Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DIB-FNO: Input-dependent spectral gating in Fourier neural operators for controlled global weather forecasting

Jianan Wu
Jun Guo
junguo0407@cuit.edu.cn

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11592671/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DIB-FNO: Input-dependent spectral gating in Fourier neural operators for controlled global weather forecasting

Jianan Wu[1] and Jun Guo[1,2*]

[1]College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, 610225, China.
[2]Key Laboratory of Mathematical Meteorology of Sichuan Province Department of Education, Chengdu University of Information Technology, Chengdu, 610225, China.

*Corresponding author(s). E-mail(s): junguo0407@cuit.edu.cn

Abstract

Spectral neural operators are computationally attractive for weather prediction, but conventional Fourier layers use an input-independent spectral allocation. We investigate whether a soft gate conditioned on the current atmospheric state changes the behaviour of a compact Fourier neural operator in a controlled global forecasting experiment. The resulting dynamic information-bottleneck Fourier neural operator (DIB-FNO) is evaluated on a **5.625°**, eight-channel ERA5-derived dataset using a common 15-epoch training budget, three random seeds, and 5,844 common six-hourly test initializations from 2019–2022. Relative to a same-width Fixed FNO, the clearest and most persistent improvement occurs for 500-hPa geopotential height (Z500): at 240 h, the RMSE difference is **−25.81** gpm with a pointwise 95% interval of **−40.67** to **−12.82** gpm. Differences for T850, U850, and V850 at 240 h are not resolved from zero. Freezing each trained gate at its train-period mean, or training a near-parameter-matched static gate, causes pronounced closed-loop error growth, consistent with a role for input dependence but not isolating it as the sole causal factor. The mean gate geometry is highly reproducible across seeds, whereas smaller input-conditioned variations remain measurable. Importantly, DIB-FNO is worse than persistence and climatology for Z500, T850, and U850 at 240 h, and worse than climatology for V850. The evidence therefore supports a bounded operator-architecture finding rather than operational medium-range forecast skill. The source code and evaluation resources are available at https://github.com/wujianan-win/DIB-FNO.

Keywords: Fourier neural operator, spectral gating, information bottleneck, weather forecasting, autoregressive stability, ERA5

1 Introduction

Global numerical weather prediction has **steadily** advanced through sustained improvements in physical modelling, data assimilation, **remote sensing** observations, and computing infrastructure (Bauer et al. 2015). Its computational cost, however, grows rapidly with spatial resolution, ensemble size, and the scale of modern computing systems (Neumann et al. 2019). Data-driven forecasting has consequently become an active complement to numerical modelling. Early medium-range and subseasonal studies established the feasibility of learned atmospheric prediction (Rasp and Thuerey 2021; Weyn et al. 2021), while subsequent graph, three-dimensional, cascaded, and neural-operator systems demonstrated increasingly accurate and computationally efficient global forecasts (Pathak et al. 2022; Bi et al. 2023; Lam et al. 2023; Chen et al. 2023).

Fourier neural operators (FNOs) learn **complex** mappings between functions by parameterizing integral kernels in spectral space (Li et al. 2021). Their global receptive field and efficient transforms are **particularly** attractive for multiscale geophysical dynamics. Adaptive Fourier neural operators further mix frequency-domain features with learnable spectral parameters (Guibas et al. 2022). Recent work has also explored dynamic spectral modes and frequency-adaptive forecasting (Dang et al. 2025; Tang et al. 2026). Nevertheless, spectral allocation in most of these architectures is fixed after training or controlled by mechanisms that are not explicitly conditioned on each evolving atmospheric state. This design may be restrictive because the spectral content relevant to the next forecast step can vary **substantially** with the input state.

We study input-dependent spectral gating in a deliberately compact setting. DIB-FNO applies a continuous gate to the real two-dimensional Fourier spectrum of the current standardized atmospheric state. The gate is produced by a small convolutional network acting on log spectral amplitudes and is optimized jointly with a one-step forecast loss and a soft sparsity penalty. The construction is motivated by the information-bottleneck principle (Tishby et al. 1999; Alemi et al. 2017): a predictive representation should retain information useful for the target while suppressing unnecessary input detail. The implemented loss is an information-bottleneck-inspired surrogate; it does not directly estimate mutual information.

The central question is narrower than whether DIB-FNO is a competitive operational weather model. We ask whether input-dependent gating changes the accuracy and closed-loop stability of a small FNO when data, training budget, initialization dates, and evaluation procedures are controlled. To address confounding by parameter count and by post-training intervention, we compare DIB-FNO with (i) a same-width gate-free Fixed FNO, (ii) a separately trained near-parameter-matched static-gate FNO, and (iii) a frozen-mean-gate inference variant that preserves every trained DIB-FNO backbone weight while replacing the dynamic gate by its train-period mean.

Our contributions are threefold. First, we formulate continuous, input-dependent spectral gating and state its optimization objective explicitly. Second, we evaluate model

and physical baselines on common 2019–2022 initializations using paired temporal block resampling and three training seeds. Third, we distinguish an architecture-level effect from useful forecast skill: the dynamic model is more stable than the static controls in this experiment, but all learned models lose competitiveness against climatology at longer leads. The resulting conclusions are therefore limited to controlled operator behaviour rather than operational meteorological capability.

2 Related work

2.1 Data-driven global weather prediction

Data-driven medium-range and subseasonal prediction has progressed from ResNet and deep-ensemble demonstrations to a broad family of global forecast systems (Rasp and Thuerey 2021; Weyn et al. 2021; de Burgh-Day and Leeuwenburg 2023). FourCastNet introduced AFNO-based global forecasts and was subsequently developed for accelerated high-resolution prediction (Pathak et al. 2022; Kurth et al. 2023). Pangu-Weather, GraphCast, and FuXi demonstrated strong medium-range performance with three-dimensional, graph-based, and cascaded architectures (Bi et al. 2023; Lam et al. 2023; Chen et al. 2023). More recent systems include data-to-forecast pipelines, FengWu, hybrid dynamical–neural frameworks, neural general circulation models, and Earth-system foundation models (Sun et al. 2025; Chen et al. 2025; Hu et al. 2026; Kochkov et al. 2024; Bodnar et al. 2025). Probabilistic machine-learning forecasts further extend this development from deterministic trajectories to predictive distributions (Price et al. 2025). These systems operate at scales and under protocols that are not directly comparable with the compact DIB-FNO configuration.

2.2 Spectral neural operators and adaptive computation

FNOs approximate solution operators by alternating Fourier-domain transformations and pointwise nonlinear maps (Li et al. 2021). AFNO replaces a dense spectral operator with efficient token mixing and has been used in large-scale forecasting (Guibas et al. 2022; Pathak et al. 2022). Subsequent neural-operator research has addressed dynamic spectral modes, irregular meshes, PDE foundation models, geometry-aware operators, scalable physics transformers, and regional ocean modelling (Dang et al. 2025; Shi et al. 2025; Herde et al. 2024; Wu et al. 2024; Alkin et al. 2024; Li et al. 2023; Jahanmard et al. 2026). Time-dependent operator learning has also emphasized temporal structure and memory (Dama et al. 2025; Buitrago Ruiz et al. 2025), while frequency-domain forecasting methods provide complementary evidence that spectral representations can be adapted to non-stationary signals (Fan et al. 2024; Tang et al. 2026).

In parallel, differentiable relaxations and conditional-computation mechanisms allow neural networks to allocate flexible representation or computation according to an input (Maddison et al. 2017; Li et al. 2022). Sparse-attention and adaptive-depth models illustrate related strategies for controlling computational pathways (Zaheer et al. 2020; Banino et al. 2021), whereas equilibrium and reversible architectures address depth and memory through different structural constraints (Bai et al. 2019; Zhao et al. 2024). Diffusion and flow-matching models provide broader examples of

learned dynamic state-dependent transformations (Ho et al. 2020; Lipman et al. 2023). DIB-FNO applies conditional computation specifically to a shared spectral gate over the input channels. Unlike hard top-C selection, normal inference uses all represented frequencies with continuous weights between zero and one.

2.3 Information-bottleneck motivation

The information bottleneck formalizes a trade-off between target-relevant information and representation complexity (Tishby et al. 1999). Variational formulations make this principle tractable in deep models (Alemi et al. 2017), and conditional and glocal bottleneck formulations extend it to time-series imputation (Choi and Lee 2024; Yang et al. 2025). For DIB-FNO, the principle serves as a design motivation rather than a directly estimated mutual-information objective. The practical regularizer is the mean soft-gate weight and should be interpreted as a differentiable sparsity proxy rather than an information-theoretic quantity.

3 Methodology

DIB-FNO combines a one-step forecasting operator with an input-conditioned spectral bottleneck. Its optimization objective combines forecast error with a soft gate penalty, and the gate acts directly on the Fourier representation. The penalty is motivated by the information bottleneck but does not estimate an information-theoretic quantity.

3.1 Forecasting problem

Let $X_t \in \mathbb{R}^{C\times H\times W}$ denote a standardized atmospheric state and X_{t+1} the state six hours later. Here $C = 8$, $H = 32$, and $W = 64$. A learned operator f_θ produces a one-step forecast,

$$\widehat{X}_{t+1} = f_\theta(X_t). \tag{1}$$

At inference, the model is iterated autoregressively for 40 steps to obtain forecasts through 240 h. Training uses only the one-step objective; no multi-step stability loss is applied.

3.2 Input-Conditioned Soft Mask Construction

Let $\mathcal{F}_r$ denote the orthonormal two-dimensional real fast Fourier transform. DIB-FNO forms

$$S_t = \mathcal{F}_r(X_t), \tag{2}$$
$$A_t = \log(1 + |S_t|)\,, \tag{3}$$
$$G_t = \sigma(h_\phi(A_t)/\tau)\,, \tag{4}$$

where h_ϕ is a three-layer convolutional gate network, σ is the logistic sigmoid, and $\tau > 0$ is a temperature. The gate has shape $1 \times 32 \times 33$ for each sample and is broadcast

across the eight channels. The zero-frequency coefficient has a separately learned logit to discourage trivial collapse. The filtered input is

$$\widetilde{X}_t = \mathcal{F}_r^{-1}(G_t \odot S_t)\,. \tag{5}$$

The filtered state is projected to a hidden representation, processed by one residual Fourier block, and mapped back to the eight forecast channels. The Fourier block transforms the lowest eight positive and eight negative meridional bands and 17 zonal real-FFT modes. The resulting mask is not prescribed as a fixed list of coefficients: it is the state-conditioned gate learned jointly from forecast error and the soft complexity penalty over the training distribution.

3.3 Training Objective

The training criterion is

$$\mathcal{L}(\theta,\phi) = \frac{1}{BCHW}\sum_{i=1}^{B}\left\| f_{\theta,\phi}(X_t^{(i)}) - X_{t+1}^{(i)} \right\|_2^2 + \beta \frac{1}{BH(W/2+1)}\sum_{i,j,k} G_{t,j,k}^{(i)}, \tag{6}$$

with $\beta = 10^{-4}$. The first term is the one-step normalized prediction loss, whereas the second term encourages smaller gate weights without imposing a fixed cardinality. Equation (6) is an information-bottleneck-inspired predictive compression surrogate, not an estimator of $I(X;Z)$ or $I(Z;Y)$.

3.4 Continuous Relaxation of Mask Selection

Direct optimization of a binary spectral subset would be discontinuous and would require a separate rule for cardinality. DIB-FNO instead uses sigmoid weights in $[0,1]$. Temperature is annealed from 1.0 to 0.2 over the first five epochs using a cosine schedule and is then held at 0.2. In normal inference G_t remains continuous; the descriptive threshold $G_t \geq 0.5$ and the hard top-C variants are used only for diagnostics and sensitivity analysis.

3.5 Generalization in Practice

The theoretical motivation concerns adaptive predictive compression, but the empirical object is an autoregressive operator. Generalization must therefore be evaluated not only at the trained six-hour horizon but also under distribution shift induced by repeated self-feeding. The experiments consequently distinguish one-step checkpoint selection, 40-step closed-loop verification, comparisons with physical baselines, and post-training gate interventions. This design tests whether the soft mask remains useful when its own forecast replaces the analysis as the next input.

4 Experimental Setup

The architecture, comparison models, chronological data splits, optimization budget, and evaluation metrics are fixed across the three random seeds. Validation error selects

checkpoints, whereas the 2019–2022 test trajectories are reserved for the reported forecast comparisons.

4.1 Overall Architecture and Intended Comparisons

The comparisons quantify four properties: **forecast** error relative to a gate-free FNO with the same hidden width and optimization budget; autoregressive behaviour relative to input-independent gate controls; competitiveness with persistence and a training-only calendar climatology; and gate reproducibility across random seeds together with measurable input dependence. All comparisons **consistently** use common initialization dates and a common evaluation pipeline.

4.1.1 Network architecture and comparison models

All learned models map eight standardized atmospheric channels to the same eight channels six hours later. DIB-FNO first evaluates a three-layer convolutional gate generator on the log-amplitude real-FFT representation. The resulting 32×33 soft mask is shared across channels, multiplied with the input spectrum, and transformed back to physical space. A pointwise lifting layer then maps the filtered state to hidden width 16. One residual Fourier block transforms the lowest eight positive and eight negative meridional bands together with 17 zonal real-FFT modes, after which a pointwise projection returns the eight forecast variables. This compact architecture supports low-cost long-rollout experiments but is not directly comparable with high-resolution production systems.

The Fixed FNO uses the same hidden width (16) and one-block spectral backbone without **any** input gate. It contains 1,896 trainable parameters and is therefore **strictly** a same-width, not parameter-matched, baseline. The Static gate FNO learns one input-independent 32×33 gate. Its hidden width is 17, yielding 3,173 parameters, compared with 3,106 **parameters** for DIB-FNO (a 2.2% difference).

The frozen-mean-gate control is a post-training intervention. For each DIB-FNO seed, we evaluate the gate on 384 evenly spaced training analyses, compute the elementwise mean gate, and use that fixed gate at every autoregressive step. The forecasting backbone and all other checkpoint weights remain unchanged. This control tests whether the trained rollout depends on gate variation, whereas the separately trained Static gate FNO tests whether an input-independent gated architecture can be learned under the same optimization budget. Neither comparison alone identifies dynamic gating as the sole causal mechanism because replacing a gate changes subsequent autoregressive inputs and hence the rollout distribution.

The controlled model configurations are summarized in Table 1.

4.1.2 Data, preprocessing, and chronological splits

The source consists of four ERA5-derived variables at 500 and 850 hPa: geopotential, temperature, zonal wind, and meridional wind (Hersbach et al. 2020). These variables form eight channels (Z500, Z850, T500, T850, U500, U850, V500, and V850) on a 5.625°, 32×64 latitude–longitude grid. The principal figures report Z500, T850, U850, and V850. Geopotential in $\mathrm{m^2\,s^{-2}}$ is divided by 9.80665 $\mathrm{m\,s^{-2}}$, and geopotential height

Table 1 Controlled model configurations.

Model	Hidden width	Gate	Parameters
DIB-FNO	16	Input-dependent soft	3,106
Static gate FNO	17	Learned input-independent	3,173
Fixed FNO	16	None	1,896

is reported in geopotential metres (gpm). Each channel is standardized with the training-period mean and standard deviation.

The channel definitions, pressure levels, units, temporal splits, and data sources are listed in Table 2.

Table 2 Eight-channel atmospheric input–output data contract. The period column gives training/validation/test intervals, respectively; WB2 denotes WeatherBench2.

Channel	Variable	Level	Unit	Period split	Source
Z500	Geopotential height	500 hPa	gpm	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
Z850	Geopotential height	850 hPa	gpm	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
T500	Air temperature	500 hPa	K	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
T850	Air temperature	850 hPa	K	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
U500	Zonal wind	500 hPa	$\mathrm{m\,s^{-1}}$	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
U850	Zonal wind	850 hPa	$\mathrm{m\,s^{-1}}$	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
V500	Meridional wind	500 hPa	$\mathrm{m\,s^{-1}}$	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset
V850	Meridional wind	850 hPa	$\mathrm{m\,s^{-1}}$	1959–2015; 2016–2018; 2019–2022	ERA5; WB2-style subset

Chronological splits are 1959–2015 for training, 2016–2018 for validation, and 2019–2022 for test initializations. The 5,844 test starts occur every six hours; the latest 240-h verification target is 10 January 2023 at 18 UTC. The local source arrays do not contain a time coordinate. Calendar timestamps are therefore inferred from the record length, the six-hour sampling interval, and the documented WeatherBench-style record origin; the inference is recorded in the repository data manifest.

Climatology is calculated only from 1959–2015. It is the unsmoothed mean for each calendar month–day and UTC-hour slot (00, 06, 12, and 18 UTC) on a leap-year template. Persistence repeats the initial analysis at every lead. All learned models, persistence, and climatology are evaluated on identical initializations.

The representative maps verify variable identity, grid orientation, units, and broad spatial organization (Figures 1–3). Zonal profiles and annual cycles independently confirm the latitude-dependent and seasonal signals inherited by the standardized

data. All climatological diagnostics are computed exclusively from the training period, without validation- or test-period information.

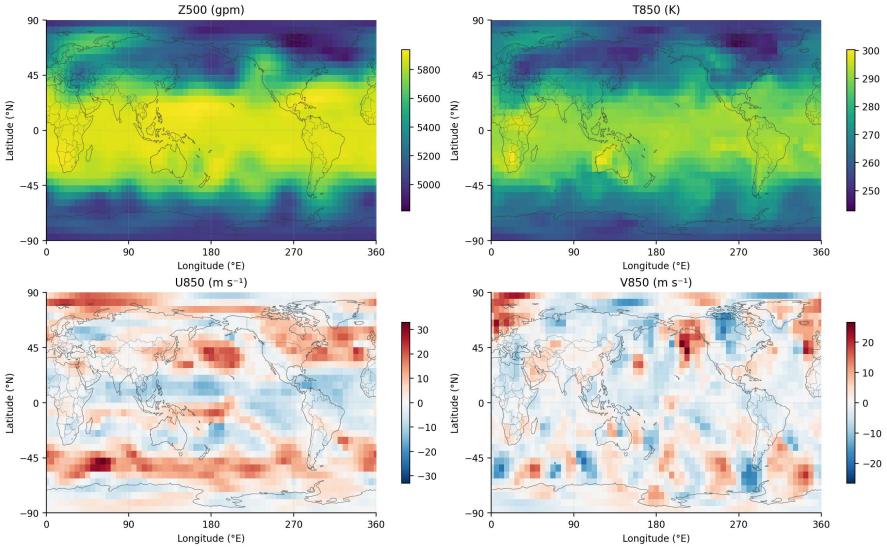

Fig. 1 Representative analysis fields. ERA5-derived Z500, T850, U850, and V850 on the 5.625° grid.

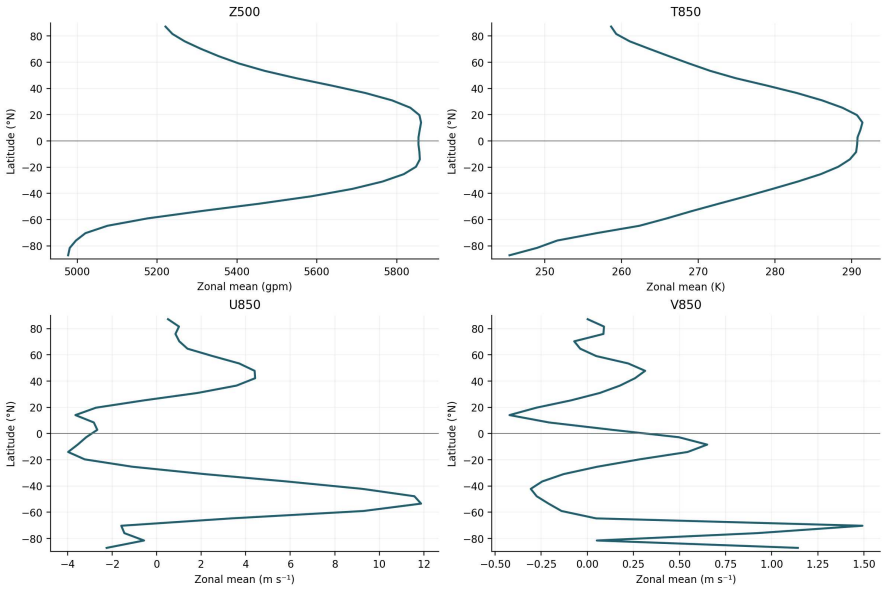

Fig. 2 Training-period zonal climatology. Zonal means of Z500, T850, U850, and V850 for 1959–2015.

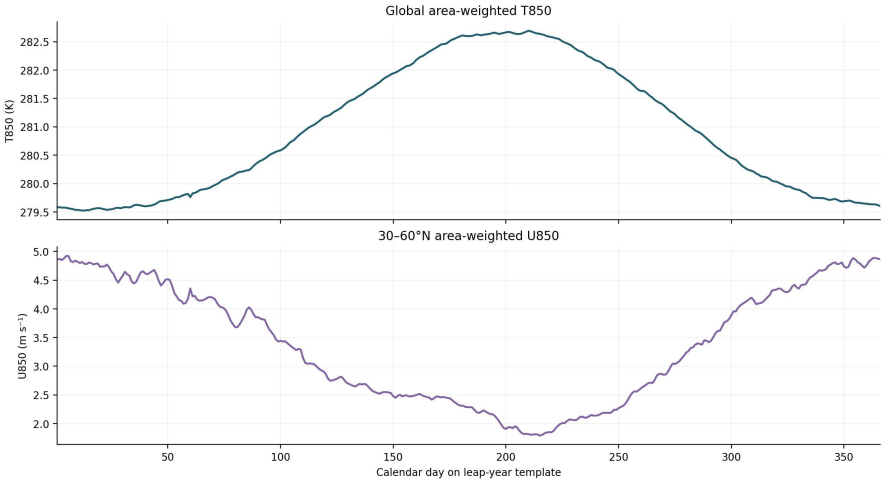

Fig. 3 Training-period annual cycle. Area-weighted T850 and Northern midlatitude U850 climatologies for 1959–2015.

4.1.3 Training and autoregressive evaluation protocol

Table 3 Numerical experiment protocol.

Item	Setting
Grid and sampling	32×64, 5.625°, 6-hourly
Prognostic channels	Z500, Z850, T500, T850, U500, U850, V500, V850
Training period	1959–2015
Validation period	2016–2018
Test initializations	2019–2022; 5,844 common starts
Training target	One six-hour step
Test rollout	40 autoregressive steps (6–240 h)
Prespecified reported leads	24, 72, 120, and 240 h
Random seeds	0, 1, and 2
Optimizer	AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}
Training ceiling	15 epochs; convergence not established
GPU model (reported runs)	NVIDIA GeForce RTX 5060 Laptop GPU (8 GB)
CPU model (reported runs)	Intel Core i9-14900HX (24 cores, 32 threads)
Mean wall-clock time per epoch	32.162 s (training plus validation; 15-epoch mean)
Wall-clock time for 15 epochs	482.882 s (8 min 02.9 s)
One 40-step rollout inference	0.1044 s (batch 1; median of 100 repetitions)

We train seeds 0, 1, and 2 for DIB-FNO, Static gate FNO, and Fixed FNO. AdamW uses a learning rate of 3×10^{-4} and weight decay of 10^{-4}. The selected checkpoint

minimizes the one-step validation mean of the per-channel normalized RMSE values. Each of the nine runs reaches the 15-epoch ceiling; the recorded validation trajectories are still changing near the end of training. We therefore describe the comparison as a common-budget experiment and do not claim convergence or asymptotic superiority.

Models are optimized only for a six-hour target but are evaluated recursively for 40 steps (Table 3). This mismatch tests whether the spectral mechanism changes error accumulation without assistance from a stability-aware training loss.

The authors confirm that the benchmark host was the machine used for the reported experiments. Because the ERA5/Zarr arrays were not retained locally, the benchmark used the DIB-FNO configuration (3,106 parameters), tensor dimensions, batch size, and batch counts with reusable synthetic tensors. Each epoch comprised 2,603 training batches (83,275 samples) and 137 validation batches (4,383 samples). The measurement included CPU-to-GPU tensor transfer, forward propagation, latitude-weighted loss, backward propagation, gradient clipping, AdamW updates, and the complete validation computation, but excluded Zarr reads and preprocessing. Results in Table 3 were obtained with Python 3.12.14, PyTorch 2.11.0+cu128, and CUDA 12.8; they are compute-path timings on the experimental hardware rather than end-to-end data-pipeline times.

5 Numerical Experiments

Forecast performance is quantified by physical-unit RMSE, anomaly correlation, reference-relative MSE skill, and paired uncertainty across lead times. Training trajectories characterize optimization, static-gate controls test the contribution of input conditioning, spatial error maps assess geographical heterogeneity, and hard top-C masks measure sensitivity to post-training spectral truncation.

5.1 Evaluation Metrics and Statistical Reporting

The evaluation combines absolute forecast error metrics, anomaly-pattern agreement, relative MSE skill, and paired uncertainty. Each metric answers a different question. RMSE measures physical-unit error magnitude; ACC measures spatial anomaly-pattern agreement; MSE skill expresses DIB-FNO error relative to a specified reference; and paired block resampling evaluates whether differences persist when temporally correlated initialization dates and the three model seeds are resampled. The reference model is always named explicitly because a positive skill value alone has no meaning without its denominator.

5.1.1 Area-weighted RMSE and anomaly correlation

Let $\widehat{y}^{(s)}_{n,\ell,c,i,j}$ and $y_{n,\ell,c,i,j}$ denote the forecast and verifying analysis for seed s, initialization n, lead ℓ, channel c, latitude index i, and longitude index j. With latitude ϕ_i,

the normalized cosine weight and reported RMSE are

$$
\begin{aligned}
w_i &= \frac{\cos\phi_i}{\sum_{r=1}^{H}\cos\phi_r}, \\
\mathrm{RMSE}_{\ell,c}^{(s)} &= \left[\frac{1}{N_\ell W}\sum_{n=1}^{N_\ell}\sum_{i=1}^{H}\sum_{j=1}^{W} w_i\left(\widehat{y}_{n,\ell,c,i,j}^{(s)} - y_{n,\ell,c,i,j}\right)^2\right]^{1/2}, \\
\overline{\mathrm{RMSE}}_{\ell,c} &= \frac{1}{S}\sum_{s=1}^{S}\mathrm{RMSE}_{\ell,c}^{(s)}.
\end{aligned}
\tag{7}
$$

Here $H = 32$, $W = 64$, N_ℓ is the number of common initialization dates at lead ℓ, and $S = 3$ is the number of seeds. Thus squared errors are averaged over dates and the cosine-weighted grid before the square root is taken for each seed; the three seed-level RMSE values are averaged only afterward.

For ACC, let $\mu_{n,\ell,c,i,j}$ be the training-period calendar climatology matched to the verifying month–day and UTC-hour slot, and define forecast and analysis anomalies $p' = \widehat{y} - \mu$ and $a' = y - \mu$. The per-initialization spatial ACC and the reported date-then-seed mean are

$$
\begin{aligned}
\mathrm{ACC}_{n,\ell,c}^{(s)} &= \frac{\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W} w_i p'^{(s)}_{n,\ell,c,i,j} a'_{n,\ell,c,i,j}}{\left[\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W} w_i\left(p'^{(s)}_{n,\ell,c,i,j}\right)^2\right]^{1/2}\left[\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W} w_i\left(a'_{n,\ell,c,i,j}\right)^2\right]^{1/2}}, \\
\overline{\mathrm{ACC}}_{\ell,c} &= \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{N_\ell}\sum_{n=1}^{N_\ell}\mathrm{ACC}_{n,\ell,c}^{(s)}\right).
\end{aligned}
\tag{8}
$$

ACC is therefore calculated spatially for each initialization, averaged over common dates within each seed, and then averaged across seeds. Climatology ACC is undefined because the climatological forecast anomaly is identically zero. The horizontal ACC value 0.6 is an empirical reference and not a formal predictability boundary.

5.1.2 Reference-relative MSE skill

For reference R, MSE skill is

$$
\mathrm{Skill}_R = 1 - \frac{\mathrm{meanMSE(DIB\text{-}FNO)}}{\mathrm{meanMSE}(R)},
\tag{9}
$$

where each mean is formed before taking the ratio. Positive values favour DIB-FNO.

5.1.3 Paired temporal-block uncertainty

Uncertainty is estimated using 2,000 paired circular block-bootstrap draws. A block contains 60 consecutive six-hour starts (15 complete days) beginning at 00 UTC. Each draw samples 98 blocks and truncates the last block to retain 5,844 starts. All models and physical baselines share the same date blocks. One common three-ID resample with replacement is applied to seeds 0–2 across model architectures. RMSE and skill are recomputed inside every draw. Reported intervals are pointwise percentile 95% intervals; they are not simultaneous confidence bands and are not adjusted for multiple comparisons. Leads of 24, 72, 120, and 240 h were specified before the final aggregation. Sensitivity was evaluated with 10- and 30-day blocks.

Changing the circular block length from 15 days to 10 or 30 days leaves the qualitative conclusions unchanged. In particular, the 240-h Z500 DIB-FNO–Fixed FNO interval remains strictly negative under all three choices, whereas the corresponding T850, U850, and V850 intervals include zero. These checks assess sensitivity to one resampling choice; they do not compensate for the small number of training seeds or for reuse of an already inspected test period.

5.2 Controlled Global Forecast Skill

5.2.1 Forecast references and aggregate lead-time performance

The primary experiment compares DIB-FNO with the same-width Fixed FNO, persistence, and the training-only calendar climatology. The Fixed FNO comparison isolates the effect of input-conditioned spectral gating under a matched hidden width and training budget, while persistence and climatology provide physically interpretable forecast references. Paired DIB-FNO-minus-reference differences are reported at every lead (Figure 4), together with absolute RMSE and anomaly-correlation trajectories (Figures 9 and 10).

5.2.2 Paired differences between DIB-FNO and forecast references

Figure 4 summarizes DIB-FNO minus Fixed FNO and persistence. Negative values favour DIB-FNO. The most consistent difference is for Z500. At 24 h, the DIB-FNO-minus-Fixed-FNO difference is -5.80 gpm, with a pointwise 95% interval of -6.40 to -5.31 gpm. At 120 h, the difference is -13.45 gpm (-16.49 to -10.76 gpm), and at 240 h it is -25.81 gpm (-40.67 to -12.82 gpm). The T850, U850, and V850 differences are smaller. At 240 h, all three intervals against Fixed FNO cross zero, so the experiment does not establish a general ten-day advantage across variables.

The comparison with persistence changes sign as lead increases. DIB-FNO is better than persistence at short leads for all four reported variables. At 240 h, however, it is worse for Z500, T850, and U850; V850 remains slightly lower in point estimate, but its paired interval crosses zero. Thus, model-to-model improvements cannot be interpreted independently of physical baselines.

The relative MSE skill curves in Figure 5 express the same comparisons on a dimensionless scale. Skill relative to Fixed FNO is positive and most persistent for Z500, whereas skill relative to persistence becomes negative at longer leads for Z500, T850,

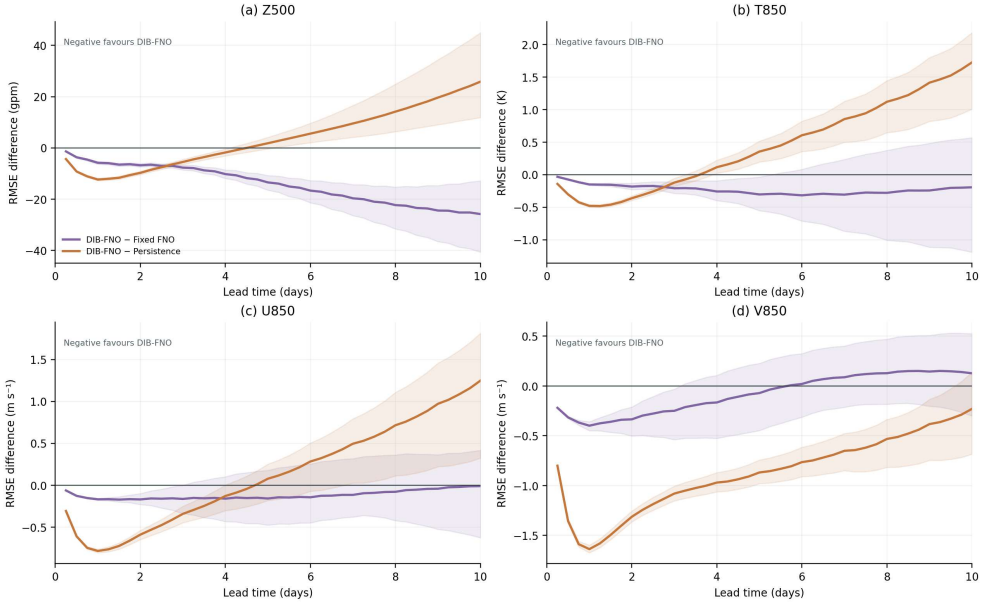

Fig. 4 Paired RMSE differences. DIB-FNO minus Fixed FNO and persistence; negative values favour DIB-FNO. Shading denotes pointwise 95% paired bootstrap intervals.

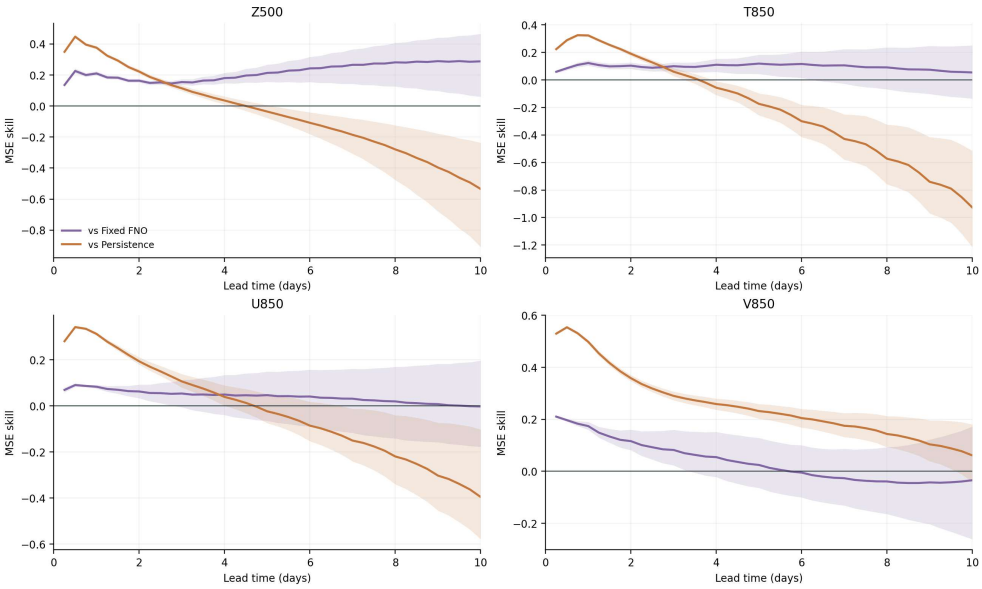

Fig. 5 Test MSE skill. Positive values favour DIB-FNO; shading denotes pointwise 95% paired bootstrap intervals.

and U850. The widening pointwise intervals at late lead also show why a favourable mean difference should not be treated as uniform evidence across variables or seeds. Relative skill is used here only to normalize a named pairwise comparison; it does not replace the absolute RMSE and ACC diagnostics below.

Table 4 provides the exact prespecified lead-time comparisons underlying the DIB-FNO–Fixed FNO curves. It complements Figure 4 by distinguishing differences whose pointwise intervals exclude zero from those that become unresolved as seed-to-seed dispersion grows at longer lead times.

Table 4 DIB-FNO minus Fixed FNO RMSE at prespecified leads; negative values favour DIB-FNO.

Variable	Lead (h)	RMSE difference	Pointwise 95% interval
Z500	24	-5.80	[-6.40, -5.31]
	72	-7.69	[-8.61, -6.70]
	120	-13.45	[-16.49, -10.76]
	240	-25.81	[-40.67, -12.82]
T850	24	-0.150	[-0.164, -0.138]
	72	-0.209	[-0.288, -0.111]
	120	-0.304	[-0.553, -0.037]
	240	-0.194	[-1.191, 0.570]
U850	24	-0.168	[-0.180, -0.156]
	72	-0.164	[-0.338, -0.001]
	120	-0.158	[-0.476, 0.178]
	240	-0.011	[-0.626, 0.417]
V850	24	-0.399	[-0.449, -0.359]
	72	-0.249	[-0.539, -0.033]
	120	-0.071	[-0.448, 0.226]
	240	0.127	[-0.300, 0.525]

5.2.3 Long-lead uncertainty and reference dependence

Figure 4 establishes an architecture-level advantage most clearly for Z500: the DIB-FNO–Fixed FNO interval remains below zero from short lead through 240 h, including under the 10- and 30-day block-length sensitivity checks. The ten-day T850, U850, and V850 intervals against Fixed FNO include zero, and the comparison with persistence becomes unfavourable for three variables. Input-conditioned spectral filtering therefore changes the closed-loop behaviour of the tested FNO and produces a robust Z500 improvement without establishing ten-day forecast superiority across variables.

5.3 Optimization Dynamics and Computational Behavior

5.3.1 Common-budget training histories

The nine training histories are reported in Figure 6. Every DIB-FNO, Static gate FNO, and Fixed FNO run reaches the common 15-epoch ceiling, and several validation trajectories continue to change near the final epoch. The saved checkpoints therefore

represent a controlled compute budget rather than demonstrated convergence; a lower test error under this fixed ceiling does not establish an asymptotic optimization advantage.

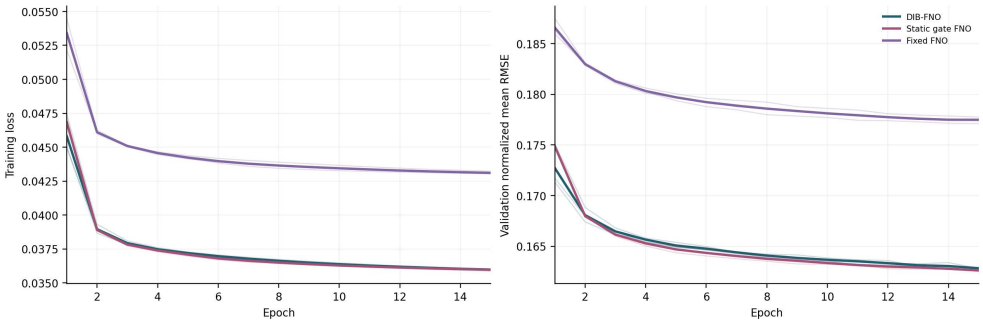

Fig. 6 Training histories. Training loss and one-step validation RMSE for all nine runs.

5.3.2 Architecture-dependent complexity

The dynamic gate adds a small convolutional controller to a one-block spectral backbone. DIB-FNO contains 3,106 trainable parameters, the near-parameter-matched Static gate FNO contains 3,173, and the same-width Fixed FNO contains 1,896. These counts permit two complementary comparisons but not a single perfectly matched one: DIB-FNO versus Fixed FNO holds hidden width constant, whereas DIB-FNO versus Static gate FNO approximately holds parameter count constant. The original-hardware replay in Table 3 supplies compute-path wall-clock timing for DIB-FNO; hardware-normalized energy was not measured. The computational evidence therefore comprises parameter counts, common epoch ceilings, recorded training histories, and the explicitly scoped runtime replay.

5.4 Lead-Time-Dependent Skill and Autoregressive Stability

5.4.1 Dynamic, frozen-mean, and static spectral gates

This experiment distinguishes three ways of using spectral weights. DIB-FNO recomputes a continuous gate from every current input; the frozen-mean control retains the trained DIB-FNO backbone but replaces its gate by the seed-specific mean computed from 384 training analyses; and the Static gate FNO learns one input-independent gate jointly with a near-parameter-matched backbone. The frozen comparison isolates a post-training intervention, whereas the independently trained static model tests whether a fixed spectral filter can be learned under the same 15-epoch budget. Figure 7 reports absolute RMSE on a logarithmic axis because the static controls diverge strongly at later leads.

5.4.2 Long-rollout error growth under static gates

Figure 7 compares absolute test RMSE for the dynamic model, two static controls, the gate-free baseline, and the physical baselines. At 24 h, Z500 DIB-FNO minus frozen-mean-gate RMSE is −0.30 gpm (−0.44 to −0.15 gpm), and DIB-FNO minus the separately trained static gate is −0.71 gpm (−0.93 to −0.47 gpm). By 72 h these differences reach −3.29 and −4.96 gpm, respectively.

The static controls become unstable at longer leads. At 240 h, their Z500 RMSE values are 830.51 gpm for the frozen mean gate and 479.80 gpm for the learned static gate, compared with 137.32 gpm for DIB-FNO. Similar divergence appears in the other variables. The large late-lead gaps mainly quantify closed-loop failure of the controls and are not proportional improvements in useful weather forecast skill. Freezing a gate also changes every subsequent autoregressive input, so this ablation cannot separate the immediate benefit of input conditioning from downstream distribution shift.

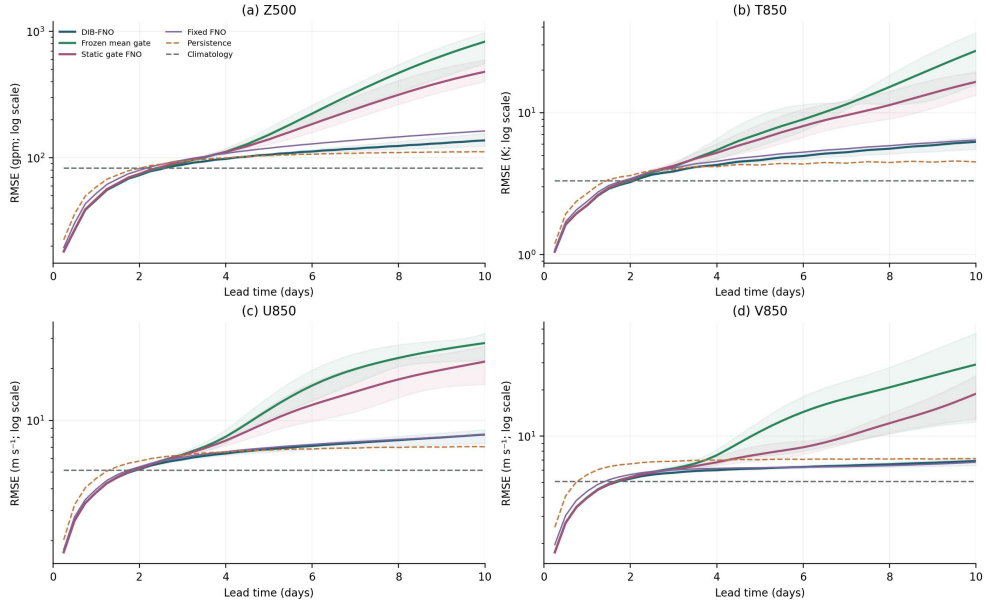

Fig. 7 Gate-ablation RMSE. Absolute test RMSE for DIB-FNO, static-gate controls, Fixed FNO, persistence, and climatology; vertical axes are logarithmic.

Figure 8 complements the logarithmic absolute curves with paired differences in physical units. Negative values favour DIB-FNO throughout most leads, but the rapidly increasing late-lead magnitude is driven largely by divergence of the static controls. The curves identify the onset of separation but do not provide evidence of operational forecast skill.

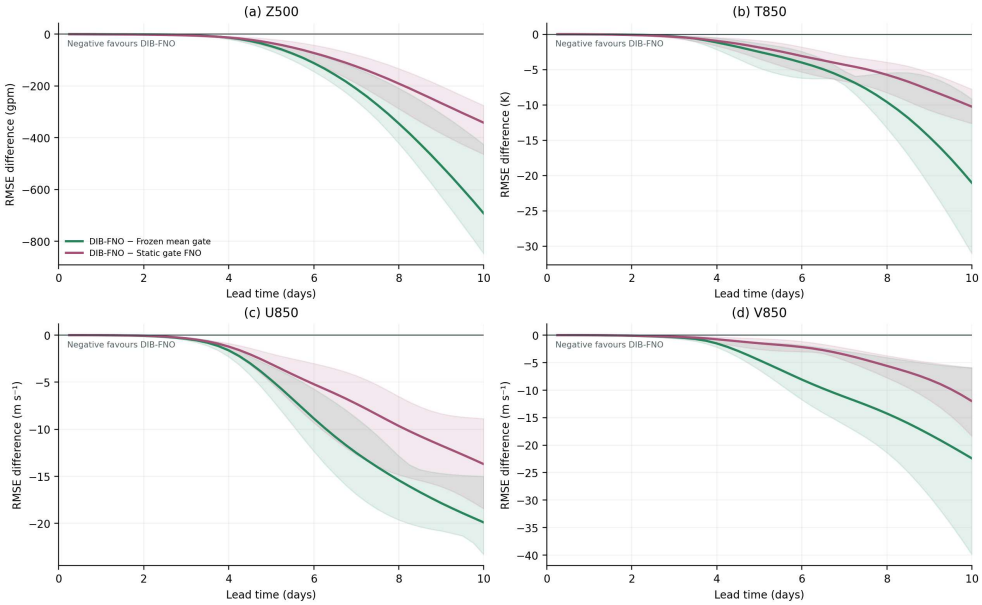

Fig. 8 Gate-control RMSE differences. DIB-FNO minus frozen-mean and static-gate controls; negative values favour DIB-FNO.

5.4.3 Lead-dependent differences in rollout stability

The early-lead differences show that the dynamic computation is not merely a late-rollout artifact: DIB-FNO is already lower than the frozen and separately trained static controls for Z500 at 24 and 72 h. The rapidly widening differences after 120 h, however, mainly indicate closed-loop instability in the controls. They should not be reported as proportional gains in operational forecast quality. Because freezing the gate changes the first prediction and hence every subsequent autoregressive input, the experiment supports dependence of the rollout on input-conditioned filtering but does not identify the gate variation as the sole causal mechanism.

5.5 Physical Baselines and Absolute Skill

5.5.1 Comparison with persistence and climatology

Absolute forecast skill is evaluated by comparing DIB-FNO and Fixed FNO with persistence and the training-only climatology in physical units. Figure 9 gives the complete RMSE trajectories, Table 5 reports the 240-h endpoint and uncertainty for each variable, and Figure 10 measures the retention of spatial anomaly patterns as amplitude errors increase.

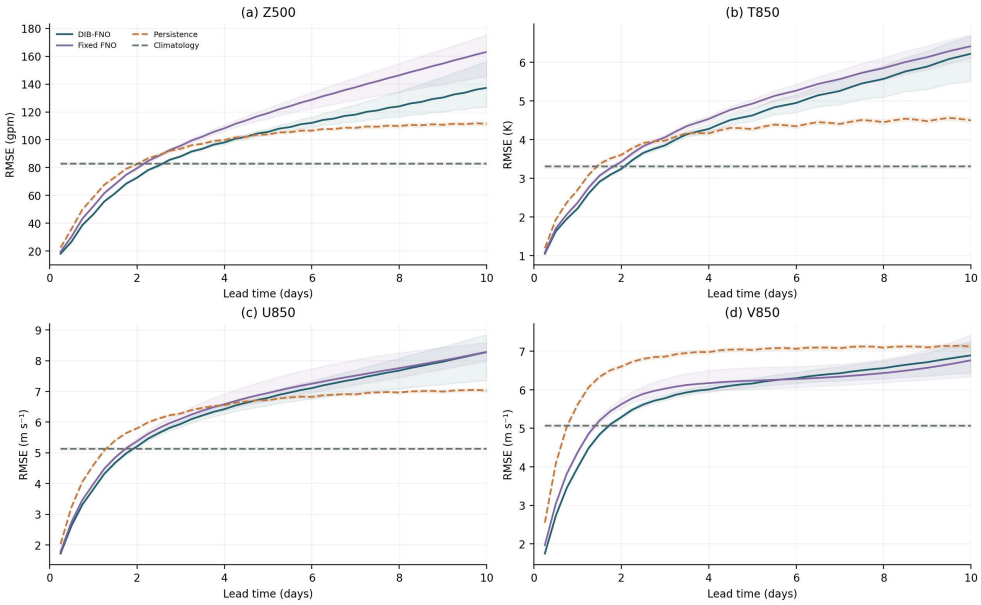

Fig. 9 Test RMSE. Three-seed means for DIB-FNO and Fixed FNO with persistence and training-only climatology; shading denotes pointwise 95% paired bootstrap intervals.

5.5.2 Physical-unit RMSE through 240 h

The absolute errors in Figure 9 qualify the paired comparisons. At 240 h, Z500 RMSE is 137.32 gpm for DIB-FNO, 163.13 gpm for Fixed FNO, 111.46 gpm for persistence, and 82.80 gpm for climatology (Table 5). DIB-FNO also exceeds both physical baselines for T850 and U850 and exceeds climatology for V850. Climatology becomes preferable to DIB-FNO between 24 and 72 h for all four reported variables, so the present configuration does not provide useful ten-day forecast skill.

5.5.3 Anomaly correlation through 240 h

ACC gives the same broad ordering at short lead (Figure 10). DIB-FNO generally retains a higher spatial anomaly correlation than Fixed FNO during the first few days, with the largest separation for V850. Nevertheless, ACC declines rapidly and approaches zero at later leads. Persistence can exceed the learned models after their early advantage has decayed. The 0.6 reference is crossed within the first few days for every variable and should not be interpreted as a universal predictability limit.

Figure 11 separates the pooled signed-error component from total RMSE as a diagnostic of systematic displacement. The fraction varies by variable, model, and lead, indicating that the aggregate RMSE curves cannot be attributed to a single uniform offset. This quantity is descriptive: it is not a dynamical error decomposition and does not identify the physical source of forecast failure.

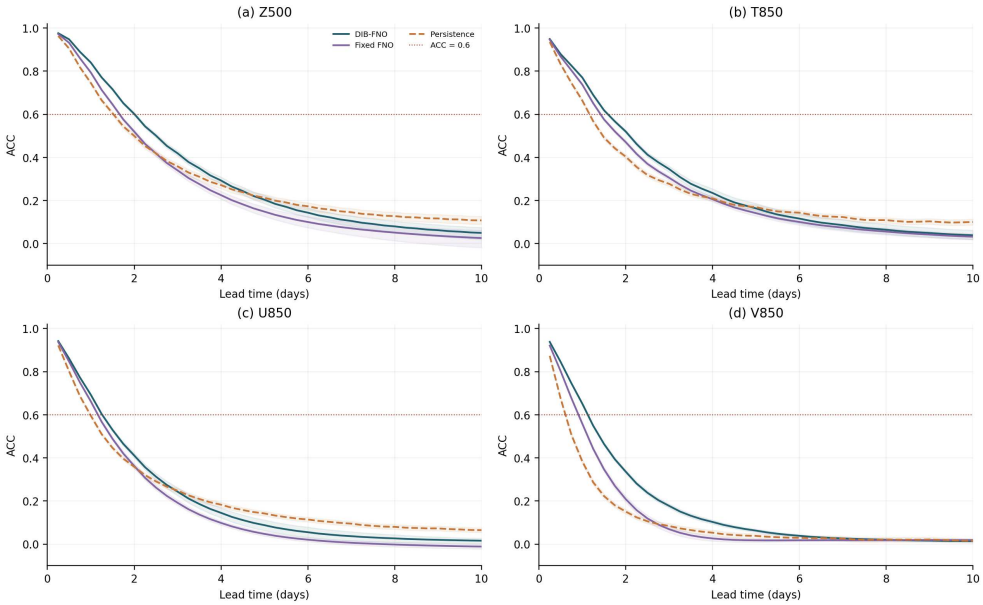

Fig. 10 Test ACC. Anomalies are defined relative to the training-only calendar climatology; shading denotes pointwise 95% paired bootstrap intervals.

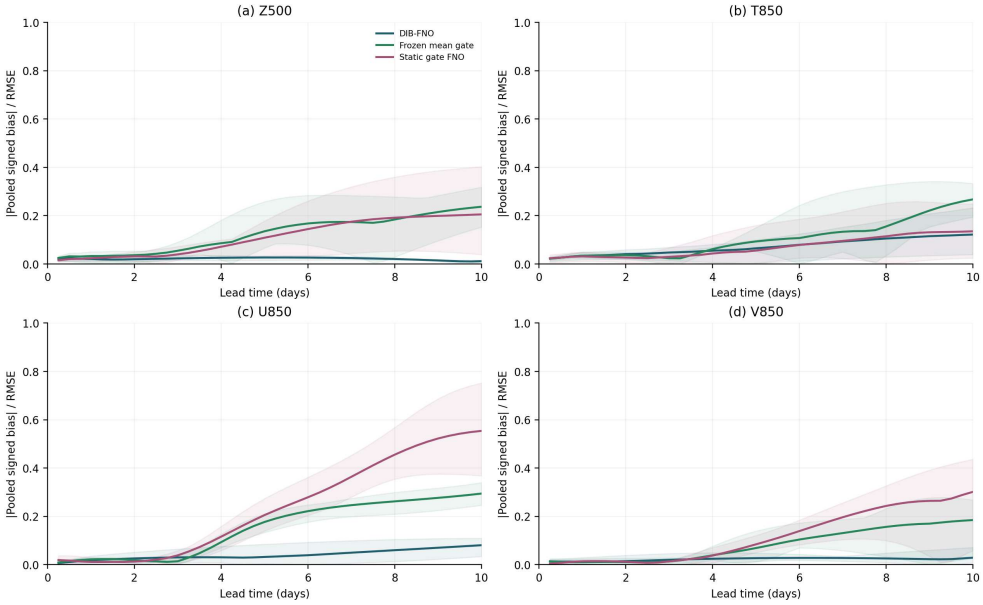

Fig. 11 Signed-bias fraction. Fraction of RMSE associated with the pooled signed-error component.

Table 5 RMSE at 240 h; brackets show pointwise 95% bootstrap intervals.

Variable	Unit	Model	RMSE	Pointwise 95% interval
Z500	gpm	DIB-FNO	137.32	[123.42, 156.21]
		Fixed FNO	163.13	[145.66, 175.38]
		Persistence	111.46	[109.92, 112.92]
		Climatology	82.80	[81.91, 83.72]
T850	K	DIB-FNO	6.22	[5.51, 6.69]
		Fixed FNO	6.42	[6.10, 6.71]
		Persistence	4.50	[4.44, 4.57]
		Climatology	3.31	[3.27, 3.35]
U850	$\mathrm{m\,s^{-1}}$	DIB-FNO	8.28	[7.35, 8.85]
		Fixed FNO	8.29	[7.99, 8.59]
		Persistence	7.03	[6.98, 7.08]
		Climatology	5.13	[5.10, 5.17]
V850	$\mathrm{m\,s^{-1}}$	DIB-FNO	6.89	[6.44, 7.26]
		Fixed FNO	6.77	[6.34, 7.43]
		Persistence	7.12	[7.07, 7.17]
		Climatology	5.07	[5.04, 5.10]

5.5.4 Ten-day skill relative to persistence and climatology

At 240 h, DIB-FNO is worse than climatology for all four reported variables and worse than persistence for Z500, T850, and U850. V850 is slightly better than persistence in point estimate, but the paired uncertainty includes zero. ACC similarly decays toward zero. Consequently, the results support a controlled comparison of operator architectures rather than a competitive medium-range forecasting system.

5.6 Spatial Structure of Forecast Errors

5.6.1 Representative 240-h test cases

Figure 12 displays Z500 forecast-minus-analysis maps for three common 240-h initializations. The rows are selected from the most favourable, median, and least favourable parts of the three-seed DIB-FNO–Fixed FNO ranking; the maps themselves show seed 0 for DIB-FNO and Fixed FNO together with persistence. This case collection is a spatial diagnostic, not an additional population estimator, and each row uses its own common colour scale.

5.6.2 Spatial structure of 240-h errors

The examples show that large ten-day errors are organized primarily in the extratropical waveguide and Southern Ocean rather than appearing as spatially uniform offsets. DIB-FNO can reduce broad error structures in favourable cases yet can also amplify regional anomalies in unfavourable cases. The maps therefore explain why an aggregate Z500 improvement can coexist with substantial case-to-case risk. Because cases were ranked after viewing model differences, they cannot support event-frequency or regional-skill claims; those remain governed by the full-date paired statistics.

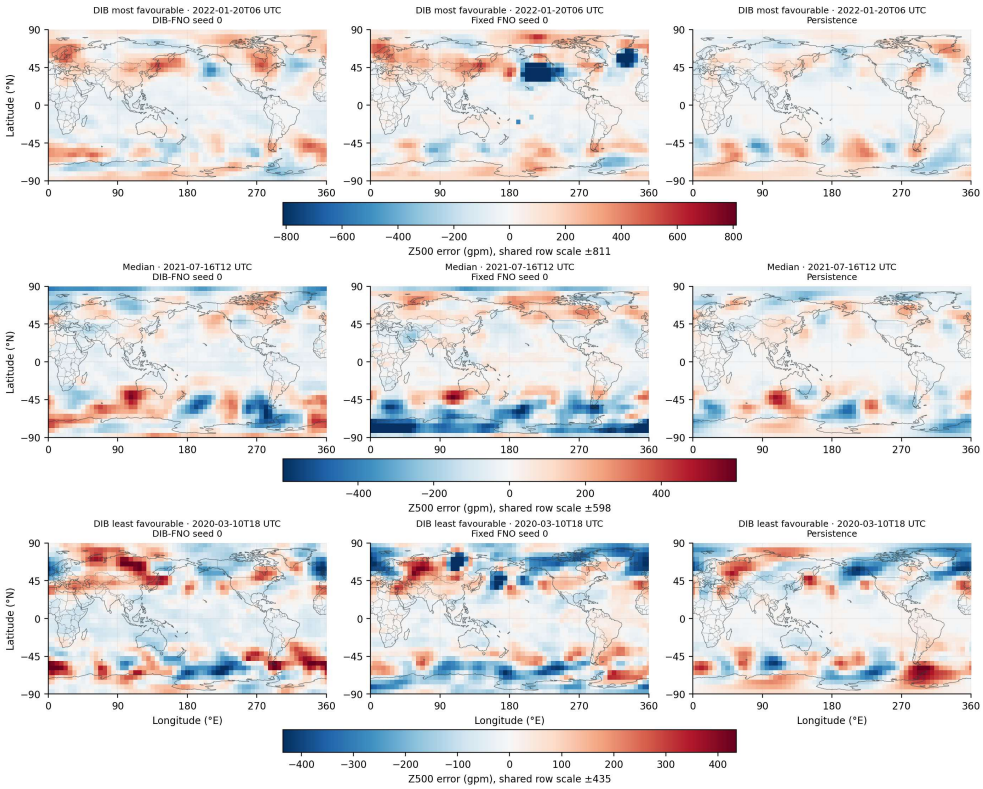

Fig. 12 Z500 errors at 240 h. DIB-FNO, Fixed FNO, and persistence for favourable, median, and unfavourable cases; each row uses a separate colour scale.

5.7 Interpretability and State Dependence of the Spectral Mask

5.7.1 Average mask geometry

The final main experiment analyzes the gate itself rather than only forecast error. For each seed, the soft gate is evaluated on the same 384 evenly spaced training analyses and averaged elementwise. Figure 13 compares the three mean maps and their between-seed standard deviation. The figure is designed to answer whether training discovers a reproducible spectral geometry, while the across-input standard deviations in Figure 14 test whether the continuous weights nevertheless vary with the atmospheric input.

5.7.2 Reproducibility across seeds

The three train-period mean gates have highly similar spatial-frequency geometry (Figure 13). Pairwise Pearson correlation **coefficients** are 0.9951–0.9972, while mean absolute gate differences range from 0.031 to 0.072. Their **respective** total soft-weight sums are 461.26, 535.56, and 493.32 out of 1,056 represented gate elements. Counts

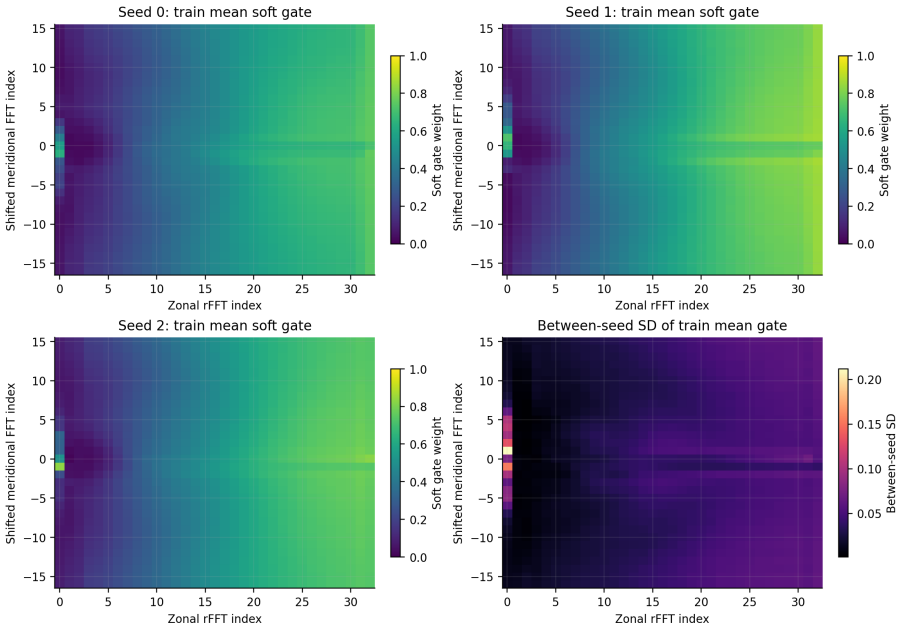

Fig. 13 Mean spectral gates. Seed-specific means over 384 training analyses and their between-seed standard deviation.

above the descriptive threshold 0.5 are 487, 575, and 526. The common **spatial** pattern is therefore more reproducible than the overall gate amplitude.

Input dependence is measurable but modest relative to this shared template (Figure 14). Across the same 384 training analyses, the mean per-frequency across-input standard deviation is 0.023, 0.027, and 0.026 for seeds 0–2. These values establish that the implemented gate changes with the input, but they do not identify adaptation to a specific meteorological regime. A validation-only hard top-C experiment further shows that a continuous gate cannot be reduced to a simple count of retained modes: the dynamic gate has a mean weight sum of 461.44, whereas 1,056 denotes the full binary support. Hard top-64, hard top-256, thresholded, and binary all-pass variants alter both support and amplitude and are therefore sensitivity tests rather than clean causal equivalents.

5.7.3 State-conditioned allocation

The correlations above 0.995 show that the dominant frequency geometry is reproducible across seeds. At the same time, nonzero across-input variation demonstrates that the gate is not static. The appropriate interpretation is a strong shared spectral prior with smaller state-conditioned corrections. The evidence does not justify assigning particular frequencies to named weather regimes, because no regime-labelled analysis was conducted. Likewise, the hard top-C validation variants change both support and amplitude and therefore provide sensitivity information rather than a causal budget-matched replacement for the continuous gate.

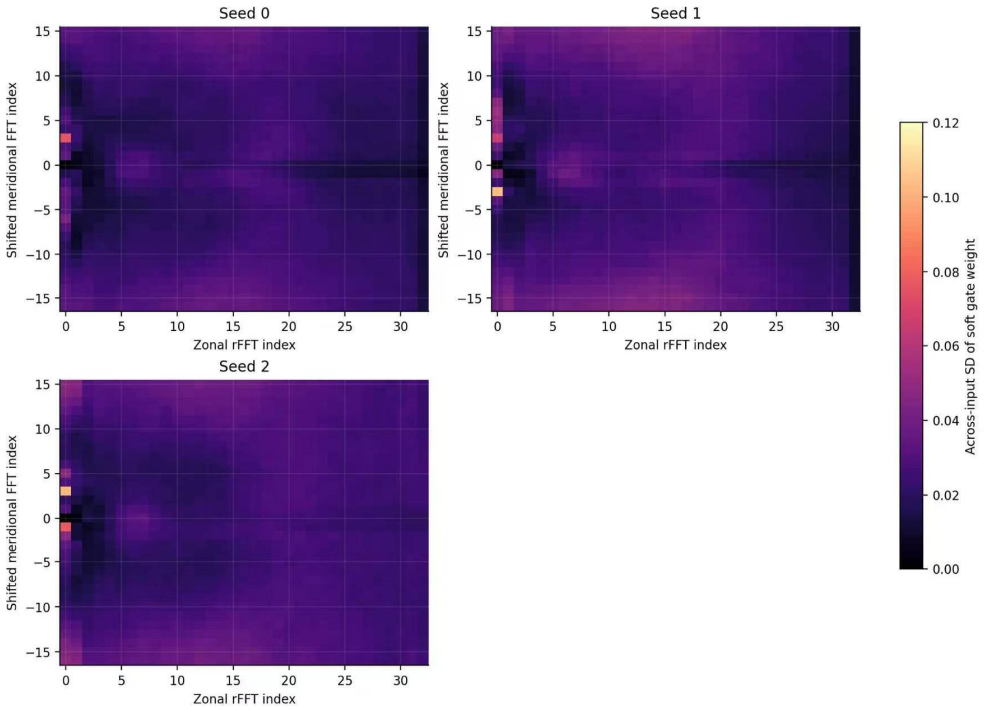

Fig. 14 Input-dependent gate variability. Across-input standard deviation of the continuous soft gate for each seed.

5.8 Sparsity Ablation and the Accuracy–Compression Trade-Off

5.8.1 Selection of the sparsity coefficient

The principal models use $\beta = 10^{-4}$, selected as a **relatively** weak regularization setting that preserves continuous spectral transmission while discouraging uniformly open gates. The three learned mean gates retain soft-weight sums of 461.26–535.56 over 1,056 represented elements, **clearly** showing that the regularizer changes amplitude without producing an explicit binary budget. No multi-β, multi-seed test sweep was performed, so 10^{-4} is a fixed comparison setting rather than an established optimum.

5.8.2 Failure under excessive compression

Figure 15 evaluates seed 0 on the validation period using hard top-64, hard top-256, hard top-1056, a 0.5 threshold, and the normal dynamic soft gate. Hard top-64 and top-256 reduce the represented support sharply and degrade the 120-h persistence-relative score; the all-pass binary variant also performs poorly despite retaining every coefficient. These results show that coefficient count alone does not explain the learned gate. Support, amplitude, and input conditioning change together, so the top-C study is a sensitivity analysis rather than a causal accuracy–compression frontier. Its

validation-only, single-seed results are exploratory rather than confirmatory test-set evidence.

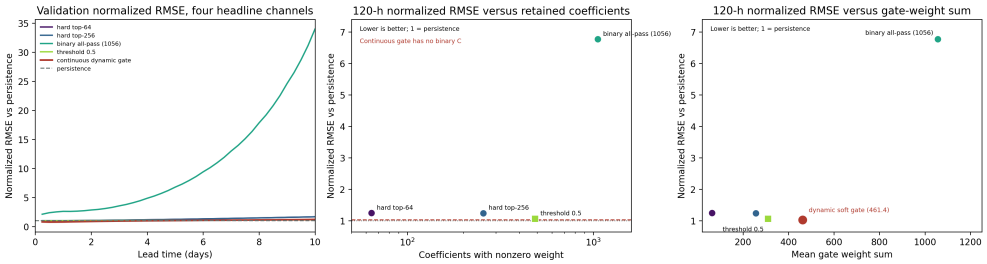

Fig. 15 Hard-gate sensitivity. Validation RMSE for hard top-C, thresholded, all-pass, and dynamic soft-gate variants for seed 0.

5.9 Integrated Interpretation of the Numerical Evidence

DIB-FNO has a robust Z500 advantage over the same-width Fixed FNO, whereas the 240-h differences for T850, U850, and V850 are unresolved. Replacing the input-conditioned gate with frozen or learned static controls produces severe long-rollout instability, indicating that gate adaptation is associated with closed-loop stability. The mean spectral template is highly reproducible across seeds, but its input-conditioned component is modest and post-training hard truncation degrades accuracy. Despite these architectural differences, DIB-FNO loses competitiveness against climatology at ten days for every reported variable. Dynamic spectral allocation therefore changes the autoregressive behaviour of a compact FNO without establishing broadly useful medium-range forecast skill under the present training and resolution constraints.

6 Discussion and Conclusion

6.1 Controlled Result

The strongest result is a reproducible Z500 improvement over a same-width Fixed FNO across lead time and across the tested temporal block lengths. The dynamic model also avoids the catastrophic late-lead growth observed when its gate is frozen and when a near-parameter-matched static gate is trained from scratch. Together with the measured across-input gate variability, these findings are consistent with input-conditioned filtering affecting the learned autoregressive trajectory.

The evidence does not support dynamic gating as the unique cause of the difference. The static model has a slightly different hidden width, training is limited to three seeds and 15 epochs, and the frozen intervention induces a rollout distribution shift after the first step. Moreover, the mean gate template is extremely similar across seeds. A strong shared spectral prior and smaller state-dependent corrections can coexist, and the results do not determine how much each component contributes to stability.

6.2 Forecast-skill Boundary

The absolute baselines are central to **proper** interpretation. By 240 h, DIB-FNO **performs** worse than persistence and climatology for Z500, T850, and U850 and worse than climatology for V850. Even at 72 h, climatology has lower RMSE than the learned models for all four reported variables. The experiment therefore evaluates an operator mechanism under controlled conditions; it is not evidence of a competitive medium-range forecasting system. Comparisons with GraphCast, Pangu-Weather, FourCastNet, FuXi, or numerical weather prediction would require common variables, resolution, initialization dates, verification data, and metrics, none of which are **currently** available here.

6.3 Limitations

Several limitations constrain inference. First, only three training seeds are available, so uncertainty due to optimization is imprecisely estimated. Second, all nine training runs reach the 15-epoch ceiling, and convergence is not established. Third, the model is trained with one-step loss but evaluated in a 40-step rollout; no stability-aware objective is used. Fourth, the gate is global in spectral space and shared across channels, so it cannot localize a weather feature or assign variable-specific spectral weights. Fifth, the grid is coarse and includes only four variables at two pressure levels. Sixth, calendar timestamps are inferred rather than read from the local source metadata. Seventh, because the 2019–2022 period informed exploratory analyses, ablations on the same period are not a fresh blind confirmation.

6.4 Future Work

A confirmatory study should introduce a newly locked or external test period, use more training seeds, train to a demonstrably stable stopping criterion, compare one-step and multi-step objectives, and separate gate adaptation from rollout distribution shift. Localized, channel-specific, spherical, or physically constrained gates may be more appropriate for higher-resolution atmospheric forecasting. A systematic multi-β study should report both continuous gate mass and hard-support diagnostics without conflating them. Multi-step or curriculum losses should be tested against the current one-step objective, and regional or regime-labelled diagnostics should be prespecified before evaluation. For example, independently curated event archives such as IBTrACS could support a preregistered tropical-cyclone stratification rather than post hoc case selection (Knapp et al. 2010). Probabilistic verification is also necessary before translating the approach to decision-relevant prediction (Price et al. 2025).

6.5 Conclusion

We evaluated input-dependent soft spectral gating in a compact Fourier neural operator. Under a common 15-epoch, three-seed protocol, DIB-FNO produces its clearest improvement over a same-width Fixed FNO for Z500 and remains substantially more stable than frozen or learned static-gate controls during long autoregressive rollouts. The learned gates combine a highly reproducible mean frequency template with smaller

but measurable input-conditioned variation. These results support further investigation of dynamic spectral allocation. They do not demonstrate operational medium-range skill: at 240 h the model is inferior to climatology for every reported variable and inferior to persistence for three of four variables. The principal contribution is therefore a bounded architecture-level observation, accompanied by explicit uncertainty, ablations, and failure modes.

Statements and Declarations

Funding. The authors acknowledge support from the Sichuan Provincial Department of Science and Technology (No. 2026NSFSC0138), the Aba Prefecture Science and Technology Achievement Transformation Project "Construction of a Remote Sensing Monitoring and Service System for the Ecological Environment of Typical Nature Reserves in Aba Prefecture" (No. R25CGZH0005), the Key Laboratory of Numerical Simulation of Sichuan Provincial Universities (No. KLNS-2023SZFZ002), and the Key Laboratory of Mathematical Meteorology (No. 2025Z0340).

Competing interests. The authors declare no competing interests.

Author contributions. Jianan Wu: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Resources. Jun Guo: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Data and code availability. The experiment uses ERA5-derived WeatherBench-style fields. A replication package containing model checkpoints, evaluation summaries, per-initialization score arrays, bootstrap indices, gate arrays, figure source data, code snapshots, and file hashes is available at https://github.com/wujianan-win/DIB-FNO. The multi-gigabyte canonical array is not duplicated in the package. Because the local source arrays lack a time coordinate, calendar timestamps were inferred from record length and six-hour spacing; this is documented in the repository data manifest.

References

Alemi AA, Fischer I, Dillon JV, et al (2017) Deep variational information bottleneck. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=HyxQzBceg

Alkin B, Fürst A, Schmid S, et al (2024) Universal physics transformers: A framework for efficiently scaling neural operators. In: Advances in Neural Information Processing Systems, pp 25152–25194, https://doi.org/10.52202/079017-0793

Bai S, Kolter JZ, Koltun V (2019) Deep equilibrium models. In: Advances in Neural Information Processing Systems

Banino A, Balaguer J, Blundell C (2021) PonderNet: Learning to ponder. In: 8th ICML Workshop on Automated Machine Learning, URL https://openreview.net/forum?id=1EuxRTe0WN

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

Bi K, Xie L, Zhang H, et al (2023) Accurate medium-range global weather forecasting with 3D neural networks. Nature 619:533–538. https://doi.org/10.1038/s41586-023-06185-3

Bodnar C, Bruinsma WP, Lucic A, et al (2025) A foundation model for the Earth system. Nature 641:1180–1187. https://doi.org/10.1038/s41586-025-09005-y

Buitrago Ruiz R, Marwah T, Gu A, et al (2025) On the benefits of memory for modeling time-dependent PDEs. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=o9kqa5K3tB

Chen K, Han T, Ling F, et al (2025) 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

Chen L, Zhong X, Zhang F, et al (2023) 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

Choi M, Lee C (2024) Conditional information bottleneck approach for time series imputation. In: International Conference on Learning Representations, URL https://mlanthology.org/iclr/2024/choi2024iclr-conditional/

Dama S, Course K, Nair PB (2025) Shifting time: Time-series forecasting with Khatri–Rao neural operators. In: Proceedings of the 42nd International Conference on Machine Learning, pp 12402–12435, URL https://proceedings.mlr.press/v267/dama25a.html

Dang HV, Nguyen BDG, Nguyen PCH, et al (2025) AdaptFNO: Adaptive Fourier neural operator with dynamic spectral modes and multiscale learning for climate modeling. In: NeurIPS Workshop on Machine Learning and the Physical Sciences, URL https://openreview.net/forum?id=9SvLwi7z2R

de Burgh-Day CO, Leeuwenburg T (2023) Machine learning for numerical weather and climate modelling: a review. Geoscientific Model Development 16:6433–6477. https://doi.org/10.5194/gmd-16-6433-2023

Fan W, Yi K, Ye H, et al (2024) Deep frequency derivative learning for non-stationary time series forecasting. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence, pp 3944–3952, https://doi.org/10.24963/ijcai.2024/436

Guibas J, Mardani M, Li Z, et al (2022) Adaptive Fourier neural operators: Efficient token mixers for transformers. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=EXHG-A3jlM

Herde M, Raonić B, Rohner T, et al (2024) POSEIDON: Efficient foundation models for PDEs. In: Advances in Neural Information Processing Systems, pp 72525–72624, https://doi.org/10.52202/079017-2311

Hersbach H, Bell B, Berrisford P, et al (2020) The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146:1999–2049. https://doi.org/10.1002/qj.3803

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, pp 6840–6851

Hu Y, Yin F, Zhang W, et al (2026) A hybrid framework for global weather forecasting via low-resolution dynamical core and multigrid neural operator. npj Climate and Atmospheric Science 9:112. https://doi.org/10.1038/s41612-026-01374-z

Jahanmard V, Hordoir R, Ramezani-Kebrya A, et al (2026) Principled Fourier neural operators for high-resolution regional ocean modeling. Journal of Geophysical Research: Machine Learning and Computation 3:e2025JH001131. https://doi.org/10.1029/2025JH001131

Knapp KR, Kruk MC, Levinson DH, et al (2010) The international best track archive for climate stewardship (IBTrACS). Bulletin of the American Meteorological Society 91:363–376. https://doi.org/10.1175/2009BAMS2755.1

Kochkov D, Yuval J, Langmore I, et al (2024) Neural general circulation models for weather and climate. Nature 632:1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kurth T, Subramanian S, Harrington P, et al (2023) FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In: Proceedings of the Platform for Advanced Scientific Computing Conference, pp 1–11, https://doi.org/10.1145/3592979.3593412

Lam R, Sanchez-Gonzalez A, Willson M, et al (2023) Learning skillful medium-range global weather forecasting. Science 382:1416–1421. https://doi.org/10.1126/science.adi2336

Li C, Zhou A, Yao A (2022) Omni-dimensional dynamic convolution. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=DmpCfq6Mg39

Li Z, Kovachki N, Azizzadenesheli K, et al (2021) Fourier neural operator for parametric partial differential equations. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=c8P9NQVtmnO

Li Z, Kovachki NB, Choy C, et al (2023) Geometry-informed neural operator for large-scale 3D PDEs. In: Advances in Neural Information Processing Systems, pp

35836–35854, https://doi.org/10.52202/075280-1556

Lipman Y, Chen RTQ, Ben-Hamu H, et al (2023) Flow matching for generative modeling. In: International Conference on Learning Representations

Maddison CJ, Mnih A, Teh YW (2017) The concrete distribution: A continuous relaxation of discrete random variables. In: International Conference on Learning Representations

Neumann P, Düben P, Adamidis P, et al (2019) Assessing the scales in numerical weather and climate predictions: Will exascale be the rescue? Philosophical Transactions of the Royal Society A 377:20180148. https://doi.org/10.1098/rsta.2018.0148

Pathak J, Subramanian S, Harrington P, et al (2022) FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. https://doi.org/10.48550/arXiv.2202.11214, arXiv:2202.11214

Price I, Sanchez-Gonzalez A, Alet F, et al (2025) Probabilistic weather forecasting with machine learning. Nature 637:84–90. https://doi.org/10.1038/s41586-024-08252-9

Rasp S, Thuerey N (2021) Data-driven medium-range weather prediction with a ResNet pretrained on climate simulations: A new model for WeatherBench. Journal of Advances in Modeling Earth Systems 13:e2020MS002405. https://doi.org/10.1029/2020MS002405

Shi Y, Ross ZE, Asimaki D, et al (2025) Mesh-informed neural operator: A transformer generative approach. Transactions on Machine Learning Research URL https://openreview.net/forum?id=K8qAuRfv0G

Sun X, Zhong X, Xu X, et al (2025) A data-to-forecast machine learning system for global weather. Nature Communications 16:6658. https://doi.org/10.1038/s41467-025-62024-1

Tang J, Zhu Q, Yan X (2026) Frequency-adaptive deep learning for multi-horizon weather forecasting in environmental monitoring applications. Scientific Reports 16:18957. https://doi.org/10.1038/s41598-026-49641-6

Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. In: Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing

Weyn JA, Durran DR, Caruana R, et al (2021) Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. Journal of Advances in Modeling Earth Systems 13:e2021MS002502. https://doi.org/10.1029/2021MS002502

Wu H, Luo H, Wang H, et al (2024) Transolver: A fast transformer solver for PDEs on general geometries. In: Proceedings of the 41st International Conference on Machine

Learning, pp 53681–53705, URL https://proceedings.mlr.press/v235/wu24r.html

Yang J, Zhang K, Zhang G, et al (2025) Glocal information bottleneck for time series imputation. In: Advances in Neural Information Processing Systems, pp 115634–115666, https://doi.org/10.52202/085713-3489

Zaheer M, Guruganesh G, Dubey KA, et al (2020) Big Bird: Transformers for longer sequences. In: Advances in Neural Information Processing Systems, pp 17283–17297

Zhao C, Liu S, Mangalam K, et al (2024) Dr2Net: Dynamic reversible dual-residual networks for memory-efficient finetuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15835–15844, https://doi.org/10.1109/CVPR52733.2024.01499