PEAR: Equal Area Weather Forecasting on the Sphere

Hampus Linander [1, 2] **Christoffer Petersson** [2,3] **Daniel Persson** [2] **Jan E. Gerken** [2]

[1] VERSES AI
Los Angeles, USA

[2] Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg, Sweden

[3] Zenseact
Gothenburg, Sweden

Abstract

Artificial intelligence is rapidly reshaping the natural sciences, with weather forecasting emerging as a flagship AI4Science application where machine learning models can now rival and even surpass traditional numerical simulations. Following the success of the landmark models Pangu Weather and Graphcast, outperforming traditional numerical methods for global medium-range forecasting, many novel data-driven methods have emerged. A common limitation shared by many of these models is their reliance on an equiangular discretization of the sphere which suffers from a much finer grid 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, removing unphysical biases. Motivated by a growing support for this grid in meteorology and climate sciences, we propose to perform weather forecasting with deep learning models which natively operate on the HEALPix grid. To this end, we introduce Pangu Equal ARea (PEAR), a transformer-based weather forecasting model which operates directly on HEALPix-features and outperforms the corresponding model on an equiangular grid without any computational overhead.

https://github.com/hlinander/PEAR-Weather

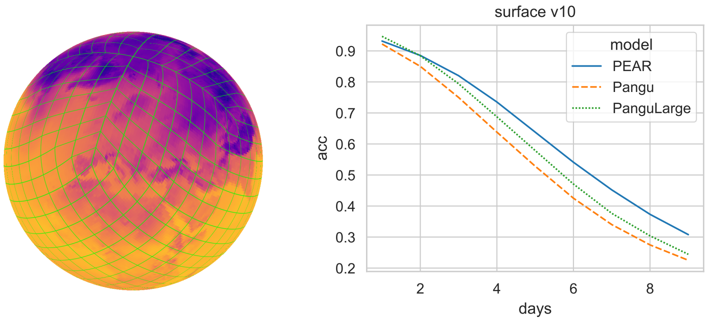

Figure 1: Left: Predicted surface level temperature from PEAR. Green lines show the HEALPix cell boundaries at 3 levels of course-graining above the model resolution. Right: Anomaly correlation coefficient (ACC) (higher is better) for surface level northward wind component with forecast horizon up to 10 days. PEAR outperforms the almost 8 times larger PanguLarge at longer forecast horizons.

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 [1, 2, 3]. Weather forecasting stands at the forefront of this shift [4, 5, 6, 7, 8], driven by advances in architectures that can capture multiscale atmospheric dynamics directly from data, while offering orders of magnitude speedups compared to traditional numerical

Published in the AI for Science workshop (NeurIPS 2025). Correspondence: hampus.linander@chalmers.se & gerken@chalmers.se.

integration methods. Yet, most existing AI-based forecasting systems inherit discretizations from the numerical models they aim to complement or replace, such as the equiangular latitude-longitude and Gaussian grids, which imposes non-uniform resolution and unphysical biases. In this paper we explore the Hierarchical Equal Area iso-Latitude Pixelization (HEALPix) as a native grid for learning-based weather forecasting, equal area modeling across the globe.

Following the publication of the landmark models Pangu-Weather [9] and Graphcast [4], the recent emergence of performant data-driven models for weather forecasting has been driven by the widespread use of the ERA5 reanalysis dataset [10], a comprehensive open access global climate data record produced by the European Center for Medium-Range Weather Forecasts (ECMWF).

ERA5 provides hourly estimates of atmospheric, land, and oceanic variables on a global scale, discretized using the regular equiangular latitude-longitude grid or a reduced Gaussian grid. Both of these grids introduce challenges when used in physics-based ML models. Specifically, both grids exhibit non-uniform resolution across latitudes, which can lead to artifacts and unphysical inductive biases in learning-based models.

To address these limitations, alternative spherical representations have been explored. One notable example is HEALPix, originally developed in astrophysics for uniform pixelization of the celestial sphere [11]. HEALPix offers equal-area pixel distributions and hierarchical resolution capabilities, making it well-suited for learning tasks on spherical domains. Recent efforts have begun to investigate the use of HEALPix for resampling next-generation weather and climate data [12, 13]. HEALPix has also been used in other machine learning contexts where spherical data is natural [14, 15].

The ECMWF are also targetting the HEALPix grid in its Destination Earth (DestinE) initiative [16, 17], which aims to develop a digital twin of the Earth for improved climate and weather predictions. By employing HEALPix, DestinE seeks to achieve high-resolution, globally consistent datasets that can better inform decision-making processes related to climate change adaptation and disaster risk management. Furthermore, ECMWF's transition from GRIB1 to GRIB2 data formats supports the integration of advanced grid systems like HEALPix [18], enabling more detailed and efficient data representation essential for next-generation forecasting models.

Motivated by these developments, we propose PEAR (Pangu Equal Area), the first machine learning weather forecasting model that operates entirely on the HEALPix grid. Unlike previous approaches that project onto planar grids or only use HEALPix for preprocessing, PEAR natively represents inputs, internal features, and outputs on the spherical HEALPix domain, enabling consistent resolution across the globe and better alignment with the underlying physical symmetries. We use ERA5 data, resampled to HEALPix, as the forecasting target and evaluate our model's performance on key atmospheric variables across various resolutions.

Our main contributions are:

- Motivated by an increased adoption of the HEALPix grid in next generation weather and climate data, we propose to use the spherical HEALPix grid as the native grid for machine learning weather predictions. This approach eliminates unphysical biases in the sampling of the sphere affecting standard equiangular based weather prediction models and removes the need for spatial weights in the loss and evaluation metrics.
- We introduce PEAR: a baseline model for neural weather simulation on HEALPix using a volumetric transformer architecture which operates at no computational overhead compared to an equivalent model on the traditional equiangular grid.
- To demonstrate the advantage of using a native HEALPix model, we show that PEAR's HEALPix-predictions outperform those produced by the same architecture operating on an equiangular grid. We evaluate all models for lead times of up to ten days and show that this advantage persists for longer forecasting horizons, even though the equiangular-based model has more than twice as many parameters. PEAR also outperforms Pangu-Large with almost eight times as many parameters at a forecast horizon of 5 days and beyond.

2 Related work

The field of machine-learning weather forecasting has received tremendous attention over the last years both for medium-range weather forecasting [19] and extreme weather prediction [20]. The first model whose performance on global medium-range forecasts surpassed that of numerical models was Pangu-Weather [9] which is based on a volumetric version of the SWIN transformer [21]. Since then, a number of models have been published which improved upon this baseline, such as GraphCast [4], FuXi [5], FengWu [6], NetMet-3 [7], Stormer [8] or the ECMWF's data-driven forecasting system [22]. Fourier neural operators have been used in a number of models in this domain [23, 24] as well. In order to take the curvature of the sphere into account, one model [25] used Spherenet. Similarly, the forecasting model CirT [26] is based on a circular transformer which takes the azimuthal circularity of the sphere into account. Probabilistic weather forecasting models allow for uncertainty estimation of the generated predictions [27]. The model GenCast creates an ensemble of stochastic predictions which outperforms the top operational medium-range weather forecast in the world, ENS, the ensemble forecast of the European Centre for Medium-Range Weather Forecasts [28]. The consideration of physical conservation laws in the training process can improve these data-driven weather prediction models [29]. Closely related to machine-learning weather forecasting systems are neural network models which predict the climate [30, 31] or general-purpose models for the earth system [32]. Contributions to the training setup include careful ablations of various aspects of the architecture [33] and a training platform for deep-learning based weather prediction models [34]. WeatherBench 2 [35] provides a well-established benchmark for machine-learning weather prediction models. Deep neural networks have also been applied in the post-processing of global weather forecasts [36]. A hybrid model of machine-learning components and a differentiable solver for circulation models was proposed in [37].

The aforementioned models have in common that they use the naive equiangular discretization of the sphere as input. The resulting grid is considerably denser towards the poles than around the equator. In contrast, the HEALPix grid is uniform over the entire sphere. The convolutional DLWP-HPX model [12] uses representations on the HEALPix grid. HEAL-ViT [13] is a vision-transformer based model which operates on the HEALPix grid and uses a learnable encoder and decoder to map the features from the equiangular grid to HEALPix and back. In contrast, our model operates entirely on HEALPix, in line with our proposal to use the HEALPix grid as the physically appropriate grid for weather forecasting.

3 Background

3.1 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.

For conversions between the ring- and nested indexing of the pixels and to retrieve the pixel positions in spherical coordinates, we use the Python bindings for the HEALPix package provided by `chealpix`.

3.2 ERA5 on HEALPix

The ERA5 dataset [10] contains the global atmospheric state in terms of a number of hydrodynamical quantities, discretized on an equiangular gridding on the sphere, over multiple vertical slices and hourly in time. Following Pangu-Weather [9], we distinguish between surface variables (temperature, wind velocity, pressure, humidity) and upper variables (temperature, wind velocity, humidity, pressure).

The raw data provided by ECWMF has an angular gridding of $0.25°$ resulting in a spatial $(\mathrm{lon}, \mathrm{lat})$ resolution of $(1440, 721)$. 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 $1.9 \cdot 10^{-5}\mathrm{rad}^2$. We target a HEALPix grid with $n_{\mathrm{side}} = 64$, 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.

3.3 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 [9]. This subset consists of the years 2007 to and including 2017, with 2019 used for validation. All samples are at 00:00UTC, resulting in a total of 4017 training samples and 365 validation samples. Together with cached normalized data, the total dataset size is about 3TB.

4 PEAR: Pangu Equal Area

The weather forecasting task is formulated as a regression problem, where the input is the global weather state at time t and the output is the global weather state at time $t + \Delta t$. Here we use a time delta of 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 [9], 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 $(12n_{\mathrm{side}}^2, 4)$ and $(12n_{\mathrm{side}}^2, 13, 5)$.

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 3 for a schematic overview of the architecture, and Appendix A Table A1 for details on the layer parameters.

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 48 channels. The patch embedded tensors have shape $(^3/_4 n_{\mathrm{nside}}^2, 48)$ and $(^3/_4 n_{\mathrm{nside}}^2, 7, 48)$. At this point the patch embedded surface variables are concatenated to the patch embedded upper variables, resulting in a single tensor of shape $(^3/_4 n_{\mathrm{nside}}^2, 8, 48)$.

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 $(^3/_4 n_{\mathrm{nside}}^2, 8, 48)$, partitioning using a window size of (W_{hp}, W_d) gives a windowed tensor of shape $(\frac{^3/_4 n_{\mathrm{nside}}^2}{W_{\mathrm{hp}}} \frac{8}{W_{\mathrm{d}}}, W_{\mathrm{hp}} W_d, 48)$. Attention is now performed over the embeddings of the $W_{\mathrm{hp}} W_d$ voxels

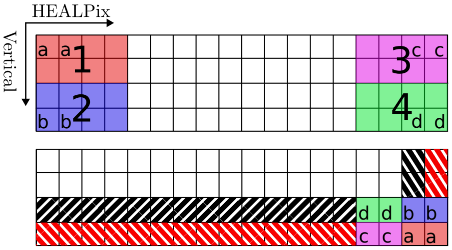

Figure 2: Shift and corresponding mask for windowed attention. Illustration of a scalar tensor with the 1d HEALPix index in the horizontal direction and the vertical direction corresponding to the discretized vertical direction above the surface. Ring indexing is used in the HEALPix direction, with the north pole to the left, and south pole to the right. The lower grid shows where the indicated voxels from the top grid end up after a negative shift of half a window in both directions. Note that the voxels from region a, b, c and d are all from spatially separated regions, and thus the window in the lower right needs to be masked accordingly. The colored striped regions indicate the mask for windows along the borders, where each window contains two regions instead of four as in the corner window.

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 A1 for an illustration of the patches and first level windows corresponding to the tensor structure at the first attention layer.

In contrast to Pangu [9], we use a simplified learned relative positional embedding. Because of the equal area grid cells in the HEALPix grid we can share relative positional embeddings between windows, and thus simply 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 embedding dimension. The simplified positional embedding accounts for most of the parameter savings compared to Pangu in Table 1.

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 [15], and a simple shift in the vertical direction. Since this shifting is performed cyclically, there 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 2 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 [38] using layer-norm and skip connection.

4.3 Downsampling and upsampling

To facilitate a bottleneck structure, we follow Pangu [9] 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.

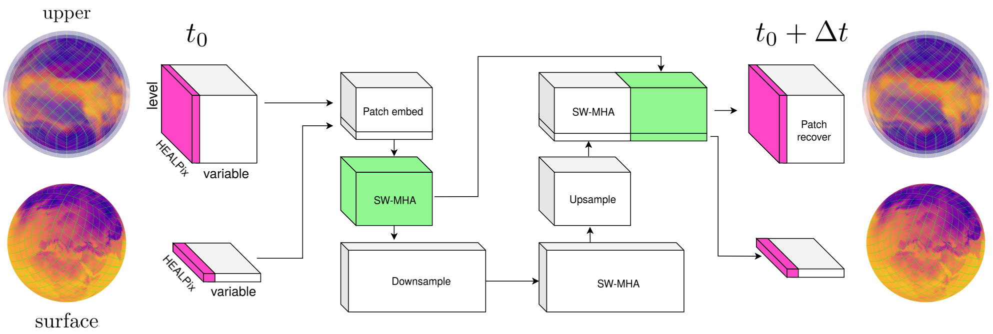

Figure 3: PEAR architecture schematic. Violet slices correspond to the variables visualized on the sphere for the input and output tensors. Each block indicates the tensor shape after the layer with the corresponding name. Patch embedding by convolution, shifted windowed multi-head attention (SW-MHA) 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 course-graining above the model resolution.

4.4 Patch recovery

To recover the surface and upper variable tensors, we use transpose convolutions on the first and remaining vertical levels correspondingly. The latent tensor x of shape $(3/4 n_{\text{side}}^2, 8, 48)$ is split along the second dimension into a surface latent x_{surface} of shape $(3/4 n_{\text{side}}^2, 1, 48)$, and an upper latent tensor x_{upper} of shape $(3/4 n_{\text{side}}^2, 7, 48)$. 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 [39] using the ERA5-lite subset described in Section 3.3. All experiments are carried out using $n_{\text{side}} = 64$ to saturate the angular resolution of the HEALPix grid using the available ERA5 data on the equiangular grid. We train all models using L1 loss, with the loss contribution from the surface variables weighted by $\frac{1}{4}$. We use the AdamW optimizer with weight decay 3×10^{-6} and learning rate 5×10^{-4}. On a single A100, PEAR converges in 20 hours and Pangu in 40 hours.

To evaluate the medium-term forecasting ability of PEAR, we perform iterated model inference up to 10 times, resulting in forward time predictions of up to 10 days. At each lead time we calculate the average RMSE and anomaly correlation coefficient (ACC) [40] of all variables over the globe. See Appendix A for details.

To evaluate baseline-predictions on an equiangular grid, we apply the latitude weighting used in prior work [9]. The equal area grid cells of HEALPix make this reweighting redundant for PEAR. The ACC measures the correlation between deviations from the climatology mean of predicted and ground truth forecasts, with a value of 1 indicating perfect agreement [10]. The climatology average is subtracted to factor out seasonal variations that would otherwise improve the raw correlation between predictions and ground truth data. To calculate the climatology mean, we average each day of the year over the 11 years in the ERA5-lite dataset for every variable.

Pangu and Pangu-Large provide baselines against the same type of architecture where an equiangular grid is used for input, latent and output tensors. Since the overall architectural blocks are the same, we can match the depth, number of attention heads and embedding dimensions for each block. Pangu and PEAR share the same architecture hyperparameters, whereas Pangu-Large has 4 times larger embedding dimension. For Pangu and Pangu-Large we use a equiangular grid with $n_{\text{lon}} = 314$ and $n_{\text{lat}} = 157$, resulting in a relative difference of the total number of pixels of 0.3%, or 146 pixels.

Table 1: Model size and inference times for PEAR, Pangu and Pangu-Large. Inference times measure a forward pass through the models with input on the GPU. Mean and standard deviation over 100 iterations after warm-up.

Model	Trainable parameters (M)	Inference time (ms)
PEAR	4.3	17 ± 0.12
Pangu	11.4	25 ± 0.05
Pangu-Large	33.7	54 ± 0.02

Table 2: Average anomaly correlation coefficient (ACC) and average root mean squared error (RMSE) for three different prediction lead times (1, 3 and 5 days). Bold face indicate best model for the corresponding metric.

Variable	Δt	ACC		RMSE		unit
	(days)	PEAR	Pangu	PEAR	Pangu	
msl surface	1	**0.984**	0.979	**84.6**	97.9	Pa
	3	**0.924**	0.860	**175**	230	
	5	**0.790**	0.675	**282**	355	
t2m surface	1	**0.886**	0.876	**0.792**	0.846	K
	3	**0.779**	0.728	**1.11**	1.31	
	5	**0.650**	0.551	**1.41**	1.77	
u10 surface	1	**0.930**	0.923	**1.05**	1.1	$\frac{m}{s}$
	3	**0.819**	0.759	**1.59**	1.8	
	5	**0.647**	0.557	**2.17**	2.42	
v10 surface	1	**0.932**	0.922	**1.07**	1.13	$\frac{m}{s}$
	3	**0.820**	0.750	**1.64**	1.87	
	5	**0.638**	0.528	**2.26**	2.54	
q upper	1	**0.824**	0.792	$\mathbf{2.87 \times 10^{-4}}$	3.1×10^{-4}	$\frac{g}{kg}$
	3	**0.706**	0.647	$\mathbf{3.69 \times 10^{-4}}$	4.09×10^{-4}	
	5	**0.584**	0.510	$\mathbf{4.52 \times 10^{-4}}$	5.03×10^{-4}	
t upper	1	**0.945**	0.938	**0.756**	0.809	K
	3	**0.877**	0.822	**1.11**	1.3	
	5	**0.763**	0.664	**1.51**	1.77	
u upper	1	**0.944**	0.937	**2.04**	2.17	$\frac{m}{s}$
	3	**0.866**	0.812	**3.12**	3.6	
	5	**0.739**	0.648	**4.31**	4.92	
v upper	1	**0.941**	0.931	**2.02**	2.16	$\frac{m}{s}$
	3	**0.859**	0.787	**3.07**	3.61	
	5	**0.717**	0.595	**4.26**	4.92	
z upper	1	**0.991**	0.985	**79**	98.8	gpm
	3	**0.949**	0.889	**175**	251	
	5	**0.850**	0.734	**291**	393	

Figure 4 shows the ACC for all surface and upper variables, with the ACC for the upper variables averaged over the 13 vertical levels. PEAR shows superior forecasting ability compared to Pangu, and is better (msl, u10, v10, t, u, v, z) or comparable (t2m) to the Pangu-Large model with 4 times larger embedding dimension. Table 1 lists model sizes and inference times. PEAR shows 1.5 times faster inference than Pangu and 3.2 times faster inference than PanguLarge.

Table 2 lists the average ACC and RMSE for three different prediction lead times. PEAR consistently outperforms Pangu at longer lead times, and is often better than PanguLarge with 4 times larger embedding dimension. Additional evaluations are provided in the appendix.

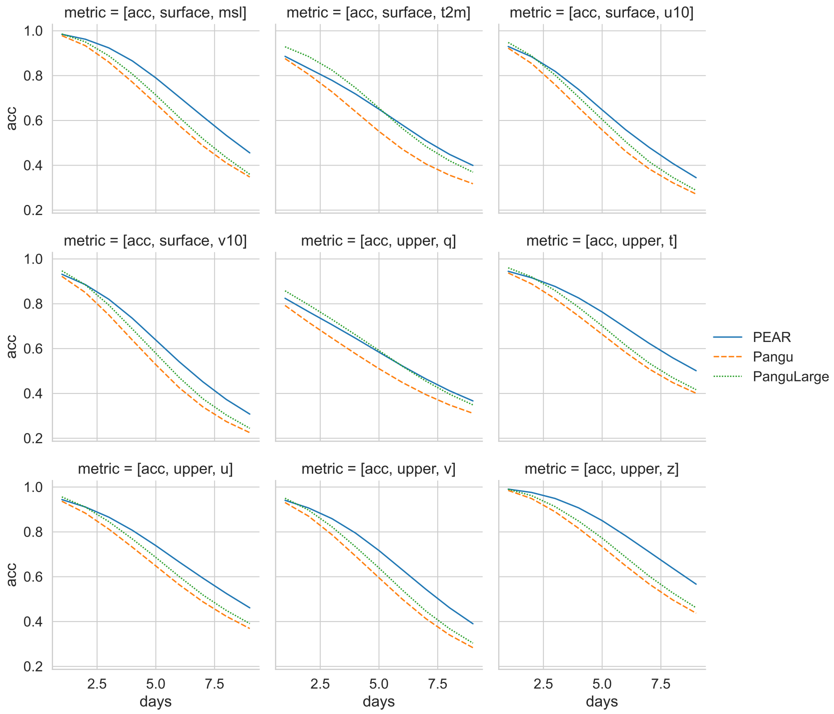

Figure 4: Mean anomaly correlation coefficient (ACC), higher is better, for the surface and upper variables after iterated model inference to perform multiple day forecasting. The upper variables are averaged over the 13 vertical levels. The metrics are mean sea level pressure (msl), temperature at 2m (t2m), eastward horizontal wind velocity at 10m (u10), northward horizontal wind velocity at 10m (v10), specific humidity (q), temperature (t), eastward wind velocity (u), northward wind velocity (v), geopotential (z). PEAR consistently outperforms Pangu at the same architecture hyperparameters, and is on par (t2m) or better (msl, u10, v10, t, u, v, z) at longer lead times compared to the almost 8 times larger PanguLarge model.

6 Limitations

Our limited compute budget restricts us to ERA5-lite, making comparisons to models trained on the full ERA5 dataset harder. Ideally we would train on ground truth data created on the HEALPix grid, but at this time the reanalysis for ERA5 is done on an equiangular grid.

Known limitations of data-driven forecasts include unphysical predictions [34], instabilities for longer forecasts [41] and lack of certain features such as the butterfly effect [42] or sub-synoptic and mesoscale weather phenomena [43]. Although we have not checked this explicitly, we suspect that our model is also subject to these limitations.

7 Conclusion

We have shown the importance of using the HEALPix grid that minimizes unphysical biases for global medium term weather prediction. The hierarchical equal area pixelation enables efficient implementation of our transformer architecture PEAR, that outperforms its counterpart on the traditional equiangular grid at no computational overhead. With forecast horizons of up to 10 days, we showed that PEAR outperforms the equiangular baseline with more than twice the number of trainable parameters. The superior performance of PEAR should also lead to more accurate extreme weather forecasting [20], a direction that would be interesting for future work.

As next generation sources for data driven weather and climate forecasting aim to include high resolution HEALPix native data [18], we hope that PEAR can pave the way for using this data in the most efficient manner.

Acknowledgments

H.L. thanks Karl Bergström for advice and support in developing custom software for model tracking and interactive visualization of spherical data. The work of J.G. and D.P. is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg (KAW) Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

[1] Manzhu Yu, Qunying Huang, and Zhenlong Li. "Deep learning for spatiotemporal forecasting in Earth system science: a review". In: *International Journal of Digital Earth* 17.1 (2024), p. 2391952.

[2] Qi Liu and Jianwei Ma. "Foundation Models for Geophysics: Review and Perspective". In: *arXiv preprint arXiv:2406.03163* (2024).

[3] Xinfeng Zhao et al. "A comprehensive review of methods for hydrological forecasting based on deep learning". In: *Water* 16.10 (2024), p. 1407.

[4] Remi Lam et al. "Learning Skillful Medium-Range Global Weather Forecasting". In: *Science* 382.6677 (Dec. 2023), pp. 1416–1421. DOI: `10.1126/science.adi2336`. arXiv: `2212.12794`.

[5] Lei Chen et al. "FuXi: A Cascade Machine Learning Forecasting System for 15-Day Global Weather Forecast". In: *npj Climate and Atmospheric Science* 6.1 (Nov. 2023), pp. 1–11. ISSN: 2397-3722. DOI: `10.1038/s41612-023-00512-1`. arXiv: `2306.12873`.

[6] Kang Chen et al. *Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead*. 2023. arXiv: `2304.02948`.

[7] Marcin Andrychowicz et al. *Deep Learning for Day Forecasts from Sparse Observations*. July 2023. arXiv: `2306.06079`.

[8] Tung Nguyen et al. "Scaling Transformer Neural Networks for Skillful and Reliable Medium-Range Weather Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 68740–68771. arXiv: `2312.03876`.

[9] Kaifeng Bi et al. "Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks". In: *Nature* 619.7970 (July 2023), pp. 533–538. ISSN: 1476-4687. DOI: `10.1038/s41586-023-06185-3`.

[10] Hans Hersbach et al. "The ERA5 global reanalysis". In: *Quarterly journal of the royal meteorological society* 146.730 (2020), pp. 1999–2049.

[11] K. M. Gorski, E. Hivon, and B. D. Wandelt. "Analysis Issues for Large CMB Data Sets". In: *Arxiv eprints arXiv:astro-ph/9812350* (1998). eprint: `arXiv:astro-ph/9812350`.

[12] Matthias Karlbauer et al. "Advancing Parsimonious Deep Learning Weather Prediction Using the HEALPix Mesh". In: *Journal of Advances in Modeling Earth Systems* 16.8 (2024), e2023MS004021. ISSN: 1942-2466. DOI: `10.1029/2023MS004021`. arXiv: `2311.06253`.

[13] Vivek Ramavajjala. *HEAL-ViT: Vision Transformers on a Spherical Mesh for Medium-Range Weather Forecasting*. Feb. 2024. arXiv: `2403.17016`.

[14] N. Perraudin et al. "DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications". In: *Astronomy and Computing* 27 (Apr. 2019), pp. 130–146. ISSN: 2213-1337. DOI: `10.1016/j.ascom.2019.03.004`. URL: `http://dx.doi.org/10.1016/j.ascom.2019.03.004`.

[15] Oscar Carlsson et al. "HEAL-SWIN: A Vision Transformer On The Sphere". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. July 2023, pp. 6067–6077. arXiv: `2307.07313`.

[16] F. J. Doblas-Reyes et al. "The Destination Earth digital twin for climate change adaptation". In: *EGUsphere* 2025 (2025), pp. 1–41. DOI: `10.5194/egusphere-2025-2198`. URL: `https://egusphere.copernicus.org/preprints/2025/egusphere-2025-2198/`.

[17] Jörn Hoffmann et al. "Destination Earth – A digital twin in support of climate services". In: *Climate Services* 30 (2023), p. 100394. ISSN: 2405-8807. DOI: `https://doi.org/10.1016/j.cliser.2023.100394`. URL: `https://www.sciencedirect.com/science/article/pii/S2405880723000559`.

[18] *Migration from GRIB1 to GRIB2: Preparing ECMWF Model Output for the Future*. https://www.ecmwf.int/en/newsletter/175/computing/migration-grib1-grib2-preparing-ecmwf-model-output-future. Text. Apr. 2023.

[19] Hailong Shu et al. *Forecasting the Future with Future Technologies: Advancements in Large Meteorological Models*. Apr. 2024. arXiv: `2404.06668`.

[20] Leonardo Olivetti and Gabriele Messori. "Advances and Prospects of Deep Learning for Medium-Range Extreme Weather Forecasting". In: *Geoscientific Model Development* 17.6 (Mar. 2024), pp. 2347–2358. ISSN: 1991-959X. DOI: `10.5194/gmd-17-2347-2024`.

[21] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Aug. 2021, pp. 10012–10022. arXiv: `2103.14030`.

[22] Simon Lang et al. *AIFS – ECMWF’s Data-Driven Forecasting System*. Aug. 2024. arXiv: `2406.01465`.

[23] Thorsten Kurth et al. “FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators”. In: *Proceedings of the Platform for Advanced Scientific Computing Conference*. Davos Switzerland: ACM, June 2023, pp. 1–11. ISBN: 9798400701900. DOI: `10.1145/3592979.3593412`.

[24] Boris Bonev et al. “Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere”. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 2806–2823. arXiv: `2306.03838`.

[25] Sebastian Scher and Gabriele Messori. “Physics-Inspired Adaptions to Low-Parameter Neural Network Weather Forecast Systems”. In: *Artificial Intelligence for the Earth Systems* 3.1 (2024), e230046. DOI: `10.1175/AIES-D-23-0046.1`. arXiv: `2008.13524`.

[26] Yang Liu et al. “CirT: Global Subseasonal-to-Seasonal Forecasting with Geometry-Inspired Transformer”. In: *The Thirteenth International Conference on Learning Representations*. 2025. arXiv: `2502.19750`.

[27] Joel Oskarsson et al. “Probabilistic Weather Forecasting with Hierarchical Graph Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 37. Dec. 2024, pp. 41577–41648. arXiv: `2406.04759`.

[28] Ilan Price et al. “Probabilistic Weather Forecasting with Machine Learning”. In: *Nature* (Dec. 2024), pp. 1–7. ISSN: 1476-4687. DOI: `10.1038/s41586-024-08252-9`.

[29] Yingkai Sha et al. *Improving AI Weather Prediction Models Using Global Mass and Energy Conservation Schemes*. Jan. 2025. arXiv: `2501.05648`.

[30] Tung Nguyen et al. “ClimaX: A Foundation Model for Weather and Climate”. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, July 2023, pp. 25904–25938.

[31] Oliver Watt-Meyer et al. “ACE: A Fast, Skillful Learned Global Atmospheric Model for Climate Prediction”. In: *Tackling Climate Change with Machine Learning: Workshop at NeurIPS*. arXiv, Oct. 2023. arXiv: `2310.02074`.

[32] Cristian Bodnar et al. *A Foundation Model for the Earth System*. Nov. 2024. arXiv: `2405.13063`.

[33] Jared D. Willard et al. “Analyzing and Exploring Training Recipes for Large-Scale Transformer-Based Weather Prediction”. In: *Artificial Intelligence for the Earth Systems* 4.2 (2025), p. 240061. DOI: `10.1175/AIES-D-24-0061.1`. arXiv: `2404.19630`.

[34] John Schreck et al. *Community Research Earth Digital Intelligence Twin (CREDIT)*. Nov. 2024. arXiv: `2411.07814`.

[35] Stephan Rasp et al. “WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models”. In: *Journal of Advances in Modeling Earth Systems* 16.6 (2024), e2023MS004019. ISSN: 1942-2466. DOI: `10.1029/2023MS004019`. arXiv: `2308.15560`.

[36] Zied Ben Bouallègue et al. “Improving Medium-Range Ensemble Weather Forecasts with Hierarchical Ensemble Transformers”. In: *Artificial Intelligence for the Earth Systems* 3.1 (Jan. 2024), e230027. DOI: `10.1175/AIES-D-23-0027.1`. arXiv: `2303.17195`.

[37] Dmitrii Kochkov et al. “Neural General Circulation Models for Weather and Climate”. In: *Nature* 632.8027 (Aug. 2024), pp. 1060–1066. ISSN: 1476-4687. DOI: `10.1038/s41586-024-07744-y`. arXiv: `2311.07222`.

[38] Ze Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 11999–12009. arXiv: `2111.09883`.

[39] Zhuoqun Li. *Lizhuoq/WeatherLearn*. Apr. 2025. URL: `https://github.com/lizhuoq/WeatherLearn`.

[40] Huug van den Dool et al. “The probability anomaly correlation and calibration of probabilistic forecasts”. In: *Weather and Forecasting* 32.1 (2017), pp. 199–206.

[41] Oliver Watt-Meyer et al. *ACE2: Accurately Learning Subseasonal to Decadal Atmospheric Variability and Forced Responses*. Nov. 2024. arXiv: `2411.11268`.

[42] T. Selz and G. C. Craig. “Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?” In: *Geophysical Research Letters* 50.20 (2023), e2023GL105747. ISSN: 1944-8007. DOI: `10.1029/2023GL105747`.

[43] Massimo Bonavita. “On Some Limitations of Current Machine Learning Weather Prediction Models”. In: *Geophysical Research Letters* 51.12 (2024), e2023GL107377. ISSN: 1944-8007. DOI: `10.1029/2023GL107377`.

Table A1: PEAR architecture overview. The windowed attention blocks contain multiple copies, indicated by the depth column, of the multi-head attention layer and layer normalization. The tensor shape column shows the shape of the output tensor from the corresponding layer. Both the input and final patch recovery layer use two separate tensors for the surface and upper variables.

Nr	Layer/Block	Depth	Tensor shape	Attention heads
	Input	1	$\begin{cases}(12n_{\text{side}}^2, 4)\\(12n_{\text{side}}^2, 13, 5)\end{cases}$	
1	Patch embed	1	$(3/4n_{\text{side}}^2, 8, 48)$	
2	Windowed attention	2	$(3/4n_{\text{side}}^2, 8, 48)$	6
3	Downsample	1	$(3/16n_{\text{side}}^2, 8, 96)$	
4	Windowed attention	12	$(3/16n_{\text{side}}^2, 8, 96)$	12
5	Upsample	1	$(3/4n_{\text{side}}^2, 8, 48)$	
6	Windowed attention	2	$(3/4n_{\text{side}}^2, 8, 48)$	6
	Concatenate 6 & 2	1	$(3/4n_{\text{side}}^2, 8, 96)$	
	Patch recovery	1	$\begin{cases}(12n_{\text{side}}^2, 4)\\(12n_{\text{side}}^2, 13, 5)\end{cases}$	

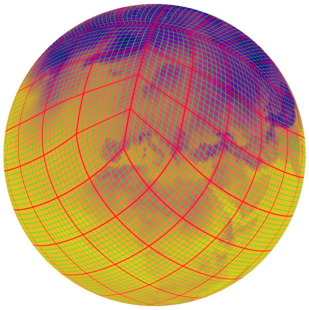

Figure A1: Patches (green) and windows (red) at the first attention layer. Each green patch contains four pixels of the input resolution $n_{\text{side}} = 64$ on the HEALPix grid, and each window contains 64 patches at this level. The window also include one vertical level above, not shown in the figure.

A Appendix

An overview of the architecture of PEAR can be found in Table. A1. The patches and windows are visualized in Figure. A1.

We use RMSE and Anomaly Correlation Coefficient (ACC) calculated on the HEALPix grid according to

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{12n_{\text{side}}^2} \sum_{i=0}^{12n_{\text{side}}^2} (y^i - \hat{y}^i)^2} \tag{1}$$

$$\text{ACC}(y, \hat{y}) = \frac{\sum_i^{12n_{\text{nside}}^2} \Delta y^i \Delta \hat{y}^i}{\sqrt{\left(\sum_i^{12n_{\text{nside}}^2} (\Delta y^i)^2\right)\left(\sum_i^{12n_{\text{nside}}^2} (\Delta \hat{y}^i)^2\right)}}, \tag{2}$$

where Δy is the difference between the predictions and the climatology average.

We include anomaly correlation coefficient (ACC) and root mean squared error (RMSE) for all upper variables and pressure levels in Fig. A4 and Fig. A5 respectively. PEAR outperforms Pangu for all variables and pressure levels for ACC in Fig. A4, and often outperforms the almost eight times larger PanguLarge.

We also include spatially resolved RMSE over the validation year for PEAR in Fig. A2 (surface) and A6 (upper), as well as the corresponding spatial RMSE for Pangu in Fig. A3 (surface) and Fig.A7 (upper).

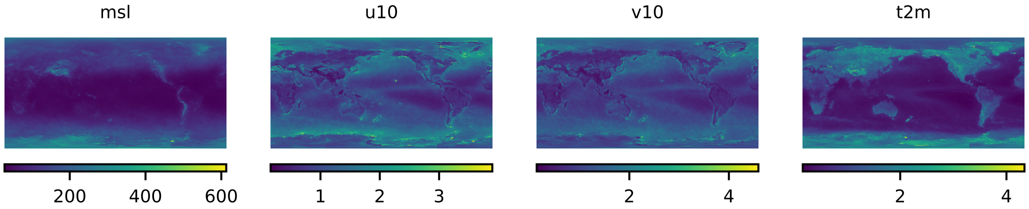

Figure A2: Spatial RMSE for PEAR predictions of surface variables with one day lead time averaged over the validation year. The HEALPix predictions are projected to a cartesian longitude (horizontal) and latitude (vertical) grid for visualization.

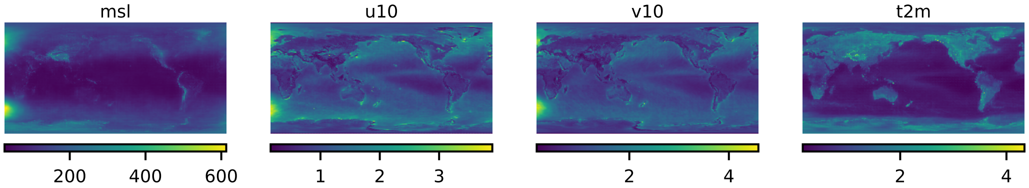

Figure A3: Spatial RMSE for Pangu predictions of surface variables with one day lead time averaged over the validation year.

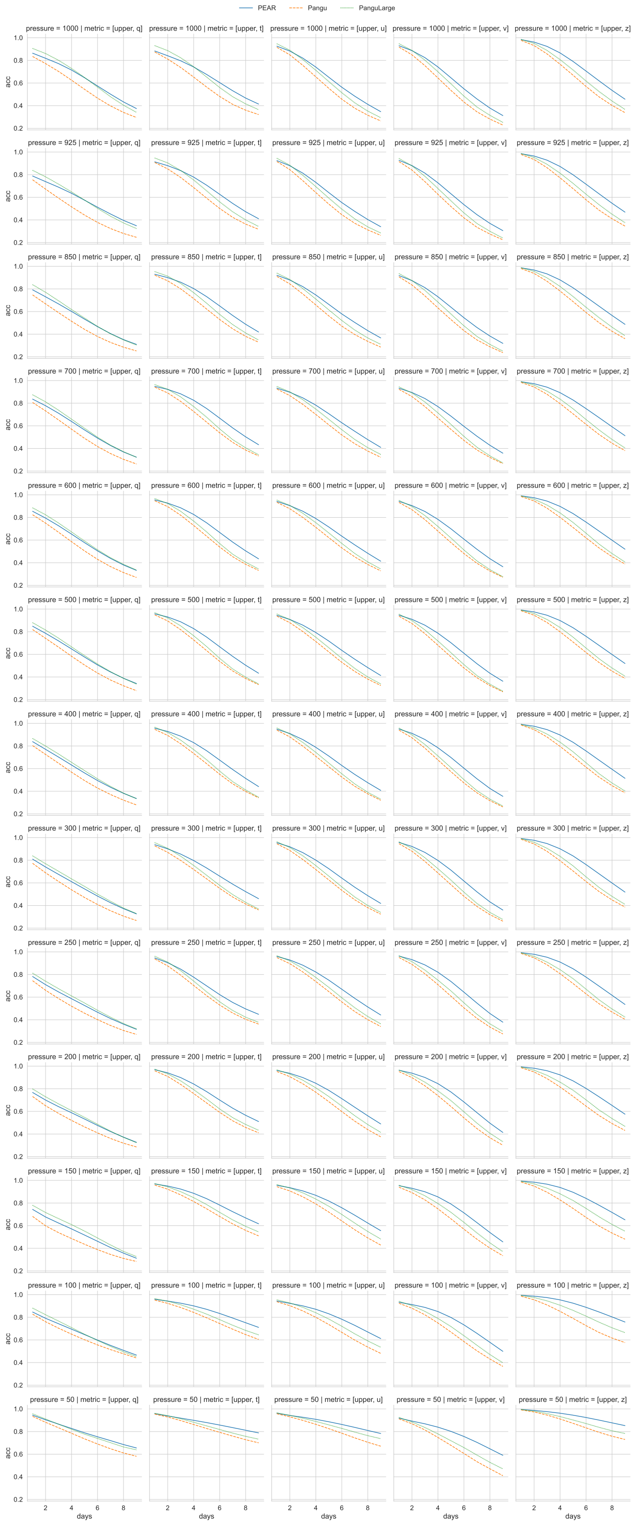

Figure A4: Anomaly correlation coefficient (ACC) for the upper variables separated by pressure level over up to 10 days lead time.

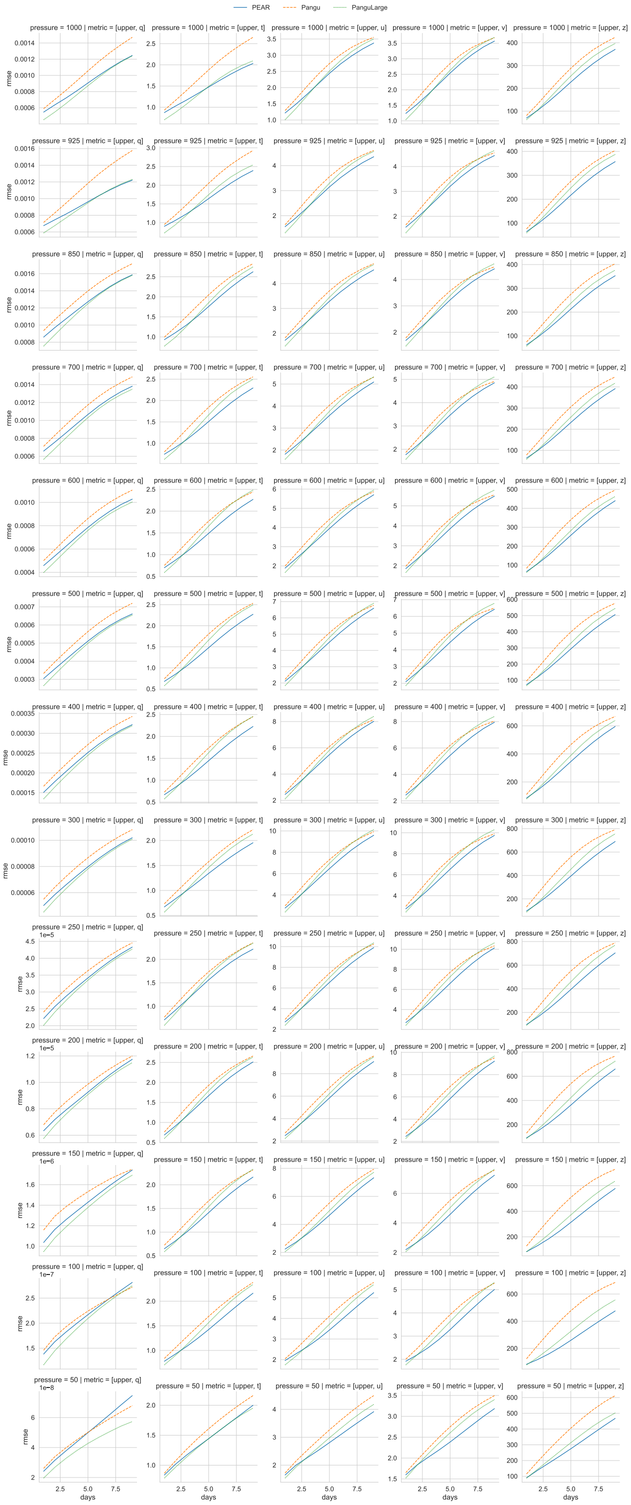

Figure A5: Root mean squared error (RMSE) for the upper variables separated by pressure level over up to 10 days lead time.

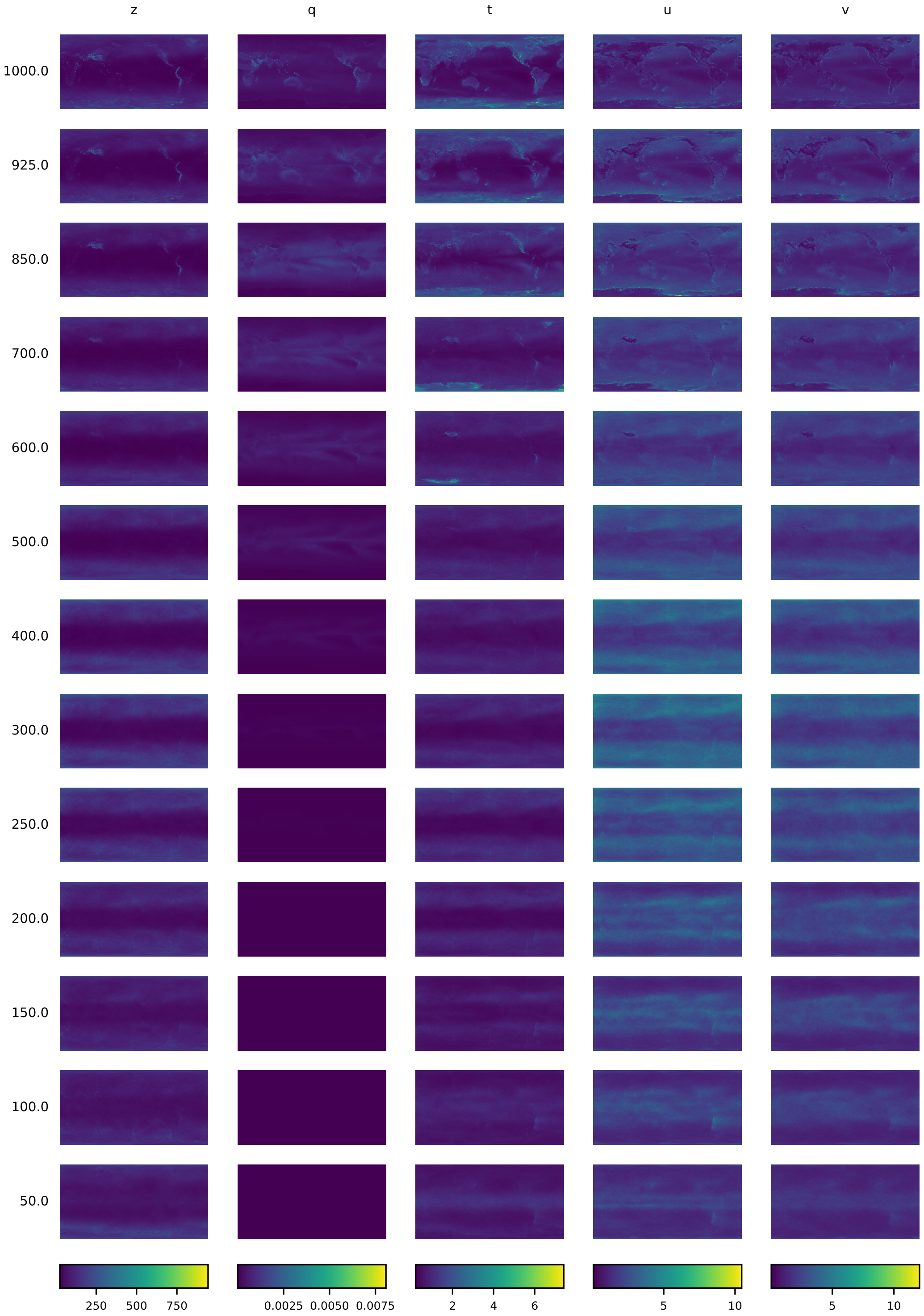

Figure A6: Spatial RMSE for PEAR predictions of upper variables with one day lead time averaged over the validation year. Pressure level (rows) and variables (columns) with joint color mapping per variable. The HEALPix predictions are projected to a cartesian longitude (horizontal) and latitude (vertical) grid for visualization.

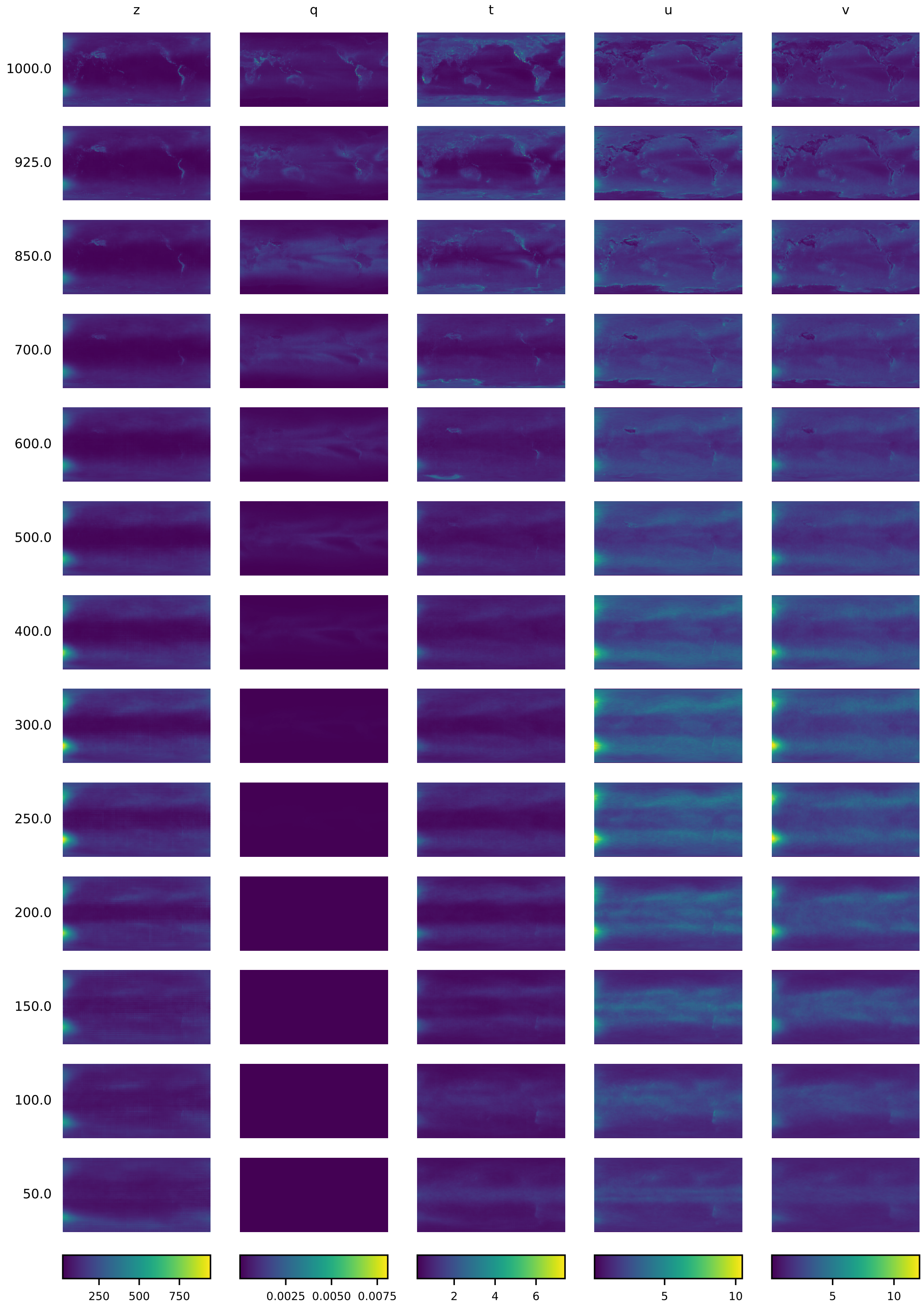

Figure A7: Spatial RMSE for Pangu predictions of upper variables with one day lead time averaged over the validation year. Pressure level (rows) and variables (columns) with joint color mapping per variable.