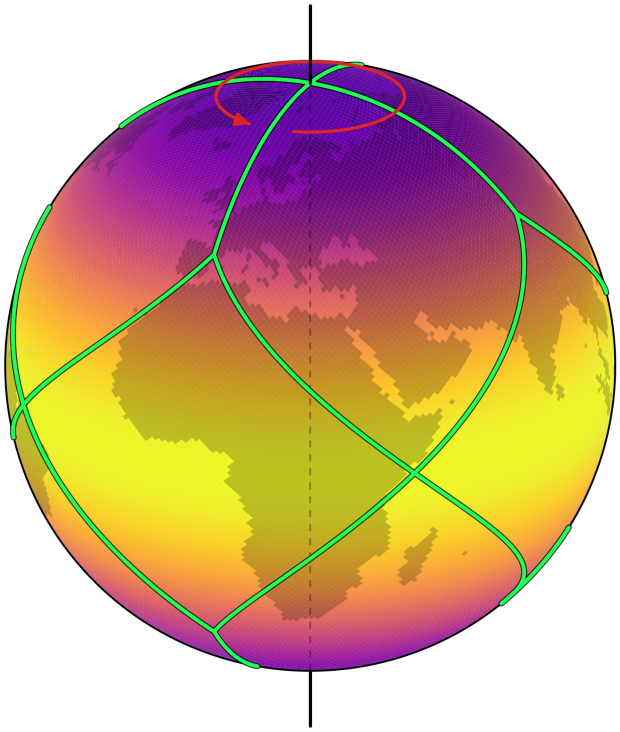


Equivariant Inductive Biases for Weather Prediction with PEAR

Investigating the exploitation of rotational symmetries for accurate transformer-based weather forecasting over the HEALPix discretisation

Degree project report in Data Science and AI

Pietro Giovanni Rosso

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Degree project report 2026

Equivariant Inductive Biases for Weather Prediction with PEAR

Investigating the exploitation of rotational symmetries for accurate transformer-based weather forecasting over the HEALPix discretisation

PIETRO GIOVANNI ROSSO

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Equivariant Inductive Biases for Weather Prediction with PEAR
Investigating the exploitation of rotational symmetries for accurate transformer-based weather forecasting over the HEALPix discretisation
Pietro Giovanni Rosso

Supervisor: Jan Gerken, Department of Mathematical Sciences
Examiner: Simon Olsson, Department of Computer Science and Engineering,

Degree project report 2026
Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden

Cover: HEALPix discretisation visualisation, with iso-latitude colour grading and red arrow that indicates the rotation around the Earth axis

Typeset in LaTeX
Gothenburg, Sweden 2026

Abstract

Weather forecasting is a complex challenge due to its intrinsically complex physical dynamics that define the evolution of the system. In recent years, deep learning weather prediction has emerged as a promising alternative to classical numerical weather prediction, matching or outperforming it on several benchmarks at a fraction of the inference time. This thesis contributes to this direction by analysing the symmetries of this system in relation to the group $SO(2)$: the rotation of the Earth around its own axis. The study builds on Pangu Equal ARea (PEAR), a transformer-based model operating on the Hierarchical Equal Area isoLatitude Pixelization (HEALPix) discretisation of the sphere, and examines whether the symmetry awareness of this architecture can be increased from two complementary perspectives: the data on which the model is trained, and the architecture itself. In the first part, starting from ERA5, the reanalysis dataset that provides global estimates of the surface and atmospheric variables, we introduce a new 2-hourly sampling, which allows a comparison of PEAR's equivariance behaviour across three configurations of dataset and forecast horizon. The analysis shows that the equivariance error is dominated by the architecture and the forecast horizon rather than by the sampling. The second part introduces two architectural modifications, an iso-latitude interspersed windowing scheme and a set of HEALPix-aware convolutions, designed to better align the model with the rotational symmetry of the sphere. These modifications successfully reduce the equivariance error at the surface level, but fail to improve it at the upper atmospheric levels, and do not translate into a forecasting advantage over the baseline. This outcome highlights the difficulty of embedding inductive biases in the case of domains that involve using high-dimensional samples, specifically in relation to window-based attention mechanisms.

Keywords: Weather Forecasting, Symmetries, Deep Machine Learning, Equivariance, ERA5, SWIN Transformers

Acknowledgements

As this beautiful experience comes to an end, I want to thank all the people who have actively helped me to reach this important milestone in my life. For starters, I reserve a huge thanks to the people who supervised my work, specifically Jan Gerken, Hampus Linander, and Daniel Persson, who guided me with helpful feedback and reliable support. I would like to thank my parents and relatives, who, despite the distance, always believed in me and supported me with a heart-warming amount of cheese. Finally, I dedicate these lines to the people that I have met during these two years. People who, in one way or another, contributed to my growth and always made me feel at home, and that, regardless of the distance that will eventually exist between us, will always be part of my life.
To all these people, I say my most sincere THANK YOU.

Pietro Giovanni Rosso, Gothenburg, June 2026

Contents

List of Figures

1
Introduction

Machine Learning research allows us to reshape the approach towards solving complex problems, by showing incredible performance in recognising complex patterns across datasets too vast for human analysis. This is not a new concept for the scientific community; the potential of these types of systems is already acknowledged by Frank Rosenblatt's 1958 paper [Rosenblatt, 1958]. Rosenblatt introduced the perceptron as a probabilistic model capable of learning to classify inputs through exposure to examples rather than explicit programming, demonstrating that a machine could autonomously organise information and generalise to previously unseen stimuli. Although the computational resources of the era severely constrained what such systems could achieve in practice, the premise that statistical learning from data could approximate human judgment established the conceptual foundation upon which machine learning is built. After decades of advances in algorithms, hardware, and the availability of large-scale datasets, this early vision has evolved into the deep neural architectures [LeCun and Hinton, 2015] that are the foundation of contemporary AI systems. As a consequence, machine learning is reshaping the way many scientific and engineering problems are approached, opening new ways of tackling them that often require less domain knowledge than conventional methods. This thesis will explore how the complex problem of weather forecasting can be tackled with data-driven models. More specifically, the research will focus on the symmetries of the Earth, approximated as a sphere, and how they can be exploited to design more efficient and accurate neural network architectures.

1.1 Weather forecasting

Weather forecasting is a crucial frontier of research for our society. Being able to formulate accurate predictions about the future state of the weather would benefit many sectors, such as agriculture, energy production, transportation, public health and many more. For more than half a century, weather forecasting has been dominated by numerical weather prediction (NWP), in which the governing equations of atmospheric dynamics are discretised and integrated forward in time on high-performance computing systems [Bauer et al., 2015]. Despite remarkable progress, this paradigm faces fundamental limitations; each forecast requires substantial computational resources, and model resolution is bounded by the cost of integration. Machine learning offers a promising alternative; by training models on a large amount of reanalysis data, it is possible to produce global-scale predictions within seconds rather than hours. Furthermore, many benchmarks showed that this

innovative type of prediction outperforms NWP [Shi et al., 2025], leading to a new generation of weather forecasting systems in which the heavy machinery of numerical integration is replaced by learned representations of atmospheric dynamics. Unlike numerical models, which are built around the conservation laws and geometric structure of the surface and atmosphere, current ML forecasters learn these properties only implicitly, from data. This setup is likely to produce inconsistent predictions, for instance, under rotations of the input field that should, in principle, leave the underlying dynamics unchanged. A natural way to mitigate this is to encode some of the relevant structure directly into the architecture, so that a portion of the symmetry is satisfied by construction rather than approximated from examples. This thesis takes a small step in that direction. It investigates a variant of an existing data-driven forecasting architecture that is made more equivariant under $SO(2)$, the group of rotations about the Earth's polar axis. Furthermore, a new dataset is implemented to enforce the rotational symmetry of the system, by exposing the model to inputs sampled at multiple times of day rather than at a single nominal hour, so that the longitudinal action of $SO(2)$ is better represented in the data the network is trained on. Together, these two contributions, an architectural modification and an investigation on a new dataset configuration, define the scope of the work. The motivation is not to build a fully physics-informed model, nor to compete with state-of-the-art operational systems, but to study, in a controlled setting, whether adding a modest geometric inductive bias offers measurable benefits.

1.2 Research Question

The goal of this thesis is to investigate whether enforcing rotational equivariance with respect to the group $SO(2)$ provides a beneficial inductive bias for the overall model performance. More specifically, if this adaptation succeeds in enhancing Pangu Equal ARea (PEAR) [Linander et al., 2026], which relies on a Hierarchical Equal Area isoLatitude Pixelization (HEALPix) [Gorski et al., 2005] for a uniform sampling on the sphere. This discretization defines the geometry of the windowing scheme that, by construction, respects only the discrete subgroup of 90° rotations corresponding to the base-quadrilateral layout of the grid. Whether the broader symmetry of the sphere can be incorporated into such an architecture, and whether doing so translates into measurable gains in forecasting accuracy, is the central question this thesis addresses.
This question is approached from two complementary perspectives. The first asks whether the equivariance behaviour of the model can be improved through the data on which it is trained: specifically, whether providing a denser, sub-daily sampling of ERA5, exposing the model to the diurnal cycle of the atmosphere, encourages a representation more consistent with the rotational symmetry of the sphere. The second asks whether the equivariance can instead be improved by acting directly on the architecture of the model, through modifications that align the attention mechanism with the geometry of the HEALPix grid. In this case, the implemented approach uses windows that aggregate pixels lying on the same iso-latitude ring through an interspersed scheme. This construction yields a partitioning that is more robust under rotations about the polar axis, which is expected to translate

into a better rotational-equivariance generalisation of the model. Furthermore, the transformer blocks are combined with HEALPix-aware convolutions, which allow the model to extract learning signals from local spatial relations between neighbouring pixels.The analysis also examines whether any equivariance improvement obtained is reflected in the model's predictive skill, since a geometric inductive bias is only practically useful if it improves the model's behaviour on the underlying forecasting task.

1.3 Main Contributions

This thesis outlines three main contributions:

1. A new 2-hour sampling of ERA5 has been tested, and the equivariance behaviour of PEAR is compared across three configurations of data and forecast horizon. The analysis establishes that the equivariance error is dominated by the architecture and the forecast horizon rather than by the sampling of the data, and forms a contribution reported in the paper Linander et al. [2026], currently under review process.
2. An Iso-latitude interspearsed windowing scheme and HEALPix-aware convolution have been implemented to align the model to the rotational symmetry of the sphere. These modifications have been trained and evaluated over the new dataset with the 2-hour prediction task and compared with the baseline PEAR.
3. Equivariance error and RMSE have been computed for each model level-by-level. This evaluation reveals a per-layer agreement between the two metrics and highlights a locality-equivariance trade-off introduced by the iso-latitude windowing that the addition of convolutions does not fully recover.

2

Background

2.1 Deep Learning for Weather Forecasting

For more than half a century, operational weather forecasting has relied on numerical weather prediction (NWP). In this framework, an initial atmospheric state is constructed through data assimilation, combining observations from radiosondes, satellites, and surface stations with a short-range model forecast to produce a physically consistent analysis. This analysis is then evolved forward in time by numerically integrating the primitive equations of atmospheric dynamics on a three-dimensional grid. Modern operational systems such as the ECMWF Integrated Forecasting System (IFS) run at resolutions of approximately 9 km and require supercomputer-class infrastructure, with a single ten-day deterministic forecast consuming on the order of one hour of wall-clock time [Bauer et al., 2015].

NWP carries important strengths: the governing equations enforce conservation of mass, energy, and momentum by construction, guaranteeing physical consistency of the forecast; the output is fully traceable to known dynamics, which facilitates diagnosis when a forecast fails; and the physics-based formulation allows the models to handle novel atmospheric regimes that have not been observed before. However, the approach also has well-documented limitations. Sub-grid processes such as convection, radiation, and boundary-layer turbulence cannot be resolved explicitly and must be approximated through parametrisation schemes, which remain a dominant source of forecast error. Furthermore, increasing resolution is expensive: the computational cost of integration grows steeply with the number of grid points, making it difficult to improve resolution without proportional increases in hardware.

Machine-learning-based weather forecasting offers a fundamentally different trade-off. Instead of integrating physical equations, a neural network is trained on decades of reanalysis data to learn a direct mapping from an input atmospheric state to a future state. At inference time, this mapping can be evaluated in seconds on a single GPU, eliminating the need for expensive numerical integration. Moreover, sub-grid processes do not need to be parametrised explicitly; they are learned implicitly from the data, bypassing a major source of NWP error. On the other hand, ML models provide no built-in guarantee of physical consistency: they can produce non-physical states such as negative humidity, and their predictions are not straightforwardly traceable to governing equations, which complicates the diagnosis of failures. Because they are trained on reanalysis datasets such as ERA5, they also inherit any biases present in the data assimilation system that produced the training set. Finally, ML models tend to struggle with extreme events that fall outside

their training distribution, whereas NWP can, in principle, extrapolate through its physical equations.
Despite these limitations, ML-based forecasting has progressed rapidly toward operational use. In early 2024, ECMWF launched its Artificial Intelligence Forecasting System (AIFS) [Lang et al., 2024] as an experimental product, marking the first time an operational centre made ML-based global forecasts publicly available. Models such as GraphCast [Lam et al., 2022] and GenCast [Price et al., 2024] have demonstrated verification scores that match or exceed those of ECMWF's high-resolution deterministic forecast (HRES) for medium-range lead times. At present, however, operational centres view ML not as a replacement for NWP but as a complement: ML models provide rapid guidance and can augment traditional ensembles, while NWP retains the role of primary operational system where physical traceability and reliability under novel conditions remain essential.
Having outlined the general landscape, we now survey the main modelling paradigms that have been explored on the ML side. The survey by Shi et al. [2025] categorises prior work on data-driven weather forecasting into three main modelling paradigms:

- **Deterministic Predictive Learning**: Models are trained to directly predict future weather states from past observations by minimising a pointwise loss function. They produce a single "best guess" forecast. They are fast and accurate for medium-range forecasts, but the outputs tend to be blurry, especially for extreme events, and they don't naturally quantify uncertainty.
- **Generative Models**: Models treat forecasting as a generative or sampling process conditioned on past observations, using architectures such as generative adversarial networks and diffusion models. Instead of one prediction, they generate an ensemble of plausible futures. They manage to capture both inherent data uncertainty and model uncertainty, and tend to handle extreme events better at the cost of being more computationally expensive to train and run.
- **Foundation Models**: Large models are first pre-trained on massive, diverse weather datasets using self-supervised objectives, then fine-tuned for specific downstream tasks like forecasting or downscaling. They offer transferable representations and adaptability to many tasks without training from scratch, but they require high pre-training costs and large parameter counts, and fine-tuning techniques for weather are still in early stages.

This thesis focuses on developing and testing systems that rely on *Deterministic Weather Forecasting* as a paradigm. Therefore, the models will learn a direct mapping from a sequence of past observations to future states:

$$(X_{t-a+1}, \ldots, X_t) \xrightarrow{F(\theta)} (Y_{t+1}, \ldots, Y_\beta)$$

where α and β are the temporal lengths of the input and output windows. Over the years, several architectures have been tested, including convolutional neural networks [Weyn et al., 2020], recurrent neural networks [Hu et al., 2023], transformer-based models [Bi et al., 2022], and graph neural networks [Keisler, 2022], in which the common objective is to learn accurate deterministic forecasts directly from historical observations. As previously mentioned, these models are typically trained using pointwise loss functions such as mean absolute error (MAE) or mean squared

error (MSE) and require a relatively small inference time. The most relevant examples include global forecasting models such as FourCastNet Pathak et al. [2022], Pangu-Weather Bi et al. [2022], GraphCast Lam et al. [2022], and FengWu Chen et al. [2023], which employ transformer or graph-based architectures to achieve competitive medium-range forecasts with substantially reduced computational cost.

2.2 PEAR

The main baseline of this project is Pangu Equal ARea (PEAR) [Linander et al., 2026], a transformer-based weather forecasting model. PEAR builds upon the Pangu-Weather architecture [Bi et al., 2022], which was the first ML model to outperform operational NWP on standard verification metrics for medium-range forecasting. The distinguishing design decision of PEAR is that it operates natively on the HEALPix grid rather than on the Driscoll-Healy grid used by Pangu-Weather and most other ML forecasting models. This eliminates the uneven spatial sampling that the Driscoll-Healy grid introduces near the poles, and enables the architecture to exploit the hierarchical and equal-area properties of HEALPix throughout its windowing, downsampling, and upsampling stages. The following sections introduce the HEALPix grid and the SWIN transformer independently, after which a detailed description of the PEAR architecture is provided.

2.3 HEALPix Mesh

Previous weather forecasting approaches relied on the Driscoll-Healy grid [Driscoll and Healy, 1994], whose sampling becomes increasingly dense near the poles relative to the equator, leading to a non-uniform distribution of pixels across the sphere. This sample density issue is related to the construction of the grid, which relies on equally spaced angles in both latitude and longitude. This construction leads to the shrinkage of the physical distance between grid points as you approach the poles, which are the convergence points of the longitude lines. Specifically, considering N the number of points in each Driscoll-Healy latitude ring, the longitude is sampled as:

$$\phi_k = \frac{2\pi k}{N}, \quad k = 0, \cdots, N-1$$

This formulation is valid regardless of where on the sphere the iso-latitude ring sits. Considering the arc length between adjacent longitude samples on a ring at colatitude θ, where R is the radius of the sphere:

$$\Delta s = R \cdot \sin(\theta) \cdot \frac{2\pi}{N}$$

As $\theta \to 0$ (getting closer to the North Pole) or $\theta \to \pi$ (getting closer to the South Pole), $\Delta s \to 0$. Therefore, the distance between adjacent points on the same latitude becomes smaller approaching the poles. The Hierarchical Equal Area iso-Latitude Pixelization (HEALPix) [Gorski et al., 2005] avoids this problem by construction. It partitions the sphere into twelve equal-area base quadrilaterals, illustrated in

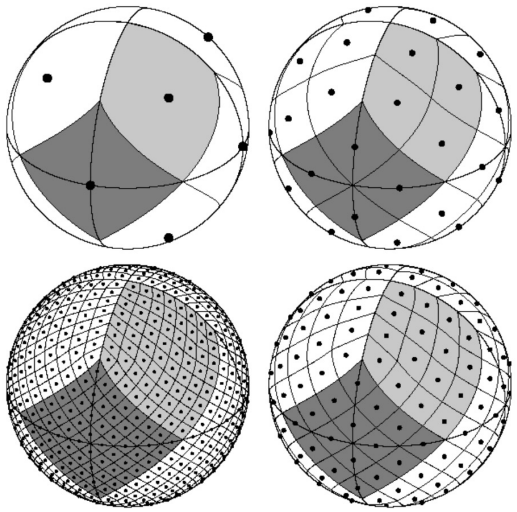


Figure 2.1: Visualization of HEALPix grid. Moving clockwise from the upper left panel the grid is hierarchically subdivided with the grid resolution parameter equal to $N_{side} = 1, 2, 4, 8$, and the total number of pixels equal to $N_{pix} = 12 \times N_{side}^2 = 12, 48, 192, 768$

Figure 2.1. Each base quadrilateral is then refined by recursively subdividing it into four equal-area sub-pixels. The resolution is controlled by the parameter N_{side}, which gives the number of pixels along each edge of a base quadrilateral, resulting in a total of $12 \cdot N_{\text{side}}^2$ pixels. This discretisation creates a sampling scheme that avoids spatial biases and offers three properties that are particularly useful for learning on the sphere:

- **Hierarchical structure**: The pixels are organised hierarchically through the nested indexing scheme; in this ordering, blocks of 4^k consecutive pixels correspond to the pixels in a quadrilateral at k division levels above the grid resolution. This property makes it straightforward to divide the surface of the sphere into attention windows and to implement coarse-graining operations by simple reshaping of the pixel index.

- **Equal Area**: Each pixel covers the same solid angle, meaning that all pixels have identical area on the sphere. This avoids spatial sampling bias (e.g., over-representing high-latitude regions) and makes global statistics and learning objectives better conditioned, since every location contributes uniformly.

- **Isolatitude**: Pixel centres are placed on a discrete set of rings of constant latitude. In the ring indexing scheme, pixels are sorted along these iso-latitude circles from north pole to south pole. Performing a roll operation on this ordering rotates the features on the sphere around the polar axis, which provides a natural mechanism for shifting attention windows along the longitudinal direction. This property will be beneficial for the iso-latitude windowing scheme described later in the thesis.

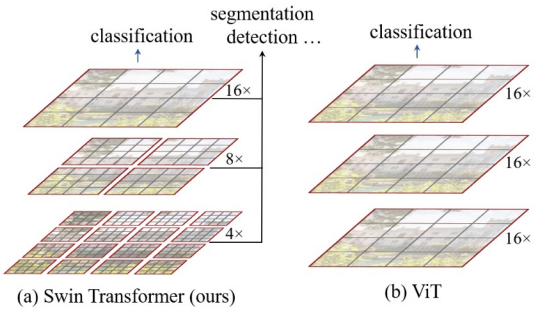


Figure 2.2: Visualisation of the SWIN hierarchical feature mapping compared to the standard implementation of Visual Transformers

2.4 SWIN Transformer

The Shifted WINdow (SWIN) transformer [Liu et al., 2021] was introduced as a general-purpose backbone for vision tasks, designed to address one of the main practical limitations of the standard Vision Transformer [Dosovitskiy et al., 2021]: the quadratic cost of global self-attention, which makes it impractical to apply directly to high-resolution images. To overcome this, SWIN partitions the input into non-overlapping local windows and computes self-attention only within each window, reducing the cost from quadratic to linear in the number of tokens. To prevent the model from being confined to fixed, isolated regions, the window partition is shifted by half a window between successive layers, so that tokens which belonged to separate windows in one layer share a window in the next. Furthermore, the architecture is organised hierarchically: as shown in Figure 2.2, tokens are progressively merged across stages, producing multi-scale feature maps, making SWIN well-suited to tasks that benefit from representations at multiple resolutions. Due to these properties, SWIN became a natural choice for weather forecasting, where the inputs are high-resolution and it is necessary to ensure even signal propagation across the globe.

2.5 PEAR Architecture

PEAR implements a transformer-based U-Net [Ronneberger et al., 2015] that takes the global atmospheric state at time t and predicts the state at time $t + \Delta t$, with a forecast lead time of 24 hours. The global volumetric weather state is discretised on the HEALPix grid along the surface and into 13 discrete pressure levels in the vertical direction. Following Pangu-Weather, the total weather state is represented as a combination of 4 surface variables and 5 upper variables at each of the 13 vertical levels. PEAR thus takes two input tensors, the surface and upper variables discretised on the spherical surface and the spherical shell correspondingly. Since the HEALPix grid covers the sphere with a one-dimensional index structure, the model input tensors have shapes $(12n_{\text{side}}^2, 4)$ and $(12n_{\text{side}}^2, 13, 5)$.
The architecture is constructed using a combination of five main layer types: patch

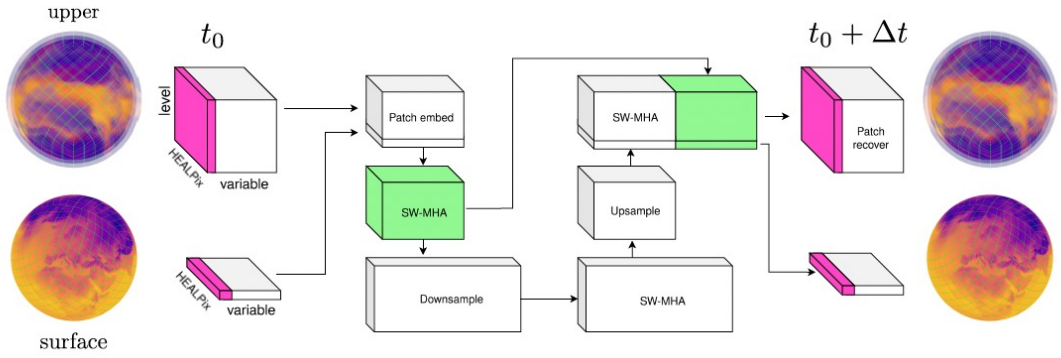


Figure 2.3: PEAR architecture schematic. Patch embedding by convolution, shifted windowed multi-head attention (SW-MHA) with learned positional embedding, downsampling and upsampling by patch merging and splitting, and patch recovery by transpose convolutions. The green block indicates the skip connection, where the output of the first attention layer is concatenated along the embedding dimension before the final patch recovery. Adapted from Linander et al. [2026].

embedding, windowed attention with alternating shifting, downsampling, upsampling, and patch recovery. These are described in the following subsections.

2.5.1 Patch Embedding and Recovery

At the input boundary, PEAR maps the raw surface and upper variables into a shared latent space using strided convolutions: a 1D convolution for the surface tensor and a 2D convolution for the upper tensor, both producing 48-channel embeddings. The two mappings are then concatenated into a single latent tensor that is passed to the attention blocks. Symmetrically, at the output boundary, the latent tensor is split back into surface and upper components and mapped to the original variable space through transpose convolutions. This design keeps the interface between physical variables and latent representations confined to the first and last layers of the network.

2.5.2 Windowed Attention

The latent tensor is partitioned into non-overlapping windows using the nested indexing of the HEALPix grid, and self-attention is computed independently within each window. Because the equal-area property of HEALPix ensures that every window covers the same solid angle, a single learned relative positional embedding can be shared across all windows, which accounts for most of the parameter savings of PEAR compared to Pangu-Weather.
To propagate information between windows, PEAR follows the SWIN shifting strategy described in the previous section, adapted to the sphere through the ring shifting mechanism introduced in HEAL-SWIN [Carlsson et al., 2024]. Every other attention layer, the features are reindexed from nested to ring ordering, rolled along the iso-latitude circles, and converted back. A simple shift is applied in the vertical direction. Since the ring shift is cyclic, pixels from opposite poles would be grouped

together; masked attention is therefore applied to prevent interaction among spatially disjoint pixels.

2.5.3 Downsampling and Upsampling

The U-Net bottleneck is built on the hierarchical structure of HEALPix: four neighbouring pixels at one resolution level correspond to a single pixel at the next coarser level. Downsampling concatenates the embeddings of each group of four pixels and linearly projects them, halving the spatial resolution while doubling the embedding dimension. In the nested indexing this reduces to a reshape followed by a linear layer. Upsampling reverses the operation, expanding each pixel back into four. A skip connection concatenates the output of the first attention stage to the upsampled tensor before the final attention stage, following the standard U-Net design.

2.6 ERA5-Lite

The training and evaluation of PEAR is carried out using a subset of ERA5 [Hersbach et al., 2020]. The full dataset consists of 60 years of hourly reanalysis data, meaning the samples are the result of blending observational data with numerical model predictions using numerical assimilation methods. This process provides a high-quality benchmark for global-scale weather forecasting. Pangu-Weather is trained over 39 years of the dataset (1979-2017), while PEAR, due to computational constraints, uses 11 years (2007-2017) with samples at 00:00UTC. The variables used for this analysis are divided into two categories:

- **Surface Variables**: two-meter temperature, u-component and v-component of 10m wind speed, and mean sea level pressure.
- **Upper Variables**: geopotential, specific humidity, temperature, u-component and v-component of wind speed. These five variables span across 13 different pressure levels: 50hPa, 100hPa, 150hPa, 200hPa, 250hPa, 300hPa, 400hPa, 500hPa, 600hPa, 700hPa, 850hPa, 925hPa, and 1000hPa

The data is discretised on a Driscoll-Healy gridding on the sphere; therefore, for PEAR and for the thesis methodology, all the samples are re-gridded into HEALPix with $N_{side} = 64$, leading to 49152 pixels per observation.

2.7 Symmetries-aware Learning

Many learning problems involve data that exhibit clear symmetries: translations of an image, rotations of a molecule, or permutations of a set should leave the underlying meaning of the data unchanged. The field of geometric deep learning Bronstein et al. [2021] provides a unified framework to expose and exploit such symmetries by embedding them directly into the architecture of neural networks, rather than leaving the model to discover them from data. This section reviews the central concepts of symmetry and equivariance that this framework relies on, before specialising them to the rotational symmetry of the sphere that is relevant to this thesis. More specifically, a system is said to be symmetric if a property of the object

is exactly the same under a specific transformation. In the context of representation learning, this notion is formalised through *equivariance* and *invariance*. Given a group of transformations G, a function $f : \mathcal{X} \to \mathcal{Y}$ is defined to be G-equivariant if:

$$f(g \cdot x) = g \cdot f(x), \quad \forall g \in G$$

Therefore, the output of the function generates a feature mapping that transforms consistently with the input for each group action $g \in G$. The invariance arises as a special case of equivariance, where the output remains unchanged under the action of G, thus:

$$f(g \cdot x) = f(x), \quad \forall g \in G$$

Providing this inductive bias can lead to significant improvement in the learning process, since the hypothesis space is restricted to functions that are consistent with the known symmetries of the problem. By enforcing equivariance or invariance at the architectural level, the model no longer needs to learn these constraints from data, thereby reducing sample complexity and improving generalisation. Specifically, the most commonly used group actions are *linear group actions*, also known as *group representation*, which map each group element to an invertible linear transformation on a vector space, preserving the group structure under composition. For example, the rotation group $SO(2)$ acts on the two-dimensional plane through the standard rotation matrices

$$\rho(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

which take a vector $\mathbf{x} \in \mathbb{R}^2$ to its rotated counterpart $\rho(\theta)\mathbf{x}$. Composing two such rotations gives $\rho(\theta_1)\rho(\theta_2) = \rho(\theta_1 + \theta_2)$, mirroring the group operation of $SO(2)$ on the level of matrices. In order to embed these equivariance properties into a neural network in a systematic way, the geometric deep learning literature identifies four building blocks that together preserve symmetry throughout the architecture:

- **Equivariant linear layers**: mapping between the feature spaces in a way that commutes with the action of the group G. These layers form the main equivariant structure and ensure that the transformation of the input is preserved throughout the network.
- **Pointwise non-linearities**: functions applied independently at each location of the feature map. Such non-linearities are equivariant with respect to the regular representation acting on the input and output: since the group action only permutes locations and the function acts independently at each location, the two operations commute by construction. This provides the non-linear expressivity required for learning without breaking the equivariance established by the linear layers.
- **Local pooling**: the aggregation of features over neighbourhoods to produce coarser representations. Pooling does not, in general, preserve the full equivariance of the network. Instead, it typically reduces the equivariance from the original group G to a subgroup that acts on the coarser representation. For example, a translation-equivariant convolutional network is no longer equivariant to translations by a single pixel after pooling with a kernel of size n; equivariance is preserved only with respect to translations that are multiples

of n, corresponding to the discrete subgroup of translations on the pooled grid. The same principle applies more generally: pooling can be designed to respect a subgroup of G that acts naturally on the coarsened domain, but it generally breaks the finer symmetries of the original input space.

- **Invariant global pooling**, applied at the end of the network to produce outputs that are invariant under G. This operation summarises the equivariant features into a representation that no longer depends on the specific group element acting on the input.

Together, these four building blocks provide a general recipe for constructing networks that respect the symmetries of a given problem. Two of the most widely used operations in modern deep learning fit naturally into this framework, each corresponding to a different symmetry group. Standard convolutions are equivariant under the group of translations: shifting the input by a given offset results in an output that is shifted by the same offset, which is what makes convolutional networks well-suited to data defined on regular grids [Cohen and Welling, 2016]. Self-attention, on the other hand, is equivariant under the group of permutations: reordering the input tokens produces a correspondingly reordered output, since attention treats its inputs as an unordered set and depends only on the pairwise relationships between elements [Lee et al., 2019]. This permutation equivariance is what makes attention a natural choice for data without an intrinsic ordering, such as sets or graphs, and it is also why positional encodings must be added explicitly when attention is applied to sequences or images, where the order of the tokens does carry meaning. This notion of equivariance for these two architectures is central to the present thesis, since the architecture described in the methodology section relies on attention as its core operation, applied to data with a well-defined spatial structure on the sphere. In this project, the relevant group transformation is defined as $SO(2)$, which represents the set of all planar rotations in 2D Euclidean space. Global-scale weather forecasting requires modelling physical processes on the surface of a rotating sphere, where the Earth rotates around its own axis. Therefore, it is possible to approximate this physical symmetry as an $SO(2)$ rotational symmetry acting along longitudinal directions. Imposing $SO(2)$-equivariance in the learned representations ensures that a rotation of the input atmospheric fields around the Earth's axis results in a corresponding rotation of the model outputs. This geometric symmetry can be generalised as a specific form of permutation of pixels in the context of HEALPix discretisation, meaning that the permutation equivariance of attention can be leveraged directly to obtain the desired equivariance, without modifying the attention mechanism itself.

3

Dataset and Horizon Effects on Model Equivariance

As described in the background, PEAR is trained on ERA5-Lite, a once-daily sampling of ERA5 in which all observations are taken at 00:00 UTC. This choice means that the training data captures day-to-day atmospheric evolution but does not represent the diurnal cycle: the model never sees how the same region of the atmosphere evolves across different times of day. From the perspective of rotational symmetry, this is a significant limitation. A rotation about the polar axis is geometrically equivalent to a shift in local solar time, so a dataset that only samples a single time of day does not expose the model to the full range of states that such rotations would produce. Therefore, a natural question arises: can richer temporal coverage in the training data improve the equivariance properties of the model?
This chapter investigates that question. Keeping the PEAR architecture fixed, three configurations are compared: the original once-daily sampling with a 24-hour forecast horizon, a new 2-hour sampling with the same 24-hour horizon, and the 2-hour sampling with a shorter 2-hour horizon. Comparing the first two isolates the effect of the sampling density at a fixed horizon, while the third allows us to study the effect of the forecast horizon itself. The evaluation is conducted entirely through the equivariance error metric, since the three setups represent different forecasting tasks for which a direct comparison of predictive skill would not be meaningful. By using a single architecture throughout, any observable difference in equivariance can be attributed to the data and the task rather than to the model. The work presented in this chapter forms part of a contribution in the paper [Linander et al., 2026] currently under review. The remainder of the chapter is organised as follows: the new dataset and the equivariance error metric are first introduced, after which the results are presented and discussed.

3.1 New Sampling of ERA5

The new dataset samples ERA5 every two hours, exposing the model to twelve initialisation times per day rather than one. This provides the temporal coverage needed to teach the network to respond when the input is shifted along the time axis, which, as discussed above, is closely related to rotations about the polar axis. To enable a fair comparison with the original PEAR setup, and given the computational constraints of this work, the new dataset is designed to contain a comparable number of training samples. Specifically, the new sampling draws observations every two

hours from the year 2012 alone, yielding 4380 samples, against PEAR's once-daily sampling spanning the eleven years from 2007 to 2017, which yields 4017 samples. The evaluation set consists of the year 2019 in both cases, with 365 samples each. For the original setup the evaluation set is unchanged, corresponding to once-daily observations at 00:00 UTC. For the new 2-hour setup, a single initialisation time per day is retained from the 2019 observations, so that the two models are evaluated on directly comparable inputs.

3.2 Equivariance Error

In theory, an architecture is said to be G-equivariant if the relation $f(g \cdot x) = g \cdot f(x)$ holds exactly for every input x and every group action $g \in G$. In practice, this property is rarely achieved. Several factors prevent a network from being strictly equivariant: continuous symmetries such as $SO(2)$ are approximated on a discrete grid, where only a finite subgroup of rotations maps grid points exactly onto other grid points; interpolation, padding, and boundary effects break equivariance in regions where the action of the group is not well-defined on the discretisation; and floating-point arithmetic introduces small numerical deviations even for operations that are equivariant in closed form. As a result, the networks discussed in this thesis, like most equivariant architectures in the literature, are more accurately described as *approximately* equivariant. Thus, they satisfy the equivariance relation up to a small but generally non-zero deviation. This motivates the introduction of an explicit measure of how far an architecture departs from exact equivariance, which we refer to as the *equivariance error* [Wang et al., 2022]. Rather than treating the equivariance as a binary property that either holds or does not, the equivariance error quantifies this property in a value that spans from zero, meaning perfect equivariance, to arbitrarily large values, indicating progressively greater departures from the equivariance relation. Concretely, we consider the group $SO(2)$ of rotations, parametrised by an angle $\theta \in [0, 2\pi)$. Let R_θ denote the rotation by the angle θ around the polar axis, and let $f : \mathcal{X} \to \mathcal{Y}$ represent the model that maps the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. We then define the *mean equivariance error* as follows:

$$\mathcal{E}(f, \theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{X}}[\|f(R_\theta \mathbf{x}) - R_\theta f(\mathbf{x})\|]$$

Where the expectation is taken over the data distribution.

3.3 Results

In all the following figures, the blue curve corresponds to the model trained on ERA5-Lite, while the orange and green curves refer to the new setups, respectively 24h and 2h horizon for 2-hour sampling over 2012. Each evaluation is run with $\theta \in (0°, 360°)$ with a step of 3°.

Figure 3.1a shows the equivariance error over the rotation angle for the surface variables. A common pattern emerges across all variables: the error peaks close to $\theta \in \{45°, 135°, 225°, 315°\}$ and reaches its minima at $\theta \in \{0°, 90°, 180°, 270°, 360°\}$. This is a direct signature of the underlying HEALPix grid, whose base resolution

partitions the sphere into 12 diamond-shaped pixels arranged with a four-fold symmetry around the polar axis. Rotations by multiples of 90° map pixel centres onto pixel centres, so the rotated field aligns with the original sampling lattice and the error reduces to its irreducible component. In contrast, rotations by odd multiples of 45° correspond to the maximal misalignment between the rotated and original grids, where each pixel centre lies halfway between two cells of the reference partition. This pattern is consistent across all three configurations and all variables analysed in this chapter. Looking at the comparison between curves, the green line (2h horizon) is consistently below the blue and orange ones, meaning that the 2h prediction achieves a significantly lower equivariance error across all surface variables. Comparing the two 24-hour configurations, the curves overlap for *msl* (mean sea level) and *t2m* (two-meter temperature), while the vectorial components (*u10*, *v10*) show a small gap that makes the original setup more equivariant. Another point worth noting is the absence of a common error magnitude across variables: each variable exhibits its own behaviour, reflecting how well the model preserves equivariance for that specific field. This comparison is made meaningful by the fact that the equivariance error is computed over normalised values, so that differences in the raw physical magnitude of each variable do not interfere with the analysis.
Figure 3.2 shows the same error computed over the upper-air variables at each pressure level. Here a clear gap emerges between the two 24-hour configurations, with the once-daily model (blue) achieving the lower equivariance error. This gap is not uniform across altitude: for specific humidity (q), temperature (t), and geopotential (z), the two configurations start from a comparable error at the lower pressure levels and progressively diverge towards the upper levels, where the once-daily model retains a lower error than the new-sampling one. As with the surface variables, the 2h horizon (green) remains consistently below both 24-hour curves.
Figure 3.1b summarises the upper-air results by showing the peak equivariance error per variable across pressure levels. The computation of this metric takes the maximum equivariance error $\mathcal{E}(f, \theta)$ parametrised by θ and averages the results over the 13 upper layers. Therefore, for a fixed channel c, the computations is:

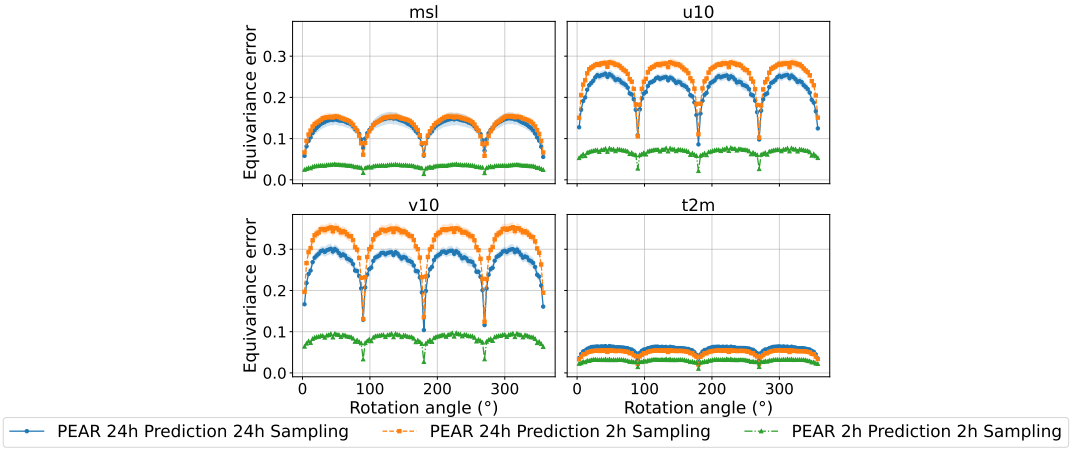
$$\text{mean max } \mathcal{E} \Rightarrow \frac{1}{13} \sum_{l=1}^{13} \left[\max_{\theta} \mathcal{E}(f, \theta) \right]_{c,l}$$

This view reveals that the equivariance of the atmospheric variables is not constant at each level. u, v (wind vectorial components) and z (geopotential) showcase a constant decrease of equivariance error from the lowest layer (1000 hPa) to the highest (50 hPa). One factor that may contribute to this trend is the asymmetry of the Earth's surface itself: the distribution of landmasses and oceans is not rotationally symmetric, so even a perfect weather prediction model would not be exactly equivariant under $SO(2)$ at levels where surface conditions strongly influence the atmospheric state. At lower pressure levels (higher altitudes), the direct influence of surface features diminishes, which could partly explain why the equivariance error tends to decrease with altitude for these variables. However, this decreasing trend is not shared by q (specific humidity) and t (temperature). t exhibits a spike of equivariance error at 200 hPa that is consistent across all three setups, while q presents a similar spike at 250 hPa, though only in some configurations. The origin of these

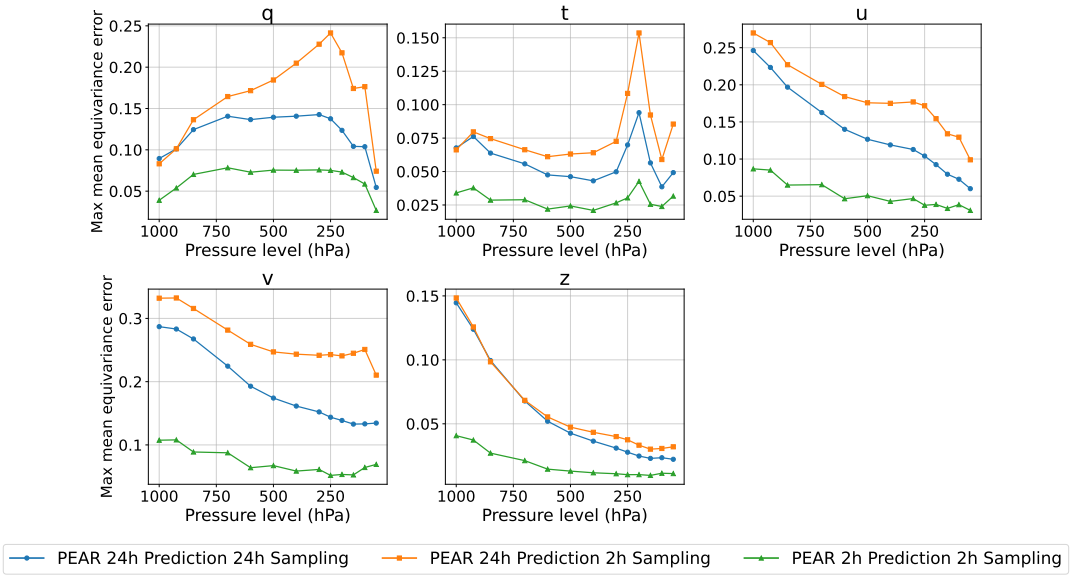
anomalies is unclear and may involve physical processes at those specific altitude ranges that are not captured by the simple landmass argument above.

3.4 Discussion

This analysis aims to determine whether denser intra-daily ERA5 samples and different prediction horizons improve the model's equivariance generalisation. The results give a clear and largely negative answer to the first question, and a more interesting one to the second. Comparing the two configurations over the same 24-hour horizon, denser sampling does not improve equivariance; on the contrary, it markedly increases the equivariance error in higher atmospheric layers. There are multiple plausible explanations for this phenomenon, which mainly relate to the diversity of the training data. Although the two setups contain a comparable number of samples, they embed a substantial difference in the prediction task. The old sampling, ERA5-Lite, contains snapshots taken at 00:00 UTC, so the training data follows a narrow distribution that captures the atmosphere at a single, fixed phase of the diurnal cycle. This makes the underlying distribution comparatively easy for the model to capture: every sample is observed at the same time of day, and the eleven-year span provides a large variety of independent atmospheric situations within this restricted setting. The 2-hour sampling, by contrast, draws observations from all phases of the diurnal cycle and therefore presents the model with a substantially broader and more complex distribution to learn. This added complexity is reinforced by the fact that the denser samples are concentrated within a single year rather than spread across eleven: consecutive 2-hour snapshots are highly correlated, so the effective number of independent training scenarios is considerably smaller than the raw sample count suggests. The model is therefore asked to capture a richer, more varied distribution while being exposed to far fewer independent atmospheric situations. The combination of a harder target distribution and reduced inter-annual diversity plausibly makes the learning problem more demanding, which would explain the higher equivariance error of the denser configuration. Specifically, on the surface level, it is interesting that the variables *u10* and *v10* present the larger gap in the equivariance error between the two 24-hour configurations, while the scalar variables *msl* and *t2m* remain largely unaffected. Although the two wind components are treated as independent channels by the model, rather than as a single vector field, they share the characteristic of being spatially more complex than the comparatively smooth pressure and temperature fields. This finer spatial structure is harder to reproduce consistently under rotation, so when the training distribution becomes more difficult to capture, as in the denser single-year sampling, the degradation in equivariance surfaces most strongly in precisely these variables.
Shifting the aim to the 2h prediction horizon, this setup achieves a lower equivariance error for each variable analysed, both at the surface and across all upper atmospheric levels. At first sight, this might appear to indicate that the 2-hour configuration is genuinely more symmetry-aware than the 24-hour ones, but the improvement is more naturally explained as a consequence of the change of prediction task itself. The atmosphere is a chaotic system, so forecasting difficulty grows rapidly with the prediction horizon: a 24-hour forecast must capture substantially



(a) Mean equivariance error vs. rotation angle for surface variables.



(b) Peak mean equivariance error vs. pressure level for each upper-level variable. Each point is the maximum over rotation angles of the epoch-mean equivariance error.

Figure 3.1: Equivariance error for the surface and upper variables.

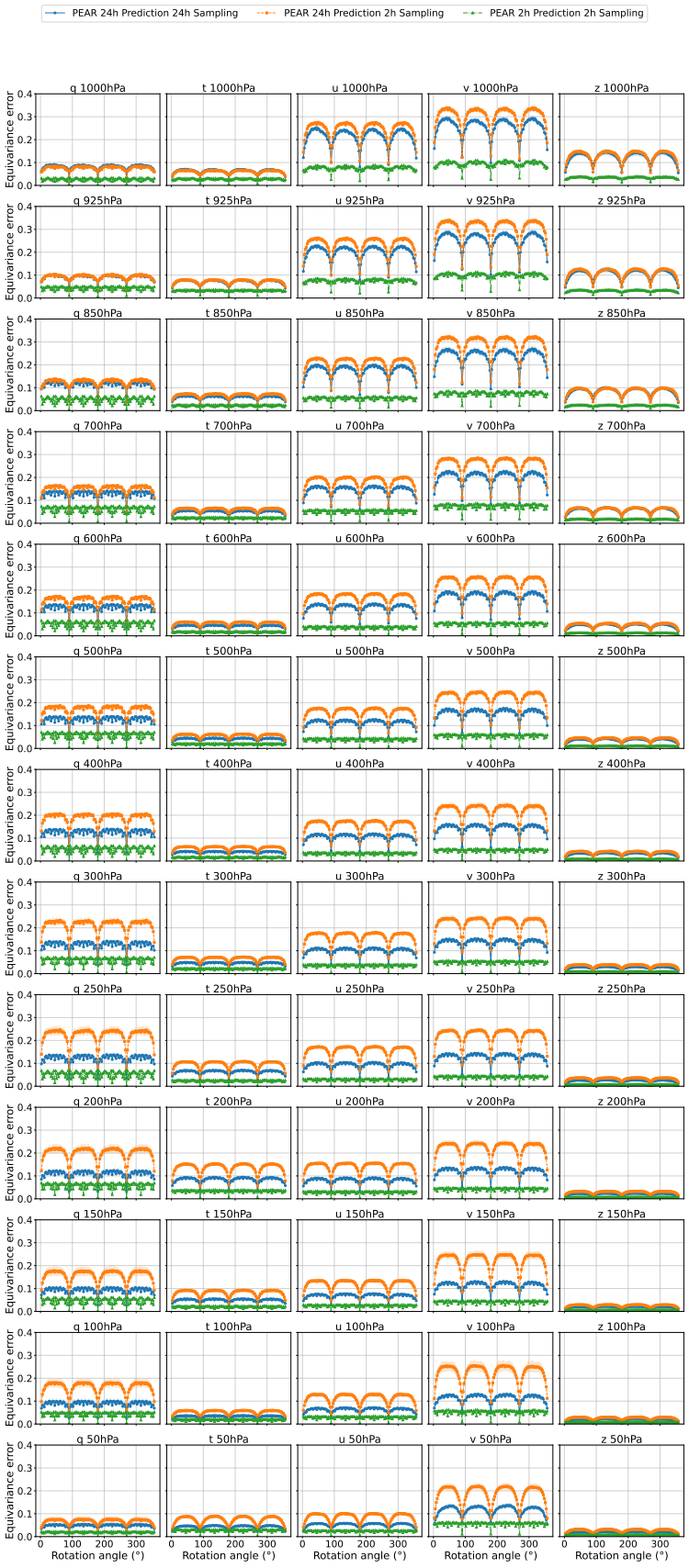


Figure 3.2: Mean equivariance error vs. rotation angle for upper-level variables at each pressure level.

more complex dynamics than a 2-hour one. A model that achieves better overall prediction performance can also be expected to exhibit lower equivariance error, since a more accurate mapping has less room for inconsistencies under rotation. At the shorter horizon the input and output states are highly correlated, so the learned mapping is closer to the identity and intrinsically easier to keep equivariant. These observations suggest that the equivariance error is strongly task-dependent: the difficulty of the forecasting problem itself influences the measured equivariance, independently of whether the model has learned any geometric structure.

The first conclusion concerns the relationship between temporal coverage of the training data and rotational equivariance. The new sampling was designed with the expectation that exposing the model to multiple phases of the diurnal cycle would encourage a representation more consistent with the rotational symmetry of the sphere, since a rotation about the polar axis is geometrically linked to a shift in local solar time. The results do not support this expectation: denser temporal sampling does not translate into improved rotational equivariance. Furthermore, the comparison across forecast horizons reveals that the equivariance error is also a task-dependent quantity, which carries relevant methodological implications. Because the difficulty of the forecasting task directly inflates the equivariance error, comparisons of equivariance are meaningful between models solving the same task. Consequently, the 2-hour prediction horizon provides a cleaner setting in which to isolate and compare the equivariance of different architectures. Because the task keeps the prediction closer to the identity, the overall equivariance error is lower and less dominated by the difficulty of the forecast, so differences between architectures are easier to observe against a smaller and more stable baseline. For this reason, the 2-hour task is adopted as the basis for the architectural study in the following chapter, where the objective shifts from the influence of the data to the influence of the model structure on rotational equivariance.

4

Architectural Modifications for Rotational Equivariance

The next step of the project is to adapt PEAR to have a better exploitation of the rotation symmetry. The limitations of the symmetry awareness of PEAR are related to the window attention mechanism inherited by the SWIN architecture. Although it operates on the HEALPix grid, a discretisation whose iso-latitude structure makes it naturally compatible with rotations about the polar axis, PEAR partitions the latent tensor into windows defined via nested indexing. Therefore, this mechanism defines a windowing system where a fixed block of pixels belonging to the same base-quadrilateral is grouped, and self-attention is computed only within each window. Because all windows share the same learned weights and the same relative positional embedding, a rotation that mapped windows to windows as a whole would simply permute the windows and leave the output unchanged up to the same permutation. However, a rotation along the longitudinal direction generally shifts pixels across window boundaries, mixing pixels that belonged to different windows in the original partition into the same window in the rotated one. Since attention is computed independently within each window, the contribution that a pixel receives from its neighbours depends on which window it falls into, and this assignment is not preserved under most rotations. Since the very first attention layer already operates on a fixed window partition, the output of that layer will differ between the original and rotated inputs whenever the rotation does not map windows onto windows. The ring-shifting performed every other layer is intended to propagate information between windows, but it does not resolve this issue: the shift is applied with respect to a fixed reference partition, so the overall sequence of attention layers remains tied to a specific alignment of windows on the sphere. In particular, as shown in Figure 3.1a and Figure 3.2, rotations by $90°, 180°, 270°$ map perfectly base-quadrilaterals onto base-quadrilaterals; therefore mapping perfectly the windows onto windows as a whole. At these specific angles, the windowing structure is preserved, the rotation acts as a permutation of complete windows, and the equivariance error is expected to be close to zero. On the other hand, at intermediate angles such as $45°, 135°, 225°, 315°$, the rotation shifts windows by approximately half a window width relative to their original positions, producing the maximum possible misalignment between the original and rotated partitions. At these angles, the equivariance error is expected to be at its highest. The following architectures will attempt to mitigate this problem and lowering the equivariant error for misaligned degrees of rotation.

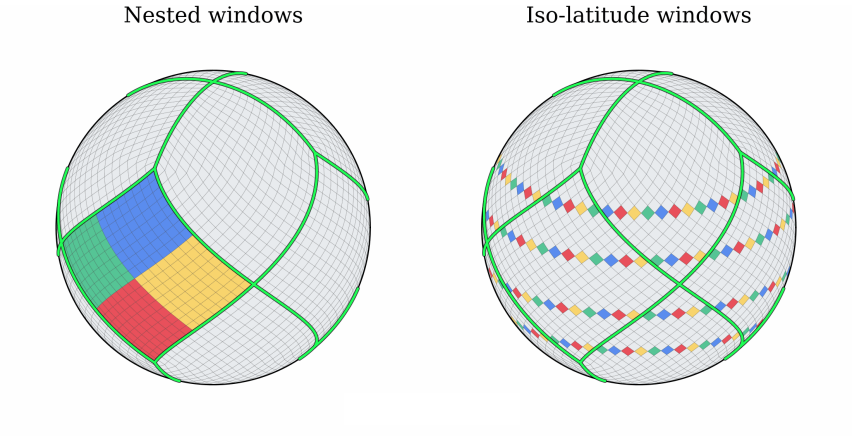


Figure 4.1: Visualization of Nested Windows (previous PEAR implementation) and Iso-latitude Windows

4.1 Iso-latitude Intersperse Windowing

The first architectural modification investigated in this thesis replaces the windowing scheme used by PEAR with one that respects the iso-latitude structure of the HEALPix grid. Instead of relying on the twelve base quadrilateral, the partition is defined ring by ring along constant-latitude circles. Furthermore, the windows in each partition are defined in an interspersed fashion, which means that each ring is split in N_w interleaved subsets. the k-th subset is obtained by selecting every N_w-th pixel starting from the index k, so that the pixels within a single subset are spaced uniformly along the iso-latitude circle. Each subset is then used as one attention window, optionally padded with repeated pixels and masked in the attention computation when its length does not match the window size exactly. More formally let $T = [T_1, T_2, \ldots, T_L] \in \mathbb{R}^{L \times C}$ denote the sequence of tokens along a ring of length L, where each $T_l \in \mathbb{R}^C$ represents one pixel. It is then possible to define a form, where each window k groups every token whose index belongs to the same equivalent class of k mod N_w, i.e., $T_k, T_{k+N_w}, T_{k+2N_w}, \ldots$ up to L :

$$W_k = \{T_l \ : \ 1 \leq l \leq L, l \equiv k \bmod N_w \}, \qquad k = 1, \ldots, N_w.$$

In this way, each window aggregates tokens from spatially separated but regularly spaced longitudinal positions, allowing the self-attention to mix information across the full longitudinal range of the ring rather than within a contiguous arc of it. It is possible to visualise the new implementation in Figure 4.1.
The key equivariance property of this scheme follows directly from the interspersed construction: because the partition is defined modulo the ring length, an axial rotation by any integer number of pixels acts as a cyclic shift of the longitudinal indices, which permutes tokens across windows but preserves the partition structure itself. Since all windows share the same learned weights and positional embeddings, this permutation leaves the output unchanged up to the corresponding permutation, making each attention layer equivariant under any discrete longitudinal shift, not only shifts that are multiples of the window size.

However, this equivariance comes at the cost of spatial locality. Because each window samples every N_w-th pixel along the ring, neighbouring pixels on the sphere are assigned to different windows and never interact directly within a single attention layer. The scheme therefore effectively operates at a coarser spatial resolution within each window, since the spacing between attended pixels is N_w times the grid spacing. The nested partition used by PEAR groups together pixels that are close on the sphere in both directions, preserving spatial locality at the expense of equivariance under longitudinal rotations; the iso-latitude scheme, conversely, achieves equivariance under such rotations but sacrifices the notion of local spatial neighbourhood that motivates windowed attention in the first place. Furthermore, the iso-latitude partition does not aggregate information across nearby latitudes: pixels belonging to adjacent rings are assigned to different windows entirely. Information across latitudes can still propagate through the architecture, but only indirectly, through the combination of patch embedding, downsampling, and the shifting mechanism applied between attention layers.

4.2 HEALPix-aware Convolutions

The second architectural implementation to be tested in this project is designed to mitigate the Iso-latitude Intersperse Windowing issues while maintaining a lower equivariance error than PEAR. The present modification aims to recover a notion of spatial locality without further usage of self-attention mechanisms. Rather than acting on the windowing scheme, the locality will be retrieved using HEALPix-aware Convolutions; these convolutions are a building block of the DLWP-HPX model [Karlbauer et al., 2023]. The key idea behind the HEALPix-aware convolution is to treat each of the twelve base faces of the HEALPix grid as a regular two-dimensional image, on which standard 2D convolutions can be applied directly, and to handle the connections between faces through a custom padding operation that fills the boundary of each face with values copied from its spherical neighbours. This allows the kernel to compute a spatially consistent receptive field across face boundaries, preserving spatial locality while respecting the global topology of the grid. The padding operation is the technical core of the method. Each of the twelve faces has four neighbouring faces on the sphere, and the alignment of these neighbours depends on whether the face belongs to the northern hemisphere, the equatorial band, or the southern hemisphere. In some cases, the neighbouring face must be rotated before being concatenated to the boundary of the central face, since adjacent faces do not share a common orientation in the 2D image representation (Figure 4.2). The architecture does not intrinsically embed $SO(2)$-equivariance, since each of the twelve faces of the HEALPix grid is treated as an individual image. As a result, the equivariance properties of the operation can be analysed locally, within a single face, but not on the global scale of the sphere. It is nevertheless possible to argue for an *approximate* equivariance under axial rotations within specific regimes. Intuitively, a rotation about the polar axis corresponds to a shift in longitude; when restricted to a sufficiently small neighbourhood, the spherical surface can be approximated as planar, and the rotation acts approximately as a translation in the flattened face coordinates. Since standard 2D convolutions are translation-equivariant away

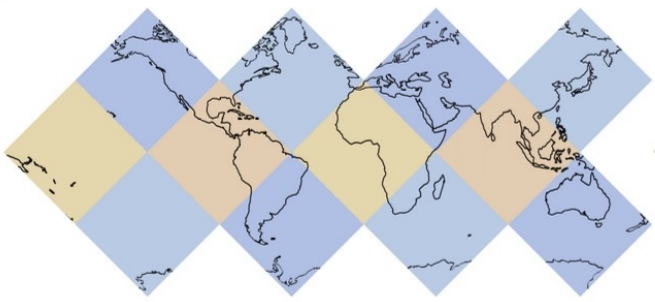


Figure 4.2: Flattened representation of the HEALPix division of the sphere [Karlbauer et al., 2023]

from boundaries, the operator can be regarded as approximately rotation-equivariant within the interior of each face, provided that the convolutional kernels remain small compared to the scale on which the local coordinate distortion varies. Within each face, the convolution is therefore exactly translation-equivariant away from the boundaries. The geometry-aware padding scheme extends this property to the global scale: by injecting values from the adjacent faces with the appropriate re-orientation, it ensures that the convolutional kernel sees a spatially consistent neighbourhood even at the boundaries between faces. The resulting operation can thus be regarded as approximately $SO(2)$-equivariant over the full sphere, with the approximation limited by the coordinate distortion inherent to the flattened face representation and by the discrete nature of the grid.
In the DLWP-HPX architecture, these convolutions are organised into a more elaborate building block, the ConvNeXt block [Liu et al., 2022] (Figure 4.3), which combines several convolutions in a structured way. The block follows an inverted bottleneck structure: given an input with embedding dimension D, a 1×1 convolution first projects it to D channels, a pair of 3×3 HEALPix-aware convolutions then process the features at an expanded depth of $4D$, and a final 1×1 convolution projects the result back to D. A residual connection [He et al., 2015] links the input of the block to its output, so that each block learns a refinement of its input rather than a full transformation.
These building blocks will be implemented and tested in different setups. Firstly, the ConvNeXt block will be used in the patch embedding, replacing the 1D and 2D convolutions of the original PEAR architecture. The original patch embedding already respects the geometry of the HEALPix grid: its kernel size and stride of 4 group consecutive blocks of 4 pixels, which correspond to a single quadrilateral one level above the grid resolution in the nested indexing, effectively reducing the resolution by one HEALPix level. The motivation for replacing it with a ConvNeXt block built on HEALPix-aware convolutions is to extend this to account for face boundaries through geometry-aware padding, so that pixels near the seams between base quadrilaterals also receive spatially consistent embeddings. Secondly, both the HEALPix-aware convolution and the ConvNeXt block will be inserted as additional layers immediately before each windowed attention block, in order to recover the spatial locality that the iso-latitude windowing scheme sacrifices. By pre-mixing

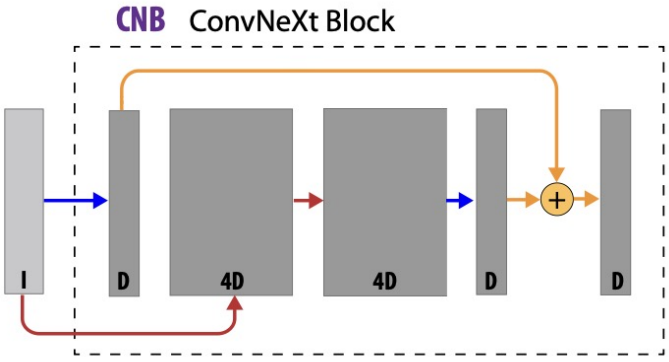


Figure 4.3: Representation of the ConvNeXt block [Karlbauer et al., 2023] . The orange arrows represent the Res-Net connection, blue arrows are the 1×1 convolution, red arrows are 3×3 convolutions, and D represent the depth or the number of channels.

each pixel with its spherical neighbours through a convolution, the attention layer no longer has to rely solely on its window partition to capture local interactions, since this information is already encoded in the convolved features it receives as input. In this second configuration, the HEALPix-aware convolution and the ConvNeXt block will be tested independently rather than jointly, so that the effect of each can be isolated. The first variant inserts a single HEALPix-aware convolution before each attention block, providing a minimal form of geometry-aware feature aggregation; the second variant replaces this single convolution with a full ConvNeXt block, which offers a deeper and more expressive local mixing through its inverted-bottleneck structure. Testing both implementations against the baseline allows us to determine whether the introduction of HEALPix-aware convolution and the additional capacity of ConvNeXt translates into a measurable improvement.

4.3 Training and Evaluation

The training of the new models follows a similar approach compared to the PEAR implementation. However, due to substantially heavier parametrisation of the models, we implemented a learning rate scheduler to attempt a more stable training.
The results are presented over two main evaluation metrics: The Equivariance Error and Root Mean Square Error (RMSE). The Equivariance Error has been described in Section 3.2, and measures the capability of generalising the rotational symmetry around the axis for an angle θ, with an error score where 0 indicates perfect equivariance. The RMSE for this type of discretisation is implemented as follows:

$$\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y^i - \hat{y}^i)^2} \tag{4.1}$$

Weather Forecasting research commonly implements the Anomaly Correlation Coefficient (ACC) as an additional type of downstream metric. This is used to measure the correlation between deviations from the climatology mean of predicted and

ground truth forecasts, with a value of 1 indicating perfect agreement. It is formulated as follows

$$\mathrm{ACC}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N} \Delta y^i \Delta \hat{y}^i}{\sqrt{\left(\sum_{i=1}^{N} \left(\Delta y^i\right)^2\right)\left(\sum_{i=1}^{N} \left(\Delta \hat{y}^i\right)^2\right)}} \tag{4.2}$$

Where Δy is the difference between the prediction and the climatology average. This metric is not reported in the following results section, since it reflects the observation from the RMSE, the full plots are reported in the Appendix Section A.2

4.4 Results

This section presents the experimental evaluation of the architectural modification presented in the previous Section 4.1 and Section 4.2. The evaluation is organised around the two central questions of this thesis: whether the proposed modifications improve the rotational equivariance of the model, and whether any such improvement comes at the cost of forecasting skill. The first part analyses the equivariance of each model directly, by measuring the equivariance error as a function of the rotation angle about the polar axis, following the methodology described in Section 3.2. The second part evaluates the forecasting performance of the same models, reporting the root mean squared error (RMSE) for the surface and upper atmospheric variables over a range of lead times. The new architecture implementations are compared to the baseline PEAR, specifically the experiment involves the following models:

- **PEAR**: the baseline model described in Section 2.2 (4,279,881 parameters).
- **PEAR-Isolatitude**: the model incorporating the iso-latitude self-attention aggregation described in Section 4.1 (1,527,369 parameters).
- **PEAR-Isolatitude + Convolution** (3×3): PEAR-Isolatitude extended with a ConvNeXt block as a spatially-aware patch embedding, together with HEALPix-aware convolutions inserted before the transformer layers. The convolution kernel size is 3×3. (2,534,361 parameters)
- **PEAR-Isolatitude + Convolution** (5×5): identical to the previous configuration, but with a larger 5×5 convolution kernel, in order to assess the effect of an enlarged receptive field (4,110,297 parameters).
- **PEAR-Isolatitude + ConvNeXt**: a variant in which the HEALPix-aware convolutions before the transformer layers are replaced by full ConvNeXt blocks, so as to compare a simple convolution against a deeper, more expressive residual block at the same position in the architecture (6,568,665 parameters).

As it follows from the previous analysis, these architectures are trained and evaluated over the new sampled dataset for the 2h prediction task. Therefore, the training set is the year 2012, and the evaluation set is the year 2019.

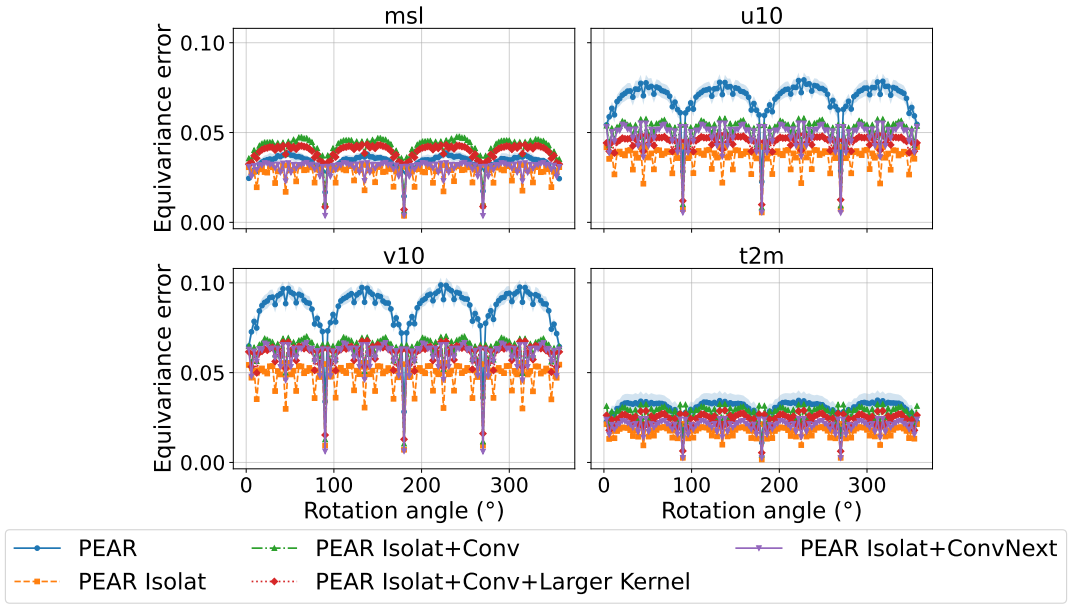
4.4.1 Equivariance Results

The Figure 4.4a shows the equivariance error over the rotation angle θ in degrees, specifically, each evaluation is run with $\theta \in (0°, 360°)$ with a step of 3°. The same idea is shown in the Figure 4.5, which represents the same error computed over the

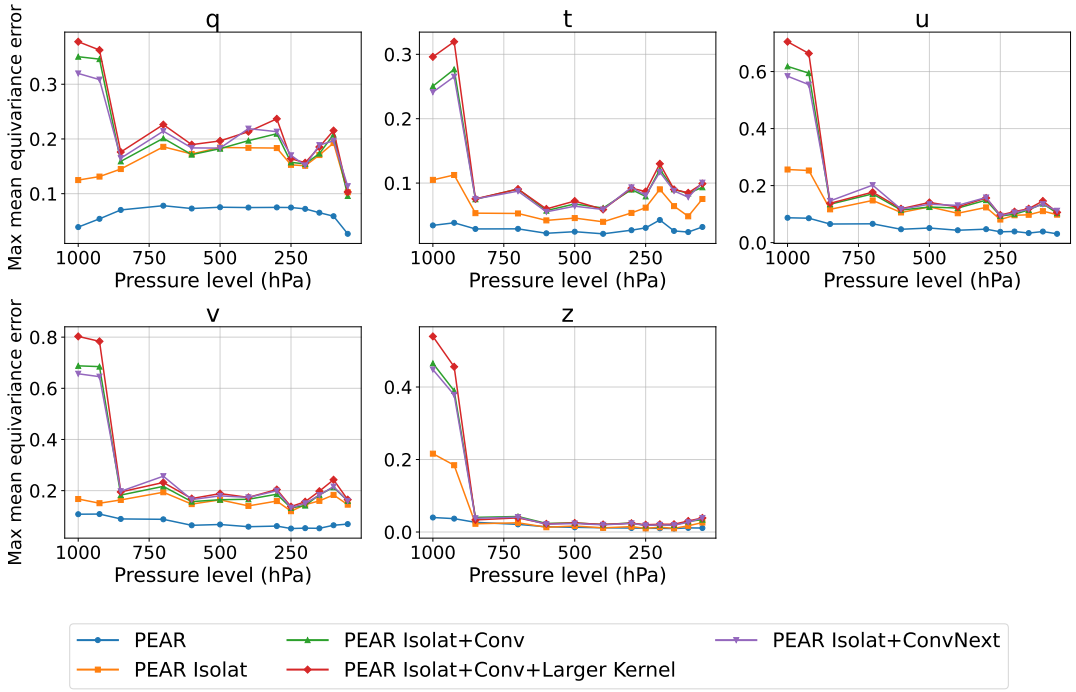
atmospheric variables taken into account during the experiments. Due to the large amount of information contained in the atmospheric figure, a secondary plot has been created (Figure 4.4b) to highlight the max equivariance error computed per variable, over the different atmospheric layers.
The first result that is possible to notice is that the HEALPix symmetry outlined by the previous analysis's results (Section 3.3) seems to be aligned with the one computed over the new architectures. The figure 4.4a shows a drop of error in relation to the angles $90°, 180°, 270°$, and it shows the maxima around $45°, 135°, 225°, 315°$ where the rotated partitions are maximally misaligned with the original one. However, this type of trend does not hold consistently for the upper variable (Figure 4.5), where for the pressure levels 1000hPa and 925hPa, the rotation angles $90°$ and $270°$ show the maximum equivariance error consistently between the two prediction tasks.
Another observation that substantially differentiates the surface and upper result is the actual comparison with the baseline (PEAR). The picture described by the surface variables outlines an overall lower error for the new architectures. Both the iso-latitude windowing and the HEALPix-aware convolution variants show a reduced equivariance error compared to PEAR across most of the rotation range, except for the convolution implementation for the *msl* (mean surface level pressure). This is consistent with the design goal of the proposed architectures, which were intended to align more closely with the rotational symmetry of the sphere. For the upper variables, however, the picture tells a different story: PEAR shows a better equivariance in every setting. The Figure 4.4b shows that the gap between the baseline and the new implementations is remarkable at higher pressure levels (1000hPa and 925hPa) and tend to reduce meaningfully across the layers.

4.4.2 Performance Results

The following results report the RMSE computed over 300 epochs of training of the models. The outputs present multiple channels, and each one reflects unique information related to the weather state analysed. In order to keep the result concise and still covering qualitatively different types of physical quantities, the values reported focus on two representative variables: the temperature and the eastward component of the wind, shown for both surface and upper levels. These two variables are chosen because they capture distinct aspects of the atmospheric state. The temperature is a scalar field, whose value at each location is independent of any choice of orientation, whereas the wind components describe the eastward component of the wind vector field. The remaining variables follow broadly similar trends and are reported in full in Appendix Section A.1. Specifically, this downstream metric is computed over the unnormalised values, so that the error is expressed in the physical units of each variable rather than in the normalised space in which the model is trained.
Figure 4.6a presents the surface results computed for two-meter temperature and ten-meter u component of the wind, while Figure 4.6b and Figure 4.6c give an overview of the error in the atmospheric variables. Figure 4.6b shows the mean RMSE computed over the different pressure levels, over the 300 epochs of training, showcasing any possible overfitting during the training. Figure 4.6c instead selects, for each model, the epoch with the best validation RMSE achieved during train-



(a) Mean equivariance error vs. rotation angle for surface variables.



(b) Peak mean equivariance error vs. pressure level for each upper-level variable. Each point is the maximum over rotation angles of the epoch-mean equivariance error.

Figure 4.4: Equivariance error for the surface and upper variables.

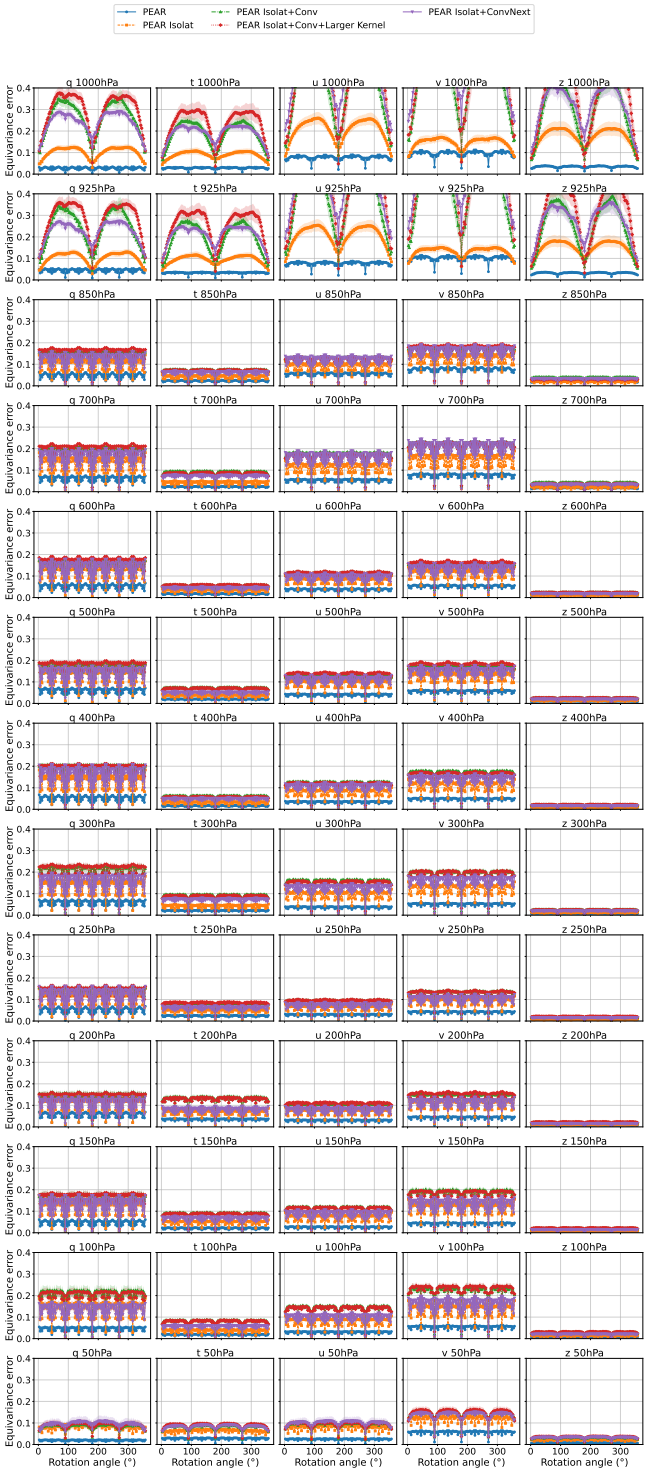


Figure 4.5: Mean equivariance error vs. rotation angle for upper-level variables at each pressure level.

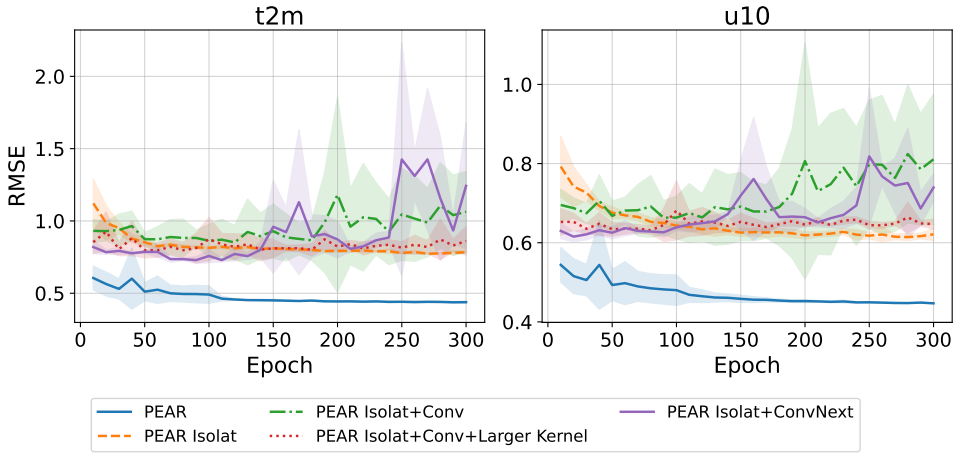
ing, and plots this error against pressure level, to characterise how the forecasting skill of each architecture varies with altitude. Together, this combination of figures separates two distinct aspects of the models' behaviour. On the one hand, the epoch-wise view in the first case concerns the optimisation dynamics, how quickly and how stably each architecture converges, and whether the additional parameters introduced by the convolutional variants lead to overfitting under the limited training budget. On the other hand, the level-wise view of the second case concerns the final predictive quality, revealing whether the differences between the architectures are uniform across the atmospheric column or concentrated at particular pressure levels.

In this 2h prediction task, the PEAR baseline achieves the lowest RMSE across both the surface and the upper variables, reaching approximately 0.4 K for the temperature on both the surface and the upper levels, and 0.4 m/s and 0.75 m/s for the u-component of the wind on the surface and upper levels, respectively. Its training curves decrease smoothly and stabilise without any major divergence between successive epochs, showing no sign of overfitting on both variables. The iso-latitude architecture also trains stably, but converges to a higher error of approx 0.8 K, 1.1 K for surface and upper and 0.6 m/s, 2.3 m/s for the eastward wind component. Adding the convolutional blocks, by contrast, introduces greater instability into the training: the convolutional variants all reach a lower error than the iso-latitude architecture alone, but they appear to overfit, with their validation error worsening in the later epochs rather than continuing to improve. Especially, the introduction of the ConvNeXt block creates high spikes in the validation loss around epochs 150 and 250.

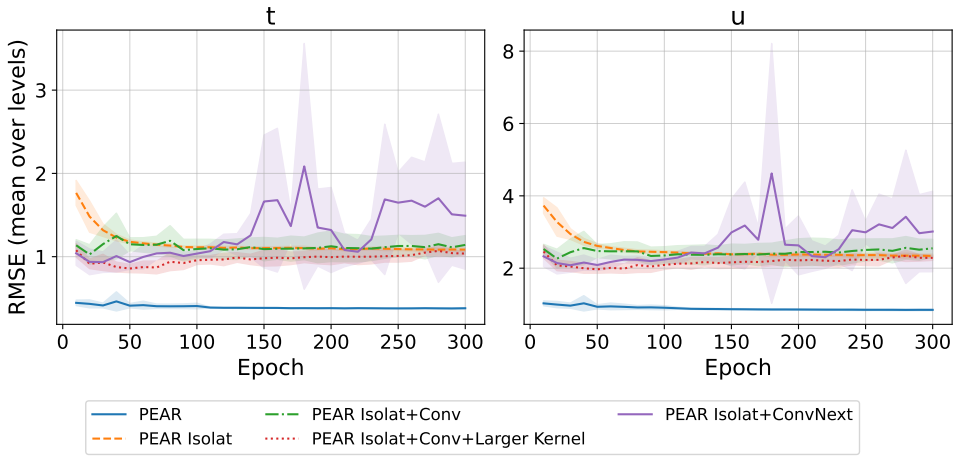
Figure 4.6c reveals patterns that reflect the vertical structure of the atmosphere. The first observation is that the results described above are largely consistent across the upper levels. PEAR performs strongly, with errors that remain stable across the pressure levels, whereas the other architectures exhibit higher errors and a clearly less uniform behaviour, with the largest deviations occurring at the high-pressure levels (1000hPa and 925hPa) and at the low-pressure levels (50hPa) for both variables. The degradation at the high-pressure levels is consistent with the elevated equivariance error measured at the same levels in Section 4.4.1, where the proposed architectures showed their largest departures from exact equivariance near 1000hPa and 925hPa.

4.4.3 Qualitative Results

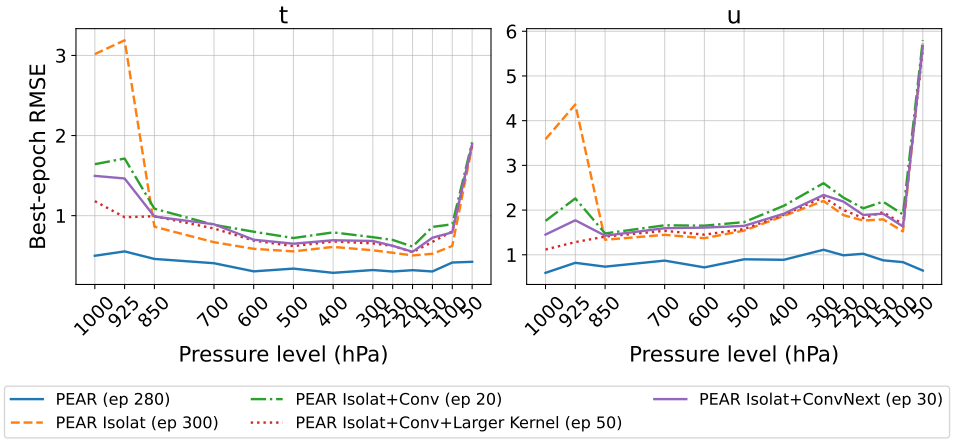
The final part of the results presents a qualitative analysis of the models' predictions. Two days are randomly selected from the validation set, and on each of them, the predicted surface temperature is examined alongside the ground truth. For both samples, three quantities are visualised on the sphere: the ground-truth field, the model's prediction, and the prediction error, the latter shown as a pointwise difference between the two. To provide a valuable comparison, the two samples are related to different parts of the year, such that it is possible to observe the model behaviour in relation to the seasonality of the samples. Specifically, Figure 4.7 represents the predictions over the 2$^{\text{nd}}$ of February 2019 at 8 a.m., and Figure 4.8 shows the same over the 8$^{\text{th}}$ of November 2019 at 6 p.m.



(a) RMSE over training epochs for the surface variables *t2m* and *u10*.



(b) Mean RMSE over training epochs for the atmospheric variables t and u.



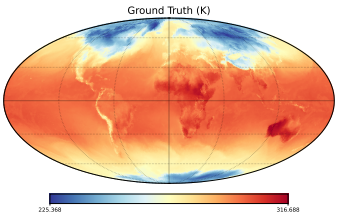
(c) Best RMSE over pressure level for the atmospheric variables t and u.

Figure 4.6: Validation RMSE for the surface and upper variables.

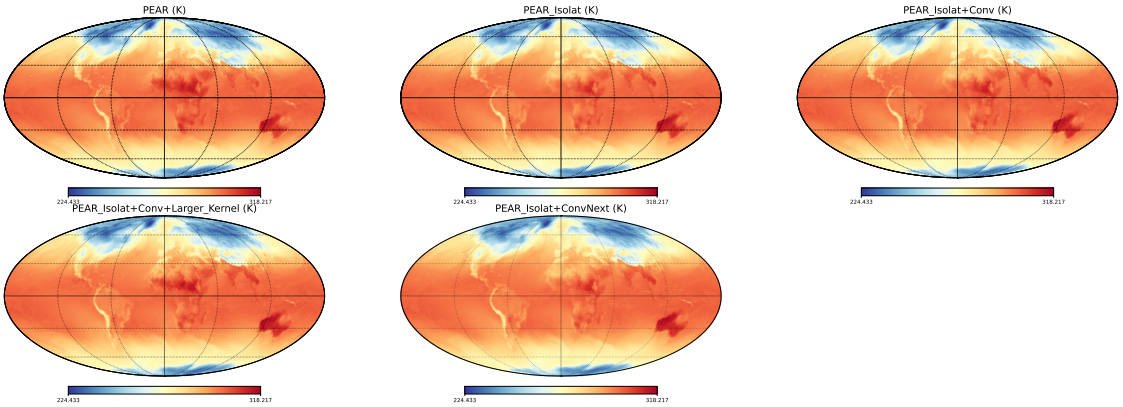
By observing the model predictions (Figure 4.7b and Figure 4.7b), all the architectures produce outputs that are visually reasonable and broadly consistent with the ground truth, reproducing the main large-scale features of the temperature field on both dates. At this level of inspection, no architecture stands out as obviously qualitatively better or worse than the others; the predicted maps look similar to one another and to the corresponding ground-truth field. Notably, none of the predictions exhibit visible artifacts from the underlying HEALPix grid structure, indicating that the discretisation does not introduce spurious spatial patterns into the forecasts. The differences between models only become visible when the error fields are examined.
The error plots (Figure 4.7c and Figure 4.8c) reveal more interesting insights into the prediction of the model tested. The first observation is that, across all models, the spatial structure of the error closely outlines the continental landmasses, with the largest deviations concentrated over land and noticeably smaller errors over the oceans. This pattern is consistent across both samples and suggests that the models capture the smoother oceanic temperature field considerably more accurately than the more variable temperatures over land. The error of the PEAR baseline is comparatively homogeneous across the globe, with no strongly localised regions of systematic over- or under-prediction. The proposed architectures, in contrast, exhibit a more pronounced regional bias: in the February sample, they tend to under-predict the temperature over Africa and over-predict it over Asia, while in the November sample, they over-predict the temperature over both Africa and Asia.

4.5 Discussion

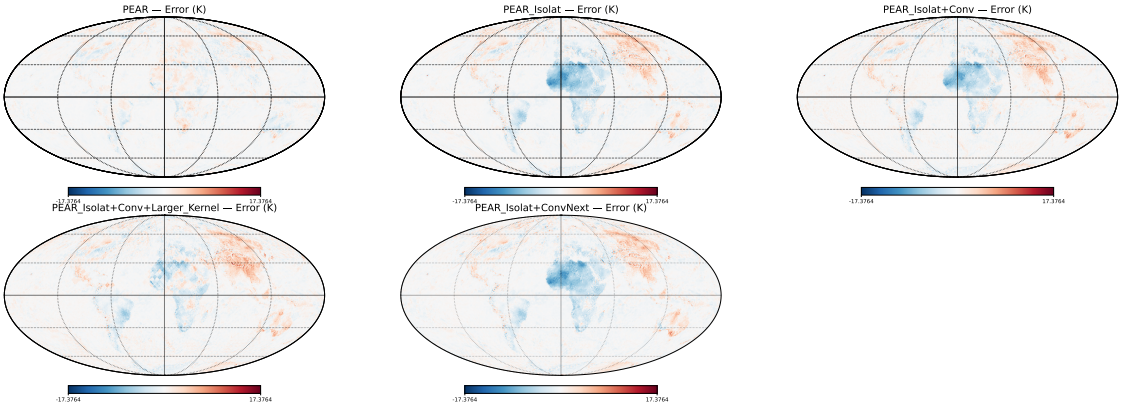
This analysis presents the comparison of four different modifications of PEAR, designed to reduce the equivariance error and therefore better respect the rotational symmetry of the sphere. The evaluation against the baseline shows mixed results: the proposed architectures achieve, on average, lower equivariance error on the surface variables, but PEAR remains the more equivariant model on the upper variables. This partial improvement does not translate into a forecasting advantage, with the proposed architectures either reaching higher RMSE than PEAR or exhibiting noisier and less stable training curves. This discussion examines these findings in more detail, addressing the contrast between surface and upper variables, the disconnect between equivariance and predictive skill, and the role of the limited training budget in shaping the observed behaviour.
Starting with the equivariance findings, the purpose of the new architecture was to bridge the gaps between the symmetric transformations discovered in the previous analysis (Section 3.3). Therefore, the goal was to flatten the angular profile of the equivariance, in particular, achieving a reasonably lower equivariance for rotations with the angles $45°, 135°, 225°, 315°$. The new architectures described in Section 4.1 and Section 4.2 manage to achieve this result on the surface level, but not on the atmospheric variable level. On the surface variables, the iso-latitude windowing alone achieves the lowest equivariance error, and the HEALPix-aware convolutions contribute to raising this error slightly, yielding higher error compared to the iso-latitude architecture alone, but still lower than the PEAR. The only exception is



(a) Ground truth.

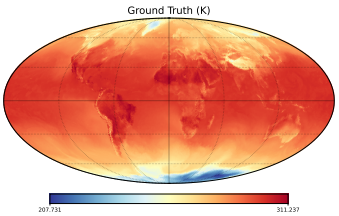


(b) Model predictions.

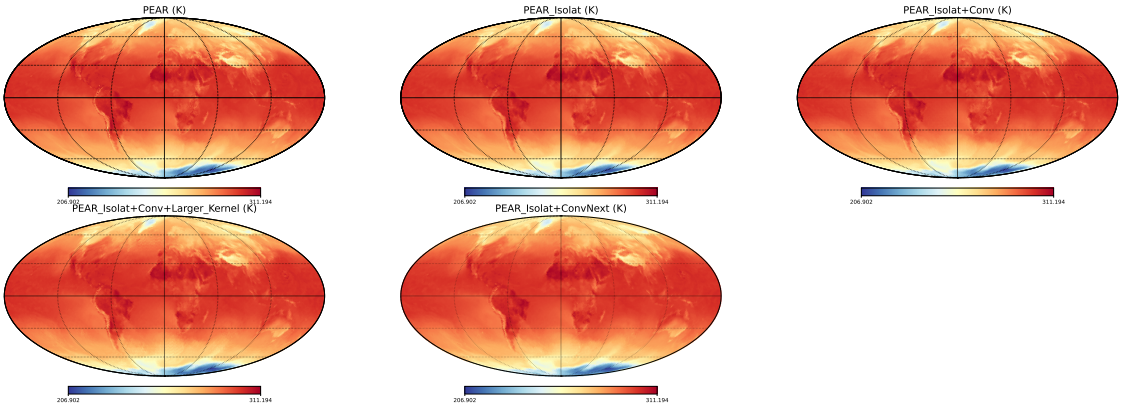


(c) Prediction error.

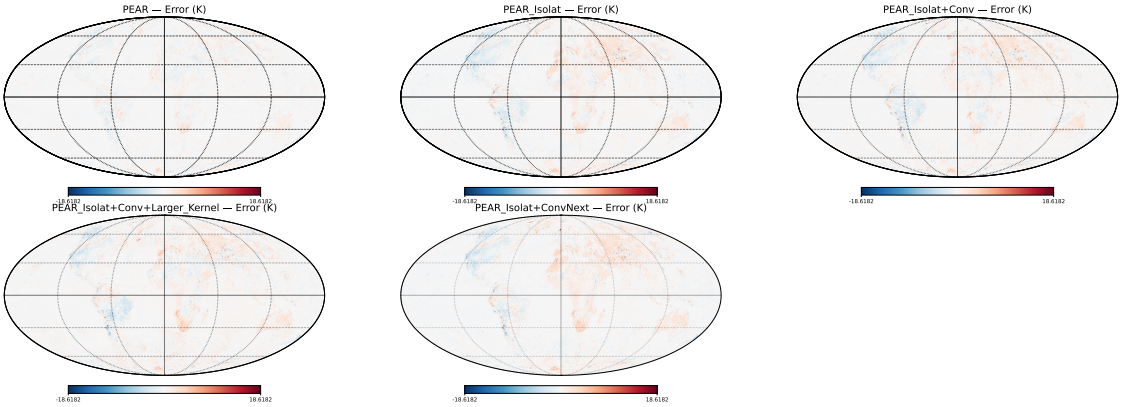
Figure 4.7: Qualitative comparison for the sample of 2 February 2019: ground truth, model predictions, and the corresponding error.



(a) Ground truth.



(b) Model predictions.



(c) Prediction error.

Figure 4.8: Qualitative comparison for the sample of 8 November 2019: ground truth, model predictions, and the corresponding error.

for the 3×3 and 5×5 convolution for the *msl* variable, where they yield a higher error than PEAR for all the rotation angles. The reason for this variable-specific behaviour is not clear from the present analysis. Overall, these results indicate that the iso-latitude modification is the principal driver of the surface-level equivariance improvement, with the convolutional variants offering at best a comparable behaviour. In the upper layers, the opposite result is outlined: PEAR consistently achieves a lower equivariance error compared to the new implementation, with the gap widening as the pressure levels approach the surface. The largest deviations occur at 1000 hPa and 925 hPa, after which the gap narrows progressively towards the higher atmospheric layers. This asymmetry is not the consequence of any transformer-level modification per se: both the iso-latitude windowing and the HEALPix-aware convolutions are applied uniformly across all tokens in the network, with no special treatment for surface or upper variables. The asymmetry must therefore arise from the interaction between the uniformly applied modifications and aspects of the architecture or the data that are themselves level-dependent. The only parts of the models that implement layer-specific operation are the Patch Embedding and the Patch Extraction, thus the operations that embed the input into the token space and vice versa. Both the old implementation of the Patch Embedding and the new one with the ConvNeXt block use convolution on the surface and upper variables independently, thereby creating two linear mappings, defined over separate weights, that do not aggregate surface and upper variables. Same for the Patch Extraction, which projects the surface and upper tokens back to the physical variables through two independent transpose convolutions, mirroring the structure of the patch embedding in reverse. As a result, surface and upper variables enter and leave the network through separate pathways, and the only stage at which they interact is the attention itself, which is volumetric and groups tokens from different vertical levels into shared windows. Therefore is difficult to pinpoint the precise block that breaks the equivariance in the new architecture; additionally, these variables represent complex physical processes that require specific domain knowledge to be interpreted rigorously. A precise mechanistic explanation would require both targeted ablation experiments, for example, isolating the effect of the patch embedding from that of the volumetric attention, or comparing configurations in which surface and upper streams are processed in parallel rather than jointly, and a deeper analysis of how the rotational symmetry interacts with the vertical structure of specific atmospheric variables. Such an investigation lies beyond the scope of this thesis, whose focus is on the architectural side of the problem rather than on the meteorological interpretation of the resulting error patterns.

The performance results of the Section 4.4.2 outline a clear answer: PEAR presents better performance compared to the new implementation for the 2h prediction task. The baseline achieves a lower error in both the surface and the upper atmospheric variables, with smooth and stable training curves that show no sign of overfitting under the chosen training budget. None of the proposed architectures matches this performance: the iso-latitude variant trains stably but converges to a higher error, while the convolutional variants reach lower errors than the iso-latitude model alone but at the cost of training instability and clear signs of overfitting in the later stages of training. This result indicates a trade-off in the iso-latitude implementation

between the inductive bias for rotational symmetry and the model's generalisation capability. By replacing the nested windowing with an iso-latitude scheme, the architecture gains a closer alignment with the rotational structure of the sphere, but loses the spatial locality of the original PEAR windows. The iso-latitude interspersed windowing proves to be ineffective in handling the complex generalisation problem posed by weather forecasting in this setting. The convolutional variants exhibit noisier training and a tendency to overfit in later stages of training. Whether this is an inherent limitation of the architectural design or a consequence of the increased parameter count and insufficiently tuned hyperparameters cannot be determined from the present experiments alone. In the first 100 epochs, the convolutional models achieve a lower error than the iso-latitude alone, meaning that the additional locality introduced by the HEALPix-aware convolutions does provide a measurable benefit in the early stages of training. By reintroducing a local receptive field on top of the iso-latitude attention, they recover some of the spatial inductive bias that the iso-latitude windowing discards. However, this initial advantage does not persist. After approximately 100 epochs, the validation error of the convolutional models begins to fluctuate and gradually rises, while the iso-latitude model alone continues to train smoothly without overfitting. Among the three convolutional variants, the ConvNeXt block exhibits the most pronounced instability, with visible spikes in the validation loss around epochs 150 and 250. This may be related to it being the largest of the tested models, with 6,568,665 total parameters, compared with 4,110,297 for the 5×5 variant, 2,534,361 for the 3×3 variant, and 4,279,881 for the PEAR baseline (the iso-latitude model alone has 1,527,369 parameters). The instability could also stem from training hyperparameters that were not specifically re-tuned for the larger model. The 3×3 and 5×5 variants follow a similar pattern of early-training improvement followed by gradual overfitting, but the effect is less severe. Interestingly, the 3×3 kernel exhibits noisier training than the 5×5, suggesting that the training dynamics may depend on factors beyond raw model capacity, such as the inductive bias provided by a larger receptive field or the sensitivity of a given configuration to the shared set of hyperparameters.

A further observation links the performance results back to the equivariance analysis of Section 4.4.1. The gap in validation error is not uniform across the surface and upper variables; it is comparatively narrow on the surface variables, where the proposed architectures achieve a lower equivariance error than PEAR, and noticeably wider on the upper variables, where PEAR retains the equivariance advantage. Specifically, the temperature exhibits a gap of $\approx$0.3 K of difference in the RMSE for the surface variables, while it averages $\approx$0.6 K of difference for the upper variable. The same idea is outlined by the wind component, it shows $\approx$0.15 m/s and $\approx$1 m/s of RMSE difference for surface and upper variables. These two metrics therefore point in the same direction on a level-by-level basis: the levels where the proposed architectures are more equivariant are also the levels where their RMSE gap with respect to PEAR is smallest, and the levels where they are less equivariant are also those where the gap is largest. As in the previous discussion, it is not possible to infer a causal relationship from this observation without further targeted experiments, but the consistency between the two metrics suggests that equivariance and forecasting accuracy may be influenced by common underlying factors at each level. Finally,

from the qualitative analysis of Section 4.4.3, it is possible to observe another point of view related to spatial prediction of the models, which completes the overall picture provided by the equivariance and RMSE analyses. Two observations stand out from the qualitative results. The first is shared by all the architectures: the spatial structure of the prediction error closely traces the continental landmasses, with the largest deviations concentrated over land and noticeably smaller errors over the oceans. It reflects the intrinsic difficulty of forecasting temperature over land, where multiple factors such as topography or diurnal cycle can introduce spatial variability. This is a property of the data rather than of the models, and it places a shared lower bound on what any of the tested architectures can achieve over the continents. The second observation is architecture-specific, and it's related to the distribution of error over the landmass. The error of PEAR is spatially homogeneous in a particular sense: within the same region, the model both over-predicts and under-predicts the temperature, with positive and negative errors appearing in close spatial proximity rather than being clustered into systematically biased areas. The proposed architectures, in contrast, exhibit pronounced regional biases that vary with season, in which entire continental-scale areas share the same sign of error rather than being a mixture of positive and negative deviations. Therefore, on top of the higher RMSE, the new architectures also present regional patterns rather than locally balanced errors, indicating that the inaccuracy is not only larger in magnitude but also more systematic in its spatial distribution.

5
Conclusion

The thesis investigates whether the rotational equivariance of PEAR can be improved, both from a new sampling of the data and from the architectural modifications, and whether these improvements could translate into better forecasting skills. The outcome of the thesis highlighted that the equivariance can be partially improved through architectural modifications, but neither the data nor the architectural analysis produces relevant improvements compared to the baseline.

The study on the dataset and the horizon effect attempted to find a correlation between the equivariance and dataset modification. The results of this analysis suggest that the equivariance error is influenced more strongly by the architecture and the forecast task than by the temporal sampling of the training data. Denser intra-day sampling did not show any improvement in the generalisation of the symmetry of the system, suggesting that a wider range of independent atmospheric scenarios may be more valuable than finer temporal resolution for equivariant generalisation. Furthermore, the forecast horizon dominates the equivariance; the 2h task was revealed to be more equivariant than the 24h task. This dominance reflects the fact that a shorter horizon brings the prediction task closer to an identity mapping, which is intrinsically easier to keep equivariant than the more complex transformation required for a full-day forecast. Therefore, in order to seek an equivariance improvement, it is necessary to apply direct modification to the model architecture, not the data, and the 2h task provides the cleanest setting in which to study how architectural choices influence the model's rotational symmetry. This conclusion directly motivates the focus of the second analysis, which examines whether architectural modifications can deliver an equivariance improvement under controlled conditions.

Given this first analysis, the new architectures rely on iso-latitude windowing and HEALPix-aware convolutions in order to embed the rotational symmetry in the model. These models succeed in generalising better the symmetry of the system on the surface level, while the upper levels lack the same type of generalisation, creating an intrinsic discrepancy in symmetry learning from the models. A parallel pattern is observed in the performance results: PEAR outperforms the proposed architectures, but the gap is narrower on the surface, where the modifications also improve the equivariance, and wider on the upper levels. The two metrics therefore agree on a level-by-level basis, even though the modifications do not deliver a forecasting advantage in aggregate. In the configurations tested, the iso-latitude windowing alone does not achieve competitive forecasting performance, and the addition of convolutions does not fully compensate for this gap.

Overall, the results of this thesis indicate that improving rotational equivariance through the proposed architectural modifications does not, by itself, translate into

better forecasting performance. Taking the two analyses together, it is possible to observe a clear picture which outlines that, first, the source of equivariance is identified as the architecture and the structure of the discretisation rather than the sampling; second, modifying the architecture along these lines partially achieves the equivariance objective but does not translate it into forecasting skill. From the combination of these two points of view, it is outlined the picture that geometric inductive biases alone may not be sufficient to improve forecasting in this setting, and that reconciling equivariance with predictive performance remains a challenging open problem. The equivariance gains delivered by the proposed modifications remain isolated from the predictive behaviour of the model, because the modifications also impose a locality cost that the available training budget cannot absorb.

5.1 Future Directions

As a consequence of the complexity of the task, many parts and details of the thesis can be further developed and extended to have a broader overview. For starters, we decided to focus on the analysis of samples 2 hours apart. It would be interesting to investigate if the same behaviours reported in this thesis are shared across different configurations of the dataset. Lowering the number of samples per day, while keeping an approximate number of samples in the dataset, would potentially allow the model to experience more varied scenarios, which would maybe exploit a better trade-off between intra-daily information and performance. Further analysis should be dedicated to addressing the asymmetry between surface and upper variable generalisation. Therefore, focusing on creating a deeper understanding of how the information propagates across atmospheric layers in the iso-latitude setting, which would allow us to understand why this type of architecture breaks the symmetry in a systematic way. More fundamentally, a new windowing scheme that preserves both equivariance and locality would directly address the trade-off that the iso-latitude variant exposed. The current scheme aligns the windows with iso-latitude rings, gaining rotational consistency at the cost of spatial neighbourhood; an alternative design might combine the two properties by partitioning the sphere into windows that respect both the local geometry around each point and the rotational symmetry of the grid. Such a scheme would likely require revisiting the way attention is partitioned over the HEALPix mesh, possibly by combining local patches with longitudinally interleaved subsets, rather than choosing one or the other.

5.2 Final Remarks

This thesis represents an initial exploration of how rotational symmetry can be incorporated into transformer-based weather prediction on the HEALPix grid. While the proposed modifications did not yield forecasting improvements, they revealed a concrete tension between equivariance and spatial locality that any future design will need to reconcile. We hope that the analysis, metrics, and architectural insights presented here provide a useful foundation for further work in this direction.

Bibliography

Hampus Linander, Tage Tykesson, Pietro Rosso, Christoffer Petersson, Daniel Persson, and Jan E. Gerken. Pear: Equal area weather forecasting on the sphere, 2026. URL https://arxiv.org/abs/2505.17720.

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. *arXiv preprint arXiv:2311.06253*, 2023. doi: 10.48550/arXiv.2311.06253. URL https://arxiv.org/abs/2311.06253.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

Yann LeCun and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

Peter Bauer, Alan J. Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015. URL https://api.semanticscholar.org/CorpusID:4451289.

Jimeng Shi, Azam Shirali, Bowen Jin, Sizhe Zhou, Wei Hu, Rahuul Rangaraj, Shaowen Wang, Jiawei Han, Zhaonan Wang, Upmanu Lall, Yanzhao Wu, Leonardo Bobadilla, and Giri Narasimhan. Deep learning and foundation models for weather prediction: A survey, 2025. URL https://arxiv.org/abs/2501.06907.

K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622 (2):759–771, April 2005. ISSN 1538-4357. doi: 10.1086/427976. URL http://dx.doi.org/10.1086/427976.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. Aifs – ecmwf's data-driven forecasting system, 2024. URL https://arxiv.org/abs/2406.01465.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Alexander Pritzel, Suman Ravuri, Timo Ewalds, Ferran Alet, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Jacklynn Stott, Oriol Vinyals, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 12 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather, 2024. URL https://arxiv.org/abs/2312.15796.

Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9), 2020. ISSN 1942-2466. doi: 10.1029/2020ms002109. URL http://dx.doi.org/10.1029/2020MS002109.

Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. Swinvrnn: A data-driven ensemble forecasting model via learned distribution perturbation. *Journal of Advances in Modeling Earth Systems*, 15(2), February 2023. ISSN 1942-2466. doi: 10.1029/2022ms003211. URL http://dx.doi.org/10.1029/2022MS003211.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, 2022. URL https://arxiv.org/abs/2211.02556.

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

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022. URL https://arxiv.org/abs/2202.11214.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead, 2023. URL https://arxiv.org/abs/2304.02948.

James R Driscoll and Dennis M Healy. Computing fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, 15(2):202–250, 1994.

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, 2021. URL https://arxiv.org/abs/2103.14030.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Heal-swin: A vision transformer on the sphere, 2024. URL https://arxiv.org/abs/2307.07313.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks, 2016. URL https://arxiv.org/abs/1602.07576.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks, 2019. URL https://arxiv.org/abs/1810.00825.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics, 2022. URL https://arxiv.org/abs/2201.11969.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

A

Appendix

A.1 RMSE Complete Results

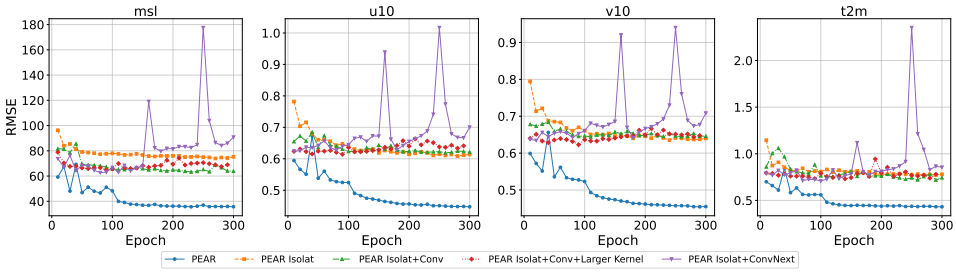


Figure A.1: RMSE vs. epochs for surface variables, using a 2 h prediction horizon.

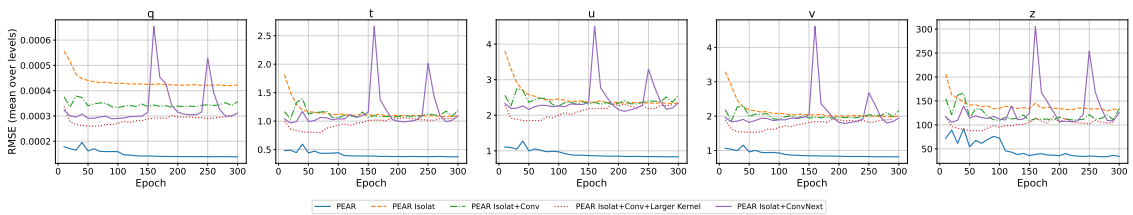


Figure A.2: Mean RMSE vs. epochs for the upper-level variables, using a 2 h prediction horizon.

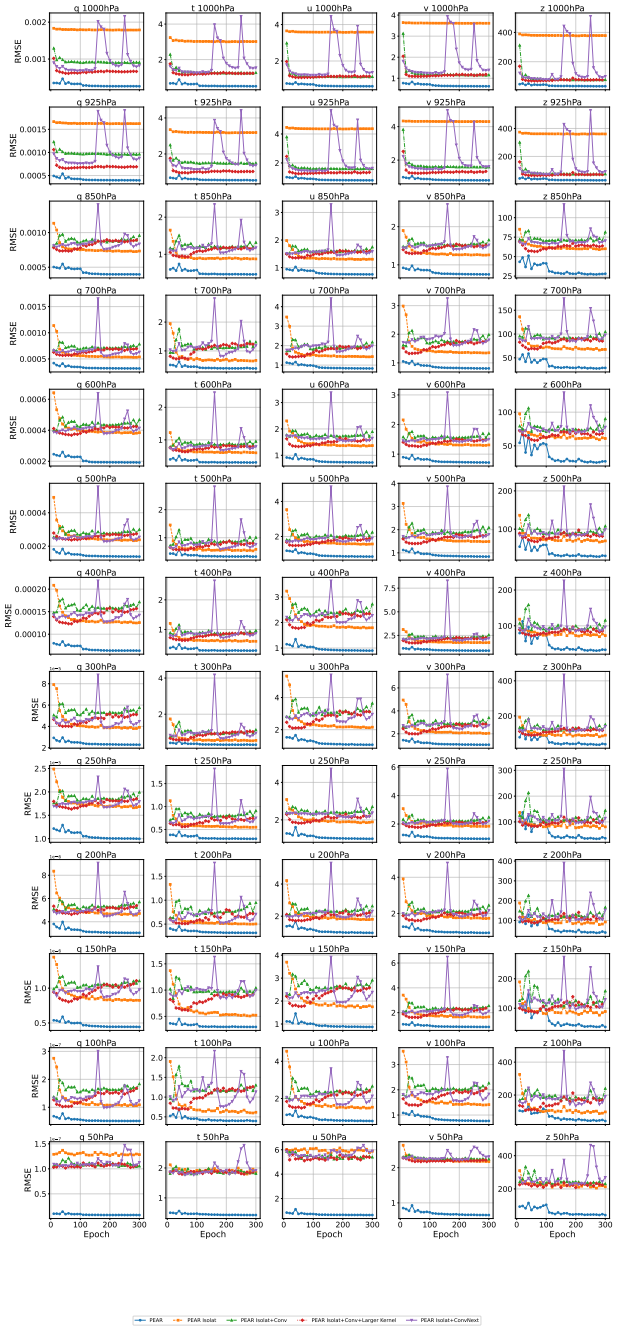


Figure A.3: RMSE vs. epochs for upper-level variables at each pressure level, using a 2 h prediction horizon.

A.2 ACC Complete Results

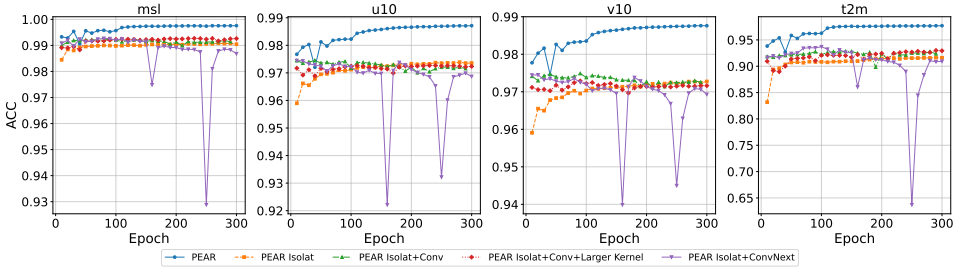


Figure A.4: ACC vs. epochs for surface variables, using a 2 h prediction horizon.

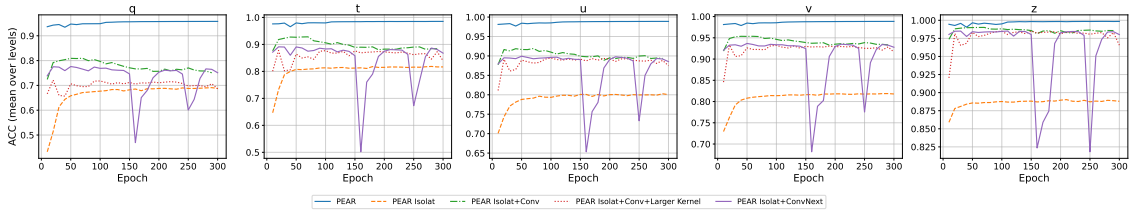


Figure A.5: Mean ACC vs. epochs for the upper-level variables, using a 2 h prediction horizon.

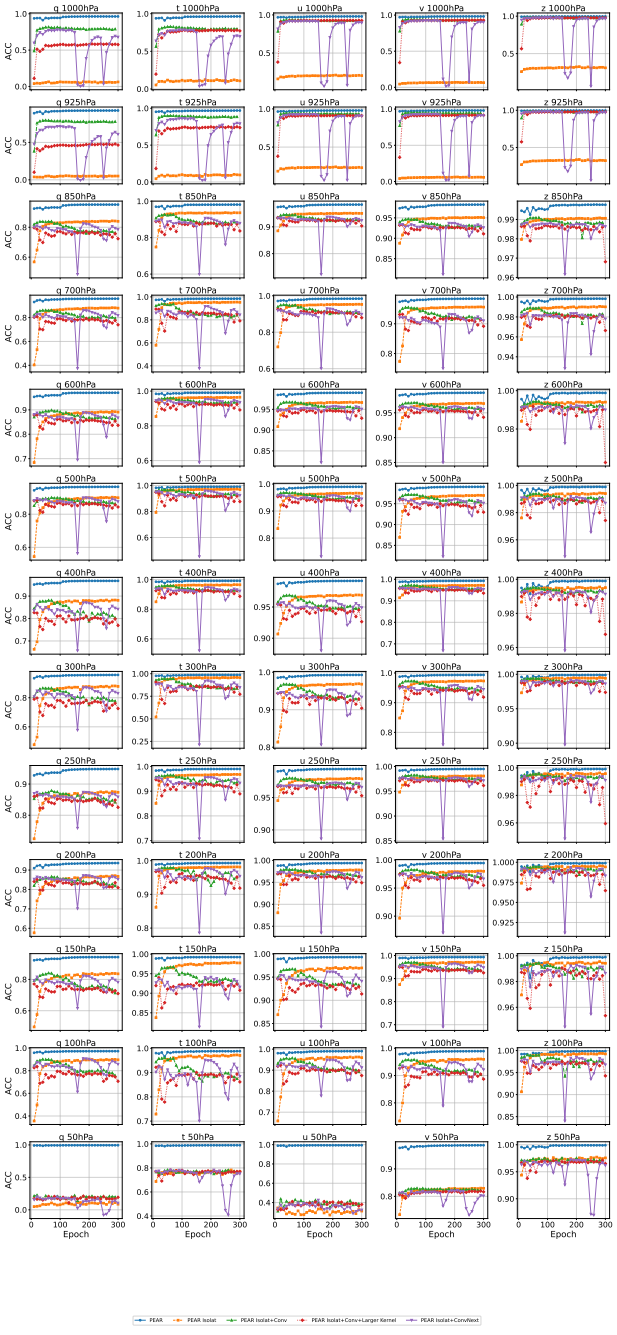


Figure A.6: ACC vs. epochs for upper-level variables at each pressure level, using a 2 h prediction horizon.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY