PEAR: EQUAL AREA WEATHER FORECASTING ON THE SPHERE

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence is rapidly reshaping the natural sciences, with weather forecasting emerging as a flagship AI4Science application where leading data-driven models now outperform traditional numerical methods across a broad range of global medium-range forecasting tasks. A common design choice in many of these models is an equiangular discretization of the sphere, which has much smaller grid cells at the poles than around the equator. In contrast, in the Hierarchical Equal Area iso-Latitude Pixelization (HEALPix) of the sphere, each pixel covers the same surface area. Motivated by the growing adoption of this grid in meteorology and climate science, we make the case for using HEALPix as the native grid for data-driven weather forecasting. To this end, we introduce Pangu Equal ARea (PEAR), an efficient volumetric transformer with equal-area inputs, intermediate representations, and outputs. We evaluate PEAR using ERA5-Lite, a reduced subset of the ERA5 atmospheric reanalysis dataset, for forecast times of one to nine days. PEAR remains competitive with the strongest evaluated baselines at short forecasting horizons and outperforms them at longer horizons. Among the evaluated models, PEAR combines the highest average forecasting accuracy with the fastest inference, while requiring less memory than its closest competitors in accuracy. Spectral analysis shows that PEAR has lower large-scale forecast errors at longer forecasting horizons compared to baselines. Furthermore, we quantify learned rotational equivariance for PEAR trained on ERA5-Lite and more densely sampled ERA5 data, and demonstrate single-step climate-model emulation, predicting surface temperature and precipitation from emission forcings.
`Code is provided in the supplementary material.`

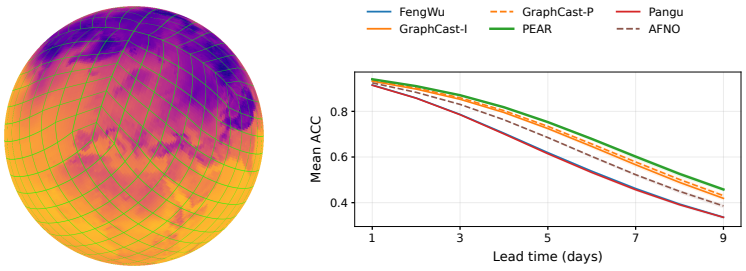


Figure 1: Left: Predicted surface level temperature from the earlier PEAR-HD8 configuration. Green lines show the HEALPix cell boundaries at 3 levels of coarse-graining above the model resolution. Right: Mean anomaly correlation coefficient (ACC, higher is better) over 69 equally weighted variable–level combinations: four surface variables and five upper variables at 13 levels. Native-grid evaluation uses spherical area weighting on 2018 test data. GraphCast-I and GraphCast-P denote baselines matched for inference speed (I) and parameter count (P). Lines show five-seed means and bands show min–max ranges.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

The convergence of artificial intelligence and the natural sciences—often referred to as AI4Science—is transforming how complex physical systems are studied and predicted. In Earth system science, machine learning approaches have achieved performance on par with or exceeding that of state-of-the-art numerical models (Yu et al., 2024; Liu & Ma, 2024; Zhao et al., 2024). Weather forecasting stands at the forefront of this shift (Lam et al., 2023; Chen et al., 2023b;a; Andrychowicz et al., 2023; Nguyen et al., 2024), driven by advances in architectures that can capture multiscale atmospheric dynamics directly from data, while offering orders of magnitude speedups compared to traditional numerical integration methods. Many of these models use regular latitude–longitude grids, where grid cells become smaller towards the poles. Consequently, treating grid points uniformly gives different geographic regions unequal weight per unit area. Here, we investigate the Hierarchical Equal Area iso-Latitude Pixelization (HEALPix) as a native grid for learning-based weather forecasting.

The emergence of data-driven weather forecasting models such as FourCastNet (Kurth et al., 2023), Pangu-Weather (Bi et al., 2023), and GraphCast (Lam et al., 2023) was enabled by ERA5 (Hersbach et al., 2020), a comprehensive global reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides hourly estimates of atmospheric, land, and oceanic variables on a global scale. Its native atmospheric grid is reduced Gaussian, while commonly used data products are interpolated onto regular latitude–longitude grids. These regular grids provide a convenient rectangular tensor representation but oversample the polar regions. Reduced Gaussian grids mitigate this oversampling, but their latitude-dependent row lengths complicate the direct application of standard convolutional and window-based attention operations.

HEALPix, originally developed for astrophysical applications (Gorski et al., 1998), combines equal-area pixels with a hierarchical discretization of the sphere. Its nested structure supports fixed-size patches, attention windows, and multiresolution processing. The Destination Earth Climate Change Adaptation Digital Twin uses HEALPix as its standard shared grid for climate-model output (Doblas-Reyes et al., 2025), providing a consistent representation of high-resolution simulations from different climate models. This development further motivates learning architectures that operate directly on HEALPix. Building on the volumetric transformer architecture of Pangu-Weather (Bi et al., 2023) and the spherical shifted-window formulation of HEAL-SWIN (Carlsson et al., 2024), we introduce PEAR, a weather forecasting model with native HEALPix inputs, internal features, and outputs.

We train and evaluate PEAR on ERA5-Lite (Bi et al., 2023), a reduced subset of ERA5 that enables systematic experimentation at substantially lower computational cost than the full dataset. We compare PEAR with leading weather-forecasting architectures on forecasting horizons from one to nine days. Across most combinations of atmospheric variables and forecasting horizons, PEAR achieves the highest accuracy while offering a favorable trade-off between accuracy and inference time. Additional experiments examine spectral errors, empirical equivariance under rotations about the polar axis, and the application of PEAR to climate-model emulation (Kaltenborn et al., 2023).

Our main contributions are:

- Motivated by the growing adoption of HEALPix in next-generation weather and climate data, we make the case for using HEALPix as the native grid for weather forecasting. To this end, we introduce PEAR, an efficient volumetric transformer with equal-area inputs, intermediate representations, and outputs.

- We evaluate all models for lead times of up to nine days and show that PEAR achieves competitive accuracy at short forecasting horizons and outperforms leading weather-forecasting baselines at longer horizons. Among the evaluated models, it combines the highest forecasting accuracy averaged across atmospheric variables with the fastest inference, while requiring less memory than its closest competitors in accuracy.

- We show that PEAR achieves lower large-scale forecast errors than the strongest evaluated baselines at longer forecasting horizons. We quantify equivariance error under rotations about the polar axis and demonstrate lower average error on single-step climate-model emulation compared to a larger convolutional baseline.

2 RELATED WORK

Pangu-Weather (Bi et al., 2023) uses a volumetric version of the Swin transformer (Liu et al., 2021). Pangu-Weather, FuXi (Chen et al., 2023b), and FengWu (Chen et al., 2023a) operate on regular latitude–longitude inputs and outputs, whereas GraphCast (Lam et al., 2023) maps latitude–longitude fields to an internal icosahedral multimesh. FourCastNet (Kurth et al., 2023) and AFNO (Guibas et al., 2022) apply adaptive Fourier operators on latitude–longitude fields, while SFNO (Bonev et al., 2023) uses spherical harmonic transforms. More broadly, learning on spherical domains includes harmonic spherical CNNs (Cohen et al., 2018), gauge-equivariant icosahedral networks (Cohen et al., 2019), graph-based approaches such as DeepSphere (Perraudin et al., 2019; Defferrard et al., 2020), and equivariant networks for pixelized spheres (Shakerinava & Ravanbakhsh, 2021). Geometry-aware forecasting approaches also include SphereNet-based convolutions that account for spherical curvature (Scher & Messori, 2024) and the circular transformer CirT, which models longitudinal periodicity (Liu et al., 2025).

HEAL-SWIN (Carlsson et al., 2024) introduced a hierarchical shifted-window transformer on HEALPix for spherical vision tasks. DLWP-HPX (Karlbauer et al., 2024) uses a convolutional recurrent architecture with native HEALPix inputs and outputs and demonstrates improvements over a cubed-sphere counterpart. HEAL-ViT (Ramavajjala, 2024) applies shifted-window attention to an internal HEALPix mesh, connected to latitude–longitude inputs and outputs through graph-based encoding and decoding. HEAL-ViT represents atmospheric levels as feature channels rather than retaining a separate vertical token dimension. PEAR combines native HEALPix inputs, intermediate representations, and outputs in a transformer architecture with an explicit vertical dimension.

3 HEALPIX

The HEALPix grid uniformly divides the sphere into four-sided polygons (quadrilaterals) of equal area. The pixels are positioned at the centers of the quadrilaterals and lie on circles of constant latitude with equal spacing in azimuthal angle. The construction of the grid starts from 12 base-quadrilaterals, 4 equally shaped quadrilaterals grouped around each pole and 4 equally shaped quadrilaterals around the equator. These base-quadrilaterals are then repeatedly divided into two equally-sized halves along their edges. After k divisions, there are $n_{\text{side}} = 2^k$ pixels along each edge of each base quadrilateral, resulting in a grid with $12 \cdot n_{\text{side}}^2$ pixels in a hierarchical structure.

For computations, the pixels are organized in a one-dimensional list. In the *nested* ordering, the pixel indices are provided by the hierarchical construction of the grid such that merging blocks of four consecutive pixels in the list coarse-grains the grid from n_{side} to $n_{\text{side}}/2$. We will use this property for coarse- and fine-graining the pixel grid. Similarly, blocks of 4^k consecutive pixels in the nested ordering correspond to the pixels in quadrilaterals k division-levels above the grid resolution. We will use this property to easily divide the surface of the sphere into attention-windows.

In the *ring* ordering of the pixel list, the pixels are sorted along the iso-latitude circles, from the north pole to the south pole. Performing a roll operation on this list rotates the features on the sphere around the polar axis. Around the poles, the features will additionally be distorted due to the decreasing number of pixels per iso-latitude circle. We will use this roll operation in the ring ordering to shift the features between windows after attention layers. For a roll by n pixels, the last n pixels around the south pole will spill over to the north pole and therefore be masked in the attention weights.

3.1 ERA5 ON HEALPIX

We use ERA5 atmospheric fields (Hersbach et al., 2020) provided hourly on a regular latitude–longitude grid with 0.25° spacing and a longitude–latitude resolution of 1440×721 points. Following Pangu-Weather (Bi et al., 2023), we select four surface variables—2-m temperature, mean sea-level pressure, and the two 10-m wind components—and five upper-air variables—temperature, specific humidity, geopotential, and the two horizontal wind components—at 13 pressure levels.

Since the equiangular grid cells are not of equal area, the effective spatial resolution on the sphere is not constant. The lowest spatial resolution can be found on the equator, where each cell has an angular area of $1.9 \cdot 10^{-5}\text{rad}^2$. We target a HEALPix grid with $n_{\text{side}} = 64$, where all of the $12 \cdot 64^2$

pixels cover an angular area of $2.6 \cdot 10^{-4}$, so that the ERA5 data has a higher spatial resolution everywhere on the globe and saturates the HEALPix pixel density.

3.2 ERA5-LITE

The full ERA5 dataset contains 40 years with hourly intervals, resulting in multiple petabytes of storage required. To facilitate research on weather models, a reduced dataset using 11 years and 24 hour intervals has been used (Bi et al., 2023). This subset consists of the years 2007 to and including 2017, with 2019 used for validation and 2018 for testing. All samples are at 00:00UTC, resulting in a total of 4017 training samples and 365 validation samples and 365 test samples. Together with cached normalized data, the total dataset size is about 3TB.

4 PEAR: PANGU EQUAL AREA

PEAR maps the global weather state at time t to the state at $t + \Delta t$, with $\Delta t = 24$ hours. The global volumetric weather state is discretized on the HEALPix grid along the surface, and into 13 discrete levels in the vertical direction. Following prior work (Bi et al., 2023), we represent the total weather state as a combination of 4 surface variables (wind speed along the surface, temperature and mean sea level pressure), and 5 upper variables (wind speed along the sphere, temperature, specific humidity and geopotential) at 13 discrete vertical levels. PEAR thus takes two input tensors, the surface and upper variables discretized on the spherical surface and the spherical shell correspondingly. Since the HEALPix grid covers the sphere with a 1d index structure, the model input tensors have shape $(12n_{\text{side}}^2, 4)$ and $(12n_{\text{side}}^2, 13, 5)$.

The architecture is constructed using a combination of 5 main layer types: patch embedding, windowed attention with alternating shifting, downsampling, upsampling and patch recovery. See Figure 2 for a schematic overview of the architecture, and Table B.1 in the appendix for details on the layer parameters. Appendix B reports ablation results for these design choices.

4.1 PATCH EMBEDDING

The initial patch embedding uses a 1d convolution with kernel size 16 and stride 16 for the surface variables, and a 2d convolution with kernel size and stride $(16, 2)$ for the upper variables. This corresponds to a patch size of 4×4 in an equiangular grid. Both convolutions output 72 channels. The patch embedded tensors have shape $(3/4 n_{\text{side}}^2, 72)$ and $(3/4 n_{\text{side}}^2, 7, 72)$. At this point the patch embedded surface variables are concatenated to the patch embedded upper variables, resulting in a single tensor of shape $(3/4 n_{\text{side}}^2, 8, 72)$.

4.2 WINDOWED ATTENTION

We partition the tensors into windows using the nested structure of the HEALPix grid. The vertical direction is also partitioned into neighbouring levels. In terms of the latent tensor shape $(3/4 n_{\text{side}}^2, 8, 72)$, partitioning using a window size of (W_{hp}, W_d) gives a windowed tensor of shape $(\frac{3/4 n_{\text{side}}^2}{W_{\text{hp}}} \frac{8}{W_{\text{d}}}, W_{\text{hp}} W_d, 72)$. Attention is now performed over the embeddings of the $W_{\text{hp}} W_d$ voxels in each window. Note that the nested index structure of the HEALPix grid naturally supports the window partitioning in terms of contiguous memory. See Figure B.1 for an illustration of the patches and first level windows corresponding to the tensor structure at the first attention layer.

In contrast to Pangu (Bi et al., 2023), we share learned relative positional embeddings between windows. We learn a relative positional embedding tensor B of shape $(1, N_{\text{heads}}, (W_d W_{\text{hp}})^2)$. The final attention computation is thus $\text{Att}(Q, K, V) = \text{SoftMax}\left(\frac{QK^\top}{\sqrt{d}} + B\right) V$, where d is the per-head embedding dimension. Sharing these embeddings reduces the positional parameter count; the complete PEAR and Pangu configurations also differ in width and head count.

To propagate information between the windows, that are otherwise disjoint in terms of attention, we shift the grid by roughly half the window size every other attention layer. Along the spherical directions of the HEALPix grid we employ the ring shifting strategy of HEAL-SWIN (Carlsson et al., 2024), and a simple shift in the vertical direction. Since this shifting is performed cyclically, there

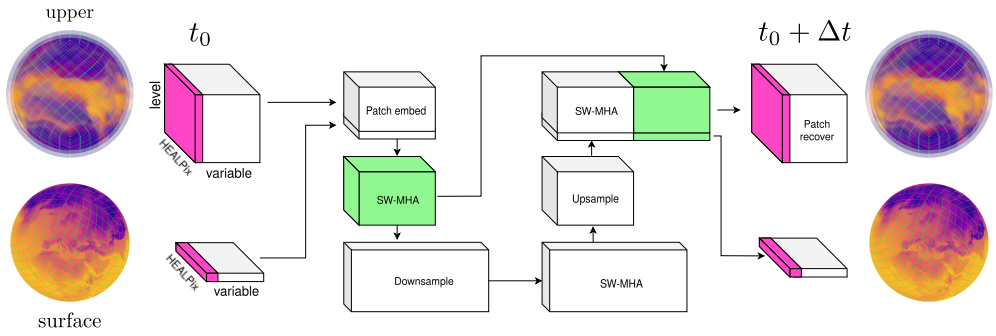


Figure 2: PEAR architecture schematic. Violet slices correspond to the variables visualized on the sphere for the input and output tensors. Each block indicates the tensor shape after the layer with the corresponding name. Patch embedding by convolution, shifted windowed multi-head attention (SW-MHA) with learned positional embedding, downsampling and upsampling by patch merging and splitting and patch recovery by transpose convolutions. Green block indicates the skip connection, where the output of the first attention layer is concatenated along the embedding dimension before the final patch recovery by transpose convolutions. Green lines on the spherical visualizations indicate the HEALPix grid at 3 levels of coarse-graining above the model resolution.

will be voxels in the polar regions that will jump from the north to the south pole, and from the lowest upper level to the highest upper level. To prevent attention among these spatially disjoint voxels, we implement masked attention. See Figure B.2 in Appendix B for an illustration of the masking pattern that arises from the above shifting strategy. Note that because of the 1d structure of the HEALPix grid, this figure is representative for the actual tensor structure used in the implementation. To facilitate the shifting in the ring indexing scheme, we precompute the index conversion from nested to ring scheme. Each windowed attention block follows the structure of SWIN-V2 (Liu et al., 2022) using layer-norm and skip connection.

4.3 Downsampling and upsampling

To facilitate a bottleneck structure, we follow Pangu (Bi et al., 2023) and perform a single downsampling along the spherical directions. The HEALPix grid has a hierarchical structure where four neighbouring grid cells combine into a single grid cell at a coarser resolution. This provides a natural downsampling by concatenating the embeddings of groups of four neighbouring pixels, and then linearly projecting to the target embedding dimension. This downsampling is efficient in the nested indexing of the HEALPix grid, where this simply corresponds to a reshaping of the tensor, followed by a linear layer for projection.

The upsampling layer follows the same logic in reverse: we first expand the embedding dimension of a voxel to four times the target embedding dimension, followed by a reshaping into four new voxels along the HEALPix grid in the nested indexing scheme.

4.4 Patch recovery

To recover the surface and upper variable tensors, we use transpose convolutions on the first and remaining vertical levels correspondingly. After concatenating the skip connection, the latent tensor x has shape $(3/4 n_{\text{side}}^2, 8, 144)$. It is split along the second dimension into a surface latent x_{surface} of shape $(3/4 n_{\text{side}}^2, 1, 144)$, and an upper latent tensor x_{upper} of shape $(3/4 n_{\text{side}}^2, 7, 144)$. To recover the output surface variables we perform a 1d transpose convolution on x_{surface} with channel count 4. The output upper variables are recovered with a 2d transpose convolution with channel count 5.

5 EXPERIMENTS

We evaluate PEAR against Pangu (Bi et al., 2023), GraphCast (Lam et al., 2023), FengWu (Chen et al., 2023a), and AFNO (Guibas et al., 2022), using the PhysicsNeMo model implementations (PhysicsNeMo Contributors, 2023). We train all models in our training framework on the ERA5-Lite subset described in Section 3.2, with five random seeds per model. PEAR uses HEALPix at $n_{\text{side}} = 64$, while the baselines use equiangular inputs and outputs. In the main comparison, all models use L1 loss on standardized outputs. Spatial weights are uniform except for AFNO, which uses cosine-latitude weighting. Appendix A provides the training details, and Appendix C reports an additional Pangu control trained with latitude-weighted loss.

All models are trained to predict the atmospheric state 24 hours ahead. We generate forecasts at horizons from one to nine days by applying each model iteratively, using the predicted state as input for the next step. At each horizon, we calculate global root mean squared error (RMSE) and anomaly correlation coefficient (ACC) (van den Dool et al., 2017) for each of the 69 variable–level combinations. Appendix C provides the metric definitions and evaluation details.

Our primary comparison evaluates forecasts on the native output grid of each model. PEAR is evaluated on 49,152 equal-area HEALPix cells, while the baselines use a 157×314 equiangular grid. Spatial averages use uniform weights on HEALPix and normalized cosine-latitude weights on the equiangular grid. ACC measures the correlation between forecast and ERA5 deviations from the same climatological mean. For each variable and grid point, this mean is estimated from 2007–2017 for the calendar day being predicted. For each seed, we select the checkpoint with the lowest mean RMSE on the 2019 validation data, using the same grid-specific spatial weights. Appendix C provides detailed native-grid results, common-grid comparisons, and the climatology construction.

Pangu uses a similar volumetric transformer architecture with equiangular inputs, internal representations, and outputs. We retain the standard Pangu block depths and attention-head counts, with a reduced embedding width, resulting in 11.4M parameters. Unlike PEAR, Pangu retains separate positional-bias tensors for different latitude–pressure window types, while sharing positional biases across longitude windows, see Appendix Table B.3 for an ablation with shared positional bias for Pangu. See Appendix B for PEAR architectural ablations.

Our scaled-down GraphCast configurations use a level-4 icosahedral multimesh, giving a horizontal processing resolution comparable to the finer internal resolution of PEAR. GraphCast-I uses 144 hidden features and seven processor layers, resulting in 1.51M parameters. GraphCast-P uses 200 hidden features and twelve processor layers, resulting in 4.30M parameters and matching the parameter count of PEAR within 1%. This configuration requires approximately twice the inference time and 3.3 times the peak allocated GPU memory of PEAR. GraphCast-I and AFNO have measured inference times approximately 15% and 10% longer than PEAR, respectively, providing some leeway in favor of the baselines. See Appendix Table A.2 for details. Appendix A provides the architecture-selection details and inference-benchmarking protocol.

Figure 3 shows ACC (higher is better) at forecasting horizons from one to nine days for the four surface variables and five upper-air variables, with upper-air results averaged over 13 pressure levels. PEAR achieves higher mean ACC and lower mean RMSE than GraphCast-I for all 69 variable–level combinations at days 3–9. Against GraphCast-P, PEAR achieves higher mean ACC for all 69 combinations at days 6–9 and lower mean RMSE for 68 of the 69 combinations at day 9. Across the six models in the native-grid comparison, PEAR achieves the highest mean ACC in 502 of the 621 combinations of variables and forecasting horizons, and the lowest mean RMSE in 492. GraphCast-P leads the remaining comparisons. Appendix Tables C.2 and C.3 provide detailed results.

Table 1 summarizes ACC separately for the four surface variables and five upper-air variables. Each group gives equal weight to its variables, with upper-air ACC first averaged over the 13 pressure levels. PEAR achieves the highest average upper-air ACC at days 1, 5, and 9 with no seed range overlap. For surface variables, GraphCast-P leads at day 1, while PEAR leads at days 5 and 9, although the five-seed ranges of PEAR and GraphCast-P overlap at these two horizons.

Figure 4 compares forecasting accuracy and inference time at days 1 and 9, with marker area indicating peak allocated GPU memory. With ACC averaged equally over the 69 variable–level combinations, PEAR achieves higher mean ACC than both GraphCast variants at every tested forecasting horizon,

Table 1: Native-grid ACC on 2018 test data using spherical area weighting. Surface averages msl, t2m, u10, and v10 equally. Upper averages q, t, u, v, and z equally after averaging over 13 levels. Sub- and superscripts give offsets to the five-seed min–max range. Higher is better. Bold entries have the highest mean or an overlapping range.

Group	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
Surface	1	$0.928^{+0.001}_{-0.002}$	$0.948^{+0.000}_{-0.001}$	$\mathbf{0.951^{+0.001}_{-0.001}}$	$0.946^{+0.001}_{-0.002}$	$0.924^{+0.002}_{-0.002}$	$0.945^{+0.000}_{-0.000}$
	5	$0.584^{+0.004}_{-0.004}$	$\mathbf{0.697^{+0.006}_{-0.009}}$	$\mathbf{0.706^{+0.006}_{-0.006}}$	$\mathbf{0.712^{+0.003}_{-0.010}}$	$0.576^{+0.003}_{-0.005}$	$0.654^{+0.001}_{-0.002}$
	9	$0.287^{+0.004}_{-0.007}$	$0.357^{+0.011}_{-0.014}$	$\mathbf{0.369^{+0.009}_{-0.006}}$	$\mathbf{0.387^{+0.005}_{-0.010}}$	$0.278^{+0.003}_{-0.002}$	$0.332^{+0.004}_{-0.005}$
Upper	1	$0.915^{+0.001}_{-0.001}$	$0.931^{+0.000}_{-0.001}$	$0.935^{+0.002}_{-0.001}$	$\mathbf{0.940^{+0.001}_{-0.002}}$	$0.914^{+0.001}_{-0.001}$	$0.924^{+0.000}_{-0.000}$
	5	$0.622^{+0.004}_{-0.007}$	$0.726^{+0.005}_{-0.007}$	$0.736^{+0.005}_{-0.010}$	$\mathbf{0.756^{+0.003}_{-0.007}}$	$0.617^{+0.003}_{-0.005}$	$0.687^{+0.004}_{-0.005}$
	9	$0.339^{+0.006}_{-0.010}$	$0.423^{+0.007}_{-0.009}$	$0.434^{+0.009}_{-0.013}$	$\mathbf{0.462^{+0.007}_{-0.010}}$	$0.339^{+0.003}_{-0.003}$	$0.389^{+0.009}_{-0.011}$

with non-overlapping five-seed ranges, while requiring less inference time. Appendix Table A.2 provides parameter counts, inference times, and peak allocated GPU memory for all six models.

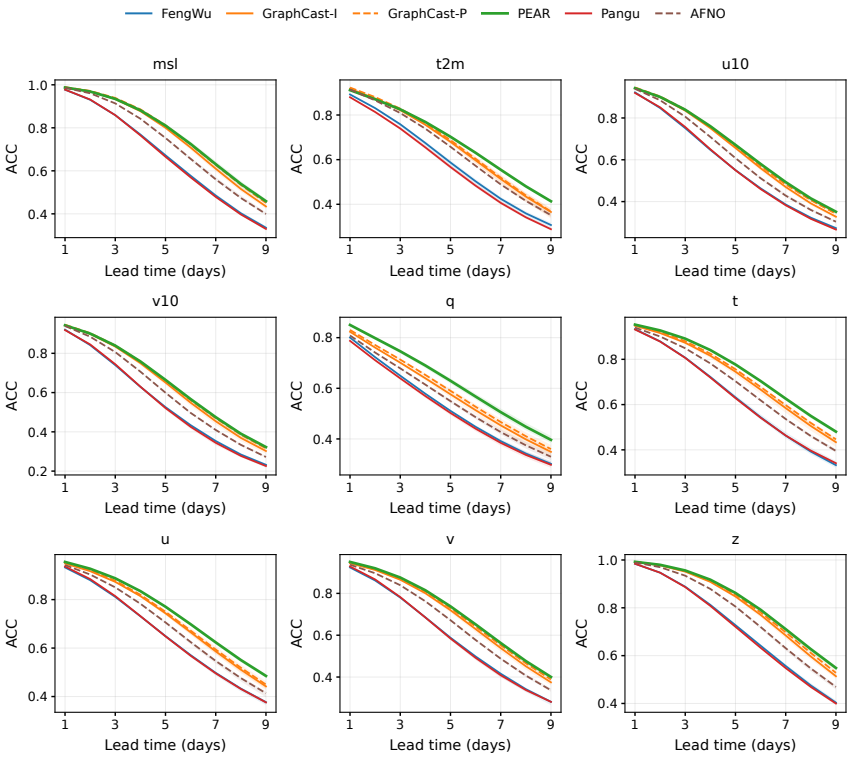


Figure 3: Mean ACC (higher is better) on 2018 test data, with each model evaluated on its native grid using spherical area weighting. Upper variables are averaged over 13 pressure levels. Bands show five-seed min–max ranges.

5.1 Spectral results

Figure 5 shows how forecast error is distributed across spatial scales at days 1 and 9. We compute spherical-harmonic error spectra and average equally over t2m, msl, z500, t850, and q700 after normalizing by the total ERA5 variance of each field. At day 9, PEAR has lower large-scale error than both GraphCast variants. Fine-scale errors are of similar magnitude in the five-field average, with lower error for both GraphCast variants. These comparisons use a common HEALPix grid,

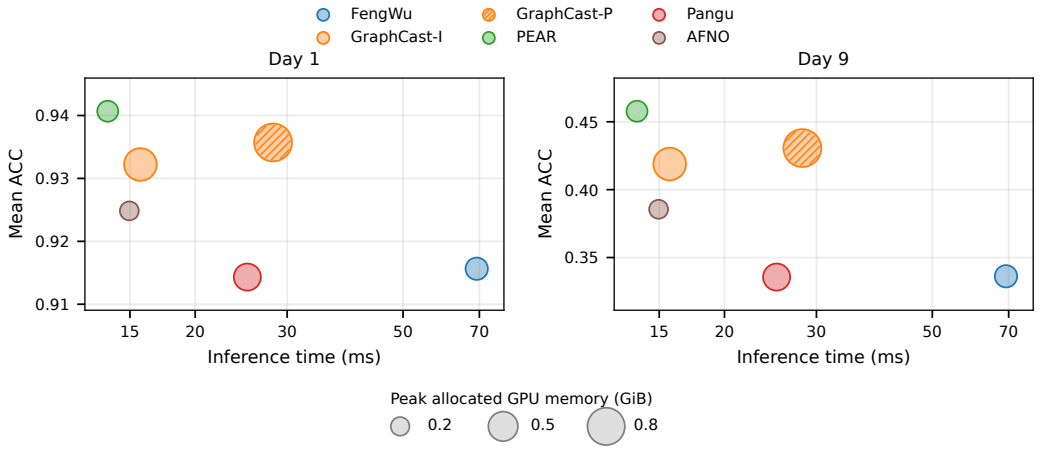


Figure 4: Native-grid mean ACC over 69 equally weighted variables versus inference time (logarithmic) at days 1 and 9. Points show five-seed means; circle area is proportional to seed-0 peak allocated GPU memory during one-step inference. GraphCast-P uses a striped fill. See Appendix Table A.2 for details.

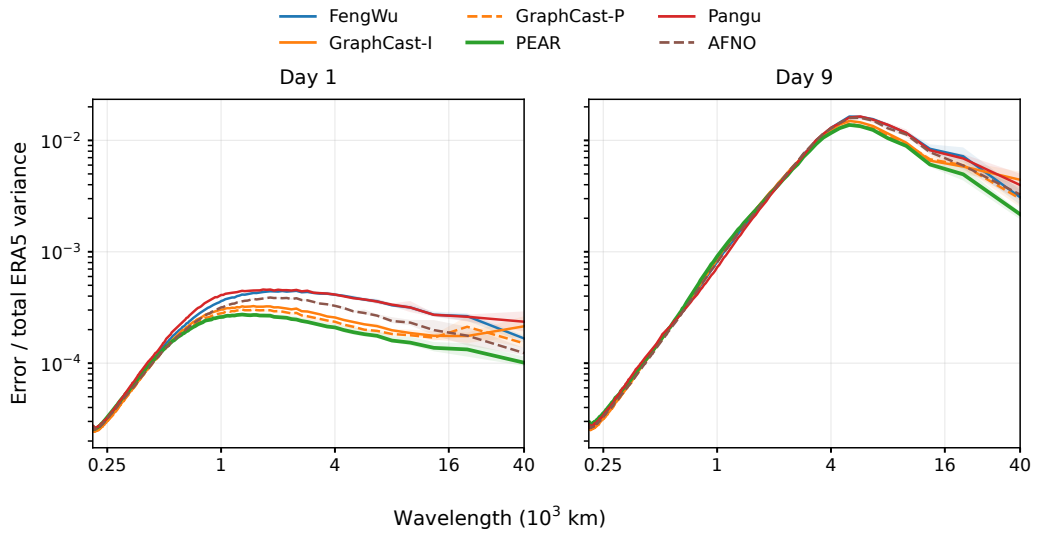


Figure 5: Forecast-error spectra on common HEALPix at days 1 and 9 (lower is better). Error variance per degree is normalized by each field's total ERA5 variance over degrees 1–191, then averaged equally over five fields. Lines and shading show five-seed means and min–max ranges. Small scales are on the left. See Appendix E for forecast spectra and details.

with equiangular forecasts remapped before evaluation. Appendix E provides the spectral definitions, normalization, and individual-field spectra.

5.2 Equivariance and climate-model emulation

We compare the rotational equivariance learned under the ERA5-Lite sampling regime with that learned from more densely sampled ERA5 data. Specifically, we compare PEAR trained on ERA5-Lite with PEAR trained on two-hourly ERA5 data from 2012. For rotations about the polar axis, we measure the discrepancy between rotating a forecast and forecasting from a rotated input. Figure 6a shows the maximum error over rotation angles, averaged across upper-air variables at each pressure level. Both trained models have similar errors, substantially below the untrained reference, with

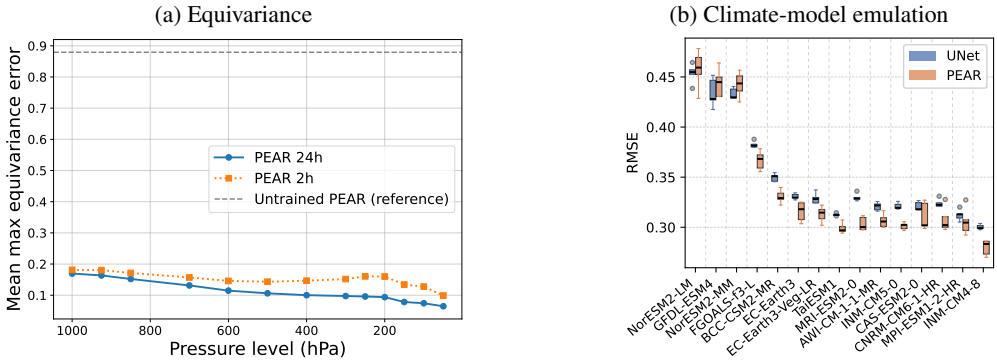


Figure 6: (a) Maximum mean absolute equivariance error over polar-axis rotations, averaged across upper variables (2019 validation). (b) Five-seed box plots of one-step ClimateSet RMSE, averaged over normalized temperature and precipitation outputs (Kaltenborn et al., 2023).

slightly lower error for the ERA5-Lite model. Error generally decreases with altitude, although not monotonically for every variable. These results indicate that PEAR trained on ERA5-Lite exhibits learned rotational equivariance comparable to that observed with a more densely sampled, single-year ERA5 dataset. Appendix F describes these separately trained configurations and their evaluation protocols.

To evaluate PEAR beyond weather forecasting, we consider single-step climate-model emulation, approximating outputs of Earth system models (ESMs) from emission forcings (Watson-Parris, 2021). We use ClimateSet (Kaltenborn et al., 2023) to predict monthly surface temperature and precipitation from emission maps of CO_2, SO_2, CH_4, and BC. For each of 15 climate models, we train PEAR with five random seeds on HEALPix at $n_{\text{side}} = 32$ and evaluate on a held-out emission scenario. Figure 6b compares PEAR with the ClimateSet UNet baseline using RMSE averaged over the normalized output variables. PEAR achieves lower overall RMSE for 13 of the 15 climate models, with improvements in precipitation and comparable performance for temperature. The broader comparison in Appendix G includes ClimaX, which has 106.6M parameters compared with 4.3M for PEAR, approximately 25 times as many. ClimaX achieves the lowest overall RMSE for 10 of the 15 climate models, while PEAR achieves the lowest precipitation RMSE for 8.

6 Limitations

Our experiments use ERA5-Lite and a single held-out test year, so the conclusions are limited to this data and model-size regime. Ideally we would train on ground truth data created on the HEALPix grid, but the ERA5 data used here are provided on a regular latitude–longitude grid. Known limitations of data-driven forecasts include unphysical predictions (Schreck et al., 2024), instabilities for longer forecasts (Watt-Meyer et al., 2024) and lack of certain features such as the butterfly effect (Selz & Craig, 2023) or sub-synoptic and mesoscale weather phenomena (Bonavita, 2024).

7 Conclusion

We introduced PEAR, an efficient volumetric transformer for global weather forecasting with equal-area inputs, intermediate representations, and outputs on the HEALPix grid. Experiments on the ERA5-Lite weather dataset show competitive accuracy at short forecasting horizons and improvements over leading baselines at longer horizons. Among the evaluated models, PEAR combines the highest forecasting accuracy averaged across atmospheric variables with the fastest inference, while requiring less memory than its closest competitors in accuracy. At nine days, spectral analysis shows improved prediction of large-scale structure, with comparable fine-scale errors. Further experiments quantify equivariance error under rotations about the polar axis and evaluate single-step climate-model emulation, predicting surface temperature and precipitation from emission forcings. Together, these results support HEALPix-based volumetric transformers as an effective approach to accurate and computationally efficient weather forecasting.

AI USE STATEMENT

In this work, we used LLMs to assist in method implementation, ablation design, compute cluster orchestration, literature searches, manuscript drafting and editing, and preparation of tables and figures. We initially implemented the training and evaluation framework manually, and subsequently used LLM assistance to integrate additional baselines, improve cluster management and overall code robustness. All output from LLMs have been reviewed or reworked in detail and we take full responsibility for the final manuscript and associated artifacts.

ETHICS STATEMENT

This work uses atmospheric reanalysis and climate-model simulation data. Forecast errors can affect safety-critical decisions, and the reported benchmark results do not establish suitability for operational deployment.

REPRODUCIBILITY STATEMENT

We provide code for model implementations, training configurations, and evaluation for the weather models and ablations in the supplementary material to support reproducibility. Appendix A documents data preparation, training objectives, model configurations, checkpoint selection, and inference benchmarking. Appendix B provides architectural details and ablations. Appendices C–G describe the evaluation protocols and additional results for forecasting accuracy, regional and spectral errors, rotational equivariance, and climate-model emulation.

REFERENCES

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep Learning for Day Forecasts from Sparse Observations, July 2023.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3.

Massimo Bonavita. On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. ISSN 1944-8007. doi: 10.1029/2023GL107377.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2806–2823. PMLR, July 2023.

Salva Rühling Cachay, Venkatesh Ramesh, Jason NS Cole, Howard Barker, and David Rolnick. Climart: A benchmark dataset for emulating atmospheric radiative transfer in weather and climate models. *arXiv preprint arXiv:2111.14671*, 2021.

Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. HEAL-SWIN: A Vision Transformer On The Sphere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6067–6077, June 2024.

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

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):1–11, November 2023b. ISSN 2397-3722. doi: 10.1038/s41612-023-00512-1.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the Icosahedral CNN. In *Proceedings of the International Conference on Machine learning (ICML)*, pp. 1321–1330. PMLR, 2019.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Michaël Defferrard, Martino Milani, Frédérick Gusset, and Nathanaël Perraudin. DeepSphere: A graph-based spherical CNN. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

F. J. Doblas-Reyes, J. Kontkanen, I. Sandu, M. Acosta, M. H. Al Turjmam, I. Alsina-Ferrer, M. Andrés-Martínez, L. Arriola, M. Axness, M. Batlle Martín, P. Bauer, T. Becker, D. Beltrán, S. Beyer, H. Bockelmann, P.-A. Bretonnière, S. Cabaniols, S. Caprioli, M. Castrillo, A. Chandrasekar, S. Cheedela, V. Correal, E. Danovaro, P. Davini, J. Enkovaara, C. Frauen, B. Früh, A. Gaya Àvila, P. Ghinassi, R. Ghosh, S. Ghosh, I. González, K. Grayson, M. Griffith, I. Hadade, C. Haine, C. Hartick, U.-U. Haus, S. Hearne, H. Järvinen, B. Jiménez, A. John, M. Juchem, T. Jung, J. Kegel, M. Kelbling, K. Keller, B. Kinoshita, T. Kiszler, D. Klocke, L. Kluft, N. Koldunov, T. Kölling, J. Kolstela, L. Kornblueh, S. Kosukhin, A. Lacima-Nadolnik, J. J. Leal Rojas, J. Lehtiranta, T. Lunttila, A. Luoma, P. Manninen, A. Medvedev, S. Milinski, A. O. A. Mohammed, S. Müller, D. Naryanappa, N. Nazarova, S. Niemelä, B. Niraula, H. Nortamo, A. Nummelin, M. Nurisso, P. Ortega, S. Paronuzzi, X. Pedruzo Bagazgoitia, C. Pelletier, C. Peña, S. Polade, H. Pradhan, R. Quintanilla, T. Quintino, T. Rackow, J. Räisänen, M. M. Rajput, R. Redler, B. Reuter, N. Rocha Monteiro, F. Roura-Adserias, S. Ruppert, S. Sayed, R. Schnur, T. Sharma, D. Sidorenko, O. Sievi-Korte, A. Soret, C. Steger, B. Stevens, J. Streffing, J. Sunny, L. Tenorio, S. Thober, U. Tigerstedt, O. Tinto, J. Tonttila, H. Tuomenvirta, L. Tuppi, G. Van Thielen, E. Vitali, J. von Hardenberg, I. Wagner, N. Wedi, J. Wehner, S. Willner, X. Yepes-Arbós, F. Ziemen, and J. Zimmermann. The destination earth digital twin for climate change adaptation. *EGUsphere*, 2025:1–41, 2025. doi: 10.5194/egusphere-2025-2198. URL `https://egusphere.copernicus.org/preprints/2025/egusphere-2025-2198/`.

K. M. Gorski, E. Hivon, and B. D. Wandelt. Analysis issues for large CMB data sets. *Arxiv eprints arXiv:astro-ph/9812350*, 1998.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=EXHG-A3jlM`.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36: 21757–21792, 2023.

Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R. Durran, Raul A. Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V. Butz. Advancing Parsimonious Deep Learning Weather Prediction Using the HEALPix Mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, 2024. ISSN 1942-2466. doi: 10.1029/2023MS004021.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. 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, Davos Switzerland, June 2023. ACM. ISBN 9798400701900. doi: 10.1145/3592979.3593412.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and

Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, December 2023. doi: 10.1126/science.adi2336.

Qi Liu and Jianwei Ma. Foundation models for geophysics: Review and perspective. *arXiv preprint arXiv:2406.03163*, 2024.

Yang Liu, Zinan Zheng, Jiashun Cheng, Fugee Tsung, Deli Zhao, Yu Rong, and Jia Li. CirT: Global subseasonal-to-seasonal forecasting with geometry-inspired transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, August 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009. IEEE, 2022.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 25904–25938. PMLR, July 2023a.

Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. *Advances in Neural Information Processing Systems*, 36:75009–75025, 2023b.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68740–68771. Curran Associates, Inc., 2024.

N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier. Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications. *Astronomy and Computing*, 27:130–146, April 2019. ISSN 2213-1337. doi: 10.1016/j.ascom.2019.03.004. URL `http://dx.doi.org/10.1016/j.ascom.2019.03.004`.

PhysicsNeMo Contributors. NVIDIA PhysicsNeMo: An open-source framework for physics-based deep learning in science and engineering, 2 2023. URL `https://github.com/NVIDIA/physicsnemo`. GitHub repository.

Vivek Ramavajjala. HEAL-ViT: Vision Transformers on a spherical mesh for medium-range weather forecasting, February 2024.

Sebastian Scher and Gabriele Messori. Physics-inspired adaptions to low-parameter neural network weather forecast systems. *Artificial Intelligence for the Earth Systems*, 3(1):e230046, 2024. doi: 10.1175/AIES-D-23-0046.1.

John Schreck, Yingkai Sha, William Chapman, Dhamma Kimpara, Judith Berner, Seth McGinnis, Arnold Kazadi, Negin Sobhani, Ben Kirk, and David John Gagne II. Community Research Earth Digital Intelligence Twin (CREDIT), November 2024.

T. Selz and G. C. Craig. Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters*, 50(20):e2023GL105747, 2023. ISSN 1944-8007. doi: 10.1029/2023GL105747.

Mehran Shakerinava and Siamak Ravanbakhsh. Equivariant networks for pixelized spheres. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 9477–9488. PMLR, 2021.

Huug van den Dool, Emily Becker, Li-Chuan Chen, and Qin Zhang. The probability anomaly correlation and calibration of probabilistic forecasts. *Weather and Forecasting*, 32(1):199–206, 2017.

Duncan Watson-Parris. Machine learning for weather and climate are worlds apart. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10): e2021MS002954, 2022.

Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses, November 2024.

Manzhu Yu, Qunying Huang, and Zhenlong Li. Deep learning for spatiotemporal forecasting in earth system science: a review. *International Journal of Digital Earth*, 17(1):2391952, 2024.

Xinfeng Zhao, Hongyan Wang, Mingyu Bai, Yingjie Xu, Shengwen Dong, Hui Rao, and Wuyi Ming. A comprehensive review of methods for hydrological forecasting based on deep learning. *Water*, 16(10):1407, 2024.

Andrea Zonca, Leo P. Singer, Daniel Lenz, Martin Reinecke, Cyrille Rosset, Eric Hivon, and Krzysztof M. Gorski. healpy: Equal area pixelization and spherical harmonics transforms for data on the sphere in Python. *Journal of Open Source Software*, 4(35):1298, 2019. doi: 10.21105/joss.01298.

A TRAINING, CHECKPOINT SELECTION, AND INFERENCE COST

A.1 DATA AND TRAINING OBJECTIVE

The weather-forecasting experiments use ERA5-Lite (Bi et al., 2023), with daily samples at midnight UTC from 2007–2017 for training, 2019 for validation, and 2018 for testing. All models are trained for a 24-hour prediction step and applied iteratively for forecasts from one to nine days. The main comparison uses five random seeds per architecture.

Standardization. Inputs and targets are standardized separately for each of the four surface variables and each of the 65 upper-air variable–pressure-level combinations. For field v, we use $z_v = (y_v - \mu_v)/\sigma_v$, with means and standard deviations computed from the 2007–2017 HEALPix training data. The same statistics are used for the HEALPix and equiangular models.

Training loss. The weather models minimize an L1 loss on standardized fields. Let $\mathcal{U}$ and $\mathcal{S}$ denote the 65 upper-air fields and four surface fields, and let w_i be spatial weights with $\sum_i w_i = 1$. For a training example, the objective is

$$\mathcal{L}(\widehat{z}, z) = \underbrace{\frac{1}{65}\sum_{v\in\mathcal{U}}\sum_i w_i|\widehat{z}_{v,i} - z_{v,i}|}_{\mathcal{L}_{\text{upper}}} + \underbrace{\frac{1}{4}\frac{1}{4}\sum_{v\in\mathcal{S}}\sum_i w_i|\widehat{z}_{v,i} - z_{v,i}|}_{\mathcal{L}_{\text{surface}}}. \tag{A.1}$$

Upper-air and surface errors are averaged separately before applying the surface coefficient $1/4$. The loss is then averaged over the batch. PEAR, PEAR-HD8, uniform-loss Pangu, GraphCast-I, GraphCast-P, and FengWu use uniform spatial weights, following the Pangu-Weather loss (Bi et al., 2023). AFNO and the cosine-weighted Pangu control use weights proportional to $\cos\phi_i$, normalized over the equiangular grid. Uniform weights on HEALPix give equal weight per unit area.

All models use AdamW with a constant learning rate of 5×10^{-4} and weight decay 3×10^{-6}. Our training implementation uses float32 and one optimizer update per batch.

A.2 VALIDATION AND CHECKPOINT SELECTION

For checkpoint selection, predictions and targets are converted back to physical units and evaluated on the native output grid of each model. Spatial RMSE uses equal-area weights on HEALPix and normalized cosine-latitude weights on the equiangular grid.

For each training seed, we select the saved candidate with the smallest 2019 validation score

$$Q(e) = \frac{1}{69\cdot 9}\sum_{v=1}^{69}\sum_{k=1}^{9}\frac{1}{N_k^{\text{val}}}\sum_{n=1}^{N_k^{\text{val}}}\left[\sum_i w_i\left(\widehat{y}_{e,v,k,n,i} - y_{v,k,n,i}\right)^2\right]^{1/2}, \tag{A.2}$$

where e indexes checkpoint epoch, k is forecast lead day, and n is forecast start date. The score gives equal weight to the 69 fields and nine leads.

Table A.1 reports the training budgets and selected checkpoints.

Table A.1: Completed weather-model training budgets and validation-selected epochs. Batch size is global across the listed GPUs, with batch size one per GPU except for FengWu, which uses eight on one GPU. Each row has five seeds. PEAR-HD8 and cosine-weighted Pangu are additional controls.

Model	Epochs	Global batch	GPUs	Selected epochs (seeds 0–4)
PEAR	1,000	1	1	590, 770, 950, 720, 960
GraphCast-I	1,000	1	1	930, 770, 980, 980, 670
GraphCast-P	500	1	1	480, 290, 400, 370, 370
FengWu	500	8	1	320, 400, 400, 470, 480
Pangu	300	1	1	170, 270, 250, 270, 260
AFNO	400	2	2	180, 180, 230, 220, 230
Pangu (cosine-weighted)	300	1	1	300, 300, 200, 200, 300
PEAR-HD8	1,000	1	1	910, 970, 830, 660, 750

Figure A.1 shows training loss, checkpoint-selection RMSE, and normalized validation RMSE. We calculate normalized RMSE by dividing the validation RMSE of each field by its training-data standard deviation, then averaging over the 69 fields and nine leads. The PEAR-HD8 training-loss curve uses four seeds through epoch 200 and five thereafter.

A.3 Model configurations

PEAR and PEAR-HD8. The main PEAR model uses $n_{\text{side}} = 64$, stage widths $[72, 144, 144, 72]$, attention-head counts $[2, 4, 4, 2]$, and stage depths $[2, 6, 6, 2]$, giving head dimension 36 and 4,316,201 parameters. PEAR-HD8 uses widths $[48, 96, 96, 48]$, head counts $[6, 12, 12, 6]$, the same depths, and 4,279,881 parameters. Appendix B gives the layer structure and ablations.

Pangu and FengWu. We use the PhysicsNeMo implementations in our training framework (PhysicsNeMo Contributors, 2023). Pangu has 11.386M parameters and the same block depths, attention-head counts, and embedding dimensions as PEAR-HD8, with the Pangu-Weather positional-embedding structure. FengWu (Chen et al., 2023a) has 16.357M parameters. Both architectures use the 157×314 equiangular grid.

AFNO. The main AFNO model uses eight layers, embedding dimension 128, eight Fourier channel blocks, patch size $(1, 2)$, and 4,317,952 parameters on the 157×314 equiangular grid (Guibas et al., 2022; Kurth et al., 2023). We select its architecture to have inference time within 5% of PEAR-HD8. The paired benchmark measures 14.96 ± 0.03 ms for AFNO and 15.36 ± 0.19 ms for PEAR-HD8. The parameter-matched AFNO-P control uses four layers, embedding dimension 144, patch size $(1, 2)$, and 4,301,856 parameters. Appendix B.4 reports its results.

GraphCast. The published GraphCast configuration uses 512 hidden features, 16 processor layers, and a level-6 icosahedral multimesh (Lam et al., 2023). An icosahedral mesh at refinement level r has $10 \cdot 4^r + 2$ nodes. Levels 4, 5, and 6 contain 2,562, 10,242, and 40,962 nodes, respectively. Both GraphCast configurations here use level 4, retaining coarser-level edges for multiscale message passing. This gives a similar number of horizontal locations to the 3,072 PEAR locations after patch embedding, before downsampling to 768. Level 5 has approximately 3.3 times as many horizontal locations as the finer PEAR representation. PEAR additionally retains eight vertical token positions, whereas GraphCast places atmospheric variables and levels in node features.

Using the PhysicsNeMo implementations, we sweep ten width–depth configurations with hidden dimensions 128 and 144 and four to seven processor layers, and dimension 160 with four or five layers, holding the mesh and equiangular input and output resolution fixed. GraphCast-I has 144 hidden features, seven processor layers, and 1,508,325 parameters. We select it by AdamW-step time within 5% of PEAR-HD8 on an A100-40GB at batch size one, with a measured step time 2.74% lower. GraphCast-P has 200 hidden features, 12 processor layers, and 4,298,069 parameters, within 1% of either PEAR configuration.

A.4 Inference benchmarking

Table A.2 reports seed-0 inference measurements on an A100-40GB using float32, batch size one, and inputs already stored on the GPU. We perform 20 warm-up passes followed by 100 synchronized forward passes. Times are means and standard deviations across the measured passes. Peak memory is the process-wide PyTorch allocator peak, covering weights, the on-device batch, outputs, and temporary tensors created during the forward pass.

A paired benchmark measures 13.60 ± 0.09 ms for PEAR and 15.56 ± 0.39 ms for PEAR-HD8. GraphCast-I and AFNO have approximately 15% and 10% more measured inference time than PEAR. GraphCast-P uses approximately $2.07\times$ the inference time and $3.26\times$ the peak allocated memory of PEAR. GraphCast-I uses $2.42\times$ the memory.

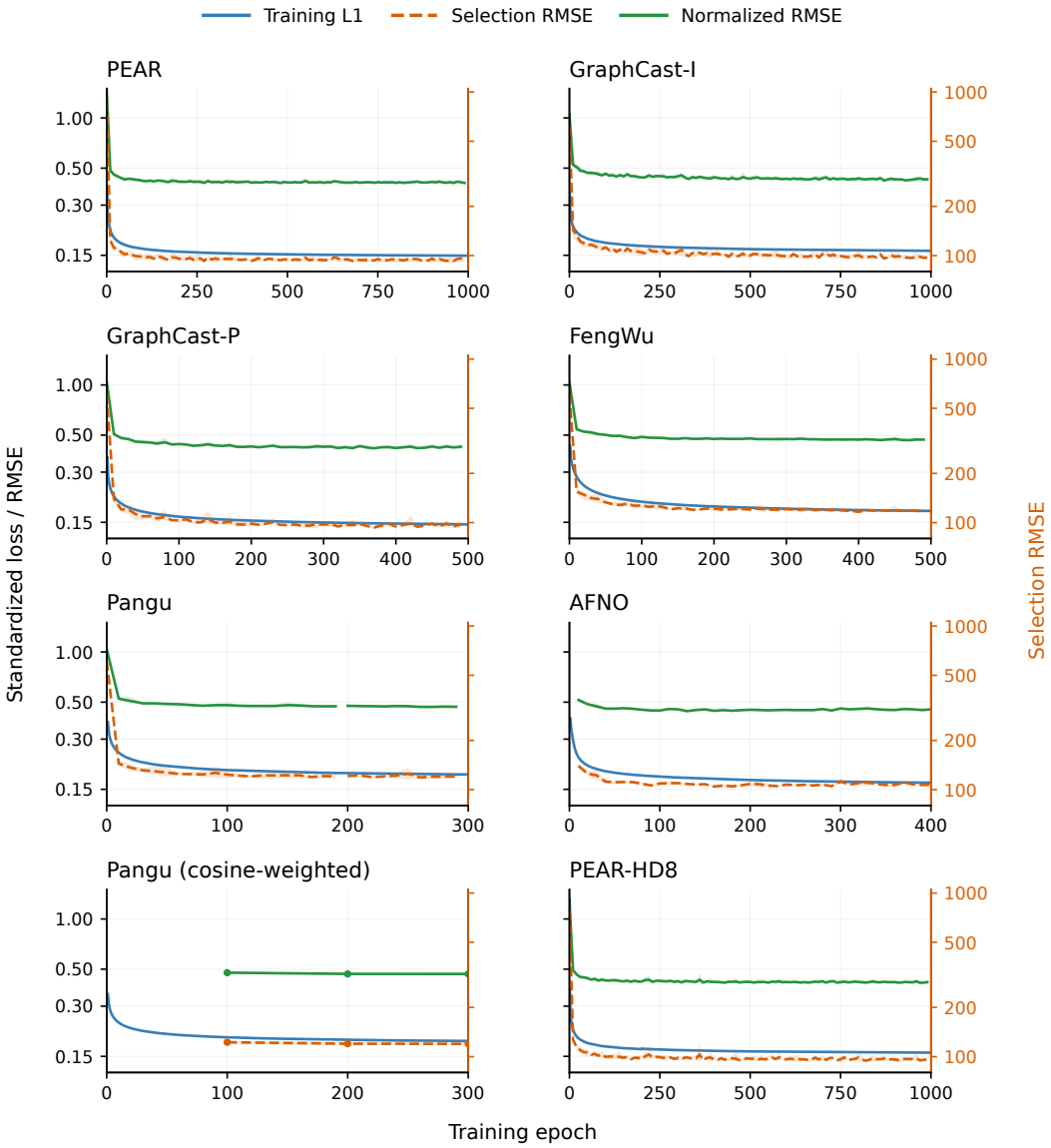


Figure A.1: Training and validation histories. Lines show means over available seeds and bands show min–max ranges. The left axis shows standardized L1 training loss (Equation A.1) and validation RMSE divided by the training standard deviation of each field. The right axis shows the physical-unit selection score (Equation A.2).

Table A.2: Model size, one-step inference time, and peak allocated GPU memory. Measurements follow Appendix A.4. PEAR and AFNO are each paired with PEAR-HD8 in separate benchmark sessions. Ratios to other models are approximate across sessions.

Model	Parameters (M)	Inference time (ms)	Peak memory (GiB)
PEAR	4.316	13.60 ± 0.09	0.253
Pangu	11.386	25.17 ± 0.11	0.420
GraphCast-I	1.508	15.71 ± 0.06	0.610
GraphCast-P	4.298	28.19 ± 0.10	0.823
FengWu	16.357	69.18 ± 1.95	0.286
AFNO	4.318	14.96 ± 0.03	0.205

B Model architecture and ablations

B.1 Layer structure, windows, and masking

The main PEAR configuration is summarized in Table B.1. Patch embedding groups 16 nested HEALPix pixels and two neighboring upper-air levels, with a separate surface embedding. Their concatenation gives eight vertical token positions. Attention windows contain 64 horizontal tokens and two vertical positions. Figures B.1 and B.2 illustrate the spatial grouping and the mask used to prevent cyclic shifts from joining physically separated regions.

Table B.1: Main PEAR layer structure. Tensor shapes refer to layer outputs. The two central stages together contain 12 attention layers. Surface and upper-air tensors are separate at input and output.

Layer/block	Depth	Output tensor shape	Heads
Input	1	$(12n_{\text{side}}^2, 4)$ and $(12n_{\text{side}}^2, 13, 5)$	–
Patch embedding	1	$(\frac{3}{4}n_{\text{side}}^2, 8, 72)$	–
Windowed attention	2	$(\frac{3}{4}n_{\text{side}}^2, 8, 72)$	2
Downsampling	1	$(\frac{3}{16}n_{\text{side}}^2, 8, 144)$	–
Windowed attention	12	$(\frac{3}{16}n_{\text{side}}^2, 8, 144)$	4
Upsampling	1	$(\frac{3}{4}n_{\text{side}}^2, 8, 72)$	–
Windowed attention	2	$(\frac{3}{4}n_{\text{side}}^2, 8, 72)$	2
Skip concatenation	–	$(\frac{3}{4}n_{\text{side}}^2, 8, 144)$	–
Patch recovery	1	$(12n_{\text{side}}^2, 4)$ and $(12n_{\text{side}}^2, 13, 5)$	–

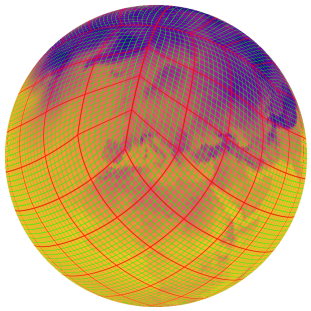


Figure B.1: Nested HEALPix grouping used to form spatial patches (green) and attention windows (red). The illustration shows the horizontal arrangement. Attention windows also span two neighboring vertical token positions. The main configuration uses 16 input pixels per patch and 64 horizontal tokens per window.

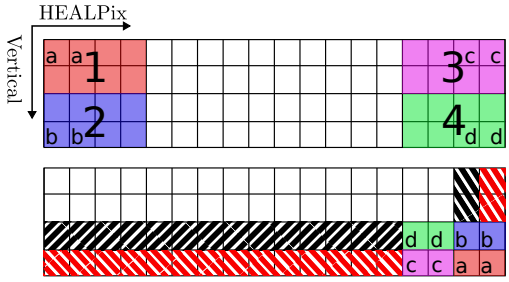


Figure B.2: Shift and mask for windowed attention. The horizontal axis is the one-dimensional HEALPix ring index, from north to south, and the vertical axis indexes atmospheric levels. The lower grid shows the shifted positions of the upper-grid regions. The mask prevents attention between physically separated regions brought together by cyclic wrapping. Corner windows require four regions and edge windows two.

B.2 PEAR AND PEAR-HD8

Using five validation-selected seeds and the evaluation protocol in Appendix C, PEAR has higher mean ACC than PEAR-HD8 in all 621 variable–lead comparisons and lower mean RMSE in 619. Table B.2 reports the relative RMSE comparison. Width and head count change together, as specified in Appendix A.3.

Table B.2: Relative RMSE changes (%) for PEAR-HD8 against PEAR, using the evaluation protocol in Appendix C. Entries average equally over 69 fields and all 25 seed combinations. Positive values indicate higher error for PEAR-HD8. Sub- and superscripts give offsets to the min–max range.

Reference grid	1 day	5 days	9 days
Native	$4.9^{+1.3}_{-2.8}$	$2.9^{+1.4}_{-2.8}$	$1.0^{+2.1}_{-2.6}$
Common DH	$4.3^{+1.1}_{-2.4}$	$2.8^{+1.4}_{-2.8}$	$1.0^{+2.1}_{-2.6}$

B.3 POSITIONAL-BIAS SHARING IN PANGU

Pangu shares positional biases across longitude windows but retains separate positional-bias tensors for different latitude–pressure window types. Here we ablate a version of Pangu with shared positional biases across all window types within each attention head and layer, keeping the within-window positional indices, widths, depths, head counts, and shifts unchanged.

Both Pangu controls use cosine-weighted L1 loss, five seeds, and 300 training epochs. The shared-position control selects checkpoints using the validation criterion in Appendix A.2. Table B.3 shows that sharing slightly increases one-day error and modestly reduces longer-horizon error, but does not close the gap to PEAR on either evaluation grid. Under native-grid evaluation, PEAR has higher mean ACC in all 621 variable–lead comparisons and lower mean RMSE in 616.

Table B.3: Relative RMSE reductions (%) on 2018 test data for the positional-bias sharing control. Positive values favor the first model named in each block. Entries average equally over 69 fields and all 25 seed combinations. Sub- and superscripts give offsets to the min–max range. Both Pangu controls use cosine-weighted loss. Common equiangular comparisons use four-neighbor interpolation or nearest-cell mapping for PEAR, following Appendix C.

Evaluation grid	1 day	5 days	9 days
Shared-position Pangu versus original Pangu			
Native equiangular	$-1.7^{+1.7}_{-3.5}$	$+0.5^{+3.7}_{-4.7}$	$+1.6^{+5.5}_{-6.3}$
Common HEALPix	$-1.8^{+1.7}_{-3.5}$	$+0.5^{+3.7}_{-4.6}$	$+1.5^{+5.5}_{-6.2}$
PEAR versus shared-position Pangu			
Native grids	$+18.4^{+2.8}_{-2.7}$	$+18.6^{+3.4}_{-3.0}$	$+6.4^{+5.4}_{-4.2}$
Common HEALPix	$+17.9^{+2.7}_{-2.7}$	$+18.4^{+3.4}_{-3.0}$	$+6.2^{+5.4}_{-4.2}$
Common equiangular, four-neighbor	$+18.7^{+2.7}_{-2.6}$	$+18.9^{+3.4}_{-3.0}$	$+6.7^{+5.4}_{-4.1}$
Common equiangular, nearest	$+14.2^{+2.9}_{-2.6}$	$+18.2^{+3.4}_{-3.0}$	$+6.3^{+5.5}_{-4.2}$

B.4 ARCHITECTURAL ABLATIONS AND PARAMETER-MATCHED AFNO

The architectural ablations and AFNO-P use PEAR-HD8 as their reference. They use a separate common-equiangular remapping protocol and an ACC climatology shared across forecast leads for each start date. Comparisons are therefore within these tables. Appendix C describes the valid-date climatology and remapping used for the main comparison.

No shift and stochastic depth preserve the 4,279,881 parameters of PEAR-HD8. Removing relative-position bias gives 1,527,369 parameters, a 64.3% reduction, while vertical patch size one gives

4,268,361. No shift, no relative bias, and vertical patch size one use 200-epoch budgets. Stochastic depth is compared at epoch 200. Vertical patch size one has three evaluated seeds, and the other controls have five. Each row compares the ablation with its corresponding reference checkpoint under these budgets.

Table B.4: Relative RMSE changes (%) against PEAR-HD8 for architectural controls on the common equiangular grid. Entries average equally over 69 fields and all seed combinations. Positive values indicate higher error. Sub- and superscripts give offsets to the min–max range. Reference checkpoints and budgets can differ between rows.

HD8 variant	Params. (M)	Seeds	1 day	5 days	9 days
No shift	4.280	5	$+53.2^{+3.3}_{-3.0}$	$+53.9^{+2.9}_{-5.2}$	$+15.7^{+5.4}_{-4.3}$
No relative bias	1.527	5	$+6.8^{+1.8}_{-1.1}$	$+7.6^{+6.0}_{-4.8}$	$+4.4^{+11.8}_{-7.0}$
Vertical patch 1	4.268	3	$+1.2^{+2.2}_{-2.0}$	$+4.4^{+3.3}_{-1.8}$	$+4.5^{+3.3}_{-5.0}$
Stochastic depth 0.2	4.280	5	$+3.8^{+1.4}_{-1.7}$	$+5.3^{+4.9}_{-5.7}$	$+3.7^{+7.2}_{-6.5}$

Removing shifts increases mean RMSE in all 621 variable–lead comparisons, with non-overlapping seed ranges in 613 and a median increase of 40.7%. Removing relative-position bias increases RMSE in 620 comparisons, stochastic depth 0.2 in 616, and vertical patch size one in 588. Their median increases are 6.3%, 3.4%, and 3.6%, respectively.

Table B.5: RMSE on 2018 test data at days 1, 5, and 9, using validation-selected checkpoints and the control-specific common-equiangular protocol in Appendix B.4. The PEAR (HD36) column also uses this protocol. Vertical patch size one has three seeds and the other models five. Lower is better. Sub- and superscripts show seed-range offsets. Bold entries have the best mean or a range overlapping that of the best mean.

Variable	Lead	PEAR-HD8	No shift	No rel. bias	PEAR (HD36)	Vertical 1
t2m (K)	1	$1.52^{+0.00}_{-0.00}$	$1.75^{+0.01}_{-0.01}$	$1.63^{+0.03}_{-0.02}$	$\mathbf{1.48}^{+0.01}_{-0.01}$	$1.54^{+0.01}_{-0.01}$
	5	$\mathbf{2.26}^{+0.02}_{-0.03}$	$3.23^{+0.13}_{-0.07}$	$2.58^{+0.20}_{-0.10}$	$\mathbf{2.21}^{+0.04}_{-0.03}$	$2.35^{+0.05}_{-0.03}$
	9	$\mathbf{3.06}^{+0.05}_{-0.05}$	$3.92^{+0.30}_{-0.16}$	$3.47^{+0.44}_{-0.18}$	$\mathbf{3.03}^{+0.07}_{-0.05}$	$\mathbf{3.20}^{+0.14}_{-0.11}$
msl (Pa)	1	$\mathbf{139}^{+1}_{-2}$	332^{+5}_{-3}	158^{+6}_{-5}	$\mathbf{134}^{+7}_{-3}$	145^{+4}_{-3}
	5	$\mathbf{462}^{+3}_{-6}$	853^{+20}_{-17}	520^{+31}_{-17}	$\mathbf{448}^{+16}_{-10}$	485^{+12}_{-17}
	9	$\mathbf{755}^{+5}_{-7}$	938^{+30}_{-22}	794^{+35}_{-25}	$\mathbf{749}^{+15}_{-8}$	788^{+21}_{-12}
z500 (gpm)	1	$\mathbf{12.3}^{+0.2}_{-0.1}$	$27.6^{+0.5}_{-0.5}$	$14.3^{+0.8}_{-0.5}$	$\mathbf{11.8}^{+0.4}_{-0.1}$	$\mathbf{12.6}^{+0.4}_{-0.5}$
	5	$47.4^{+0.4}_{-0.4}$	$90.5^{+0.9}_{-0.5}$	$54.0^{+4.5}_{-2.9}$	$\mathbf{45.6}^{+0.9}_{-0.4}$	$49.9^{+1.6}_{-1.4}$
	9	$\mathbf{82.7}^{+0.7}_{-0.7}$	$105.0^{+2.3}_{-1.8}$	$\mathbf{87.8}^{+6.8}_{-6.3}$	$\mathbf{81.6}^{+0.9}_{-0.7}$	$86.9^{+2.7}_{-2.8}$
t850 (K)	1	$1.30^{+0.01}_{-0.01}$	$1.67^{+0.03}_{-0.03}$	$1.44^{+0.03}_{-0.02}$	$\mathbf{1.23}^{+0.01}_{-0.01}$	$1.31^{+0.02}_{-0.01}$
	5	$2.50^{+0.03}_{-0.02}$	$3.89^{+0.06}_{-0.06}$	$2.81^{+0.20}_{-0.12}$	$\mathbf{2.42}^{+0.03}_{-0.02}$	$2.57^{+0.05}_{-0.04}$
	9	$\mathbf{3.76}^{+0.06}_{-0.03}$	$4.50^{+0.08}_{-0.09}$	$4.08^{+0.41}_{-0.29}$	$\mathbf{3.72}^{+0.04}_{-0.04}$	$3.89^{+0.09}_{-0.11}$
q700 (10^{-4} kg/kg)	1	$9.35^{+0.03}_{-0.02}$	$10.71^{+0.17}_{-0.14}$	$10.00^{+0.14}_{-0.13}$	$\mathbf{8.96}^{+0.14}_{-0.06}$	$9.16^{+0.07}_{-0.04}$
	5	$\mathbf{14.84}^{+0.22}_{-0.11}$	$18.76^{+0.14}_{-0.17}$	$16.29^{+0.82}_{-0.41}$	$\mathbf{14.65}^{+0.29}_{-0.34}$	$15.45^{+0.17}_{-0.21}$
	9	$\mathbf{18.94}^{+0.38}_{-0.26}$	$20.92^{+0.55}_{-0.41}$	$20.33^{+1.17}_{-0.58}$	$\mathbf{18.91}^{+0.30}_{-0.45}$	$\mathbf{19.80}^{+0.69}_{-0.66}$

Table B.6: ACC for the checkpoints and evaluation grid in Table B.5, using the control-specific climatology in Appendix B.4. Comparisons are within this protocol. Higher is better. Seed coverage, ranges, and bolding follow Table B.5.

Variable	Lead	PEAR-HD8	No shift	No rel. bias	PEAR (HD36)	Vertical 1
t2m	1	$0.841^{+0.001}_{-0.001}$	$0.791^{+0.004}_{-0.003}$	$0.822^{+0.005}_{-0.006}$	$\mathbf{0.852}^{+0.001}_{-0.002}$	$0.838^{+0.003}_{-0.004}$
	5	$0.650^{+0.010}_{-0.008}$	$0.329^{+0.008}_{-0.010}$	$0.576^{+0.019}_{-0.030}$	$\mathbf{0.674}^{+0.005}_{-0.007}$	$0.637^{+0.006}_{-0.005}$
	9	$0.402^{+0.013}_{-0.012}$	$0.177^{+0.017}_{-0.014}$	$0.324^{+0.040}_{-0.048}$	$\mathbf{0.425}^{+0.006}_{-0.008}$	$0.384^{+0.016}_{-0.009}$
msl	1	$\mathbf{0.982}^{+0.000}_{-0.000}$	$0.896^{+0.001}_{-0.002}$	$0.977^{+0.001}_{-0.002}$	$\mathbf{0.984}^{+0.001}_{-0.002}$	$0.981^{+0.001}_{-0.001}$
	5	$\mathbf{0.792}^{+0.006}_{-0.003}$	$0.250^{+0.008}_{-0.012}$	$0.739^{+0.015}_{-0.026}$	$\mathbf{0.806}^{+0.009}_{-0.014}$	$\mathbf{0.777}^{+0.014}_{-0.008}$
	9	$\mathbf{0.442}^{+0.007}_{-0.006}$	$0.125^{+0.002}_{-0.002}$	$0.387^{+0.024}_{-0.029}$	$\mathbf{0.456}^{+0.010}_{-0.012}$	$0.419^{+0.014}_{-0.011}$
z500	1	$\mathbf{0.990}^{+0.000}_{-0.000}$	$0.947^{+0.002}_{-0.002}$	$0.986^{+0.001}_{-0.001}$	$\mathbf{0.990}^{+0.000}_{-0.001}$	$\mathbf{0.989}^{+0.001}_{-0.001}$
	5	$0.839^{+0.003}_{-0.004}$	$0.365^{+0.007}_{-0.008}$	$0.791^{+0.017}_{-0.028}$	$\mathbf{0.852}^{+0.002}_{-0.007}$	$0.826^{+0.011}_{-0.008}$
	9	$0.508^{+0.006}_{-0.004}$	$0.181^{+0.008}_{-0.014}$	$0.450^{+0.040}_{-0.046}$	$\mathbf{0.526}^{+0.005}_{-0.008}$	$0.484^{+0.009}_{-0.017}$
t850	1	$0.932^{+0.001}_{-0.001}$	$0.885^{+0.004}_{-0.003}$	$0.917^{+0.003}_{-0.003}$	$\mathbf{0.939}^{+0.001}_{-0.002}$	$0.931^{+0.001}_{-0.002}$
	5	$0.748^{+0.004}_{-0.005}$	$0.347^{+0.008}_{-0.007}$	$0.682^{+0.021}_{-0.026}$	$\mathbf{0.766}^{+0.004}_{-0.006}$	$0.738^{+0.009}_{-0.007}$
	9	$0.438^{+0.005}_{-0.009}$	$0.181^{+0.011}_{-0.009}$	$0.365^{+0.041}_{-0.046}$	$\mathbf{0.457}^{+0.005}_{-0.008}$	$0.420^{+0.009}_{-0.015}$
q700	1	$0.840^{+0.001}_{-0.001}$	$0.783^{+0.007}_{-0.007}$	$0.815^{+0.005}_{-0.005}$	$\mathbf{0.854}^{+0.002}_{-0.005}$	$0.847^{+0.002}_{-0.003}$
	5	$\mathbf{0.598}^{+0.004}_{-0.005}$	$0.310^{+0.009}_{-0.012}$	$0.524^{+0.016}_{-0.026}$	$\mathbf{0.612}^{+0.009}_{-0.012}$	$0.576^{+0.005}_{-0.004}$
	9	$\mathbf{0.358}^{+0.004}_{-0.004}$	$0.177^{+0.016}_{-0.007}$	$0.291^{+0.018}_{-0.029}$	$\mathbf{0.366}^{+0.013}_{-0.012}$	$0.328^{+0.003}_{-0.006}$

Table B.7: Parameter-matched AFNO (AFNO-P) against PEAR-HD8, using the control-specific ACC and remapping protocols in Appendix B.4. Common DH uses four-neighbor interpolation or nearest-cell sampling for HEALPix predictions. Positive ΔACC and RMSE reduction favor PEAR-HD8. Entries average over 69 fields and all 25 seed combinations. Sub- and superscripts give offsets to their min–max range.

Evaluation	Control	Lead	ΔACC	PEAR-HD8 RMSE reduction (%)
Native	AFNO-P	1	$+0.021^{+0.001}_{-0.001}$	$+15.2^{+1.4}_{-1.1}$
		5	$+0.084^{+0.009}_{-0.006}$	$+12.8^{+1.8}_{-0.9}$
		9	$+0.076^{+0.014}_{-0.011}$	$+5.4^{+1.6}_{-1.7}$
Common HP	AFNO-P	1	$+0.024^{+0.002}_{-0.002}$	$+16.7^{+1.3}_{-1.0}$
		5	$+0.084^{+0.009}_{-0.005}$	$+12.7^{+1.9}_{-0.8}$
		9	$+0.074^{+0.013}_{-0.011}$	$+5.2^{+1.6}_{-1.7}$
Common DH, four-neighbor	AFNO-P	1	$+0.020^{+0.001}_{-0.001}$	$+13.8^{+1.4}_{-1.1}$
		5	$+0.084^{+0.009}_{-0.006}$	$+12.8^{+1.8}_{-1.0}$
		9	$+0.075^{+0.013}_{-0.012}$	$+5.5^{+1.7}_{-1.7}$
Common DH, nearest	AFNO-P	1	$+0.016^{+0.002}_{-0.001}$	$+9.6^{+1.5}_{-1.1}$
		5	$+0.080^{+0.009}_{-0.006}$	$+12.1^{+1.8}_{-1.0}$
		9	$+0.073^{+0.013}_{-0.012}$	$+5.1^{+1.7}_{-1.7}$

C FORECAST METRICS AND GRID COMPARISONS

C.1 METRICS AND AGGREGATION

For field v, forecast lead k, and start date n, spatial RMSE is

$$\mathrm{RMSE}_{v,k,n} = \left[\sum_i w_i(\widehat{y}_{v,k,n,i} - y_{v,k,n,i})^2\right]^{1/2}. \tag{C.1}$$

We compute RMSE in physical units and average over forecast start dates. On HEALPix, $w_i = (12n_{\mathrm{side}}^2)^{-1}$. On the equiangular grid, w_i is proportional to cosine latitude and sums to one.

For a fixed field and lead, the anomaly correlation coefficient is

$$\mathrm{ACC} = \frac{\sum_{n,i} w_i \Delta y_{n,i} \Delta \widehat{y}_{n,i}}{\sqrt{\left(\sum_{n,i} w_i \Delta y_{n,i}^2\right)\left(\sum_{n,i} w_i \Delta \widehat{y}_{n,i}^2\right)}}. \tag{C.2}$$

Both anomalies subtract the same climatological mean for the forecast valid calendar day. For each field and grid point, this mean is computed from 2007–2017, aligning month and day separately in each year with explicit leap-year handling. Spatial cross-products and squared anomalies are summed over start dates before forming the ratio.

The test year has $365 - k$ forecast start dates at lead day k, giving 364, 360, and 356 at days 1, 5, and 9. Tables report seed means and min–max ranges. Upper-variable summaries average over 13 pressure levels, while surface rows are individual variables. Overall ACC averages give equal weight to all 69 variable–level combinations. Bolding marks the best mean and models whose seed ranges overlap its range.

For relative RMSE comparisons, we first calculate a ratio for each variable and pair of seeds, then average equally over variables. In Tables C.1, B.2, and B.4, excess RMSE is $100(\mathrm{RMSE}_{\mathrm{control}}/\mathrm{RMSE}_{\mathrm{reference}} - 1)$. Positive values favor the reference. Five-seed pairs give 25 combinations, and three-seed controls give 15. Regional and spectral changes use the named baseline as denominator, as specified in their captions.

C.2 NATIVE-GRID AND COMMON-GRID EVALUATION

Native-grid evaluation scores PEAR on equal-area HEALPix and the baselines on their cosine-weighted equiangular grid. Common-grid comparisons evaluate all models against targets on the same grid. For common HEALPix, equiangular forecasts are bilinearly interpolated using stored latitude and longitude coordinates and periodic longitude boundaries. For common equiangular evaluation, HEALPix predictions use either four-neighbor interpolation or nearest-cell sampling. These comparisons assess sensitivity to remapping as well as forecasting differences.

Tables C.2 and C.3 report the six-model native-grid results. Tables C.4 and C.5 use common HEALPix, while Tables C.6 and C.7 use the common equiangular grid with nearest-cell sampling. On the latter grid, PEAR has the highest mean ACC in 451 of 621 variable–lead comparisons and the lowest mean RMSE in 444. GraphCast-P leads the remaining comparisons. Some seed ranges overlap.

C.3 COSINE-WEIGHTED PANGU CONTROL

The Pangu control changes the spatial loss weights to normalized cosine latitude while keeping the architecture and training budget fixed. Relative to uniform-loss Pangu, weighting improves ACC and RMSE for 66 of 69 fields at day 1, and ACC for 67 and RMSE for 64 at day 2. At day 9, it improves ACC for 17 fields and RMSE for one.

PEAR exceeds weighted Pangu in all 621 variable–lead comparisons for both metrics under native-grid evaluation, on common HEALPix, and with four-neighbor interpolation to the equiangular grid. It leads in 620 comparisons with nearest-cell sampling. Against uniform-loss Pangu under native-grid evaluation, the mean ACC advantages at days 1, 5, and 9 are 0.026, 0.139, and 0.122. The corresponding excess RMSE values of Pangu relative to PEAR are 24.5%, 24.1%, and 6.7%. These comparisons include differences in grid, architecture, and training budget.

Table C.1: Weighted and unweighted Pangu against PEAR on 2018 test data. Positive ΔACC and excess RMSE favor PEAR. Excess RMSE is $100(\mathrm{RMSE}_{\mathrm{Pangu}}/\mathrm{RMSE}_{\mathrm{PEAR}} - 1)$. Native uses each model's output grid. Common HP and common DH use a shared HEALPix or equiangular evaluation grid. Entries average over 69 fields and all 25 seed combinations. Sub- and superscripts give offsets to their min–max range.

Evaluation	Control	Lead	ΔACC	Baseline excess RMSE (%)
Native	Pangu	1	$0.026^{+0.002}_{-0.003}$	$24.5^{+2.4}_{-3.8}$
		5	$0.139^{+0.008}_{-0.010}$	$24.1^{+2.7}_{-3.0}$
		9	$0.122^{+0.011}_{-0.013}$	$6.7^{+2.2}_{-2.9}$
	Pangu weighted	1	$0.021^{+0.002}_{-0.003}$	$21.5^{+1.7}_{-3.7}$
		5	$0.134^{+0.011}_{-0.012}$	$24.6^{+3.5}_{-3.8}$
		9	$0.127^{+0.022}_{-0.015}$	$9.1^{+4.7}_{-3.3}$
Common HP	Pangu	1	$0.025^{+0.002}_{-0.003}$	$23.6^{+2.3}_{-3.8}$
		5	$0.137^{+0.008}_{-0.010}$	$23.7^{+2.7}_{-3.0}$
		9	$0.121^{+0.011}_{-0.013}$	$6.4^{+2.2}_{-2.9}$
	Pangu weighted	1	$0.021^{+0.002}_{-0.003}$	$20.7^{+1.7}_{-3.7}$
		5	$0.133^{+0.011}_{-0.012}$	$24.3^{+3.5}_{-3.7}$
		9	$0.126^{+0.022}_{-0.015}$	$8.8^{+4.7}_{-3.3}$
Common DH (interp.)	Pangu	1	$0.027^{+0.002}_{-0.003}$	$25.2^{+2.4}_{-3.7}$
		5	$0.140^{+0.008}_{-0.009}$	$24.5^{+2.7}_{-3.0}$
		9	$0.123^{+0.011}_{-0.013}$	$7.0^{+2.2}_{-2.9}$
	Pangu weighted	1	$0.022^{+0.002}_{-0.003}$	$22.2^{+1.7}_{-3.5}$
		5	$0.135^{+0.011}_{-0.012}$	$25.1^{+3.5}_{-3.7}$
		9	$0.128^{+0.022}_{-0.015}$	$9.4^{+4.7}_{-3.3}$
Common DH (nearest)	Pangu	1	$0.023^{+0.002}_{-0.003}$	$17.8^{+2.1}_{-3.2}$
		5	$0.136^{+0.008}_{-0.010}$	$23.4^{+2.7}_{-3.0}$
		9	$0.121^{+0.011}_{-0.013}$	$6.5^{+2.2}_{-2.9}$
	Pangu weighted	1	$0.017^{+0.002}_{-0.003}$	$15.0^{+1.5}_{-3.1}$
		5	$0.132^{+0.011}_{-0.012}$	$23.9^{+3.5}_{-3.7}$
		9	$0.126^{+0.022}_{-0.015}$	$8.9^{+4.7}_{-3.3}$

C.4 DETAILED FORECAST RESULTS

Table C.2: Native-grid ACC at days 1, 5, and 9 on 2018 test data. PEAR uses equal-area HEALPix evaluation, and the other models use cosine-latitude-weighted equiangular evaluation. Upper variables are averaged over 13 levels. Higher is better. Sub- and superscripts show five-seed range offsets. Bold entries have the best mean or an overlapping range.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl	1	$0.977^{+0.001}_{-0.001}$	$0.988^{+0.000}_{-0.000}$	$\mathbf{0.989^{+0.000}_{-0.000}}$	$0.987^{+0.001}_{-0.001}$	$0.977^{+0.001}_{-0.001}$	$0.986^{+0.000}_{-0.000}$
	5	$0.672^{+0.005}_{-0.007}$	$0.800^{+0.004}_{-0.007}$	$\mathbf{0.812^{+0.006}_{-0.004}}$	$\mathbf{0.810^{+0.009}_{-0.013}}$	$0.665^{+0.003}_{-0.004}$	$0.753^{+0.003}_{-0.005}$
	9	$0.336^{+0.007}_{-0.013}$	$0.434^{+0.010}_{-0.011}$	$\mathbf{0.451^{+0.007}_{-0.006}}$	$\mathbf{0.459^{+0.010}_{-0.013}}$	$0.330^{+0.004}_{-0.003}$	$0.399^{+0.007}_{-0.004}$
t2m	1	$0.892^{+0.001}_{-0.002}$	$0.918^{+0.001}_{-0.001}$	$\mathbf{0.923^{+0.002}_{-0.002}}$	$0.911^{+0.001}_{-0.002}$	$0.880^{+0.003}_{-0.004}$	$0.914^{+0.000}_{-0.001}$
	5	$0.587^{+0.010}_{-0.015}$	$0.680^{+0.011}_{-0.010}$	$\mathbf{0.687^{+0.009}_{-0.016}}$	$\mathbf{0.703^{+0.006}_{-0.008}}$	$0.567^{+0.009}_{-0.008}$	$0.658^{+0.005}_{-0.003}$
	9	$0.307^{+0.010}_{-0.015}$	$0.366^{+0.017}_{-0.018}$	$0.368^{+0.013}_{-0.031}$	$\mathbf{0.414^{+0.006}_{-0.008}}$	$0.288^{+0.005}_{-0.006}$	$0.351^{+0.009}_{-0.008}$
u10	1	$0.922^{+0.002}_{-0.002}$	$0.943^{+0.001}_{-0.001}$	$\mathbf{0.946^{+0.002}_{-0.001}}$	$0.942^{+0.001}_{-0.003}$	$0.920^{+0.001}_{-0.002}$	$0.939^{+0.000}_{-0.000}$
	5	$0.551^{+0.005}_{-0.004}$	$0.657^{+0.004}_{-0.009}$	$\mathbf{0.665^{+0.008}_{-0.008}}$	$\mathbf{0.671^{+0.007}_{-0.010}}$	$0.550^{+0.006}_{-0.003}$	$0.608^{+0.003}_{-0.005}$
	9	$0.273^{+0.006}_{-0.006}$	$0.328^{+0.006}_{-0.010}$	$\mathbf{0.342^{+0.008}_{-0.005}}$	$\mathbf{0.351^{+0.008}_{-0.009}}$	$0.266^{+0.004}_{-0.004}$	$0.304^{+0.003}_{-0.006}$
v10	1	$0.921^{+0.002}_{-0.002}$	$0.942^{+0.001}_{-0.001}$	$\mathbf{0.946^{+0.001}_{-0.001}}$	$0.943^{+0.001}_{-0.003}$	$0.918^{+0.002}_{-0.002}$	$0.939^{+0.000}_{-0.001}$
	5	$0.526^{+0.006}_{-0.008}$	$\mathbf{0.651^{+0.005}_{-0.012}}$	$\mathbf{0.661^{+0.005}_{-0.004}}$	$\mathbf{0.665^{+0.006}_{-0.010}}$	$0.521^{+0.003}_{-0.009}$	$0.599^{+0.005}_{-0.004}$
	9	$0.232^{+0.006}_{-0.015}$	$\mathbf{0.302^{+0.010}_{-0.015}}$	$\mathbf{0.316^{+0.009}_{-0.007}}$	$\mathbf{0.322^{+0.007}_{-0.011}}$	$0.226^{+0.003}_{-0.004}$	$0.272^{+0.006}_{-0.003}$
q	1	$0.800^{+0.002}_{-0.002}$	$0.823^{+0.001}_{-0.001}$	$0.830^{+0.003}_{-0.002}$	$\mathbf{0.850^{+0.002}_{-0.005}}$	$0.787^{+0.001}_{-0.003}$	$0.808^{+0.001}_{-0.001}$
	5	$0.511^{+0.004}_{-0.005}$	$0.576^{+0.008}_{-0.012}$	$0.590^{+0.008}_{-0.009}$	$\mathbf{0.630^{+0.009}_{-0.011}}$	$0.502^{+0.007}_{-0.009}$	$0.549^{+0.007}_{-0.007}$
	9	$0.302^{+0.011}_{-0.008}$	$0.349^{+0.013}_{-0.022}$	$0.360^{+0.016}_{-0.019}$	$\mathbf{0.396^{+0.015}_{-0.013}}$	$0.298^{+0.013}_{-0.012}$	$0.330^{+0.015}_{-0.014}$
t	1	$0.933^{+0.001}_{-0.001}$	$0.948^{+0.001}_{-0.002}$	$\mathbf{0.951^{+0.002}_{-0.001}}$	$\mathbf{0.953^{+0.001}_{-0.002}}$	$0.932^{+0.001}_{-0.001}$	$0.940^{+0.000}_{-0.001}$
	5	$0.632^{+0.010}_{-0.008}$	$0.745^{+0.007}_{-0.014}$	$0.756^{+0.005}_{-0.013}$	$\mathbf{0.777^{+0.002}_{-0.005}}$	$0.628^{+0.005}_{-0.003}$	$0.703^{+0.004}_{-0.004}$
	9	$0.332^{+0.018}_{-0.012}$	$0.434^{+0.013}_{-0.021}$	$0.446^{+0.007}_{-0.019}$	$\mathbf{0.480^{+0.005}_{-0.009}}$	$0.341^{+0.003}_{-0.002}$	$0.395^{+0.009}_{-0.008}$
u	1	$0.933^{+0.001}_{-0.001}$	$0.949^{+0.001}_{-0.001}$	$0.952^{+0.001}_{-0.001}$	$\mathbf{0.955^{+0.001}_{-0.002}}$	$0.937^{+0.001}_{-0.001}$	$0.943^{+0.001}_{-0.000}$
	5	$0.650^{+0.004}_{-0.006}$	$0.744^{+0.005}_{-0.007}$	$0.750^{+0.008}_{-0.014}$	$\mathbf{0.771^{+0.003}_{-0.005}}$	$0.648^{+0.004}_{-0.004}$	$0.706^{+0.002}_{-0.006}$
	9	$0.377^{+0.006}_{-0.012}$	$0.442^{+0.007}_{-0.009}$	$0.449^{+0.013}_{-0.022}$	$\mathbf{0.485^{+0.008}_{-0.007}}$	$0.376^{+0.006}_{-0.007}$	$0.414^{+0.007}_{-0.017}$
v	1	$0.924^{+0.001}_{-0.001}$	$0.944^{+0.001}_{-0.001}$	$\mathbf{0.947^{+0.001}_{-0.001}}$	$\mathbf{0.950^{+0.001}_{-0.002}}$	$0.929^{+0.001}_{-0.001}$	$0.937^{+0.000}_{-0.000}$
	5	$0.588^{+0.005}_{-0.010}$	$0.720^{+0.003}_{-0.011}$	$\mathbf{0.730^{+0.004}_{-0.007}}$	$\mathbf{0.738^{+0.002}_{-0.007}}$	$0.584^{+0.003}_{-0.008}$	$0.672^{+0.006}_{-0.004}$
	9	$0.282^{+0.006}_{-0.016}$	$0.375^{+0.008}_{-0.015}$	$\mathbf{0.388^{+0.006}_{-0.010}}$	$\mathbf{0.400^{+0.007}_{-0.011}}$	$0.281^{+0.002}_{-0.004}$	$0.337^{+0.009}_{-0.007}$
z	1	$0.984^{+0.000}_{-0.001}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$\mathbf{0.992^{+0.000}_{-0.001}}$	$0.984^{+0.001}_{-0.001}$	$0.990^{+0.000}_{-0.000}$
	5	$0.727^{+0.004}_{-0.004}$	$0.847^{+0.003}_{-0.007}$	$\mathbf{0.855^{+0.003}_{-0.008}}$	$\mathbf{0.862^{+0.003}_{-0.006}}$	$0.721^{+0.002}_{-0.003}$	$0.805^{+0.003}_{-0.005}$
	9	$0.404^{+0.010}_{-0.008}$	$0.514^{+0.007}_{-0.011}$	$0.528^{+0.009}_{-0.012}$	$\mathbf{0.548^{+0.007}_{-0.010}}$	$0.400^{+0.006}_{-0.007}$	$0.469^{+0.007}_{-0.011}$

Table C.3: Native-grid RMSE at days 1, 5, and 9, using the checkpoints, weighting, and seed aggregation of Table C.2. Units are shown alongside the variables. Upper variables are averaged over 13 pressure levels. Lower is better.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl (Pa)	1	156^{+3}_{-2}	114^{+1}_{-1}	$\mathbf{109^{+1}_{-1}}$	117^{+7}_{-3}	156^{+3}_{-3}	123^{+1}_{-1}
	5	567^{+7}_{-5}	458^{+8}_{-5}	$\mathbf{442^{+5}_{-5}}$	$\mathbf{445^{+15}_{-10}}$	572^{+12}_{-7}	507^{+7}_{-7}
	9	788^{+6}_{-7}	773^{+10}_{-9}	$\mathbf{751^{+17}_{-12}}$	$\mathbf{747^{+15}_{-8}}$	791^{+14}_{-11}	788^{+8}_{-16}
t2m (K)	1	$1.23^{+0.01}_{-0.01}$	$1.08^{+0.01}_{-0.01}$	$\mathbf{1.05^{+0.01}_{-0.01}}$	$1.12^{+0.02}_{-0.01}$	$1.30^{+0.02}_{-0.02}$	$1.10^{+0.00}_{-0.00}$
	5	$2.41^{+0.08}_{-0.06}$	$2.20^{+0.06}_{-0.06}$	$2.17^{+0.07}_{-0.04}$	$\mathbf{2.05^{+0.04}_{-0.03}}$	$2.50^{+0.02}_{-0.04}$	$2.18^{+0.03}_{-0.02}$
	9	$3.21^{+0.14}_{-0.09}$	$3.27^{+0.12}_{-0.14}$	$3.24^{+0.15}_{-0.09}$	$\mathbf{2.93^{+0.07}_{-0.05}}$	$3.33^{+0.05}_{-0.04}$	$\mathbf{3.04^{+0.03}_{-0.04}}$
u10 ($m\,s^{-1}$)	1	$1.59^{+0.02}_{-0.02}$	$1.37^{+0.01}_{-0.01}$	$\mathbf{1.33^{+0.02}_{-0.02}}$	$1.37^{+0.03}_{-0.01}$	$1.61^{+0.02}_{-0.01}$	$1.41^{+0.01}_{-0.01}$
	5	$3.64^{+0.03}_{-0.02}$	$3.27^{+0.05}_{-0.02}$	$\mathbf{3.23^{+0.03}_{-0.04}}$	$\mathbf{3.20^{+0.04}_{-0.03}}$	$3.62^{+0.04}_{-0.05}$	$3.49^{+0.03}_{-0.03}$
	9	$4.56^{+0.04}_{-0.03}$	$\mathbf{4.60^{+0.08}_{-0.09}}$	$\mathbf{4.53^{+0.08}_{-0.06}}$	$\mathbf{4.48^{+0.04}_{-0.03}}$	$\mathbf{4.57^{+0.06}_{-0.07}}$	$4.65^{+0.01}_{-0.04}$
v10 ($m\,s^{-1}$)	1	$1.63^{+0.02}_{-0.02}$	$1.40^{+0.01}_{-0.01}$	$\mathbf{1.36^{+0.01}_{-0.01}}$	$1.40^{+0.04}_{-0.02}$	$1.65^{+0.02}_{-0.02}$	$1.44^{+0.01}_{-0.00}$
	5	$3.78^{+0.04}_{-0.05}$	$\mathbf{3.35^{+0.07}_{-0.03}}$	$\mathbf{3.32^{+0.04}_{-0.03}}$	$\mathbf{3.30^{+0.07}_{-0.03}}$	$3.79^{+0.07}_{-0.06}$	$3.61^{+0.04}_{-0.04}$
	9	$\mathbf{4.72^{+0.03}_{-0.04}}$	$\mathbf{4.76^{+0.09}_{-0.05}}$	$\mathbf{4.71^{+0.11}_{-0.07}}$	$\mathbf{4.68^{+0.06}_{-0.03}}$	$\mathbf{4.75^{+0.07}_{-0.09}}$	$4.87^{+0.06}_{-0.10}$
q ($10^{-4}\,kg\,kg^{-1}$)	1	$4.43^{+0.04}_{-0.03}$	$4.05^{+0.03}_{-0.02}$	$\mathbf{3.95^{+0.02}_{-0.03}}$	$\mathbf{3.88^{+0.07}_{-0.03}}$	$4.60^{+0.02}_{-0.02}$	$4.17^{+0.02}_{-0.01}$
	5	$7.07^{+0.16}_{-0.13}$	$6.51^{+0.08}_{-0.10}$	$6.41^{+0.06}_{-0.05}$	$\mathbf{6.22^{+0.12}_{-0.13}}$	$7.07^{+0.11}_{-0.09}$	$6.66^{+0.08}_{-0.06}$
	9	$8.59^{+0.24}_{-0.16}$	$\mathbf{8.36^{+0.23}_{-0.20}}$	$\mathbf{8.27^{+0.13}_{-0.10}}$	$\mathbf{8.11^{+0.14}_{-0.18}}$	$8.56^{+0.29}_{-0.16}$	$\mathbf{8.35^{+0.14}_{-0.12}}$
t (K)	1	$1.15^{+0.01}_{-0.01}$	$1.01^{+0.02}_{-0.01}$	$\mathbf{0.97^{+0.01}_{-0.02}}$	$\mathbf{0.96^{+0.02}_{-0.01}}$	$1.16^{+0.01}_{-0.01}$	$1.09^{+0.00}_{-0.00}$
	5	$2.65^{+0.03}_{-0.02}$	$2.24^{+0.07}_{-0.03}$	$2.19^{+0.04}_{-0.02}$	$\mathbf{2.10^{+0.03}_{-0.01}}$	$2.64^{+0.02}_{-0.03}$	$2.41^{+0.02}_{-0.01}$
	9	$3.60^{+0.06}_{-0.06}$	$3.39^{+0.09}_{-0.06}$	$3.34^{+0.04}_{-0.05}$	$\mathbf{3.21^{+0.05}_{-0.02}}$	$3.52^{+0.02}_{-0.04}$	$3.47^{+0.04}_{-0.03}$
u ($m\,s^{-1}$)	1	$3.10^{+0.03}_{-0.02}$	$2.73^{+0.01}_{-0.01}$	$2.64^{+0.03}_{-0.04}$	$\mathbf{2.51^{+0.04}_{-0.02}}$	$3.01^{+0.03}_{-0.02}$	$2.91^{+0.01}_{-0.01}$
	5	$6.89^{+0.04}_{-0.04}$	$5.99^{+0.08}_{-0.07}$	$5.93^{+0.17}_{-0.10}$	$\mathbf{5.64^{+0.06}_{-0.05}}$	$6.84^{+0.09}_{-0.08}$	$6.41^{+0.07}_{-0.04}$
	9	$9.12^{+0.07}_{-0.11}$	$8.88^{+0.05}_{-0.07}$	$8.83^{+0.19}_{-0.18}$	$\mathbf{8.55^{+0.07}_{-0.10}}$	$9.01^{+0.15}_{-0.11}$	$9.12^{+0.14}_{-0.07}$
v ($m\,s^{-1}$)	1	$3.11^{+0.03}_{-0.02}$	$2.67^{+0.03}_{-0.02}$	$\mathbf{2.58^{+0.02}_{-0.03}}$	$\mathbf{2.51^{+0.05}_{-0.02}}$	$3.03^{+0.02}_{-0.02}$	$2.86^{+0.01}_{-0.01}$
	5	$7.08^{+0.11}_{-0.08}$	$5.97^{+0.10}_{-0.06}$	$\mathbf{5.88^{+0.02}_{-0.05}}$	$\mathbf{5.80^{+0.10}_{-0.03}}$	$7.05^{+0.11}_{-0.09}$	$6.50^{+0.07}_{-0.06}$
	9	$9.15^{+0.13}_{-0.10}$	$\mathbf{8.95^{+0.09}_{-0.12}}$	$\mathbf{8.87^{+0.12}_{-0.14}}$	$\mathbf{8.84^{+0.16}_{-0.07}}$	$\mathbf{9.04^{+0.15}_{-0.16}}$	$9.28^{+0.18}_{-0.25}$
z (gpm)	1	$15.31^{+0.32}_{-0.19}$	$\mathbf{10.85^{+0.14}_{-0.15}}$	$\mathbf{10.53^{+0.37}_{-0.15}}$	$\mathbf{10.36^{+0.40}_{-0.12}}$	$15.07^{+0.30}_{-0.33}$	$12.40^{+0.15}_{-0.19}$
	5	$61.42^{+0.35}_{-0.54}$	$46.95^{+0.96}_{-0.61}$	$45.63^{+0.95}_{-0.39}$	$\mathbf{44.22^{+0.79}_{-0.47}}$	$62.25^{+1.01}_{-0.42}$	$52.98^{+0.86}_{-0.49}$
	9	$89.40^{+0.42}_{-0.55}$	$84.63^{+0.90}_{-0.90}$	$82.81^{+0.76}_{-0.39}$	$\mathbf{80.89^{+1.12}_{-1.01}}$	$89.97^{+0.81}_{-0.79}$	$88.33^{+2.05}_{-1.32}$

Table C.4: Common-HEALPix ACC at days 1, 5, and 9 on 2018 test data, at $n_{\text{side}} = 64$. Equiangular forecasts use bilinear interpolation to HEALPix cell centers. Checkpoints, upper-level averaging, seed ranges, and bolding follow Table C.2. Higher is better.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl	1	$0.977^{+0.001}_{-0.001}$	$0.988^{+0.000}_{-0.000}$	$\mathbf{0.989^{+0.000}_{-0.000}}$	$0.987^{+0.001}_{-0.001}$	$0.978^{+0.001}_{-0.001}$	$0.985^{+0.000}_{-0.001}$
	5	$0.673^{+0.005}_{-0.007}$	$0.801^{+0.004}_{-0.007}$	$\mathbf{0.812^{+0.006}_{-0.004}}$	$\mathbf{0.810^{+0.009}_{-0.013}}$	$0.666^{+0.003}_{-0.004}$	$0.753^{+0.003}_{-0.006}$
	9	$0.336^{+0.007}_{-0.013}$	$0.435^{+0.010}_{-0.011}$	$\mathbf{0.452^{+0.007}_{-0.006}}$	$\mathbf{0.459^{+0.010}_{-0.013}}$	$0.330^{+0.004}_{-0.003}$	$0.399^{+0.007}_{-0.004}$
t2m	1	$0.875^{+0.001}_{-0.001}$	$0.897^{+0.001}_{-0.000}$	$0.902^{+0.002}_{-0.002}$	$\mathbf{0.911^{+0.001}_{-0.002}}$	$0.864^{+0.003}_{-0.003}$	$0.894^{+0.001}_{-0.001}$
	5	$0.581^{+0.009}_{-0.015}$	$0.670^{+0.010}_{-0.008}$	$0.678^{+0.009}_{-0.016}$	$\mathbf{0.703^{+0.006}_{-0.008}}$	$0.564^{+0.010}_{-0.010}$	$0.646^{+0.005}_{-0.003}$
	9	$0.306^{+0.010}_{-0.015}$	$0.364^{+0.017}_{-0.018}$	$0.365^{+0.015}_{-0.032}$	$\mathbf{0.414^{+0.006}_{-0.008}}$	$0.290^{+0.005}_{-0.007}$	$0.346^{+0.009}_{-0.008}$
u10	1	$0.921^{+0.001}_{-0.002}$	$0.940^{+0.001}_{-0.001}$	$\mathbf{0.944^{+0.002}_{-0.001}}$	$\mathbf{0.942^{+0.001}_{-0.003}}$	$0.919^{+0.001}_{-0.002}$	$0.937^{+0.000}_{-0.000}$
	5	$0.553^{+0.005}_{-0.004}$	$\mathbf{0.659^{+0.003}_{-0.009}}$	$\mathbf{0.668^{+0.007}_{-0.008}}$	$\mathbf{0.671^{+0.007}_{-0.010}}$	$0.552^{+0.006}_{-0.003}$	$0.610^{+0.003}_{-0.005}$
	9	$0.275^{+0.006}_{-0.006}$	$0.330^{+0.007}_{-0.010}$	$\mathbf{0.344^{+0.008}_{-0.005}}$	$\mathbf{0.351^{+0.008}_{-0.009}}$	$0.268^{+0.004}_{-0.004}$	$0.305^{+0.003}_{-0.006}$
v10	1	$0.920^{+0.002}_{-0.002}$	$0.941^{+0.001}_{-0.001}$	$\mathbf{0.944^{+0.002}_{-0.001}}$	$\mathbf{0.943^{+0.001}_{-0.003}}$	$0.918^{+0.002}_{-0.002}$	$0.937^{+0.000}_{-0.000}$
	5	$0.527^{+0.006}_{-0.008}$	$\mathbf{0.653^{+0.005}_{-0.012}}$	$\mathbf{0.663^{+0.005}_{-0.004}}$	$\mathbf{0.665^{+0.006}_{-0.010}}$	$0.522^{+0.003}_{-0.009}$	$0.600^{+0.005}_{-0.004}$
	9	$0.233^{+0.006}_{-0.015}$	$\mathbf{0.304^{+0.010}_{-0.014}}$	$\mathbf{0.317^{+0.009}_{-0.007}}$	$\mathbf{0.322^{+0.007}_{-0.011}}$	$0.227^{+0.003}_{-0.004}$	$0.273^{+0.006}_{-0.003}$
q	1	$0.802^{+0.002}_{-0.002}$	$0.823^{+0.002}_{-0.001}$	$0.830^{+0.003}_{-0.002}$	$\mathbf{0.850^{+0.002}_{-0.005}}$	$0.790^{+0.001}_{-0.003}$	$0.809^{+0.001}_{-0.001}$
	5	$0.513^{+0.004}_{-0.005}$	$0.578^{+0.008}_{-0.012}$	$0.592^{+0.008}_{-0.009}$	$\mathbf{0.630^{+0.009}_{-0.011}}$	$0.505^{+0.007}_{-0.009}$	$0.551^{+0.007}_{-0.007}$
	9	$0.304^{+0.011}_{-0.008}$	$0.351^{+0.013}_{-0.022}$	$0.362^{+0.016}_{-0.019}$	$\mathbf{0.396^{+0.015}_{-0.013}}$	$0.300^{+0.013}_{-0.012}$	$0.331^{+0.015}_{-0.015}$
t	1	$0.934^{+0.001}_{-0.001}$	$0.948^{+0.001}_{-0.001}$	$\mathbf{0.951^{+0.002}_{-0.001}}$	$\mathbf{0.953^{+0.001}_{-0.002}}$	$0.933^{+0.001}_{-0.001}$	$0.940^{+0.001}_{-0.001}$
	5	$0.633^{+0.010}_{-0.008}$	$0.746^{+0.007}_{-0.014}$	$0.757^{+0.005}_{-0.013}$	$\mathbf{0.777^{+0.002}_{-0.005}}$	$0.630^{+0.005}_{-0.003}$	$0.704^{+0.004}_{-0.004}$
	9	$0.333^{+0.018}_{-0.012}$	$0.435^{+0.013}_{-0.021}$	$0.447^{+0.007}_{-0.019}$	$\mathbf{0.480^{+0.005}_{-0.009}}$	$0.342^{+0.003}_{-0.002}$	$0.396^{+0.009}_{-0.008}$
u	1	$0.933^{+0.001}_{-0.001}$	$0.949^{+0.001}_{-0.001}$	$0.952^{+0.001}_{-0.001}$	$\mathbf{0.955^{+0.001}_{-0.002}}$	$0.937^{+0.001}_{-0.001}$	$0.943^{+0.001}_{-0.001}$
	5	$0.651^{+0.004}_{-0.006}$	$0.745^{+0.005}_{-0.007}$	$0.751^{+0.008}_{-0.014}$	$\mathbf{0.771^{+0.003}_{-0.005}}$	$0.650^{+0.004}_{-0.004}$	$0.707^{+0.002}_{-0.006}$
	9	$0.378^{+0.006}_{-0.012}$	$0.443^{+0.007}_{-0.009}$	$0.451^{+0.013}_{-0.022}$	$\mathbf{0.485^{+0.008}_{-0.007}}$	$0.377^{+0.006}_{-0.007}$	$0.415^{+0.007}_{-0.017}$
v	1	$0.925^{+0.001}_{-0.001}$	$0.944^{+0.001}_{-0.001}$	$0.947^{+0.001}_{-0.001}$	$\mathbf{0.950^{+0.001}_{-0.002}}$	$0.929^{+0.001}_{-0.001}$	$0.937^{+0.000}_{-0.000}$
	5	$0.589^{+0.005}_{-0.010}$	$0.721^{+0.003}_{-0.011}$	$\mathbf{0.731^{+0.004}_{-0.007}}$	$\mathbf{0.738^{+0.002}_{-0.007}}$	$0.585^{+0.003}_{-0.008}$	$0.673^{+0.006}_{-0.004}$
	9	$0.282^{+0.006}_{-0.016}$	$0.376^{+0.008}_{-0.015}$	$\mathbf{0.389^{+0.006}_{-0.010}}$	$\mathbf{0.400^{+0.007}_{-0.011}}$	$0.282^{+0.002}_{-0.004}$	$0.337^{+0.009}_{-0.007}$
z	1	$0.984^{+0.000}_{-0.001}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$\mathbf{0.992^{+0.000}_{-0.001}}$	$0.985^{+0.001}_{-0.001}$	$0.990^{+0.000}_{-0.000}$
	5	$0.728^{+0.004}_{-0.004}$	$0.848^{+0.003}_{-0.007}$	$\mathbf{0.856^{+0.003}_{-0.008}}$	$\mathbf{0.862^{+0.003}_{-0.006}}$	$0.721^{+0.002}_{-0.003}$	$0.805^{+0.003}_{-0.005}$
	9	$0.404^{+0.010}_{-0.008}$	$0.514^{+0.007}_{-0.011}$	$0.528^{+0.009}_{-0.012}$	$\mathbf{0.548^{+0.007}_{-0.010}}$	$0.400^{+0.006}_{-0.007}$	$0.469^{+0.007}_{-0.011}$

Table C.5: Common-HEALPix RMSE using the evaluation protocol of Table C.4. Units are shown alongside the variables. Upper variables are averaged over 13 pressure levels. Lower is better.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl (Pa)	1	156^{+3}_{-2}	116^{+1}_{-1}	$\mathbf{111}^{+1}_{-1}$	117^{+7}_{-3}	156^{+2}_{-3}	126^{+2}_{-1}
	5	566^{+7}_{-5}	457^{+8}_{-5}	$\mathbf{442}^{+5}_{-5}$	$\mathbf{445}^{+15}_{-10}$	571^{+12}_{-7}	507^{+8}_{-7}
	9	787^{+6}_{-7}	772^{+10}_{-9}	$\mathbf{750}^{+16}_{-12}$	$\mathbf{747}^{+15}_{-8}$	790^{+14}_{-11}	788^{+8}_{-16}
t2m (K)	1	$1.33^{+0.01}_{-0.01}$	$1.22^{+0.01}_{-0.00}$	$1.19^{+0.01}_{-0.01}$	$\mathbf{1.12}^{+0.02}_{-0.01}$	$1.39^{+0.02}_{-0.02}$	$1.23^{+0.01}_{-0.00}$
	5	$2.44^{+0.08}_{-0.06}$	$2.24^{+0.05}_{-0.06}$	$2.22^{+0.07}_{-0.05}$	$\mathbf{2.05}^{+0.04}_{-0.03}$	$2.51^{+0.01}_{-0.04}$	$2.23^{+0.03}_{-0.02}$
	9	$3.22^{+0.14}_{-0.09}$	$3.28^{+0.12}_{-0.13}$	$3.25^{+0.15}_{-0.10}$	$\mathbf{2.93}^{+0.07}_{-0.05}$	$3.32^{+0.05}_{-0.04}$	$3.07^{+0.03}_{-0.04}$
u10 ($m\,s^{-1}$)	1	$1.60^{+0.02}_{-0.01}$	$1.39^{+0.01}_{-0.01}$	$\mathbf{1.35}^{+0.02}_{-0.02}$	$\mathbf{1.37}^{+0.03}_{-0.01}$	$1.62^{+0.02}_{-0.01}$	$1.43^{+0.01}_{-0.00}$
	5	$3.62^{+0.03}_{-0.02}$	$\mathbf{3.25}^{+0.05}_{-0.02}$	$\mathbf{3.21}^{+0.03}_{-0.04}$	$\mathbf{3.20}^{+0.04}_{-0.03}$	$3.60^{+0.04}_{-0.05}$	$3.47^{+0.03}_{-0.03}$
	9	$\mathbf{4.54}^{+0.04}_{-0.03}$	$\mathbf{4.57}^{+0.08}_{-0.08}$	$\mathbf{4.50}^{+0.08}_{-0.05}$	$\mathbf{4.48}^{+0.04}_{-0.03}$	$\mathbf{4.55}^{+0.06}_{-0.07}$	$4.63^{+0.02}_{-0.04}$
v10 ($m\,s^{-1}$)	1	$1.64^{+0.02}_{-0.02}$	$1.42^{+0.01}_{-0.01}$	$\mathbf{1.38}^{+0.01}_{-0.02}$	$\mathbf{1.40}^{+0.04}_{-0.02}$	$1.66^{+0.02}_{-0.02}$	$1.46^{+0.01}_{-0.00}$
	5	$3.77^{+0.04}_{-0.05}$	$\mathbf{3.34}^{+0.07}_{-0.03}$	$\mathbf{3.30}^{+0.03}_{-0.03}$	$\mathbf{3.30}^{+0.07}_{-0.03}$	$3.78^{+0.07}_{-0.05}$	$3.59^{+0.04}_{-0.04}$
	9	$\mathbf{4.71}^{+0.03}_{-0.04}$	$\mathbf{4.73}^{+0.09}_{-0.05}$	$\mathbf{4.68}^{+0.11}_{-0.07}$	$\mathbf{4.68}^{+0.06}_{-0.03}$	$\mathbf{4.74}^{+0.07}_{-0.09}$	$4.85^{+0.06}_{-0.10}$
q ($10^{-4}\,kg\,kg^{-1}$)	1	$4.41^{+0.04}_{-0.03}$	$4.06^{+0.03}_{-0.02}$	$\mathbf{3.97}^{+0.02}_{-0.03}$	$\mathbf{3.88}^{+0.07}_{-0.03}$	$4.55^{+0.02}_{-0.02}$	$4.17^{+0.02}_{-0.01}$
	5	$7.03^{+0.16}_{-0.12}$	$6.47^{+0.08}_{-0.10}$	$\mathbf{6.37}^{+0.06}_{-0.04}$	$\mathbf{6.22}^{+0.12}_{-0.13}$	$7.02^{+0.11}_{-0.09}$	$6.62^{+0.08}_{-0.07}$
	9	$8.55^{+0.23}_{-0.15}$	$\mathbf{8.31}^{+0.22}_{-0.19}$	$\mathbf{8.22}^{+0.13}_{-0.10}$	$\mathbf{8.11}^{+0.14}_{-0.18}$	$8.51^{+0.29}_{-0.16}$	$\mathbf{8.30}^{+0.14}_{-0.12}$
t (K)	1	$1.14^{+0.01}_{-0.01}$	$1.01^{+0.02}_{-0.01}$	$\mathbf{0.98}^{+0.01}_{-0.02}$	$\mathbf{0.96}^{+0.02}_{-0.01}$	$1.15^{+0.01}_{-0.00}$	$1.09^{+0.00}_{-0.01}$
	5	$2.64^{+0.02}_{-0.02}$	$2.24^{+0.07}_{-0.03}$	$2.19^{+0.04}_{-0.02}$	$\mathbf{2.10}^{+0.03}_{-0.01}$	$2.63^{+0.02}_{-0.03}$	$2.40^{+0.02}_{-0.01}$
	9	$3.60^{+0.06}_{-0.06}$	$3.38^{+0.09}_{-0.06}$	$3.33^{+0.04}_{-0.05}$	$\mathbf{3.21}^{+0.05}_{-0.02}$	$3.51^{+0.02}_{-0.04}$	$3.46^{+0.04}_{-0.03}$
u ($m\,s^{-1}$)	1	$3.09^{+0.03}_{-0.02}$	$2.73^{+0.01}_{-0.01}$	$2.65^{+0.03}_{-0.04}$	$\mathbf{2.51}^{+0.04}_{-0.02}$	$2.99^{+0.02}_{-0.02}$	$2.91^{+0.01}_{-0.01}$
	5	$6.86^{+0.04}_{-0.04}$	$5.96^{+0.08}_{-0.07}$	$5.90^{+0.16}_{-0.10}$	$\mathbf{5.64}^{+0.06}_{-0.05}$	$6.82^{+0.09}_{-0.08}$	$6.38^{+0.07}_{-0.04}$
	9	$9.09^{+0.07}_{-0.11}$	$8.85^{+0.05}_{-0.07}$	$\mathbf{8.80}^{+0.19}_{-0.18}$	$\mathbf{8.55}^{+0.07}_{-0.10}$	$8.98^{+0.15}_{-0.11}$	$9.09^{+0.14}_{-0.07}$
v ($m\,s^{-1}$)	1	$3.10^{+0.02}_{-0.02}$	$2.67^{+0.03}_{-0.02}$	$\mathbf{2.58}^{+0.02}_{-0.03}$	$\mathbf{2.51}^{+0.05}_{-0.02}$	$3.02^{+0.02}_{-0.02}$	$2.85^{+0.00}_{-0.01}$
	5	$7.07^{+0.11}_{-0.08}$	$\mathbf{5.95}^{+0.10}_{-0.06}$	$\mathbf{5.86}^{+0.02}_{-0.05}$	$\mathbf{5.80}^{+0.10}_{-0.03}$	$7.04^{+0.11}_{-0.09}$	$6.48^{+0.07}_{-0.06}$
	9	$9.14^{+0.13}_{-0.10}$	$\mathbf{8.93}^{+0.09}_{-0.12}$	$\mathbf{8.85}^{+0.12}_{-0.13}$	$\mathbf{8.84}^{+0.16}_{-0.07}$	$\mathbf{9.02}^{+0.15}_{-0.16}$	$9.26^{+0.18}_{-0.25}$
z (gpm)	1	$15.18^{+0.32}_{-0.19}$	$\mathbf{10.81}^{+0.15}_{-0.15}$	$\mathbf{10.51}^{+0.38}_{-0.16}$	$\mathbf{10.36}^{+0.40}_{-0.12}$	$14.86^{+0.27}_{-0.33}$	$12.43^{+0.17}_{-0.19}$
	5	$61.30^{+0.35}_{-0.54}$	$46.87^{+0.96}_{-0.61}$	$45.55^{+0.96}_{-0.39}$	$\mathbf{44.22}^{+0.79}_{-0.47}$	$62.12^{+1.00}_{-0.40}$	$53.01^{+0.84}_{-0.53}$
	9	$89.30^{+0.42}_{-0.55}$	$84.51^{+0.91}_{-0.89}$	$82.70^{+0.76}_{-0.39}$	$\mathbf{80.89}^{+1.12}_{-1.01}$	$89.87^{+0.80}_{-0.77}$	$88.32^{+2.02}_{-1.13}$

Table C.6: Common-equiangular ACC at days 1, 5, and 9 on 2018 test data, using the 157×314 cosine-weighted grid and nearest-cell sampling for HEALPix forecasts. Upper-level averaging, seed ranges, and bolding follow Table C.2. Higher is better.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl	1	$0.977^{+0.001}_{-0.001}$	$0.988^{+0.000}_{-0.000}$	$\mathbf{0.989^{+0.000}_{-0.000}}$	$0.985^{+0.001}_{-0.001}$	$0.977^{+0.001}_{-0.001}$	$0.986^{+0.000}_{-0.000}$
	5	$0.672^{+0.005}_{-0.007}$	$0.800^{+0.004}_{-0.007}$	$\mathbf{0.812^{+0.006}_{-0.004}}$	$\mathbf{0.809^{+0.009}_{-0.013}}$	$0.665^{+0.003}_{-0.004}$	$0.753^{+0.003}_{-0.005}$
	9	$0.336^{+0.007}_{-0.013}$	$0.434^{+0.010}_{-0.011}$	$\mathbf{0.451^{+0.007}_{-0.006}}$	$\mathbf{0.458^{+0.010}_{-0.013}}$	$0.330^{+0.004}_{-0.003}$	$0.399^{+0.007}_{-0.004}$
t2m	1	$0.892^{+0.001}_{-0.002}$	$0.918^{+0.001}_{-0.001}$	$\mathbf{0.923^{+0.002}_{-0.002}}$	$0.870^{+0.001}_{-0.002}$	$0.880^{+0.003}_{-0.004}$	$0.914^{+0.000}_{-0.001}$
	5	$0.587^{+0.010}_{-0.015}$	$\mathbf{0.680^{+0.011}_{-0.010}}$	$\mathbf{0.687^{+0.009}_{-0.016}}$	$\mathbf{0.682^{+0.006}_{-0.008}}$	$0.567^{+0.009}_{-0.008}$	$0.658^{+0.005}_{-0.003}$
	9	$0.307^{+0.010}_{-0.015}$	$0.366^{+0.017}_{-0.018}$	$0.368^{+0.013}_{-0.031}$	$\mathbf{0.404^{+0.005}_{-0.007}}$	$0.288^{+0.005}_{-0.006}$	$0.351^{+0.009}_{-0.008}$
u10	1	$0.922^{+0.002}_{-0.002}$	$0.943^{+0.001}_{-0.001}$	$\mathbf{0.946^{+0.002}_{-0.001}}$	$0.934^{+0.001}_{-0.003}$	$0.920^{+0.001}_{-0.002}$	$0.939^{+0.000}_{-0.000}$
	5	$0.551^{+0.005}_{-0.004}$	$\mathbf{0.657^{+0.004}_{-0.009}}$	$\mathbf{0.665^{+0.008}_{-0.008}}$	$\mathbf{0.667^{+0.006}_{-0.010}}$	$0.550^{+0.006}_{-0.003}$	$0.608^{+0.003}_{-0.005}$
	9	$0.273^{+0.006}_{-0.006}$	$0.328^{+0.006}_{-0.010}$	$\mathbf{0.342^{+0.008}_{-0.005}}$	$\mathbf{0.350^{+0.007}_{-0.009}}$	$0.266^{+0.004}_{-0.004}$	$0.304^{+0.003}_{-0.006}$
v10	1	$0.921^{+0.002}_{-0.002}$	$0.942^{+0.001}_{-0.001}$	$\mathbf{0.946^{+0.001}_{-0.001}}$	$0.935^{+0.001}_{-0.003}$	$0.918^{+0.002}_{-0.002}$	$0.939^{+0.000}_{-0.001}$
	5	$0.526^{+0.006}_{-0.008}$	$0.651^{+0.005}_{-0.012}$	$\mathbf{0.661^{+0.005}_{-0.004}}$	$\mathbf{0.661^{+0.006}_{-0.010}}$	$0.521^{+0.003}_{-0.009}$	$0.599^{+0.005}_{-0.004}$
	9	$0.232^{+0.006}_{-0.015}$	$\mathbf{0.302^{+0.010}_{-0.015}}$	$\mathbf{0.316^{+0.009}_{-0.007}}$	$\mathbf{0.321^{+0.007}_{-0.011}}$	$0.226^{+0.003}_{-0.004}$	$0.272^{+0.006}_{-0.003}$
q	1	$0.800^{+0.002}_{-0.002}$	$0.823^{+0.001}_{-0.001}$	$0.830^{+0.003}_{-0.002}$	$\mathbf{0.846^{+0.002}_{-0.005}}$	$0.787^{+0.001}_{-0.003}$	$0.808^{+0.001}_{-0.001}$
	5	$0.511^{+0.004}_{-0.005}$	$0.576^{+0.008}_{-0.012}$	$0.590^{+0.008}_{-0.009}$	$\mathbf{0.627^{+0.009}_{-0.011}}$	$0.502^{+0.007}_{-0.009}$	$0.549^{+0.007}_{-0.007}$
	9	$0.302^{+0.011}_{-0.008}$	$0.349^{+0.013}_{-0.022}$	$0.360^{+0.016}_{-0.019}$	$\mathbf{0.395^{+0.015}_{-0.012}}$	$0.298^{+0.013}_{-0.012}$	$0.330^{+0.015}_{-0.014}$
t	1	$0.933^{+0.001}_{-0.001}$	$0.948^{+0.001}_{-0.002}$	$\mathbf{0.951^{+0.002}_{-0.001}}$	$\mathbf{0.950^{+0.001}_{-0.002}}$	$0.932^{+0.001}_{-0.001}$	$0.940^{+0.000}_{-0.001}$
	5	$0.632^{+0.010}_{-0.008}$	$0.745^{+0.007}_{-0.014}$	$0.756^{+0.005}_{-0.013}$	$\mathbf{0.775^{+0.002}_{-0.005}}$	$0.628^{+0.005}_{-0.003}$	$0.703^{+0.004}_{-0.004}$
	9	$0.332^{+0.018}_{-0.012}$	$0.434^{+0.013}_{-0.021}$	$0.446^{+0.007}_{-0.019}$	$\mathbf{0.479^{+0.006}_{-0.009}}$	$0.341^{+0.003}_{-0.002}$	$0.395^{+0.009}_{-0.008}$
u	1	$0.933^{+0.001}_{-0.001}$	$0.949^{+0.001}_{-0.001}$	$\mathbf{0.952^{+0.001}_{-0.001}}$	$\mathbf{0.952^{+0.001}_{-0.002}}$	$0.937^{+0.001}_{-0.001}$	$0.943^{+0.001}_{-0.000}$
	5	$0.650^{+0.004}_{-0.006}$	$0.744^{+0.005}_{-0.007}$	$0.750^{+0.008}_{-0.014}$	$\mathbf{0.769^{+0.003}_{-0.005}}$	$0.648^{+0.004}_{-0.004}$	$0.706^{+0.002}_{-0.006}$
	9	$0.377^{+0.006}_{-0.012}$	$0.442^{+0.007}_{-0.009}$	$0.449^{+0.013}_{-0.022}$	$\mathbf{0.484^{+0.008}_{-0.007}}$	$0.376^{+0.006}_{-0.007}$	$0.414^{+0.007}_{-0.017}$
v	1	$0.924^{+0.001}_{-0.001}$	$0.944^{+0.001}_{-0.001}$	$\mathbf{0.947^{+0.001}_{-0.001}}$	$\mathbf{0.948^{+0.001}_{-0.002}}$	$0.929^{+0.001}_{-0.001}$	$0.937^{+0.000}_{-0.000}$
	5	$0.588^{+0.005}_{-0.010}$	$0.720^{+0.003}_{-0.011}$	$\mathbf{0.730^{+0.004}_{-0.007}}$	$\mathbf{0.737^{+0.002}_{-0.006}}$	$0.584^{+0.003}_{-0.008}$	$0.672^{+0.006}_{-0.004}$
	9	$0.282^{+0.006}_{-0.016}$	$0.375^{+0.008}_{-0.015}$	$\mathbf{0.388^{+0.006}_{-0.010}}$	$\mathbf{0.400^{+0.007}_{-0.011}}$	$0.281^{+0.002}_{-0.004}$	$0.337^{+0.009}_{-0.007}$
z	1	$0.984^{+0.000}_{-0.001}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$\mathbf{0.992^{+0.000}_{-0.000}}$	$0.990^{+0.000}_{-0.001}$	$0.984^{+0.001}_{-0.001}$	$0.990^{+0.000}_{-0.000}$
	5	$0.727^{+0.004}_{-0.004}$	$0.847^{+0.003}_{-0.007}$	$\mathbf{0.855^{+0.003}_{-0.008}}$	$\mathbf{0.860^{+0.003}_{-0.006}}$	$0.721^{+0.002}_{-0.003}$	$0.805^{+0.003}_{-0.005}$
	9	$0.404^{+0.010}_{-0.008}$	$0.514^{+0.007}_{-0.011}$	$0.528^{+0.009}_{-0.012}$	$\mathbf{0.547^{+0.007}_{-0.010}}$	$0.400^{+0.006}_{-0.007}$	$0.469^{+0.007}_{-0.011}$

Table C.7: Common-equiangular RMSE using the evaluation protocol of Table C.6. Units are shown alongside the variables. Upper variables are averaged over 13 pressure levels. Lower is better.

Variable	Lead	FengWu	GraphCast-I	GraphCast-P	PEAR	Pangu	AFNO
msl (Pa)	1	156^{+3}_{-2}	114^{+1}_{-1}	$\mathbf{109^{+1}_{-1}}$	128^{+6}_{-2}	156^{+3}_{-3}	123^{+1}_{-1}
	5	567^{+7}_{-5}	458^{+8}_{-5}	$\mathbf{442^{+5}_{-5}}$	$\mathbf{447^{+15}_{-9}}$	572^{+12}_{-7}	507^{+7}_{-7}
	9	788^{+6}_{-7}	773^{+10}_{-9}	$\mathbf{751^{+17}_{-12}}$	$\mathbf{748^{+15}_{-8}}$	791^{+14}_{-11}	788^{+8}_{-16}
t2m (K)	1	$1.23^{+0.01}_{-0.01}$	$1.08^{+0.01}_{-0.01}$	$\mathbf{1.05^{+0.01}_{-0.01}}$	$1.37^{+0.01}_{-0.01}$	$1.30^{+0.02}_{-0.02}$	$1.10^{+0.00}_{-0.00}$
	5	$2.41^{+0.08}_{-0.06}$	$\mathbf{2.20^{+0.06}_{-0.06}}$	$\mathbf{2.17^{+0.07}_{-0.04}}$	$\mathbf{2.16^{+0.04}_{-0.03}}$	$2.50^{+0.02}_{-0.04}$	$\mathbf{2.18^{+0.03}_{-0.02}}$
	9	$3.21^{+0.14}_{-0.09}$	$3.27^{+0.12}_{-0.14}$	$3.24^{+0.15}_{-0.09}$	$\mathbf{2.99^{+0.07}_{-0.05}}$	$3.33^{+0.05}_{-0.04}$	$\mathbf{3.04^{+0.03}_{-0.04}}$
u10 ($m\,s^{-1}$)	1	$1.59^{+0.02}_{-0.02}$	$1.37^{+0.01}_{-0.01}$	$\mathbf{1.33^{+0.02}_{-0.02}}$	$1.47^{+0.03}_{-0.01}$	$1.61^{+0.02}_{-0.01}$	$1.41^{+0.01}_{-0.01}$
	5	$3.64^{+0.03}_{-0.02}$	$\mathbf{3.27^{+0.05}_{-0.02}}$	$\mathbf{3.23^{+0.03}_{-0.04}}$	$\mathbf{3.22^{+0.04}_{-0.03}}$	$3.62^{+0.04}_{-0.05}$	$3.49^{+0.03}_{-0.03}$
	9	$\mathbf{4.56^{+0.04}_{-0.03}}$	$\mathbf{4.60^{+0.08}_{-0.09}}$	$\mathbf{4.53^{+0.08}_{-0.06}}$	$\mathbf{4.50^{+0.03}_{-0.03}}$	$\mathbf{4.57^{+0.06}_{-0.07}}$	$4.65^{+0.01}_{-0.04}$
v10 ($m\,s^{-1}$)	1	$1.63^{+0.02}_{-0.02}$	$1.40^{+0.01}_{-0.01}$	$\mathbf{1.36^{+0.01}_{-0.01}}$	$1.48^{+0.03}_{-0.01}$	$1.65^{+0.02}_{-0.02}$	$1.44^{+0.01}_{-0.00}$
	5	$3.78^{+0.04}_{-0.05}$	$\mathbf{3.35^{+0.07}_{-0.03}}$	$\mathbf{3.32^{+0.04}_{-0.03}}$	$\mathbf{3.32^{+0.06}_{-0.03}}$	$3.79^{+0.07}_{-0.06}$	$3.61^{+0.04}_{-0.04}$
	9	$\mathbf{4.72^{+0.03}_{-0.04}}$	$\mathbf{4.76^{+0.09}_{-0.05}}$	$\mathbf{4.71^{+0.11}_{-0.07}}$	$\mathbf{4.69^{+0.06}_{-0.03}}$	$\mathbf{4.75^{+0.07}_{-0.09}}$	$4.87^{+0.06}_{-0.10}$
q ($10^{-4}\,kg\,kg^{-1}$)	1	$4.43^{+0.04}_{-0.03}$	$4.05^{+0.03}_{-0.02}$	$\mathbf{3.95^{+0.02}_{-0.03}}$	$\mathbf{3.96^{+0.07}_{-0.03}}$	$4.60^{+0.02}_{-0.02}$	$4.17^{+0.02}_{-0.01}$
	5	$7.07^{+0.16}_{-0.13}$	$6.51^{+0.08}_{-0.10}$	$\mathbf{6.41^{+0.06}_{-0.05}}$	$\mathbf{6.26^{+0.12}_{-0.13}}$	$7.07^{+0.11}_{-0.09}$	$6.66^{+0.08}_{-0.06}$
	9	$8.59^{+0.24}_{-0.16}$	$\mathbf{8.36^{+0.23}_{-0.20}}$	$\mathbf{8.27^{+0.13}_{-0.10}}$	$\mathbf{8.13^{+0.14}_{-0.17}}$	$8.56^{+0.29}_{-0.16}$	$\mathbf{8.35^{+0.14}_{-0.12}}$
t (K)	1	$1.15^{+0.01}_{-0.01}$	$1.01^{+0.02}_{-0.01}$	$\mathbf{0.97^{+0.01}_{-0.02}}$	$\mathbf{0.99^{+0.02}_{-0.01}}$	$1.16^{+0.01}_{-0.01}$	$1.09^{+0.00}_{-0.00}$
	5	$2.65^{+0.03}_{-0.02}$	$2.24^{+0.07}_{-0.03}$	$2.19^{+0.04}_{-0.02}$	$\mathbf{2.11^{+0.03}_{-0.01}}$	$2.64^{+0.02}_{-0.03}$	$2.41^{+0.02}_{-0.01}$
	9	$3.60^{+0.06}_{-0.06}$	$3.39^{+0.09}_{-0.06}$	$3.34^{+0.04}_{-0.05}$	$\mathbf{3.22^{+0.05}_{-0.02}}$	$3.52^{+0.02}_{-0.04}$	$3.47^{+0.04}_{-0.03}$
u ($m\,s^{-1}$)	1	$3.10^{+0.03}_{-0.02}$	$2.73^{+0.01}_{-0.01}$	$\mathbf{2.64^{+0.03}_{-0.04}}$	$\mathbf{2.61^{+0.04}_{-0.02}}$	$3.01^{+0.03}_{-0.02}$	$2.91^{+0.01}_{-0.01}$
	5	$6.89^{+0.04}_{-0.04}$	$5.99^{+0.08}_{-0.07}$	$5.93^{+0.17}_{-0.10}$	$\mathbf{5.67^{+0.06}_{-0.05}}$	$6.84^{+0.09}_{-0.08}$	$6.41^{+0.07}_{-0.04}$
	9	$9.12^{+0.07}_{-0.11}$	$8.88^{+0.05}_{-0.07}$	$8.83^{+0.19}_{-0.18}$	$\mathbf{8.56^{+0.07}_{-0.10}}$	$9.01^{+0.15}_{-0.11}$	$9.12^{+0.14}_{-0.07}$
v ($m\,s^{-1}$)	1	$3.11^{+0.03}_{-0.02}$	$2.67^{+0.03}_{-0.02}$	$\mathbf{2.58^{+0.02}_{-0.03}}$	$\mathbf{2.57^{+0.04}_{-0.02}}$	$3.03^{+0.02}_{-0.02}$	$2.86^{+0.01}_{-0.01}$
	5	$7.08^{+0.11}_{-0.08}$	$\mathbf{5.97^{+0.10}_{-0.06}}$	$\mathbf{5.88^{+0.02}_{-0.05}}$	$\mathbf{5.82^{+0.10}_{-0.03}}$	$7.05^{+0.11}_{-0.09}$	$6.50^{+0.07}_{-0.06}$
	9	$9.15^{+0.13}_{-0.10}$	$\mathbf{8.95^{+0.09}_{-0.12}}$	$\mathbf{8.87^{+0.12}_{-0.14}}$	$\mathbf{8.84^{+0.16}_{-0.07}}$	$\mathbf{9.04^{+0.15}_{-0.16}}$	$9.28^{+0.18}_{-0.25}$
z (gpm)	1	$15.31^{+0.32}_{-0.19}$	$\mathbf{10.85^{+0.14}_{-0.15}}$	$\mathbf{10.53^{+0.37}_{-0.15}}$	$11.95^{+0.34}_{-0.11}$	$15.07^{+0.30}_{-0.33}$	$12.40^{+0.15}_{-0.19}$
	5	$61.42^{+0.35}_{-0.54}$	$46.95^{+0.96}_{-0.61}$	$\mathbf{45.63^{+0.95}_{-0.39}}$	$\mathbf{44.57^{+0.78}_{-0.46}}$	$62.25^{+1.01}_{-0.42}$	$52.98^{+0.86}_{-0.49}$
	9	$89.40^{+0.42}_{-0.55}$	$84.63^{+0.90}_{-0.90}$	$82.81^{+0.76}_{-0.39}$	$\mathbf{81.06^{+1.12}_{-1.01}}$	$89.97^{+0.81}_{-0.79}$	$88.33^{+2.05}_{-1.32}$

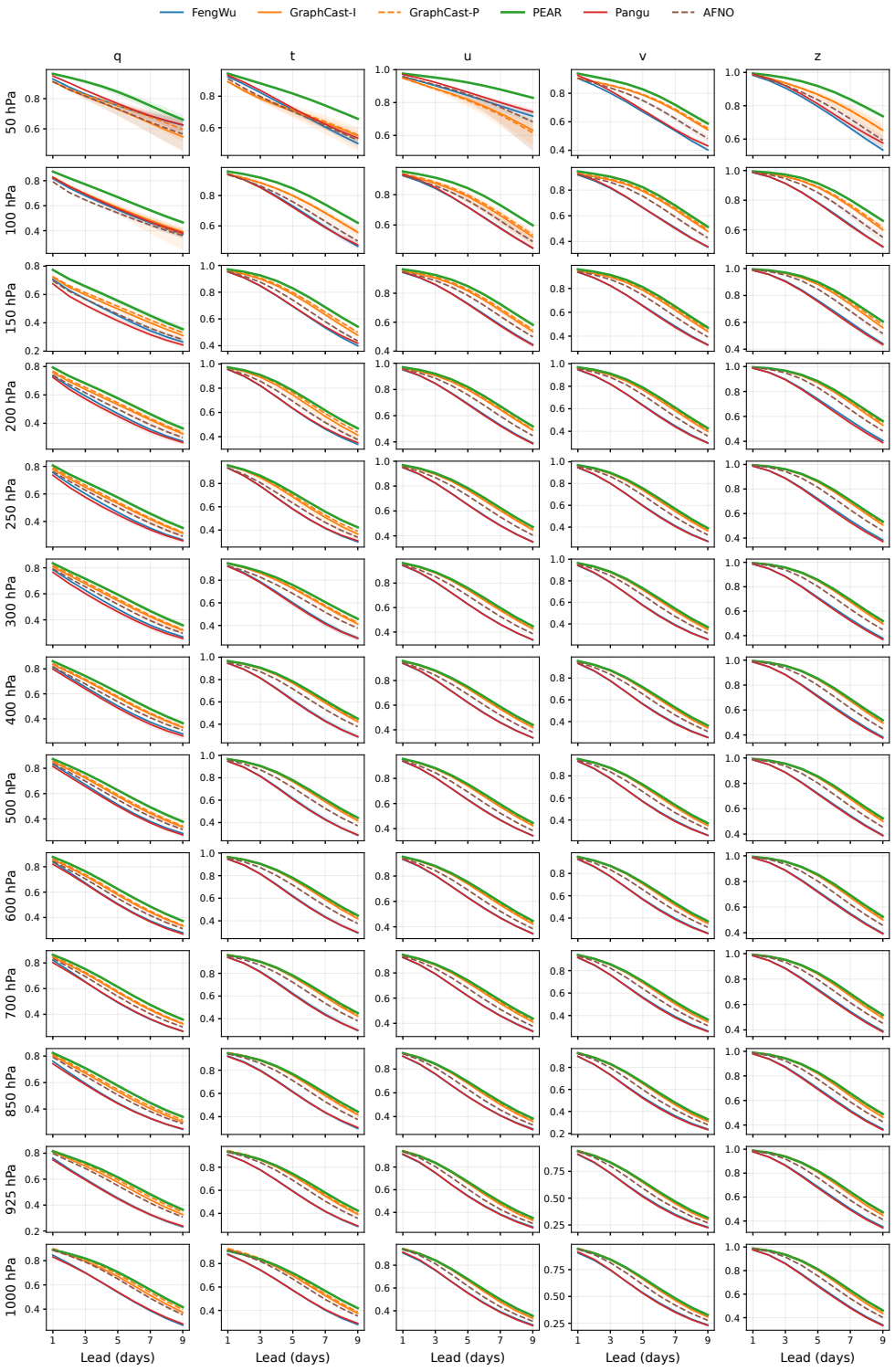


Figure C.1: Native-grid ACC across all upper variables (columns) and pressure levels (rows) on 2018 test data. Curves show five-seed means and bands show min–max ranges. HEALPix and equiangular evaluation use equal-area and cosine-latitude weights, respectively. Higher is better.

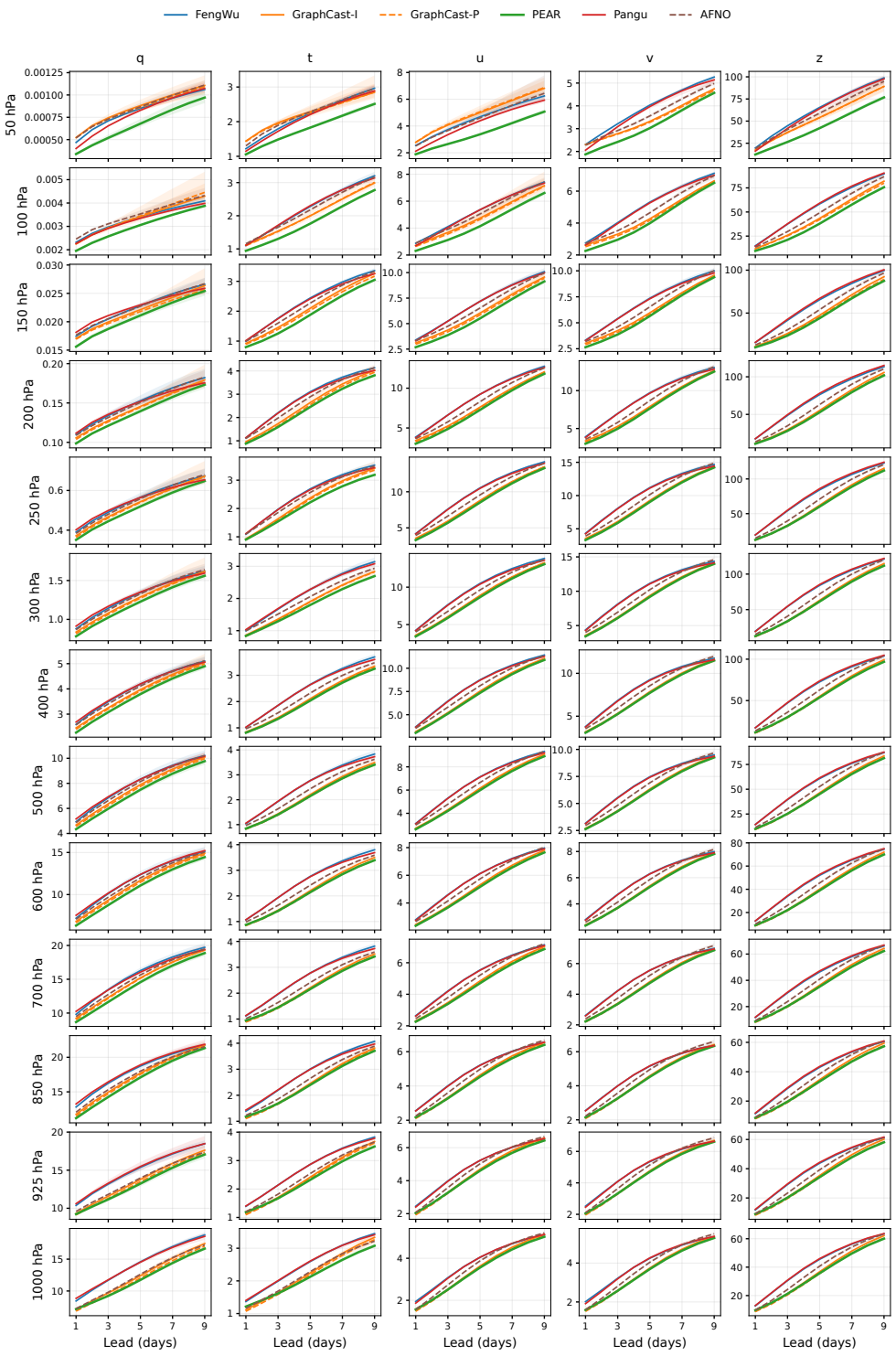


Figure C.2: Native-grid RMSE for the same models, variables, and levels as Figure C.1. Curves show five-seed means and bands min–max ranges. Specific humidity is shown in $10^{-4}\,\mathrm{kg\,kg^{-1}}$ and geopotential height in gpm. Lower is better.

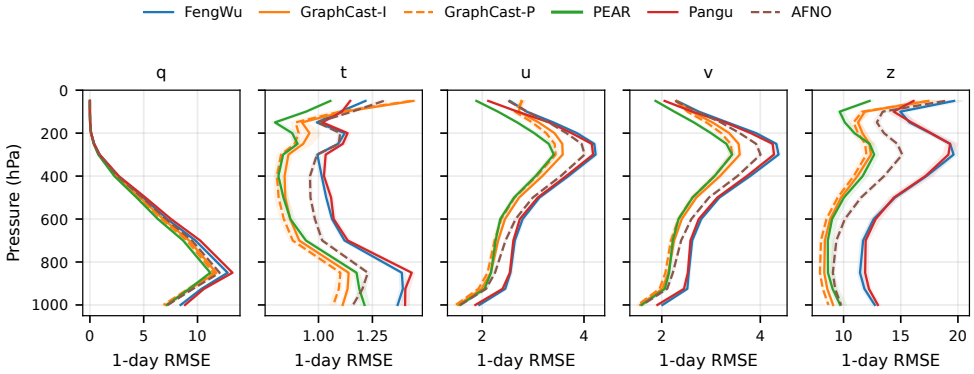


Figure C.3: One-day native-grid RMSE over pressure levels for q, t, u, v, and z. Curves are five-seed means and bands show min–max ranges. Specific humidity is shown in $10^{-4}\,\mathrm{kg\,kg^{-1}}$ and geopotential height in gpm.

D REGIONAL ERRORS

We compare PEAR with cosine-weighted Pangu on common HEALPix at $n_{\text{side}} = 64$. Both training losses weight locations by represented area. We evaluate t2m, msl, z500, t850, and q700 at days 1, 5, and 9, using 364, 360, and 356 forecast start dates in 2018 and five validation-selected seeds per model. Equiangular forecasts use the bilinear interpolation described in Appendix C.

The regions are the tropics ($|\phi| < 30°$), mid-latitudes ($30° \le |\phi| < 60°$), and polar regions ($|\phi| \ge 60°$), combining both hemispheres. For start date t and region R,

$$r_{t,R} = \left[\frac{1}{|R|} \sum_{i \in R} (\widehat{y}_{t,i} - y_{t,i})^2 \right]^{1/2}. \tag{D.1}$$

For each seed, we average over start dates, $\overline{r}_{M,R} = N^{-1} \sum_t r_{t,R}^{(M)}$. Regional ACC restricts the spatial sums in Equation (C.2) to R, retaining the sum over start dates before taking the ratio.

Table D.1 first computes $100(\overline{r}_{\text{PEAR},R}/\overline{r}_{B,R} - 1)$ for each field and seed pairing, with B denoting cosine-weighted Pangu, then averages equally over the five fields and all 25 seed pairs. Negative values favor PEAR. PEAR has lower mean regional RMSE and higher mean ACC in all 45 variable–lead–region comparisons.

Table D.1: Regional RMSE changes (%) for PEAR relative to cosine-weighted Pangu on common HEALPix, averaged equally over five fields. Negative values favor PEAR. Sub- and superscripts give offsets to the min–max range across all 25 seed combinations.

Region	1 day	5 days	9 days
Tropics	$-19.7^{+1.8}_{-1.1}$	$-23.6^{+3.2}_{-2.2}$	$-15.2^{+3.8}_{-3.1}$
Mid-latitudes	$-22.6^{+2.5}_{-1.4}$	$-21.4^{+2.8}_{-2.1}$	$-8.7^{+2.6}_{-2.5}$
Polar	$-13.1^{+3.3}_{-2.3}$	$-14.6^{+3.6}_{-2.7}$	$-6.8^{+4.2}_{-5.5}$

E SPECTRAL QUANTITIES AND NORMALIZATION

We evaluate the six main models on common HEALPix using the five fields, five seeds, and forecast start dates specified in Appendix D. Figure E.1 also includes cosine-weighted Pangu and PEAR-HD8. All spectral calculations use fields in physical units, with z500 denoting geopotential.

Power spectra describe the distribution of variation across spatial scales (Bonavita, 2024). We subtract the spatial mean from each forecast and target separately and denote the resulting fields by $\widehat{y}'_{M,s,t}$ and y'_t, with model M, seed s, and forecast start date t. For each field and lead,

$$\begin{aligned} F^M_{\ell,s} &= \alpha_\ell \left\langle C_\ell(\widehat{y}'_{M,s,t}) \right\rangle_t, & T_{\ell,s} &= \alpha_\ell \left\langle C_\ell(y'_t) \right\rangle_t, \\ D^M_{\ell,s} &= \alpha_\ell \left\langle C_\ell(\widehat{y}'_{M,s,t} - y'_t) \right\rangle_t, & \alpha_\ell &= \frac{2\ell+1}{4\pi}, \end{aligned} \tag{E.1}$$

where $C_\ell = (2\ell+1)^{-1} \sum_m |a_{\ell m}|^2$ and $\langle\cdot\rangle_t$ averages over start dates. The quantities F, T, and D are forecast, target, and error contributions to area-mean variance. We compute D by transforming the forecast-minus-target field. The transform uses `healpy.anafast` (Zonca et al., 2019) with $\ell_{\max} = m_{\max} = 191$, three iterations, uniform pixel weights, and smoothing and pixel-window deconvolution disabled. We show degrees 1–191.

Normalization across fields. For field v, define total represented ERA5 variance $S_{v,s} = \sum_{\ell=1}^{191} T_{\ell,v,s}$. The main-text error spectrum is the equal-field average $\frac{1}{5}\sum_v D^M_{\ell,v,s}/S_{v,s}$, summarized over seeds. Figure E.1 additionally shows $\frac{1}{5}\sum_v F^M_{\ell,v,s}/S_{v,s}$ and the corresponding ERA5 reference. This normalization gives dimensionless values while preserving each degree's contribution to the total. Summing over degrees gives the mean normalized error variance represented by the transform.

Per-field spectra. Figures E.2–E.6 show forecast-power agreement F/T in the top row, absolute forecast-error variance D in the middle row, and normalized error D/T in the bottom row. Values closer to one are better in the top row, while lower values are better in the middle and bottom rows. The error quantities also capture spatial mismatches, and $D/T = 1$ means error power equals target power. Ratios are formed from date-averaged powers within each seed before averaging across seeds. Lines and bands show five-seed means and min–max ranges. We label each degree by the approximate wavelength $2\pi R_\oplus/\ell$, with $R_\oplus = 6371\,\mathrm{km}$, placing small scales on the left.

Summing D over degrees 1–191 gives the forecast-error variance over these spatial scales after removing each field's spatial mean, up to numerical error in the transform. Coarse, intermediate, and fine bands cover degrees 1–20, 21–80, and 81–191.

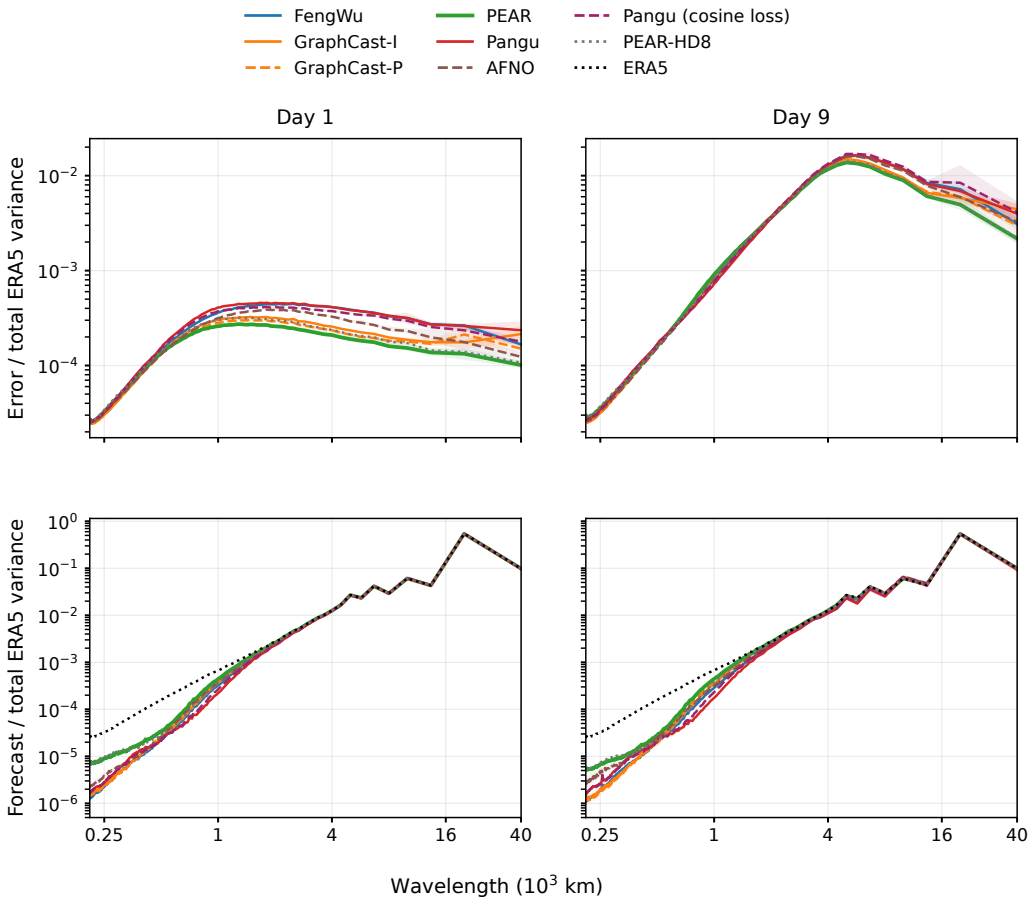


Figure E.1: Spherical spectra on common HEALPix at days 1 and 9 for the six main models and the cosine-weighted Pangu and PEAR-HD8 controls. The top row shows forecast-error variance (lower is better). The bottom row shows forecast variance and the ERA5 reference. Each field is normalized by its total ERA5 variance over degrees 1–191, then the five fields are averaged equally. Lines and bands show five-seed means and min–max ranges. Small scales are on the left.

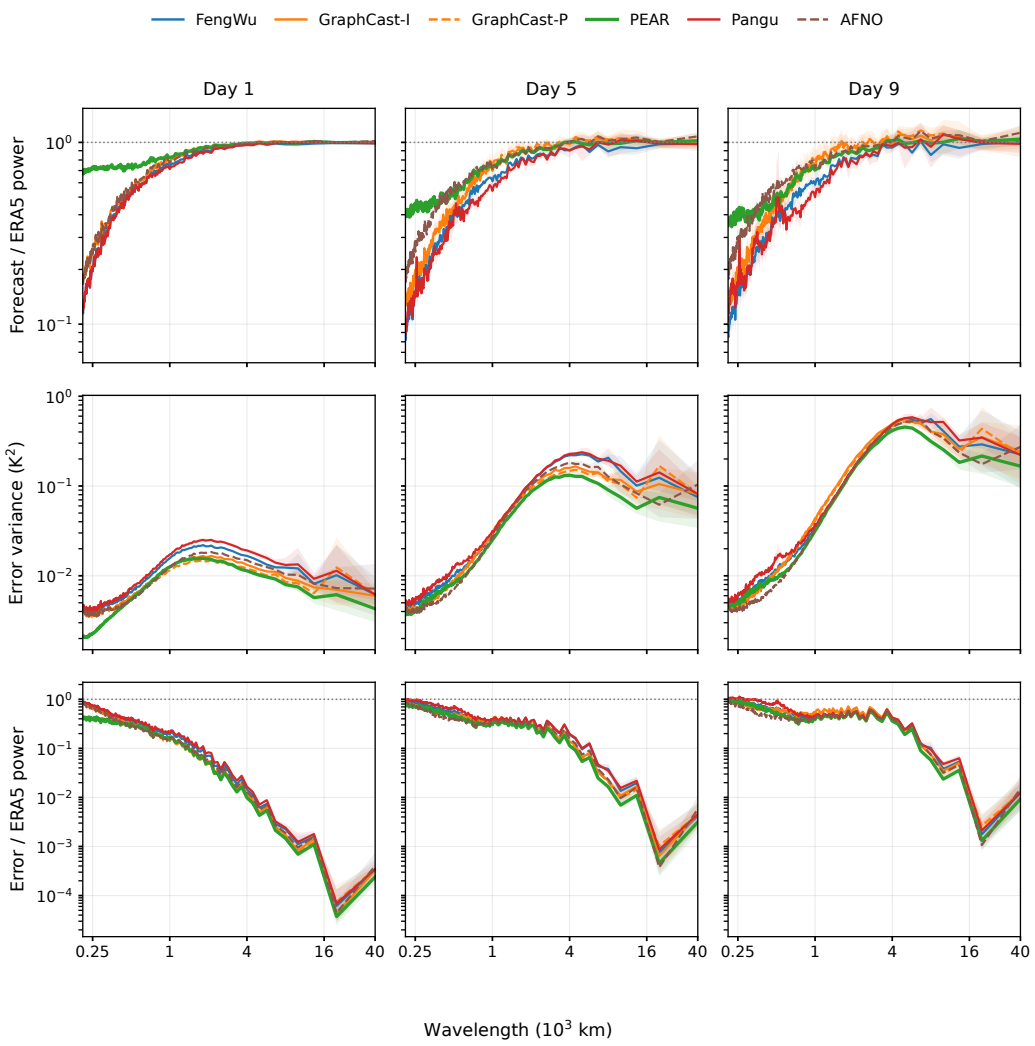


Figure E.2: Full spherical spectra for 2-m temperature at days 1, 5, and 9. Rows show forecast/ERA5 power (top), forecast-error variance in K^2 (middle), and error/ERA5 power (bottom). Values closer to one are better in the top row, while lower values are better in the middle and bottom rows. Each seed averages over 364/360/356 start dates at the respective leads. Lines and bands show five-seed means and ranges. The spatial mean is subtracted from each forecast and target before computing spectra on the common HEALPix grid over degrees 1–191. Wavelength increases to the right, with row limits shared across leads. The same processing applies to Figures E.3–E.6.

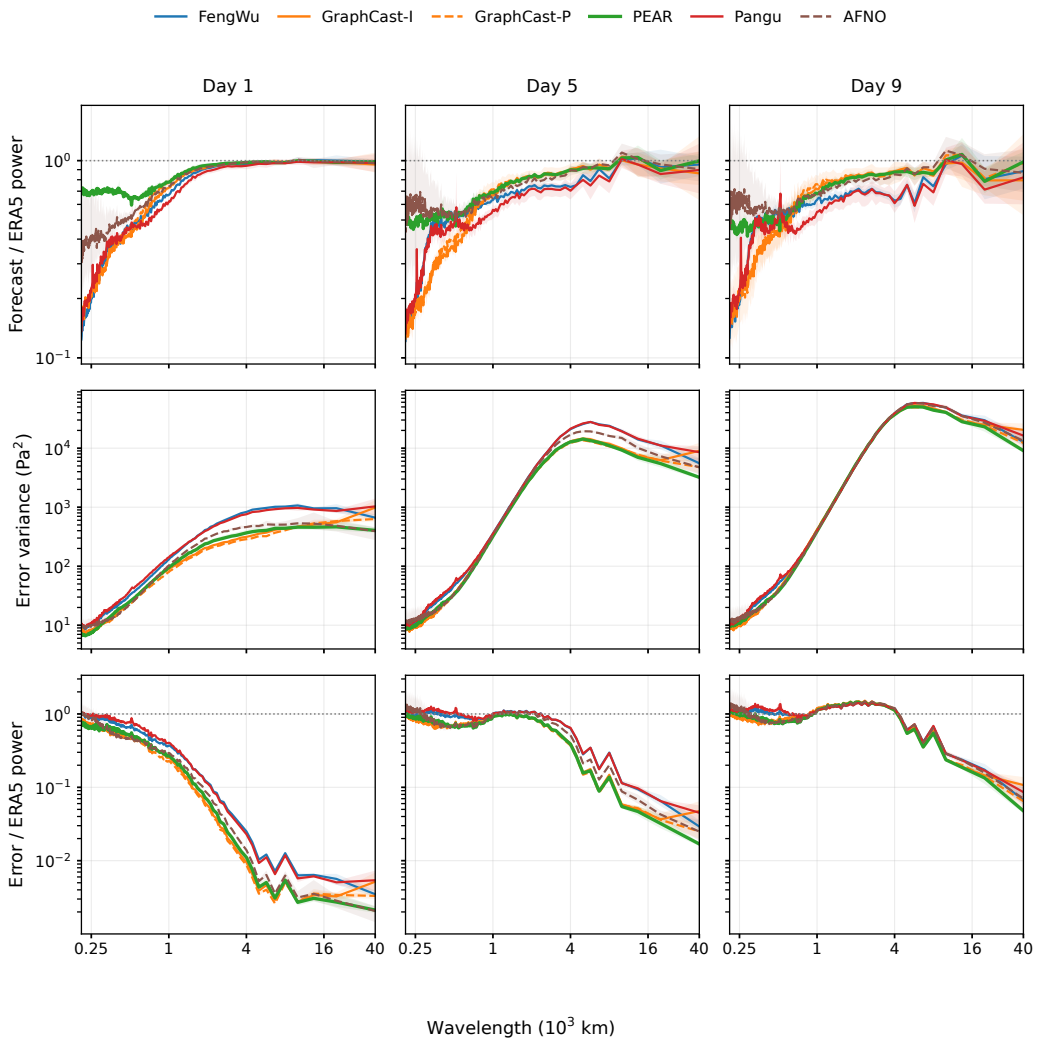


Figure E.3: Mean sea-level pressure spectra with the layout and processing of Figure E.2. Absolute error variance is in Pa^2. Values closer to one are better in the top row, while lower values are better in the middle and bottom rows.

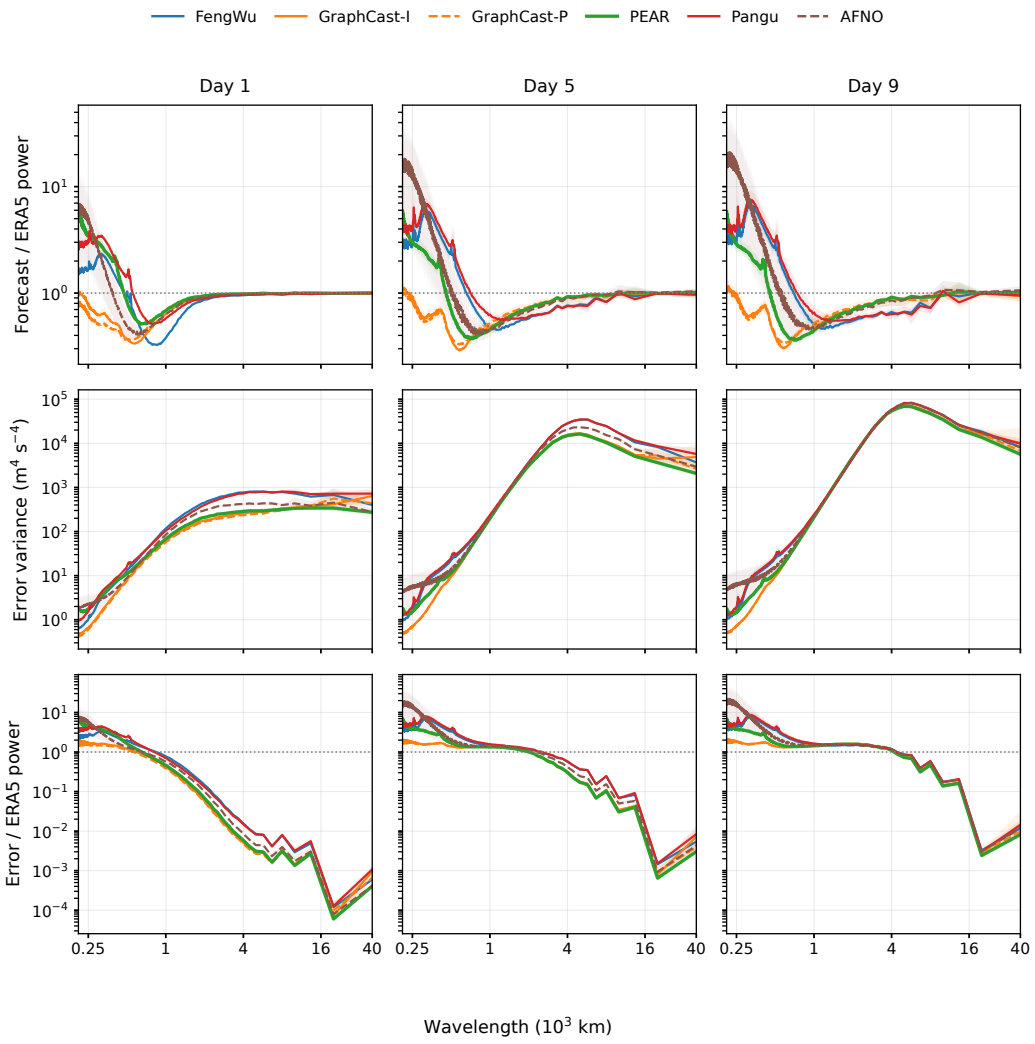


Figure E.4: 500-hPa geopotential spectra with the layout and processing of Figure E.2. Absolute error variance is in $\mathrm{m}^4\,\mathrm{s}^{-4}$. The forecast-power ratio uses a logarithmic axis to retain its full range. Values closer to one are better in the top row, while lower values are better in the middle and bottom rows.

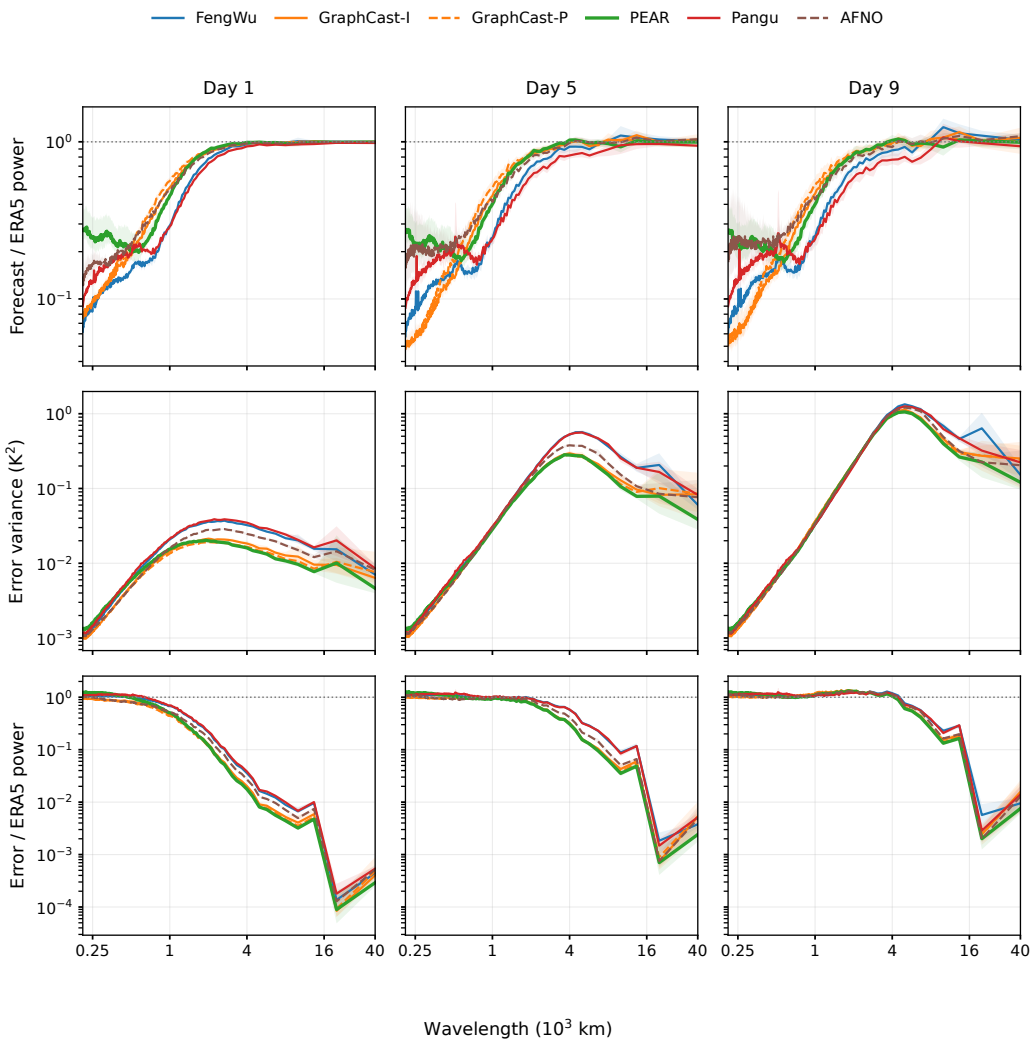


Figure E.5: 850-hPa temperature spectra with the layout and processing of Figure E.2. Absolute error variance is in K^2. Values closer to one are better in the top row, while lower values are better in the middle and bottom rows.

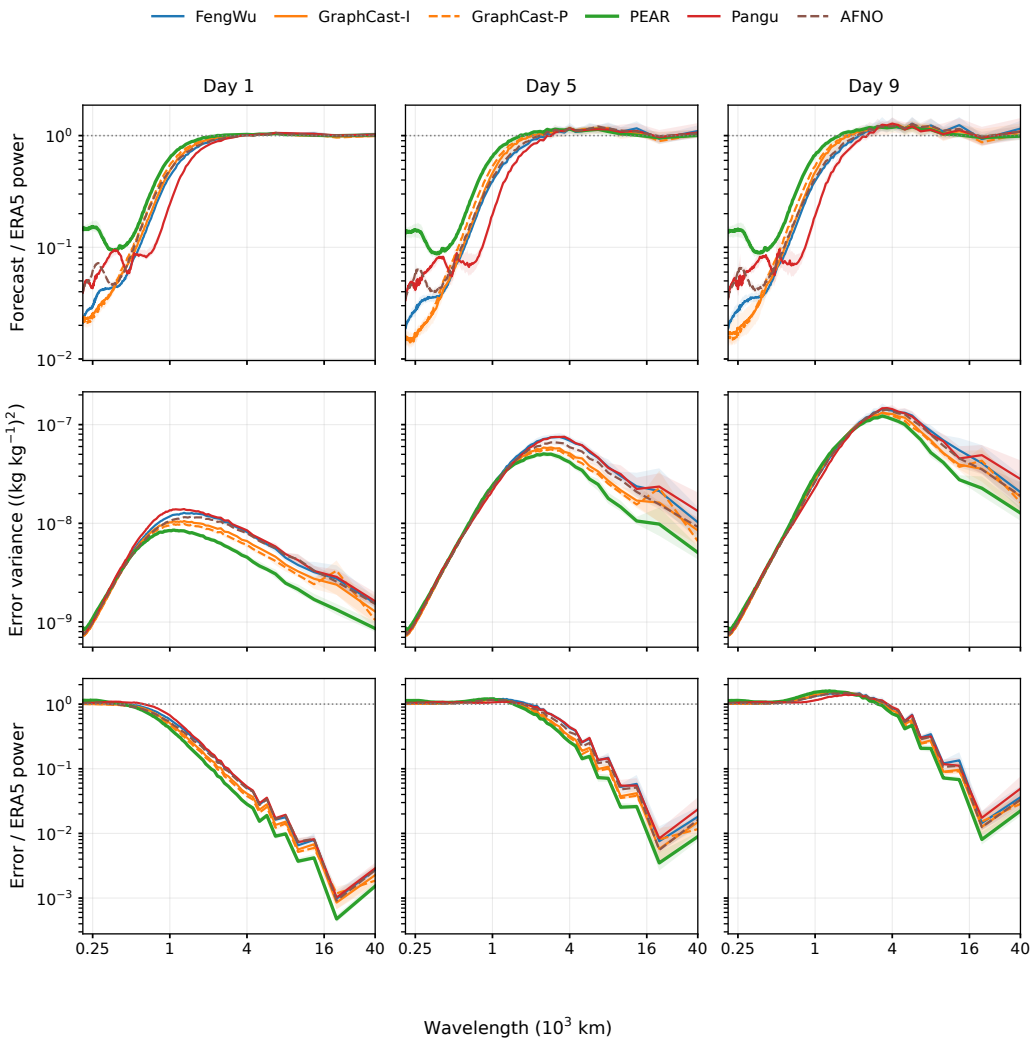


Figure E.6: 700-hPa specific-humidity spectra with the layout and processing of Figure E.2. Absolute error variance is in $(\mathrm{kg\,kg^{-1}})^2$. Values closer to one are better in the top row, while lower values are better in the middle and bottom rows.

F EMPIRICAL ROTATIONAL EQUIVARIANCE

F.1 TRAINING SETS AND DIAGNOSTIC

ERA5-Lite samples the atmosphere at a fixed time of day, whereas two-hourly sampling covers the diurnal cycle. We investigate the learned rotational equivariance under these two sampling regimes.

We compare PEAR configurations trained on ERA5-Lite (midnight UTC samples from 2007–2017, 4,017 samples) and two-hourly ERA5 data from 2012 (4,380 samples). Both predict the state 24 hours ahead. We evaluate the 200-epoch checkpoints on 2019 validation data sampled in the same way as the corresponding training set.

For rotation R_θ about the polar axis, exact equivariance would require $f(R_\theta x) = R_\theta f(x)$. We measure the discrepancy

$$\mathcal{E}_v(f,\theta) = \mathbb{E}_{x\in\text{validation}}\left[\frac{1}{N_{\text{pix}}}\sum_i |f(R_\theta x)_{v,i} - (R_\theta f(x))_{v,i}|\right], \tag{F.1}$$

separately for each field, sweeping θ from 0° to 360° in 3° increments.

F.2 RESULTS AND INTERPRETATION

Figures F.1 and F.2 show the angle-dependent errors. The two sampling setups give similar errors for surface temperature and mean sea-level pressure, while the ERA5-Lite model has lower errors for the surface wind components. Upper-air differences vary with field and level (Figure F.3). Error generally decreases with altitude, with variable-specific departures from this trend. Both trained models lie well below the untrained reference in the upper-variable summary in Figure 6a.

Equivariance error is lowest near multiples of 90° and highest near odd multiples of 45°, consistent with the fourfold polar-axis structure of HEALPix sampling. At arbitrary angles, the diagnostic includes the effects of representing rotated fields on the grid. The two data setups also differ in year coverage, temporal correlation, and validation sampling.

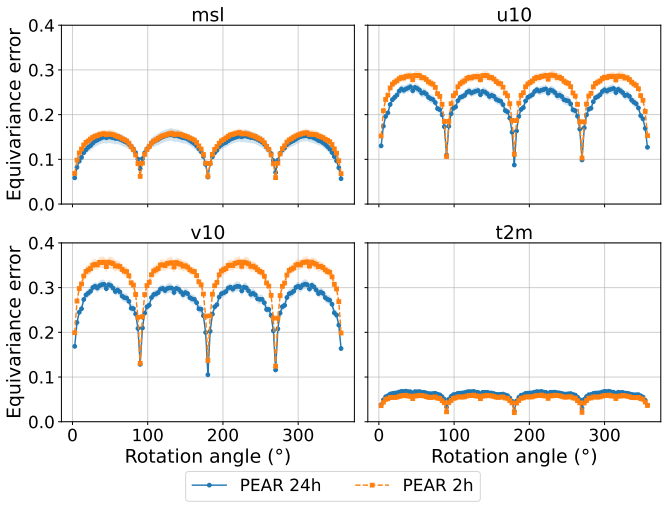


Figure F.1: Mean equivariance error versus rotation angle for the four surface variables after 200 training epochs. The curves compare ERA5-Lite with two-hourly sampling of 2012.

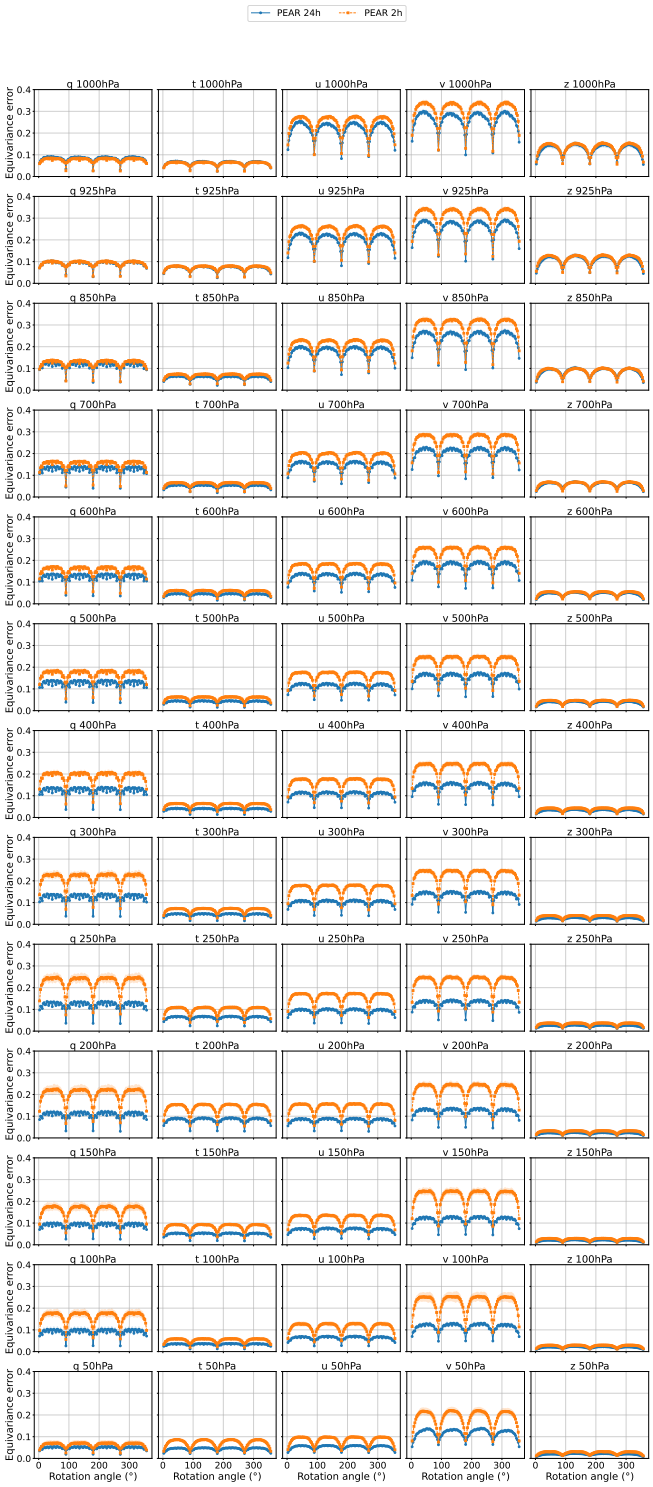


Figure F.2: Mean equivariance error versus rotation angle for upper-air variables at each pressure level, comparing the same two configurations as Figure F.1.

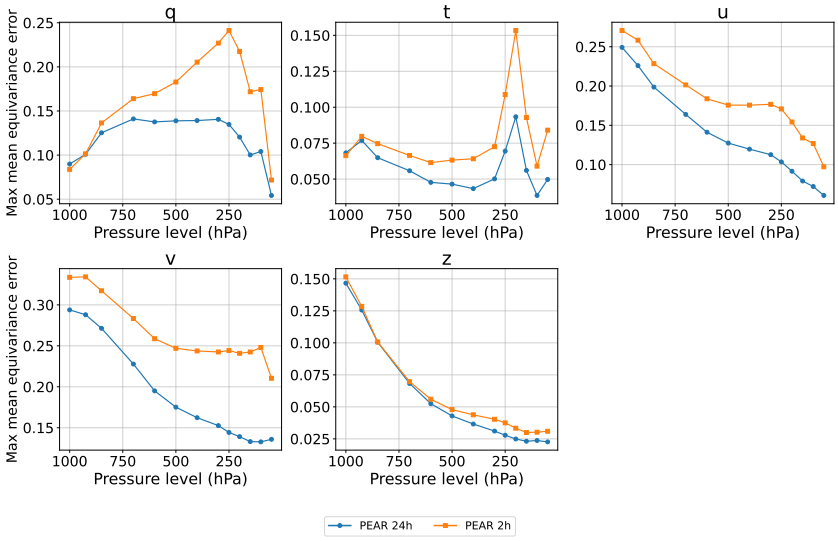


Figure F.3: Maximum mean equivariance error over rotation angles, shown separately for each upper-air variable and pressure level.

G SINGLE-STEP CLIMATE-MODEL EMULATION

A related but distinct challenge that has recently attracted machine learning interest is climate model emulation. Unlike medium-range weather forecasting, where the task is to predict the future atmospheric state from initial conditions, emulation aims to approximate the outputs of computationally intensive Earth System Models (ESMs) for a given set of inputs. ESMs require substantial computational resources and individual runs can take months, severely limiting the number of scenarios that can be explored (Watson-Parris, 2021). To support machine learning research in this area, several dedicated datasets have been developed such as ClimateBench (Watson-Parris et al., 2022), ClimateLearn (Nguyen et al., 2023b), Climart (Cachay et al., 2021) and ClimateSet (Kaltenborn et al., 2023). On the modeling side, a range of architectures have been explored, including Convolutional LSTMs (Watson-Parris et al., 2022), UNet models (Kaltenborn et al., 2023) and transformer architectures such as ClimaX (Nguyen et al., 2023a).

We evaluate PEAR on single-step climate-model emulation, predicting surface temperature and precipitation from emission maps. We use ClimateSet, a large-scale, machine-learning-ready dataset containing processed outputs for 21 different climate models from the CMIP6 ensemble paired with a shared set of forcing inputs (Input4MIPS). The input dataset is a processed collection of monthly emission maps for four primary climate forcing agents (CO2, SO2, CH4, and BC), and the output variables are global, monthly averaged maps of surface air temperature and precipitation at a resolution of $2.5° \times 1.875°$— significantly coarser than the ERA5 data on which the original PEAR model was trained. The data covers a historical scenario (1850–2014) and four future Shared Socioeconomic Pathways (SSPs): SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5, covering the period 2014–2100. For use with PEAR, the dataset has been discretized onto the HEALPix grid with $n_{\text{side}} = 32$.

PEAR, UNet, and ClimaX treat each time step independently in this one-step comparison.

The PEAR architecture has been adapted to handle the single set of surface-level variables, with 4 input channels for the climate forcing agents. The architecture is simplified to a single-branch patch embedding using only a 1D convolution, collapsing the depth dimension to $D = 1$ throughout all transformer stages. A single output head then projects back to the desired number of output channels. In this configuration, the model still has 4.3M trainable parameters.

We evaluate PEAR against two baselines from the ClimateSet benchmark, a UNet with 18.3M trainable parameters and a ClimaX model with 106.6M trainable parameters. Each model is trained separately from scratch on 15 of the 21 available climate models with 5 random seeds, and evaluated on the held-out test set of the corresponding climate model. Training follows the original ClimateSet experimental setup, using the SSP1-2.6, SSP3-7.0, and SSP5-8.5 scenarios with 10% of the data held out for validation, and testing on the SSP2-4.5 scenario. The reported metric is the RMSE computed on normalized labels, averaged over temperature and precipitation.

Figure G.1 shows the per-variable and overall RMSE for each model across the 15 climate models and 5 seeds. In terms of overall RMSE, ClimaX achieves the lowest error on 10 of the 15 climate models, followed by UNet on 3 and PEAR on the remaining 2. Despite being roughly four times smaller than UNet, PEAR outperforms it on 13 of the 15 models in overall RMSE. Separating the results by variable, ClimaX performs best on surface temperature for 12 of the 15 climate models, with UNet performing best on the remaining 3. For precipitation, however, the trend differs with PEAR achieving the lowest RMSE on 8 of the 15 climate models, compared with 4 for ClimaX and 3 for UNet.

It is unsurprising that ClimaX outperforms the other models given its considerably larger parameter count. More notable is that PEAR generally matches or exceeds UNet across both variables despite being roughly four times smaller, and achieves the lowest precipitation RMSE for a majority of the climate models. Although PEAR does not achieve the same relative advantage for surface temperature as it does for precipitation, its strong performance relative to its size leaves open the possibility that further experimentation with model capacity and architectural refinements could improve results on temperature as well. Overall, these results demonstrate that PEAR is well-suited for the task of climate model emulation, achieving competitive performance despite having a substantially smaller parameter count.

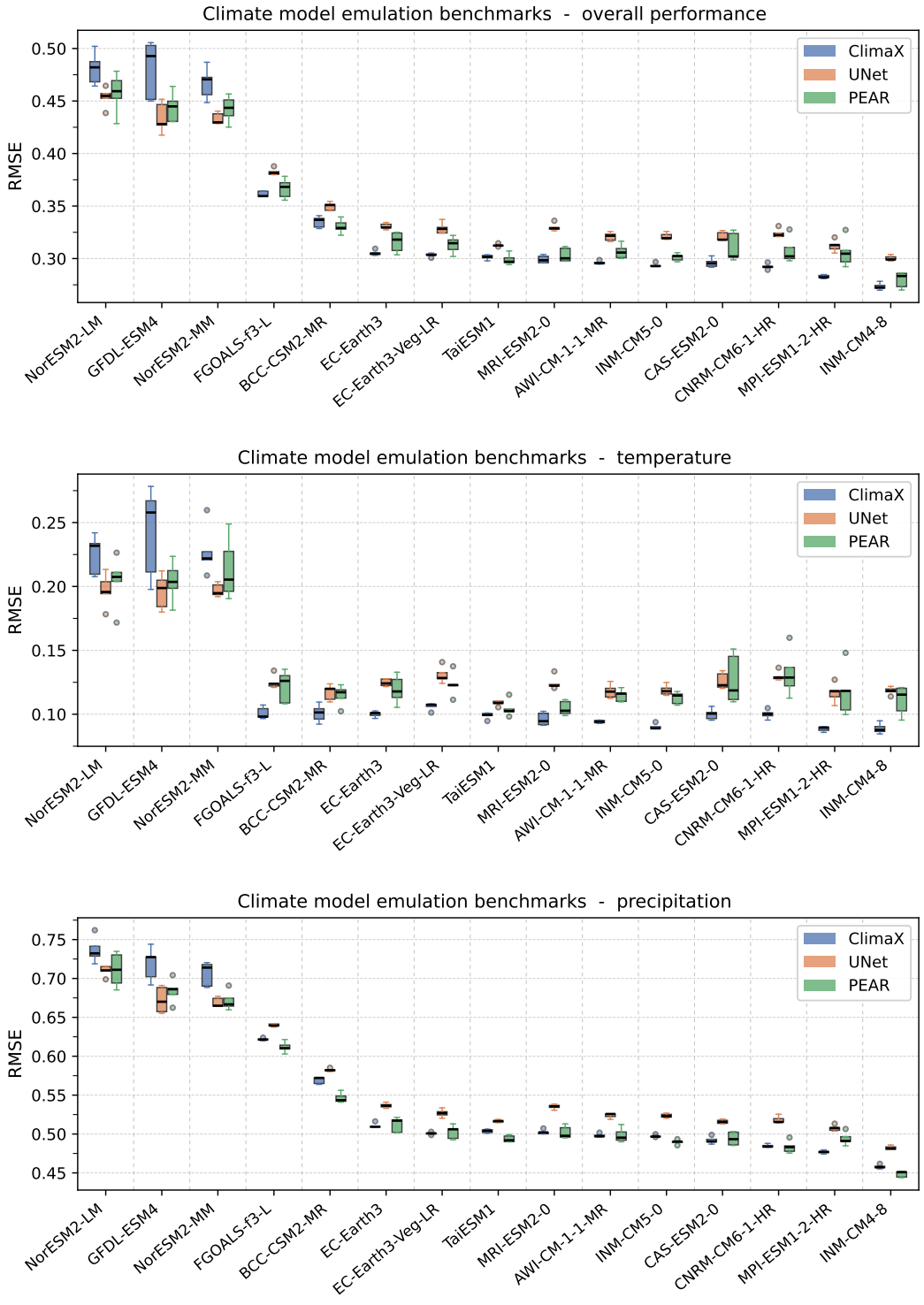


Figure G.1: Benchmarking results for climate model emulation. Box plots showing RMSE on normalized outputs for ClimaX, UNet, and PEAR across 15 ClimateSet climate models. Plot from top to bottom showing (a) Overall performance. (b) Surface air temperature. (c) Precipitation.