$

The conditioning linear layer and its bias are initialized to zero, ensuring that each BiMamba block behaves approximately as an identity at initialization, consistent with the design of the original adaLN-Zero Transformer block. Importantly, the lead-time conditioning mechanism is preserved without architectural modification, requiring no modifications to the timestep embedder or the training objective.

The complete architecture comparison is shown in Fig. 1.

3.4 Training Strategy

Training follows the two-phase strategy of the original Stormer design (Nguyen et al., 2023b). After one-step pretraining, both models undergo multi-step finetuning to improve rollout stability and reduce long-range error accumulation. The finetuning phase unrolls four consecutive 6-hour steps during training, giving a 24-hour optimization window.

One-step pretraining. The model is trained to predict the atmospheric state at a single future time step. Lead times are randomly sampled from $\{6, 12, 24\}$ hours at each iteration. Pretraining runs for 100 epochs from scratch, and the best checkpoint is selected by validation loss at the 72h lead time.

Multi-step finetuning. Finetuning initializes from the best pretraining checkpoint for each backbone and runs for 20 epochs with 4-step autoregressive unrolling. Validation is performed at 6h, 72h, 120h, and 168h, and the final checkpoint is selected by validation loss at the 168h lead time.

Optimization and learning-rate schedule. Both phases use AdamW (Kingma and Ba, 2014; Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, weight decay 10^{-5}, cosine decay, and gradient clipping at norm 0.5. During one-step pretraining, the Transformer uses a peak learning rate of 5×10^{-4} with $\beta_2 = 0.95$, while BiMamba uses a reduced peak learning rate of 2×10^{-4} with $\beta_2 = 0.999$ for SSM stability. During multi-step finetuning, both models use a peak learning rate of 5×10^{-6} with $\beta_2 = 0.95$ and a 2-epoch warm-up.

Training objective. The training objective is a latitude-weighted mean-squared error on normalized per-variable prediction residuals. Specifically, the model predicts the normalized difference $\Delta\hat{\mathbf{x}}/\sigma_v$ for each variable v, where σ_v is the training-period standard deviation of that variable (computed globally per variable, not per grid point). The loss is the average over all $V = 69$ variables, weighted within each variable by $w_h = \cos\phi_h/\bar{w}$ (normalized to unit mean over latitude rows), consistent with the WeatherBench 2 evaluation protocol (Rasp et al., 2024).

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

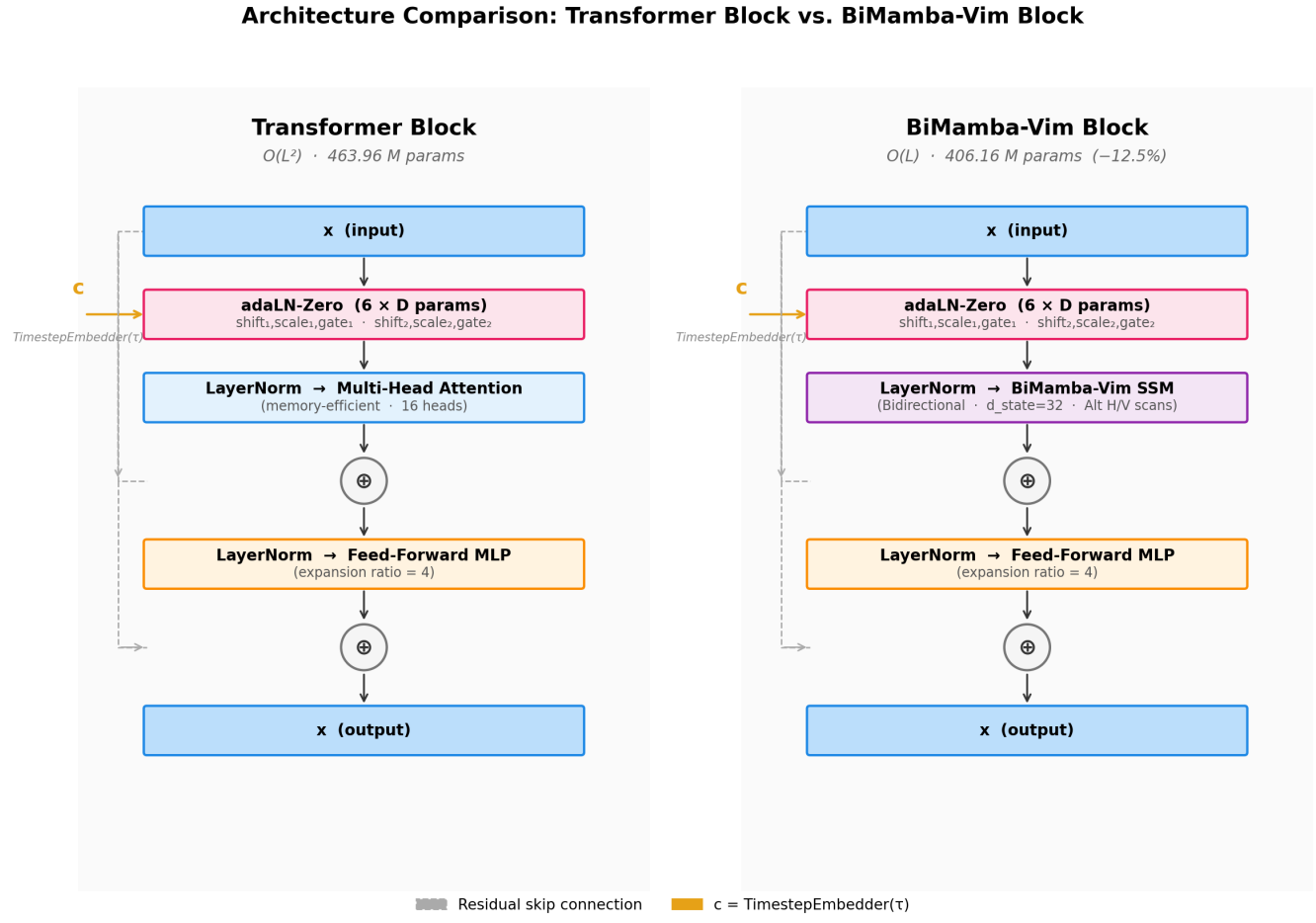

Figure 1. Architecture comparison between the original Transformer block (left) and the proposed BiMamba block (right). Both blocks receive the same conditioning vector $\mathbf{c}$ from the timestep embedder and maintain identical input/output shapes $(B, L, D) \to (B, L, D)$. The BiMamba block preserves the two-sub-block adaLN-Zero residual skeleton but replaces multi-head self-attention with a bidirectional selective SSM; the feed-forward MLP is retained. The scan direction (H or V) alternates by layer index.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

4 Model Evaluation

4.1 Data

We use the ERA5 atmospheric reanalysis dataset (Hersbach et al., 2020) from the WeatherBench 2 benchmark (Rasp et al., 2024) at approximately 1.5° resolution (240 longitude x 121 latitude grid points, 6-hourly). The model predicts a total of 69 atmospheric variables, partitioned into surface and pressure-level variables as follows:

- **Surface variables** (4): 2-m temperature (T2m), 10-m zonal wind velocity (U10), 10-m meridional wind velocity (V10), and mean sea level pressure (MSLP).
- **Pressure-level variables** (65): geopotential height (Z), temperature (T), zonal wind velocity (U), meridional wind velocity (V), and specific humidity (Q) at 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa).

All variables are normalized by subtracting the temporal climatological mean and dividing by the standard deviation. Both normalization statistics are computed independently for each variable over the 1979–2018 training interval using all spatial locations jointly, rather than separately at each grid point.

The dataset is partitioned chronologically into mutually exclusive training, validation, and test subsets: 1979–2018 for training (approximately 58,000 six-hourly atmospheric states), 2019 for validation, and 2020–2023 for testing (out-of-sample evaluation).

4.2 Training Setup

Both Transformer-Stormer and Mamba-Stormer (BiMamba) are pretrained from scratch and then multi-step finetuned on 4x NVIDIA H200 NVL GPUs using PyTorch DDP within the PyTorch Lightning framework (Falcon et al., 2020). One-step pretraining uses per-GPU batch size 4 (global batch 16). Multi-step finetuning uses per-GPU batch size 2 for both models. Both models use identical hardware, data splits, normalization files, and random seed (42) to ensure a fair comparison.

Both models share the same hidden dimension ($D = 1024$), patch size ($p = 4$), and gradient clipping (norm, 0.5). BiMamba uses 14 layers versus the Transformer's 24, chosen to keep total model size comparable given the higher per-block parameter count of bidirectional Mamba. Full hyperparameter details for both training phases are provided in Table 4 (Appendix).

All final accuracy numbers reported below come from the multi-step finetuned checkpoints. Full software environment details are provided in Table 4 (Appendix).

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

4.3 Evaluation Metrics

We report latitude-weighted root mean square error (RMSE) at 6h, 72h, and 120h lead times. Final manuscript results are computed on the full 2020–2023 test set (4404 samples) using the multi-step finetuned checkpoints, ensemble-mean rollout evaluation, and WeatherBench 2 RMSE aggregation. Predictions are first denormalized to physical units before error computation. Tables report physical-unit RMSE averaged over the 69 per-variable RMSEs; because variables span heterogeneous physical units, these averages are not directly comparable in magnitude across variables, but percentage differences are unit-consistent within each variable and constitute the primary accuracy comparison. Figures report climatology-normalized RMSE (per-variable RMSE divided by the corresponding training-period standard deviation) to place variables on a common scale.

The per-variable latitude-weighted RMSE is defined as:

$$\mathrm{RMSE}_v = \sqrt{\frac{\sum_{h,w} w_h \cdot \overline{(\hat{x}_{v,h,w} - x_{v,h,w})^2}}{W \sum_h w_h}}, \qquad w_h = \frac{\cos\phi_h}{\bar{w}} \tag{8}$$

where ϕ_h is the latitude at row h, W is the number of longitude columns, weights w_h are normalized to unit mean ($\bar{w} = \frac{1}{H}\sum_h \cos\phi_h$), and $\overline{(\cdot)}$ denotes the mean over test samples. Predictions $\hat{x}_{v,h,w}$ are in physical units (post-denormalization). For visualization only, RMSE_v is divided by the corresponding climatological standard deviation before averaging across variables. The aggregate RMSE improvement reported in Table 1 is computed from the equal-weighted mean RMSE over all $V = 69$ variables:

$$\overline{\mathrm{RMSE}}^{(m)} = \frac{1}{V}\sum_{v=1}^{V} \mathrm{RMSE}_v^{(m)}, \qquad \Delta\mathrm{RMSE} = \left(1 - \frac{\overline{\mathrm{RMSE}}^{\mathrm{BiMamba}}}{\overline{\mathrm{RMSE}}^{\mathrm{Transformer}}}\right) \times 100\%. \tag{9}$$

Forecast generation at 72h and 120h. The model is trained on lead times $\tau \in \{6, 12, 24\}$h only. In the final multi-step finetuning evaluation, multi-step forecasts are produced with an **ensemble-mean rollout** over all valid base intervals dividing the requested lead time. For 72h, we average rollouts of $(12 \times 6\text{h})$, $(6 \times 12\text{h})$, and $(3 \times 24\text{h})$; for 120h, we average $(20 \times 6\text{h})$, $(10 \times 12\text{h})$, and $(5 \times 24\text{h})$. Each rollout repeatedly denormalizes the predicted difference, adds it to the previous state in physical space, and renormalizes before the next step.

Statistical significance assessment. We perform a paired t-test comparing per-variable RMSE values at each forecast lead time, treating each atmospheric variable as a paired observation. However, the assumption of statistical independence is only approximately satisfied, since adjacent pressure levels exhibit substantial physical correlation and therefore share variance structure. Consequently, the effective sample size is smaller than the nominal value of 69

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

variables, implying that the resulting p-values should be interpreted cautiously rather than as exact measures of statistical significance.

Nevertheless, the paired test remains informative as a summary statistic for evaluating whether the observed improvements are directionally consistent across variables, which constitutes the primary claim of this study. This interpretation is further supported by the direct per-variable comparison counts — for example, BiMamba achieves lower RMSE on 69/69 variables at the 72-hour forecast horizon.

4.4 Local Baseline Configuration and Comparison Protocol

Our locally trained Transformer baseline exhibits minor discrepancies relative to the officially reported Stormer WeatherBench 2 results for certain variables. These differences likely arise from implementation-level factors, including variations in HDF5 preprocessing pipelines, normalization statistics, or the 240x121 spatial regridding procedure. To eliminate such confounding factors, all comparative evaluations in this study are conducted against our locally trained Transformer baseline rather than against the originally published Stormer results.

Both architectures were pretrained, fine-tuned, and evaluated within an identical experimental pipeline using the same data splits, normalization files, hardware configuration, and random seed. However, selected optimization hyperparameters — including peak learning rate and AdamW β_2 — were tuned independently for each backbone in order to ensure stable and architecture-appropriate training dynamics. Consequently, the reported results should be interpreted as a comparison between practically optimized implementations of the Transformer and BiMamba backbones, rather than as a strictly controlled isolation of optimizer-invariant architectural effects.

4.5 Architecture Ablation and Stabilization Analysis

Replacing self-attention with selective state space layers required additional architecture-specific stabilization mechanisms. In preliminary validation experiments, a unidirectional Mamba replacement performed substantially worse than the Transformer baseline. Across two independent training runs, the normalized latitude-weighted validation MSE at the 72-hour ensemble-mean forecast horizon increased by approximately 141–189% relative to the Transformer baseline. These results suggest that a purely causal one-dimensional scan is insufficient for modeling the unordered spatial structure of two-dimensional atmospheric fields. Consequently, these preliminary findings motivated the adoption of bidirectional spatial scanning. Importantly, these diagnostic experiments were exploratory only and are not included in the final benchmark evaluation.

However, naïvely introducing bidirectional scanning did not, by itself, resolve the optimization difficulties. An early bidirectional implementation initially achieved competitive validation performance by epoch 30, reaching a validation MSE of 0.0655, but subsequently exhibited catastrophic training instability. Specifically, the validation MSE increased sharply to 0.4211 at epoch 31 — corresponding to an increase of approximately +543% relative to the previous epoch

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

— and remained elevated at 0.490 by epoch 39, representing a total increase of approximately +648% relative to epoch 30.

These instability diagnostics motivated the final stabilized BiMamba architecture used throughout all reported experiments. In the final design, bidirectional scan outputs are fused only after an intermediate pre-fusion LayerNorm operation, the original internal Mamba initialization scheme is explicitly preserved, and the one-step pretraining learning rate is reduced to 2×10^{-4} to improve optimization stability for selective state space dynamics.

5 Results

5.1 Forecasting Accuracy

Table 1 reports mean physical-unit latitude-weighted RMSE at 6h, 72h, and 120h lead times for the final multi-step finetuned checkpoints.

Table 1: Forecasting accuracy on the 2020–2023 ERA5 test set using the full 4404-sample multi-step finetuning evaluation (`ensemble_mean`, WB2-style RMSE aggregation). Reported RMSE values are physical-unit means over the 69 per-variable RMSEs; RMSE Improvement is the percentage reduction in this equal-weighted mean RMSE, $(1 - \overline{\text{RMSE}}_{\text{BiMamba}}/\overline{\text{RMSE}}_{\text{Transformer}})$. Lower RMSE is better.

Lead Time	Transformer RMSE	BiMamba RMSE	RMSE Improvement	p-value
6 hours	7.615	7.407	**+2.73%**	1.18×10^{-4}
72 hours	32.082	31.333	**+2.34%**	1.51×10^{-4}
120 hours	62.881	62.210	**+1.07%**	4.93×10^{-3}

Mamba-Stormer achieves consistent improvements over the Transformer at all evaluated lead times. All three gains are statistically significant under a paired t-test across variables: $p = 1.18 \times 10^{-4}$ at 6h, 1.51×10^{-4} at 72h, and 4.93×10^{-3} at 120h. The strongest consistency occurs at 72h, where BiMamba improves on all 69 variables. These results suggest that the bidirectional scanning pattern provides a spatial inductive bias well suited to atmospheric modeling, and that this advantage is reinforced rather than erased by the multi-step finetuning regime.

Figure 2 shows climatology-normalized RMSE as a function of lead time for both models.

5.2 Per-Variable Analysis

Mamba-Stormer's improvement is systematic across atmospheric variables, as illustrated in Fig. 3. At the 72h lead time, BiMamba achieves lower RMSE on all 69 variables. At 6h, two upper-tropospheric variables — temperature at

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

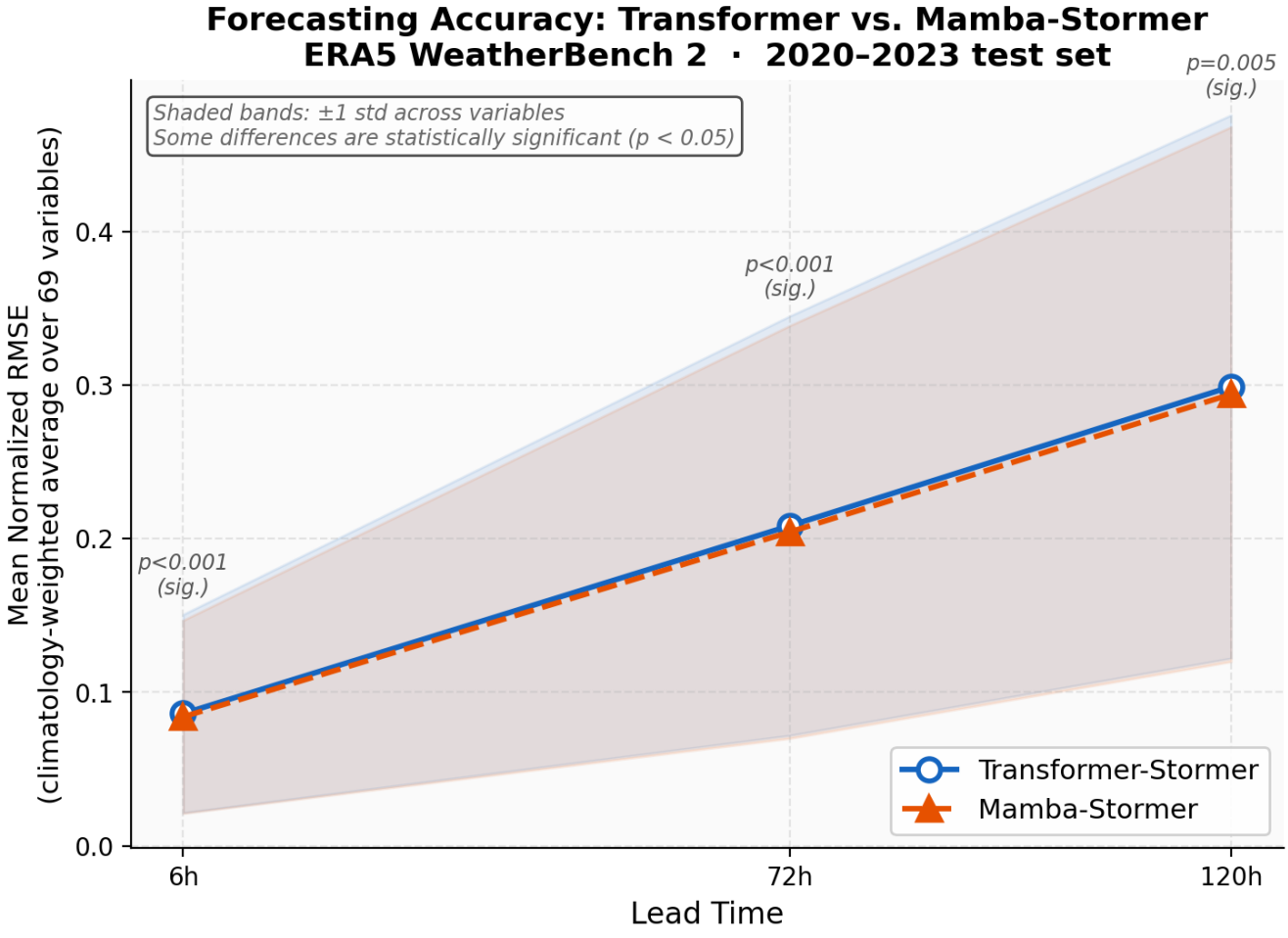

Figure 2. Climatology-normalized RMSE across lead times for Transformer-Stormer (blue) and Mamba-Stormer (orange). Shaded regions indicate ±1 standard deviation across 69 atmospheric variables. BiMamba consistently achieves lower error than the Transformer baseline.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

100 hPa and zonal wind at 100 hPa — favor the Transformer by less than 1%, suggesting that the very highest pressure levels (where stratospheric dynamics dominate) may be the regime least improved by the BiMamba inductive bias. At 120h, the three exceptions are all upper-stratospheric geopotential fields (Z50, Z100, Z150), where Transformer RMSE is lower by 0.1–1.7%; the remaining 66 of 69 variables still favor BiMamba. The consistent pattern across the four surface variables and the lower-tropospheric pressure levels is unchanged across all lead times.

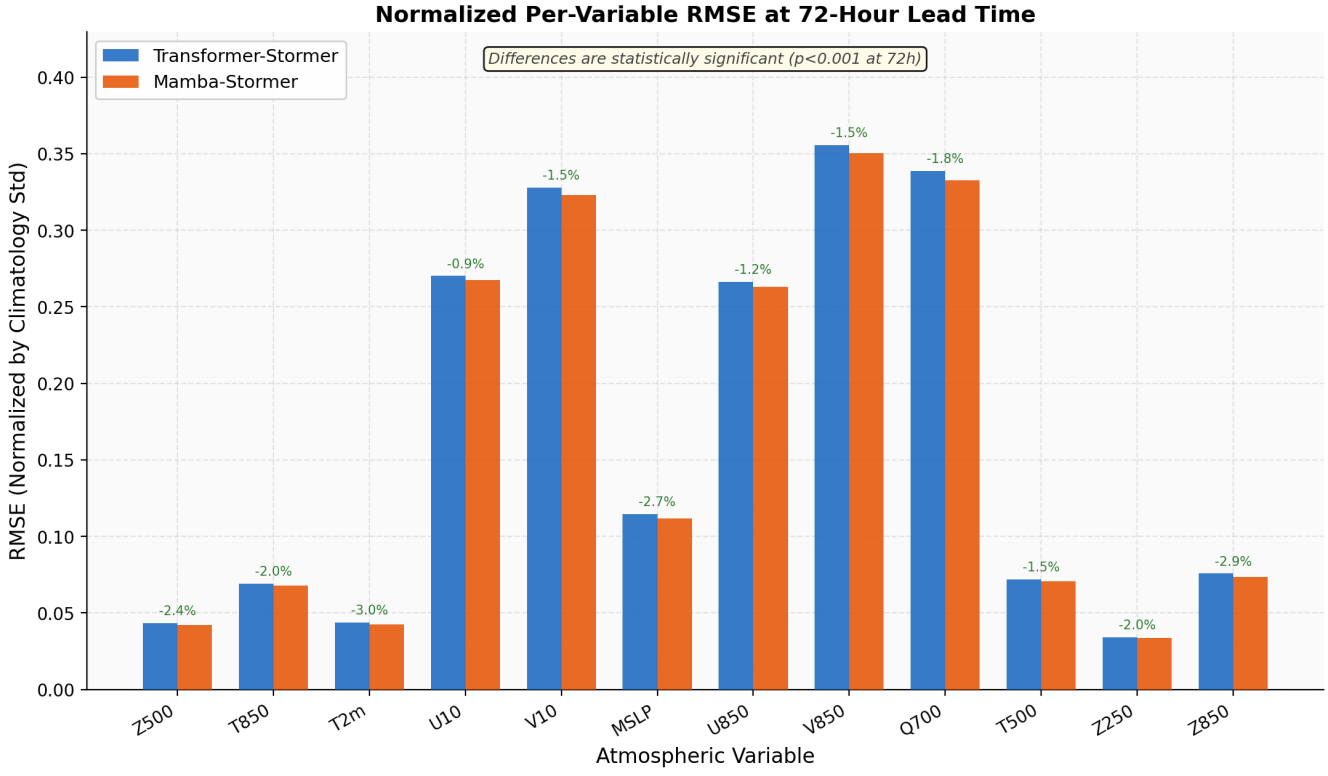

Figure 3. Comparison of climatology-normalized per-variable RMSE at 72-hour lead time. Mamba-Stormer (orange) outperforms Transformer-Stormer (blue) on all 69 variables, demonstrating consistent gains across surface and pressure-level fields after multi-step finetuning.

5.3 Computational Efficiency

Table 2 compares model size and throughput at the WeatherBench 2 1.5° training resolution (121x240, $L \approx 1{,}860$, batch size 4) on a single NVIDIA H200 NVL GPU in fp32.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Table 2: Computational efficiency at training resolution (121x240, $L \approx 1860$, batch=4, NVIDIA H200 NVL). Measurements reflect the full model (backbone + weather embedding + output head), based on 5 warm-up and 50 measurement forward passes in fp32.

Metric	Transformer-Stormer	Mamba-Stormer (BiMamba)	Change
Parameters	463.96M	406.16M	**-12.5%**
Peak GPU memory (GB)	15.74	15.56	-1.1%
Time per sample (ms)	240.4	218.7	-9.0%
Throughput (samples/s)	4.159	4.572	—
Speedup	1.00x	**1.099x**	—

At the 1.5° training resolution, Mamba-Stormer achieves a 1.10x throughput speedup with ~12% fewer parameters and 1.1% less peak memory.

Backbone-only throughput. The full-model measurements above include the shared weather embedding (variable aggregation cross-attention) and output head, which add overhead to both models equally. To isolate the $O(L)$ backbone contribution, we conducted a separate backbone-only benchmark on the same NVIDIA H200 NVL hardware (fp32, 10 warm-up + 100 runs), using a synthetic $(B, L, D) = (4, 1860, 1024)$ token sequence as input. Under this setup, the BiMamba backbone (14 layers) achieves 1.12x throughput over the Transformer backbone (24 layers; 148.1 ms vs. 166.5 ms per batch of 4), confirming that the $O(L)$ advantage is present at the backbone level independent of the shared embedding and head cost.

Scaling with resolution. Figure 4 plots throughput speedup across resolutions. At $L \approx 8{,}000$ (256x512), BiMamba achieves 1.72x throughput. At the highest resolution tested, 512x512 ($L = 16{,}384$), Mamba-Stormer delivers a 2.59x speedup. While both models consume similar peak memory (~123 GB), the $O(L)$ backbone complexity allows BiMamba to maintain significantly higher throughput as the sequence length increases.

Table 3 provides the complete multi-resolution breakdown.

Table 3: Multi-resolution throughput benchmark on NVIDIA H200 NVL (batch=4, fp32, 141 GB VRAM). Bold row (121x240) denotes the standard training resolution; bold speedup value denotes the maximum observed gain.

Resolution (HxW)	Seq. L	Transformer (ms)	BiMamba (ms)	Speedup
32x64	128	20.5	18.5	1.107x
64x128	512	60.0	63.8	0.941x
121x240	**1,860**	**240.4**	**218.7**	**1.099x**

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Resolution (HxW)	Seq. L	Transformer (ms)	BiMamba (ms)	Speedup
128x256	2,048	255.9	238.2	1.074x
256x256	4,096	609.6	469.2	1.299x
512x256	8,192	1,606.3	933.4	1.721x
256x512	8,192	1,606.6	932.6	1.723x
512x512	16,384	4,804.9	1,854.8	**2.591x**

6 Discussion

6.1 Hypotheses for BiMamba's Accuracy Advantage

Our results show that BiMamba achieves consistently superior forecasting accuracy relative to the Transformer baseline, rather than merely matching it. Given that weather forecasting requires capturing complex multi-scale spatial dependencies, including global teleconnections, such performance improvements cannot be attributed solely to architectural novelty. We offer three hypotheses that may explain BiMamba's advantage, noting that while our experiments support these interpretations, we did not conduct ablations isolating each factor.

Hypothesis 1: Bidirectionality reduces directional bias. The causal (unidirectional) Mamba imposes a sequential ordering on the token sequence, meaning each token only integrates context from prior positions. In two-dimensional atmospheric fields, however, no intrinsic directional ordering exists: pressure anomalies correlate with neighbors in all directions. Bidirectional scanning removes this artificial asymmetry. Combined with the alternating horizontal/vertical scan pattern, each token integrates row and column context within two layers, yielding structured global spatial coverage.

Hypothesis 2: Increased state capacity. With $d_{\text{state}} = 32$, each SSM channel retains richer hidden memory, allowing better integration of long-range context. This enhanced capacity may be particularly beneficial for capturing large-scale patterns, such as Rossby waves or jet streams, which necessitate spatial context over extended ranges.

Hypothesis 3: Parameter-normalized architectural fit. The accuracy improvements occur even though BiMamba uses roughly 12% fewer parameters (~405M vs. ~464M) than the Transformer. While this suggests a better architectural alignment with atmospheric modeling, we note that the models differ not only in parameter count but also in depth, training dynamics, and optimizer configurations. Thus, parameter count alone does not fully explain the observed accuracy differences.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

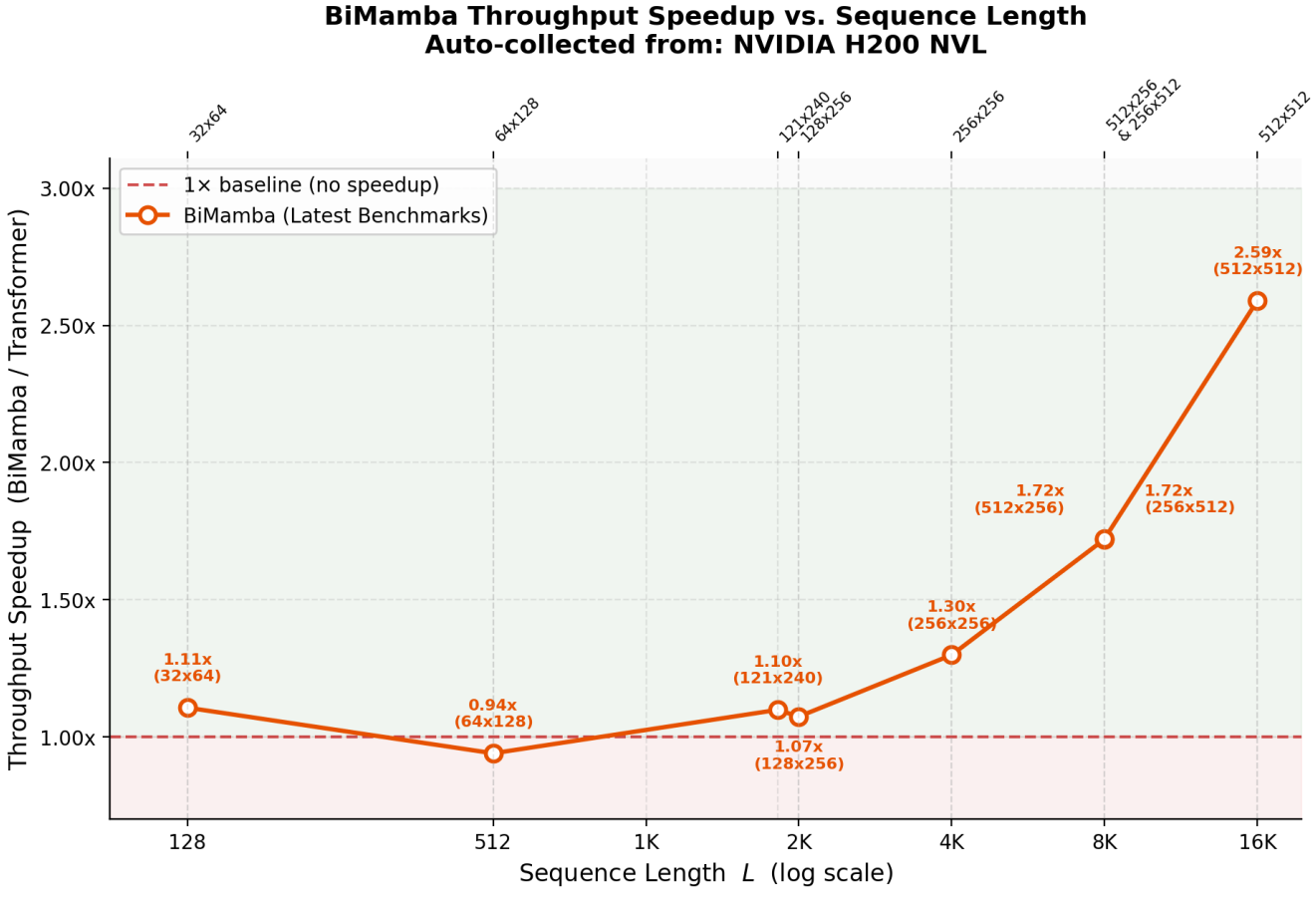

Figure 4. Throughput speedup of BiMamba over Transformer-Stormer across spatial resolutions, measured on NVIDIA H200 NVL (batch=4, fp32). The advantage grows monotonically with sequence length, reaching 2.59x at 16,384 tokens (512x512).

6.2 Meteorological Interpretation of Performance Discrepancies

The variable-specific and height-dependent accuracy differences between Mamba-Stormer and the Transformer baseline (Section 5.2) are consistent with hypotheses based on atmospheric dynamics, though we note these are post-hoc interpretations that would require targeted diagnostic experiments — such as spectral decomposition or teleconnection analysis — to validate.

In the boundary layer and lower troposphere (e.g., surface variables such as 2-m temperature, 10-m winds, and pressure-level variables at 850 hPa and 1000 hPa), atmospheric processes are characterized by highly localized, anisotropic dynamics, turbulent mixing, and advection. For these fields, we hypothesize that the alternating horizontal–vertical bidirectional scanning of BiMamba may serve as an effective spatial propagation mechanism, capturing localized thermal and momentum gradients along primary axes of advection. This would be consistent with the structured,

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

axis-aligned spatial coverage of the alternating scan design, which builds two-dimensional context progressively rather than through global pairwise interactions at every layer.

Conversely, at the highest pressure levels (e.g., geopotential height Z and wind fields at 50 hPa and 100 hPa), the dynamics are dominated by planetary-scale Rossby waves, the polar night jet, and large-scale stratospheric circulation. These processes are inherently global, exhibiting teleconnections and phase coherence over planetary distances. We hypothesize that the Transformer's direct pairwise global receptive field may be better suited to representing such long-range wave structures within a single layer. In contrast, Mamba's recurrent state bottleneck ($d_{\text{state}} = 32$) may constrain the propagation of planetary-wave phase information across the full token sequence, which could contribute to the slightly lower performance observed at high altitudes over longer forecast horizons (such as 120h) where global wave phase coherence is most critical. Confirming these interpretations would require, at minimum, spectral error decomposition by wavenumber and height.

6.3 Simultaneous Gains in Accuracy and Computational Efficiency

BiMamba yields improvements in both forecasting accuracy and computational throughput under this controlled comparison. Architecturally, this is notable because the accuracy gains do not stem from increased computational scale or parameter count; instead, they occur alongside a 12% reduction in parameters and a widening throughput advantage as sequence length increases. The synergy of bidirectional context integration, alternating two-dimensional scanning, and increased state capacity represents a deliberate adaptation of the Mamba framework to two-dimensional spatial modeling. In a manner analogous to how bidirectionality improved Vision Mamba's performance over unidirectional Mamba in image classification, these adaptations appear to be well suited to atmospheric forecasting.

6.4 Efficiency Scaling Behavior

At training resolution ($L \approx 1{,}860$), the 1.10x speedup is modest — xFormers' memory-efficient attention is already highly optimized and the sequence length is short enough that kernel overhead and projection FLOPs dominate both models. The advantage grows substantially with L: at 256x512 ($L = 8{,}192$), BiMamba is 1.72x faster. This is consistent with the $O(L)$ vs. $O(L^2)$ asymptotic difference becoming dominant beyond $L \approx 2{,}000$, where the attention computation overtakes the fixed overhead.

At 512x512 ($L = 16{,}384$), the throughput divergence is most pronounced: BiMamba is 2.59x faster than the Transformer (0.54 vs. 0.21 samples/sec on H200). Both models consume approximately 123 GB of peak VRAM (Transformer: 123.34 GB, BiMamba: 123.20 GB) under batch=4 fp32 conditions on the NVIDIA H200 NVL (141 GB VRAM). Profiling indicates that the dominant memory consumer is the shared variable aggregation cross-attention regardless of backbone.[1] Resolving this shared bottleneck through memory-efficient variable aggregation would unblock further

[1]At 512x512 with 69 variables, the variable aggregation cross-attention allocates approximately 35 GB in a single activation — comparable in scale to the backbone.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

scaling for both architectures, after which the BiMamba backbone's $O(L)$ efficiency would provide an even more decisive advantage at operational 0.25° resolutions ($L \approx 16{,}000$).

One exception in Table 3 is the 64x128 resolution ($L = 512$), where BiMamba is 6% slower than the Transformer (63.8 ms vs. 60.0 ms). This is consistent with the fixed overhead of Mamba's selective scan dominating at short sequences, where the $O(L)$ asymptotic advantage has not yet taken hold; the anomaly is absent at 32x64 ($L = 128$) and is fully amortized at training resolution ($L = 1{,}860$) and above. This edge case does not affect practical weather modeling, where training resolutions correspond to $L \geq 1{,}860$.

6.5 Implications for Ensemble Forecasting and Operational Deployment

The observed 1.10–2.59× throughput improvement (depending on resolution) has direct implications for computational efficiency in operational ensemble forecasting systems. Probabilistic weather prediction relies on generating multiple forecast realizations (ensemble members) in order to quantify forecast uncertainty and characterize the distribution of possible atmospheric trajectories (Price et al., 2024; Lang et al., 2024). Under a fixed computational budget, the higher inference throughput of the BiMamba backbone enables either the generation of larger ensembles — potentially improving uncertainty quantification and tail-risk estimation for extreme events — or more rapid delivery of probabilistic forecasting products. The latter is particularly important in operational medium-range forecasting environments, where forecast latency directly affects downstream decision-making and warning dissemination timelines.

6.6 Limitations

The present study has several limitations worth noting.

512x512 shared bottleneck. At 512x512, peak VRAM reaches ~123 GB, dominated by the shared variable aggregation cross-attention rather than the backbone. Memory-efficient variable aggregation would be needed to demonstrate the backbone-level $O(L)$ advantage at this resolution on hardware with less than 141 GB VRAM.

More modest gains at 120h. The 120h margin remains positive and statistically significant after multi-step finetuning (+1.07%, $p = 4.93 \times 10^{-3}$), but it is smaller than the gains at 6h and 72h. This suggests that longer-horizon forecasts are still the most sensitive regime for separating backbone quality.

Single training resolution. All accuracy results are for models trained at 240x121. Higher-resolution training may shift the relative accuracy balance; this remains a key direction for future research.

Limited unroll depth. Our final results use a 4-step (24h) multi-step finetuning curriculum. Longer unrolls, scheduled sampling, or horizon-aware curricula may further change the long-range accuracy balance between the two backbones.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Inference at batch size 1. Our benchmarks use batch size 4. At batch size 1, Mamba's recurrent nature means the selective scan has lower GPU utilization. Operational inference throughput may differ from training throughput benchmarks.

6.7 Future Directions

Several directions warrant future investigation.

High-resolution scaling. Training Mamba-Stormer at higher spatial resolutions, such as 0.7° or 0.35°, is a key step toward assessing operational scalability and practical deployment feasibility. At such resolutions, both the computational efficiency advantages and the representational benefits of selective state space models are expected to become increasingly pronounced as sequence length L grows.

Extended autoregressive training curricula. Alternative multi-step training strategies — including longer autoregressive unrolling horizons, scheduled sampling, and horizon-aware fine-tuning curricula — may substantially influence long-range forecast stability and accuracy. In particular, these approaches could alter the relative performance balance between transformer-based and SSM-based architectures at extended forecast lead times and therefore warrant systematic evaluation in future work.

7 Conclusions

We have presented Mamba-Stormer — an alternative to the Stormer transformer backbone based on bidirectional Vision Mamba (BiMamba). The principal architectural contributions include alternating horizontal–vertical bidirectional scan operations, which provide structured two-dimensional spatial coverage without incurring the quadratic computational complexity of self-attention; integration within Stormer's zero-initialized adaLN residual framework through replacement of the self-attention module with BiMamba; and several training stabilization strategies — including pre-fusion LayerNorm, a state dimension of $d_{\text{state}} = 32$, and preservation of the original Mamba initialization scheme — that collectively enable stable and reliable SSM training on atmospheric forecasting data.

In a controlled local comparison using the ERA5 WeatherBench 2 benchmark at 1.5° spatial resolution (69 variables with multi-step fine-tuning), Mamba-Stormer achieves consistent forecasting improvements relative to the 24-layer Transformer baseline: +2.73% mean per-variable RMSE improvement at 6-hour lead time, +2.34% at 72 hours, and +1.07% at 120 hours, with all differences statistically significant at $p < 0.01$. Moreover, Mamba-Stormer achieves lower RMSE on 67/69, 69/69, and 66/69 variables at 6-hour, 72-hour, and 120-hour lead times, respectively.

These forecasting improvements are accompanied by superior computational throughput, with Mamba-Stormer achieving a 1.10x throughput advantage at the 1.5° training resolution and a 2.59x advantage at 512x512 resolution on NVIDIA H200 NVL hardware. This efficiency gap widens as sequence length increases, reflecting the increasingly

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

dominant computational burden of quadratic self-attention complexity, $O(L^2)$. At present, both architectures remain constrained at 512x512 resolution by the shared variable aggregation module rather than by the sequence-modeling backbone itself. Addressing this bottleneck represents an important avenue for future work and may further amplify the scaling advantages of BiMamba at operational weather forecasting resolutions approaching 0.25°.

8 Appendix

8.1 Model Configuration and Training Hyperparameters

Table 4: Full model configuration and training hyperparameters for both models across the one-step pretraining and multi-step finetuning phases. Parameter counts are from the PyTorch Lightning model summary at 121x240 resolution.

Hyperparameter	Transformer	BiMamba
Depth N	24 layers	14 layers
Hidden size D	1024	1024
Patch size p	4	4
Attention heads	16	—
d_{state}	—	32
d_{conv}	—	4
expand	—	2
Parameters	~463M	~405M
Pretraining		
Peak learning rate	5×10^{-4}	2×10^{-4}
LR schedule	cosine + 10-epoch warmup	cosine + 10-epoch warmup
β_1 / β_2	0.9 / 0.95	0.9 / 0.999
Weight decay	10^{-5}	10^{-5}
Gradient clip (norm)	0.5	0.5
Per-GPU batch size	4	4
Global batch size (4x GPUs)	16	16
Max epochs	100	100
Early stopping patience	10 epochs	10 epochs
Early stopping metric	Validation weighted MSE at 72 h lead time	Validation weighted MSE at 72 h lead time
Best checkpoint epoch	99	73

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Hyperparameter	Transformer	BiMamba
Rollout steps	1	1
Multi-step Finetuning		
Peak learning rate	5×10^{-6}	5×10^{-6}
LR schedule	cosine + 2-epoch warmup	cosine + 2-epoch warmup
β_1 / β_2	0.9 / 0.95	0.9 / 0.95
Weight decay	10^{-5}	10^{-5}
Gradient clip (norm)	0.5	0.5
Per-GPU batch size	2	2
Max epochs	20	20
Checkpoint metric	Validation weighted MSE at 168 h lead time	Validation weighted MSE at 168 h lead time
Rollout steps	4	4
Software		
PyTorch / CUDA	2.10.0 / 12.8	2.10.0 / 12.8
mamba-ssm	—	2.3.1
xFormers	0.0.35	—

Code availability. The Mamba-Stormer v1.0 source code, including the model architecture, training configurations, data preprocessing scripts, and figure-generation scripts, is available under the MIT License. The code, along with the pre-trained and fine-tuned model checkpoints (both Mamba-Stormer and the Transformer-Stormer baseline) and per-variable normalization statistics, is archived as version v1.1.0 on Zenodo at https://doi.org/10.5281/zenodo.21890043 (Chao, 2026).

Data availability. The ERA5 reanalysis data used in this study are generated by the Copernicus Climate Change Service (C3S) and the European Centre for Medium-Range Weather Forecasts (ECMWF). The raw hourly datasets on single levels (Copernicus Climate Change Service (C3S), 2018b) and pressure levels (Copernicus Climate Change Service (C3S), 2018a) were obtained from the Copernicus Climate Data Store (CDS). Contains modified Copernicus Climate Change Service information [2026]. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data contained in this publication. For training and evaluation, we utilize the standardized WeatherBench 2 (WB2) repackaging hosted on Google Cloud Storage (gs://weatherbench2/datasets/era5/). Specifically, the 1.5° (240x121) equiangular grid version with poles conservative interpolation (`1959-2023_01_10-6h-240x121_equiangular_with_poles_conservative.zarr`) was accessed

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

on March 9, 2026, and preprocessed into HDF5 formats using our archived code (https://doi.org/10.1029/2023MS004019, Rasp et al., 2024).

Author contribution. JC: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review and editing. WZ: Conceptualization, Supervision, Funding acquisition, Writing – review and editing. GT: Conceptualization, Investigation, Writing – review and editing. ZL, WL, MZ, and MM: Investigation, Writing – review and editing.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the Stony Brook University High-Performance Computing center for access to the NVwulf cluster. This work was partly supported by NSF under Grant NRT-HDR 2125295 and the SBU Institute for Advanced Computational Science under Grant IACS-2024-03. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

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., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., Gupta, J. K., Tambiratnam, K., Archibald, A., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: Aurora: a foundation model of the atmosphere, arXiv preprint arXiv:2405.13063, https://doi.org/10.48550/arXiv.2405.13063, 2024.

Chao, J.: Jiayou-Chao/Mamba-Stormer: Code, model checkpoints and normalization statistics for GMD publication, https://doi.org/10.5281/zenodo.21890043, 2026.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., et al.: The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time, Communications Earth & Environment, https://doi.org/10.1038/s43247-025-02502-y, 2025.

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 Climate and Atmospheric Science, 6, 190, https://doi.org/10.1038/s41612-023-00512-1, 2023.

Copernicus Climate Change Service (C3S): ERA5 hourly data on pressure levels from 1940 to present, Data set, https://doi.org/10.24381/cds.bd0915c6, 2018a.

Copernicus Climate Change Service (C3S): ERA5 hourly data on single levels from 1940 to present, Data set, https://doi.org/10.24381/cds.adbb2d47, 2018b.

Couairon, G., Lessig, C., Charantonis, A., and Monteleoni, C.: ArchesWeather: an efficient AI weather forecasting model at 1.5° resolution, arXiv.org, https://doi.org/10.48550/arXiv.2405.14527, 2024.

Dao, T.: FlashAttention-2: faster attention with better parallelism and work partitioning, in: International Conference on Learning Representations, https://doi.org/10.48550/arXiv.2307.08691, 2023.

Dao, T. and Gu, A.: Transformers are SSMs: generalized models and efficient algorithms through structured state space duality, in: Proceedings of the International Conference on Machine Learning (ICML), https://arxiv.org/abs/2405.21060, 2024.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C.: FlashAttention: fast and memory-efficient exact attention with IO-awareness, Advances in Neural Information Processing Systems (NeurIPS), 35, 16 344–16 359, https://doi.org/10.52202/068431-1189, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations (ICLR), https://arxiv.org/abs/2010.11929, 2021.

Falcon, W., Borovec, J., Eggert, N., Bereznyuk, V., Ir1dXD, Wälchli, A., Jordan, J., Præsius, S., Murrell, T., Harris, E., Bapat, S., Schröter, H., Kulkarni, A., Haunschmid, V., Lipin, D., Singh, A., Fan, T. J., Skafte, N., Mary, H., Eyzaguirre, C., Cinjon, Bakhtin, A., ZH, Z., Jo, Y., Izsak, P., Rangel, O. A., Ling, J., Sharma, H., Waite, E., and Aydın, A.: PyTorchLightning/pytorch-lightning: TPU support & profiling, https://doi.org/10.5281/zenodo.3706330, 2020.

Gu, A. and Dao, T.: Mamba: linear-time sequence modeling with selective state spaces, arXiv.org, https://arxiv.org/abs/2312.00752, 2023.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Gu, A., Goel, K., and Ré, C.: Efficiently modeling long sequences with structured state spaces, in: Proceedings of the International Conference on Learning Representations (ICLR), https://arxiv.org/abs/2111.00396, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

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

Kingma, D. P. and Ba, J.: Adam: a method for stochastic optimization, in: International Conference on Learning Representations, https://arxiv.org/abs/1412.6980, 2014.

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., et al.: GraphCast: learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto Nemesio, A., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F.: AIFS – ECMWF's data-driven forecasting system, https://doi.org/10.48550/arXiv.2406.01465, 2024.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., and Liu, Y.: VMamba: visual state space model, ArXiv, abs/2401.10166, https://arxiv.org/abs/2401.10166, 2024a.

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), pp. 10 012–10 022, https://doi.org/10.1109/ICCV48922.2021.00986, 2021.

Liu, Z., Chen, H., Bai, L., Li, W., Ouyang, W., Zou, Z., and Shi, Z.: MambaDS: near-surface meteorological field downscaling with topography constrained selective state-space modeling, IEEE Transactions on Geoscience and Remote Sensing, 62, 1–15, https://doi.org/10.1109/TGRS.2024.3496895, 2024b.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, in: Proceedings of the International Conference on Learning Representations (ICLR), 2019.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: a foundation model for weather and climate, https://doi.org/10.48550/arXiv.2301.10343, 2023a.

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, https://doi.org/10.48550/arXiv.2312.03876, 2023b.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z.-Y., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: a global data-driven high-resolution weather model using adaptive Fourier neural operators, arXiv, https://arxiv.org/abs/2202.11214, 2022.

Peebles, W. and Xie, S.: Scalable diffusion models with transformers, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4172–4182, https://doi.org/10.1109/ICCV51070.2023.00387, 2023.

https://doi.org/10.5194/egusphere-2026-3398
Preprint. Discussion started: 13 August 2026

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2024.

Qin, H., Chen, Y., Jiang, Q., Sun, P., Ye, X., and Lin, C.: MetMamba: regional weather forecasting with spatial-temporal Mamba model, arXiv.org, https://doi.org/10.48550/arXiv.2408.06400, 2024.

Rabe, M. N. and Staats, C.: Self-attention does not need $O(n^2)$ memory, https://arxiv.org/abs/2112.05682, 2021.

Rasp, S., Dueben, P., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: a benchmark data set for data-driven weather forecasting, https://doi.org/10.1029/2020MS002203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., 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, Journal of Advances in Modeling Earth Systems, 16, e2023MS004019, https://doi.org/10.1029/2023MS004019, 2024.

Tang, Y., Dong, P., Tang, Z., Chu, X., and Liang, J.: VMRNN: integrating vision Mamba and LSTM for efficient and accurate spatiotemporal forecasting, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 5663–5673, https://doi.org/10.1109/CVPRW63382.2024.00575, 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: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, https://arxiv.org/abs/1706.03762, 2017.

Wang, C., Tsepa, O., Ma, J., and Wang, B.: Graph-Mamba: towards long-range graph sequence modeling with selective state spaces, arXiv.org, https://doi.org/10.48550/arXiv.2402.00789, 2024.

Wang, W., Yang, W., Wang, L., Wang, G., and Lei, R.: Seasonal forecasting of Pan-Arctic sea ice with state space model, npj Climate and Atmospheric Science, 8, https://doi.org/10.1038/s41612-025-01058-0, 2025.

Weyn, J. A., Durran, D., and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, Journal of Advances in Modeling Earth Systems, 12, https://doi.org/10.1029/2020MS002109, 2020.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X.: Vision Mamba: efficient visual representation learning with bidirectional state space model, in: International Conference on Machine Learning, https://doi.org/10.48550/arXiv.2401.09417, 2024.