Master Thesis

Predicting High-Risk Atmospheric Patterns Linked to Northern Hemispheric Crop Failures with Spatiotemporal Transformers

by

Mark Musa Mitrani Sabah
(mmi393)

First Supervisor: Vera Melinda Galfi
Daily Supervisor: Miltiadis Kofinas
Second Reader: Charlotte Gerritsen

December 26, 2025

Submitted in partial fulfillment of the requirements for
the VU degree of Master of Science in Artificial Intelligence

Predicting High-Risk Atmospheric Patterns Linked to Northern Hemispheric Crop Failures with Spatiotemporal Transformers

Mark Musa Mitrani Sabah[1], Vera Melinda Galfi[2], and Miltiadis Kofinas[3]

[1] Vrije Universiteit Amsterdam, Faculty BETA, The Netherlands m.m.mitrani.sabah@student.vu.nl
[2] Vrije Universiteit Amsterdam, IVM v.m.galfi@vu.nl
[3] Vrije Universiteit Amsterdam, IVM m.kofinas@vu.nl

Abstract. Quasi-stationary Rossby Waves (QSWs) are linked to teleconnected extreme weather events, some of which can trigger concurrent crop failures across multiple breadbasket regions in the Northern Hemisphere. Such events threaten agricultural productivity, social welfare, and economic stability on a global scale. Global warming is expected to amplify the impacts of concurrent extreme events, increasing their societal and economic risks. Consequently, developing robust methods to identify and predict these atmospheric patterns is crucial. The key scientific question that this thesis addresses is whether dominant large-scale circulation patterns associated with concurrent heat extremes can be robustly identified and forecast at subseasonal lead times. Specifically, we investigate how Northern Hemispheric circulation anomalies relate to temperature extremes over major agricultural regions, and whether the temporal prevalence of such patterns can be predicted beyond simple baseline approaches. Using archetypal analysis, we extract summer (June-July-August) circulation anomalies from the 100 years of the present-day slice of KNMI-LENTIS, a large ensemble climate model dataset used to study climate variability in conditions resembling 2000-2009. Having extracted our hemispheric anomalies in the form of 'archetypes'– patterns of extreme weather behavior– we obtain temperature composites associated with each of the archetypes and calculate severity scores based on temperature anomalies over breadbasket regions in North America and Europe. Lastly, we develop a predictive framework building on Earthformer (a space-time transformer for Earth System Forecasting) that aims to forecast the prevalence of a pre-selected archetype at a user-specified lead time. Findings demonstrate that our framework can predict a 7-day rolling-averaged target series with 60% accuracy at a lead time of 5 days and outperform climatology and persistence baselines at lead times of up to 10 days. The code used in this work is openly available at https://github.com/markmitrani/concurrent-heatwave-prediction.

Keywords: Heatwaves · Quasi-stationary Rossby Waves · Spatiotemporal transformers · Earthformer · Deep Learning

1 Introduction

A significant portion of the world's crop supply is grown in North America, Europe, and Asia. These regions have been titled "breadbasket" regions, referring to their critical role in global food production. Simultaneous extreme weather conditions over these regions threaten disruptions that may present a dangerous vulnerability to the global food supply [14]. Mitigation strategies therefore must aim to prevent widespread food shortages by methods such as early warnings to allow farms to prepare appropriately.

1.1 Large-scale circulation and Rossby waves

To understand why distant regions can experience simultaneous extremes, also known as teleconnected risks, it is necessary to describe the large-scale circulation of the mid-latitude atmosphere. This circulation arises from the combined effects of differential heating between the equator and the poles and the rotation of the Earth. Together, these factors establish strong meridional (north-south) temperature gradients and induce large-scale atmospheric motions that are constrained by planetary rotation. In the Northern Hemisphere, excess heat in the tropics drives rising motion while cooler air subsides at higher latitudes, which establishes a meridional flow. The trajectory of this movement is altered by the Coriolis effect: due to the Earth's

rotation, meridional motions in the Northern Hemisphere are deflected to the right. The combined influence of differential heating and Coriolis deflection gives rise to large-scale planetary waves known as Rossby waves [20].

The jet stream emerges as a band of strong westerly winds in the upper troposphere that is closely associated with the same dynamical balance. The jet stream provides the flow along which these waves are expressed. Rossby waves therefore manifest as large-scale undulations of the jet, with alternating poleward and equatorward excursions respectively corresponding to ridges and troughs in the flow.

Depending on the amplitude of these waves, the upper-level circulation may appear predominantly zonal (west-east) or strongly meridional. Low-amplitude Rossby waves are associated with relatively straight, west–east oriented flow, whereas high-amplitude waves lead to pronounced north–south meanders. When such amplified waves propagate slowly or become quasi-stationary, they can give rise to blocking conditions, characterized by persistent deviations from the climatological flow that strongly influence regional weather [13].

1.2 Surface impacts of Rossby waves

Having described how wave-like disturbances develop within the jet, we now turn to their surface expression, where ridges and troughs modulate pressure systems and generate the temperature and precipitation anomalies that shape regional climate extremes. Rossby waves are defined as a succession of ridge–trough pairings over the atmosphere. Ridges are associated with high pressure systems, also known as anti-cyclones. These are characterized by descending air, reducing cloud formation, preventing precipitation, and causing higher surface temperatures. Troughs are related to low pressure systems, also called cyclones. With a mass of warm air ascending from the Earth's surface, air cools down, water vapor condenses into clouds, leading to higher chance of rain and cooler surface temperatures.

Factors such as orography and land-sea temperature gradients occasionally cause Rossby waves to slow down and undergo a period of sustained phase, resulting in Quasi-stationary Rossby Waves (QSWs) [14]. A strengthened and zonally confined jet stream can further enhance waveguidability, allowing Rossby-wave energy to remain trapped within the mid-latitudes and thereby favoring high-amplitude, quasi-stationary configurations [7]. Certain large-scale waves, such as wavenumbers 3, 5, and 7, are more prone to this behavior as their wavelengths interact more strongly with orography, land-sea contrasts, and the mean jet structure. Within a period of stationarity, amplified QSWs lead to persistent temperature anomalies under regions which are encompassed by the ridges and troughs, characterized by higher and lower temperatures respectively [23]. Persistence in QSWs does not necessarily imply a single stationary anomaly, but may also result from a sequence of transient waves recurring with similar phase [7]. Persistent ridges lead to enduring heating, whereas persistent troughs cause repeated clouds and rain. These conditions respectively lead to heatwaves and cold/dry spells, both of which can be detrimental to crop growth. QSWs are hence the drivers of many extreme events that lead to concurrent heatwaves that are particularly severe in duration and/or intensity over multiple breadbasket regions, which in turn could cause simultaneous crop failures [14]. A conceptual illustration of how a QSW, such as wavenumber 7, can simultaneously impact distant crop regions through teleconnections is shown in Figure 1.

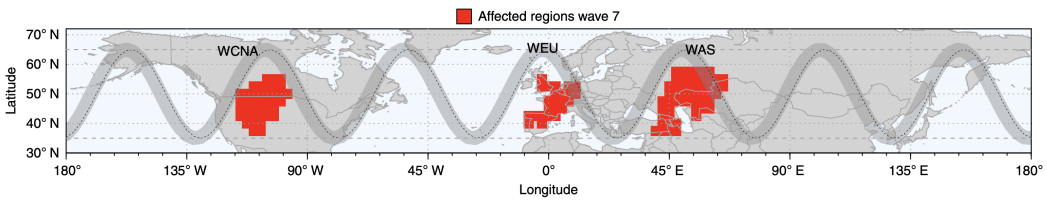


Fig. 1: Conceptual illustration of regions affected by anomalous heat during a preferred phase of a quasi-stationary Rossby wave (wavenumber 7). The schematic highlights how a single large-scale circulation pattern can induce simultaneous heat extremes across multiple Northern Hemisphere breadbasket regions (Western Central North America, Western Europe, Western Asia). Adapted from [14].

1.3 Climate change and societal relevance

We also consider how anthropogenic climate change further exacerbates QSW-like behavior. With the arctic warming faster than the global average, the pole-equator temperature gradient is weakening, which in turn slows the jet stream and leads to more meridional movement, increasing blocking and persistence. Consequently, this increases the frequency, severity, and duration of extreme conditions. Notably, the number of yearly persistent and extreme heatwaves over the Northern Hemisphere has been on the increase for the past two decades [26].

QSWs live in the upper-level circulation and behave as persistent regimes which are physically predictable configurations. There are other climate related drivers of multiple breadbasket failures, such as El-Niño Southern Oscillation, but the effects of QSWs are less well studied, reflecting a gap in current scientific literature. While QSWs present a serious threat, their successful modeling and prediction may provide a chance to anticipate concurrent heatwaves, and consequently, potential food shortages [14]. Together, these conditions emphasize the need for better forecasting capabilities to reduce the agronomic, socioeconomic, and health related impact of heatwaves. It is therefore our aim in this work to forecast the prevalence of Rossby-wave-related circulation regimes.

1.4 Identifying impactful Rossby-wave-related circulation regimes

A prerequisite to making predictions is the identification and modeling of large-scale Rossby-wave-related circulation regimes, and Archetypal Analysis (AA) is well suited to this task for several reasons. Originally proposed by Cutler and Breiman, AA "represents each individual in a data set as a mixture of individuals of pure type or archetypes" [4]. Conceptually, the technique identifies the most extreme and distinctive patterns in a dataset such that each observed state can be expressed as a combination of these patterns. In recent years, the technique's successful application to geophysical datasets has further established its utility [2] [3]. AA is particularly appropriate for identifying extreme Rossby-wave-related circulation patterns because it determines the convex hull of the dataset, thereby capturing its most amplified and atypical states. This aligns with the nature of amplified Rossby waves, which often manifest as departures from the dominant zonal flow and can, in some cases, display quasi-stationary characteristics.

Principal Component Analysis is unsuitable in this context because it identifies directions of maximum variance rather than extreme states, and its resulting modes are linear combinations that can be difficult to interpret physically [3]. Clustering methods are likewise inappropriate, as they assign each observation to a single discrete group, whereas circulation extremes do not always fall neatly into mutually exclusive categories. By expressing each observation as a mixture of several extreme states – its archetypes – AA accommodates this nuance. Following the identification of circulation archetypes, composite analysis is employed to characterize and interpret the associated large-scale circulation and surface impacts.

Following the identification of circulation archetypes, composite analysis can be used to characterize their associated surface impacts. By averaging temperature fields conditioned on the occurrence of a given archetype, composites reveal the typical temperature anomalies linked to each circulation regime [3]. Composite diagnostics enable the identification of the archetypes that are most strongly associated with concurrent heat extremes and are therefore most relevant for prediction.

1.5 Predictive framework

Having addressed the problem of identifying impactful Rossby-wave-related circulation regimes, we now turn to prediction techniques. Standard weather and climate forecasting systems often rely on Numerical Weather Prediction ensembles that simulate physical processes by solving mathematical equations which make them accurate but computationally expensive. Over the past decade, deep-learning based approaches have become increasingly more appealing: while training such models can be costly, running subsequent predictions is comparatively efficient and fast.

A limitation of simple feed-forward neural networks, however, is their inability to capture the spatial and temporal context inherent in atmospheric dynamics [22]. This limitation has motivated the development of architectures capable of modeling long-range dependencies more flexibly than traditional networks. Among such developments, attention-based architectures have

emerged as a powerful approach for learning contextual relationships within complex datasets. The original transformer mechanism achieved substantial advances in Natural Language Processing by converting input text to tokens, which are then vectorized, and the attention mechanism models the contextual dependencies between these tokens [25]. Since then, variants of the transformer have been applied to various domains such as computer vision [21] and biological sequence analysis [10].

Spatiotemporal transformers build on these advances by extending the attention mechanism to represent both spatial and temporal dependencies, thereby addressing the specific limitations encountered when modeling atmospheric dynamics and overcoming the restrictions of more naive neural-network architectures. One such architecture that has demonstrated notable Earth-system forecasting performance is Earthformer, which implements a space-time attention block named Cuboid Attention [9]. This motivates the use of spatiotemporal transformers as a core component of the predictive framework that we develop.

In this work, we investigate whether dominant large-scale circulation patterns associated with concurrent heat extremes can be robustly identified and predicted at subseasonal lead times. Specifically, we ask:

> *What rare Rossby waves could trigger co-occurring extreme weather events over several breadbasket regions with the potential to lead to multiple breadbasket failures, and how can such circulation regimes be predicted?*

To that end, we use archetypal analysis to extract extreme configurations, composite analysis to discern the regimes most strongly associated with concurrent heat extremes, and sequence prediction using a spatiotemporal transformer to predict the temporal prevalence of selected regimes. We make our implementation publicly available to support reproducibility [17].

2 Data

2.1 Dataset

The primary dataset for this study is the KNMI-LENTIS (Large Ensemble Time Slice) dataset [18] produced with EC-Earth v3, a global climate model developed by ECMWF [6]. Ensembles are collections of climate models run with slight changes in initial conditions, where the many runs of an ensemble are particularly useful in studying climate variability. In the case of LENTIS, the ensemble consists of two distinct 10-year time slices: (1) a present-day slice representative of the 2000-2009 climate, and (2) a +2 K warmer future time slice [18]. Each time slice is run with fixed external forcing representative of either the present-day or future climate state, which ensures that variability across ensemble members arises purely from internal atmospheric dynamics. For each time slice, 160 runs of the model together make up 1600 years worth of data.

We base our studies on the present-day slice, which is especially useful as it allows the investigation of a greater number of extreme conditions. Such a large sample size from a stationary climate state provides weather configurations that do not appear in historical reanalysis, but could feasibly occur in our climate. Applying our analysis to this dataset simultaneously enables us to uncover more circulation anomalies than would be possible with historical data alone and yields a larger pool of samples for the prediction stage.

Due to computational constraints, we utilize only the first 100 years of the dataset. Handling the full 1600-year dataset poses substantial I/O and preprocessing constraints: operations such as loading, regridding, and performing dimensionality reduction on the entire dataset simultaneously would require substantially more memory, bandwidth, and processing time than is practical for this study. On an NVIDIA A30 GPU with 24 GB of VRAM, training on the 100-year subset requires approximately 12 hours per experiment, which allows systematic exploration of model configurations and ablation studies. Using all 1600 years would increase training time by more than an order of magnitude and render iterative experimentation impractical. Moreover, 100 years already provide a large sample of about 9200 daily June-July-August circulation fields, which is sufficient for extracting diverse circulation regimes and training the prediction model. Because each LENTIS time slice is generated under fixed external forcing, any 100-year subset is expected to sample a broad range of internal variability, even if it may not capture the full diversity present in the entire 1600-year ensemble.

2.2 Variables

The variables used from the dataset are stream function at 250 hPa, 2m surface temperature (TAS), and outgoing longwave radiation (OLR). All of these variables are provided on a daily temporal and 0.7 degrees spatial resolution, which is later coarsened to 2.1 degrees.

Stream function The 250 hPa stream function (ψ) indicates the non-divergent, rotational component of the upper-tropospheric flow at 250hPa geopotential height, or approximately 9 kilometers above the surface. The stream function is a scalar representation of large-scale circulation patterns, where contours of constant ψ correspond to streamlines of the horizontal wind field. Furthermore, since large-scale wind motions at upper levels are dominated by rotation rather than divergence – an assumption that is well justified at upper tropospheric levels [8] – focusing on the non-divergent, rotational component isolates the balanced, quasi-horizontal flow that governs Rossby-wave dynamics and long-lived circulation patterns. Consequently, this variable is particularly relevant for identifying and characterizing quasi-stationary waves, as it effectively captures the large-scale, rotational flow patterns associated with their development and persistence. The following definitions formalize this representation and clarify how the stream function is constructed from the underlying wind field.

The precise mathematical definition of the horizontal wind uses pressure coordinates, thus the wind components (u, v) are expressed on a surface of constant pressure rather than constant altitude [8]. These components are then defined as:

$$u = -\frac{\partial \psi}{\partial\, y}, v = \frac{\partial \psi}{\partial x}$$

This formulation inherently ensures that the flow is non-divergent:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0,$$

thus ψ represents a mass-conserving rotational flow. The relationship between ψ and the relative vorticity (ζ) further links the stream function to the atmosphere's rotational dynamics:

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \nabla^2 \psi$$

From a practical standpoint, this link allows circulation anomalies to be analyzed through a single scalar field that preserves the essential rotational dynamics. In the Northern Hemisphere, positive vorticity indicates counterclockwise rotation, meaning that regions enclosed by high positive ψ contour lines correspond to cyclic (counterclockwise) flow. Conversely, negative vorticity indicates clockwise rotation, where regions enclosed by high negative ψ contour lines correspond to anticyclonic (clockwise) flow. These rotational signatures provide the physical basis for interpreting the circulation archetypes extracted from the stream function fields.

The stream function for LENTIS was computed by inverting the vorticity-stream function relationship using the Climate Data Operators [24], as explained in [12]. Within the context of this study, the stream function serves as the variable for computing archetypes associated with circulation anomalies and is subsequently employed as an input feature during prediction.

Surface air temperature TAS represents the near-surface atmospheric temperature measured at 2 meters above ground level and is a crucial indicator of surface climate variability. Anomalies in TAS, defined as deviations from the regional and temporal mean, can have significant implications for surface energy exchange and crop growth [14]. In this study, TAS is not used in the archetype computation but rather serves as a diagnostic variable for composite analyses conditioned on the identified archetypes, enabling the discernment of those archetypes associated with the most severe impacts to crop regions.

Outgoing longwave radiation OLR is a measure of the amount of infrared energy emitted back into space from Earth's atmosphere. In the context of prediction, it is useful as a supplementary variable to the stream function.

Lower OLR values indicate strong deep convection – vigorous vertical motion in the tropical atmosphere that is associated with cloud formation and large energy release [27]. This process, often described as diabatic heating, injects heat into the upper troposphere and can influence large-scale circulation patterns.

Observed relationships show that changes in OLR are tightly linked to these heating events on daily timescales [27]. Such tropical heating anomalies can precede the development of high-amplitude Rossby waves – including QSW events – by roughly 5–15 days [7].

For this reason, we include OLR as an additional predictor and lag it by 14 days, allowing the model to incorporate potentially early-stage tropical signals that may precede later circulation changes in the mid-latitudes.

3 Methodology

3.1 Research Problem and Objectives

This study aims to address the research question introduced in Section 1.5, namely how rare large-scale circulation patterns associated with concurrent heat extremes can be identified and predicted at subseasonal lead times. Addressing this question requires (i) a physically interpretable representation of large-scale atmospheric variability and (ii) a predictive framework capable of learning spatiotemporal dependencies in the circulation.

3.2 Proposed Approach

To meet our research objectives, we propose a methodology that combines archetypal analysis with deep learning-based sequence modeling.

Archetypal Analysis We apply AA to 250 hPa stream function data to extract circulation pattern archetypes. Each day is assigned a soft membership to each of the p archetypes, which enables the identification of dominant regimes for each time point. Together with composite analysis of the archetypes conditioned on near-surface air temperature, this approach reveals temperature patterns characteristic of each archetype.

For $\mathbf{X} \in \mathbb{R}^{d \times n}$, where $\mathbf{X}$ is a set of n column vectors with d features each, archetypal analysis finds vectors $z_1, \cdots, z_p \in \mathbb{R}^d$, that together represent the 'archetypal' patterns in the data.

Given a predefined number of archetypes, the archetypes are determined by minimizing the reconstruction error between the original data and its approximation using a convex combination of the archetypal patterns. Archetypal analysis finds a set of p archetypes that lie on the convex hull of the data and can best represent all data points through convex combinations of these archetypes. Concretely, this is achieved by solving the optimization problem:

$$\min_{\mathbf{C},\mathbf{S}} ||\mathbf{X} - \mathbf{XCS}||_F^2$$

with the constraints that the columns of $\mathbf{C} \in \mathbb{R}^{n \times p}$ are element-wise non-negative and sum to 1. $\mathbf{X} \in \mathbb{R}^{d \times n}$ is the input data matrix, and the matrix $\mathbf{Z} = \mathbf{XC} \in \mathbb{R}^{d \times p}$ holds the learned archetypes. For each of the n data points, their soft membership weights are held in $\mathbf{S} \in \mathbb{R}^{p \times n}$ with the constraints that the values are non-negative and each column sums to 1.

Directly applying AA to geospatial data is infeasible due to the issues of high dimensionality and large computational costs. Therefore a standard practice is to reduce dimensionality with a technique such as Truncated Singular-Value Decomposition (SVD) beforehand [2]. Given a matrix $\mathbf{X}$, SVD decomposes the matrix into $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$, respectively containing the left singular vectors, singular values, and right singular vectors. Truncation then selects the first k columns $(u_1, \ldots, u_k)$ of $\mathbf{U}$, top k singular values $(\sigma_1, \cdots, \sigma_k)$ from $\mathbf{\Sigma}$, and the first k columns $(v_1, \ldots, v_k)$ of $\mathbf{V}$, resulting in the matrices $\mathbf{U}_k$, $\mathbf{\Sigma}_k$, and $\mathbf{V}_k$.

Following this step, the archetypes of $\mathbf{\Sigma}_k \mathbf{V}_k$ are computed, and then projected back into the original space by multiplying the archetypes with the matrix $\mathbf{U}_k$.

In practice, we perform archetype analysis using the **py_pcha** package [1], which provides an implementation of the efficient solver proposed by Mørup and Hansen [19].

Composite Analysis Given a set of archetypes, composite analysis enables us to identify the temperature anomaly patterns that are characteristic of each circulation regime. Rather than forming weighted composites based on continuous archetype participation, we adopt a hard assignment approach in which each time step is associated with its dominant archetype. This argmax-based compositing emphasizes distinct regime expressions and facilitates physical interpretation.

Recalling the notation introduced for archetypal analysis, we denote by $\mathbf{S} \in \mathbb{R}^{p\times n}$ the archetype participation matrix, where p is the number of archetypes and n the number of time steps. For each timestep $t = 1, \cdots, n$, the dominant archetype is defined as

$$a(t) = \arg\max_{p} \mathbf{S}_{pt}$$

The composite associated with archetype p is then computed as the conditional mean of the temperature anomaly field over all time steps where archetype p is dominant:

$$\bar{\mathbf{X}}_p = \frac{1}{|\mathcal{T}_p|} \sum_{t\in\mathcal{T}_p} \mathbf{X}_t, \quad \mathcal{T}_p = \{\, t \mid a(t) = p \,\}$$

where $\mathbf{X}_t$ denotes the temperature anomaly field at time t and $\mathcal{T}_p$ is the set of time indices where archetype p is dominant.

Using this approach, each time step contributes to exactly one composite, allowing the resulting patterns to be interpreted as representative realizations of distinct circulation regimes. These composites are subsequently used to identify archetypes that are most strongly associated with extreme temperature anomalies over the regions of interest, and are therefore most relevant for prediction in the context of concurrent heatwaves.

Informed by the regional definitions introduced in [14], we define two mid-latitude regions of interest (ROIs) covering Europe (latitudes 40–55°, longitudes -10 to 45°) and North America (latitudes 40–55°, longitudes -110 to -75°). We then compute regionally aggregated temperature-anomaly z-scores for each archetype composite. Mean absolute ($|z|$), positive (z^+), and negative (z^-) deviations are used to quantify the magnitude and sign of associated surface temperature anomalies.

Predictive Modeling We train a regressor that wraps the Earthformer model on sequences of upper-level atmospheric fields to predict the emergence and evolution of Rossby-wave-related archetypes. The Earthformer model is particularly preferred for this task as its transformer-based architecture is designed to capture spatio-temporal dynamics with its cuboid attention mechanism [9].

Archetype Prevalence Prediction Task. The predictive model is trained to estimate the participation coefficient for a pre-selected archetype at a given lead time, representing how strongly the atmospheric circulation pattern at that time resembles the chosen archetypal pattern. Formally, let $\mathbf{x}_t \in \mathbb{R}^{H\times W\times C}$ denote the input atmospheric field at time t, where C is the number of channels (variables), and H, W are the spatial dimensions. The model receives an input sequence $\mathbf{X}_t = \{\mathbf{x}_{t-s+1}, \ldots, \mathbf{x}_t\}$ of length s days and aims to predict the archetype coefficient $y_{t+l} \in [0, 1]$ at a lead time of l days.

Base Earthformer Model. Earthformer is a transformer-based architecture designed for spatiotemporal forecasting tasks in Earth system science [9]. It takes as input a spatiotemporal tensor $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$ where T is the sequence length, $H \times W$ the spatial grid, and C the number of channels.

Rather than applying global self-attention directly to the full 4D tensor, which would require $O((THW)^2)$ operations, Earthformer first decomposes the input tensor into a set of structured spatiotemporal cuboids $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$ where each cuboid is a contiguous block of size (b_T, b_H, b_W) [9]. The model then applies Cuboid Self-Attention within each cuboid in parallel by flattening the spatiotemporal volume into a sequence of $L = b_T \times b_W \times b_H$ tokens:

$$\mathbf{z}^{(n)} = \mathrm{reshape}\big(\mathbf{x}^{(n)}\big) \in \mathbb{R}^{L\times C}$$

This drastically reduces the computationally prohibitive quadratic attention cost and enables the model to scale to large spatial domains typical of climate datasets. Queries, keys, and values are then obtained by linear projections

$$\mathbf{Q}^{(n)} = \mathbf{z}^{(n)}\mathbf{W}_Q, \quad \mathbf{K}^{(n)} = \mathbf{z}^{(n)}\mathbf{W}_K, \quad \mathbf{V}^{(n)} = \mathbf{z}^{(n)}\mathbf{W}_V$$

with parameters $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C\times d}$ shared across all cuboids, and d is the dimensionality of the queries/keys for a single head. Single-head self-attention for cuboid n is then

$$\mathbf{x}_{\text{out}}^{(n)} = \text{Attention}\big(\mathbf{z}^{(n)}\big) = \text{softmax}\left(\frac{\mathbf{Q}^{(n)}(\mathbf{K}^{(n)})^\top}{\sqrt{d}}\right)\mathbf{V}^{(n)}$$

and the multi-head version is obtained in the standard way by concatenating the outputs of several such heads and applying a final linear projection. Following self-attention, the resulting sequences are restructured back into 3D tensors with the original block shapes and then merged back into the original input shape.

The cuboid outputs are reassembled into a global feature tensor whose spatial structure matches the block layout of the input, while both its temporal dimension and channel dimension may differ from the original, depending on output specifications.

To enable communication between cuboids, Earthformer augments the local attention mechanism with a small set of learnable global vectors that each cuboid attends to [9]. This allows the model to capture large-scale spatial dependencies while maintaining the efficiency of cuboid-based attention.

EarthformerPredictor. The model architecture builds upon the Earthformer framework, retaining its spatiotemporal Cuboid Transformer backbone and appending a lightweight predictor head for scalar prediction. To align our input with the spatial configuration which is typically applied for Earthformer, we reshape our input tensor's height (latitudes) and width (longitudes) from their native [29, 170] grid to a standardized [128, 128] grid by means of interpolation and padding while maintaining the original aspect ratio.

Let the feature tensor produced by the final layer of Earthformer be $\mathbf{X}_{\text{out}} \in \mathbb{R}^{T_{\text{out}}\times H\times W\times C_{\text{out}}}$, where $T_{\text{out}} = 1$ is the output sequence length and $C_{\text{out}} = 64$ is the number of output channels. We apply a 3D adaptive max-pooling operator that reduces $\mathbf{X}_{\text{out}}$ to a fixed spatial and temporal resolution, which is flattened into a vector:

$$\mathbf{P} = \text{Pool}_{T,H,W}(\mathbf{X}_{\text{out}}) \in \mathbb{R}^{1\times 8\times 8\times 64}, \quad \mathbf{f} = \text{Flatten}(\mathbf{P}) \in \mathbb{R}^{4096}$$

The flattened vector is then passed through a two-layer MLP:

$$\mathbf{h} = \text{ReLU}(\mathbf{W_1 f} + \mathbf{b_1}), \quad \mathbf{h}' = \text{Dropout}(\mathbf{h})$$

$$\hat{y} = \sigma(\mathbf{W_2 h}' + \mathbf{b_2})$$

with $\mathbf{W}_1 \in \mathbb{R}^{512\times 4096}$ and $\mathbf{W}_2 \in \mathbb{R}^{1\times 512}$. The predictor maps the pooled Earthformer features to a single scalar value interpreted as the estimated participation of the target archetype.

All Earthformer backbone parameters are initialized from scratch using the default initialization scheme provided in the Earthformer implementation [9]. For the predictor head, weights of all linear layers are initialized with Xavier uniform initialization [11], while biases are set to zero to avoid introducing spurious offsets at the start of training. A dropout rate of 0.4 is applied during training time between the two linear layers to reduce overfitting.

Figure 2 provides a schematic overview of the EarthformerPredictor architecture, illustrating the sequence of operations applied to transform the input data into the final scalar prediction.

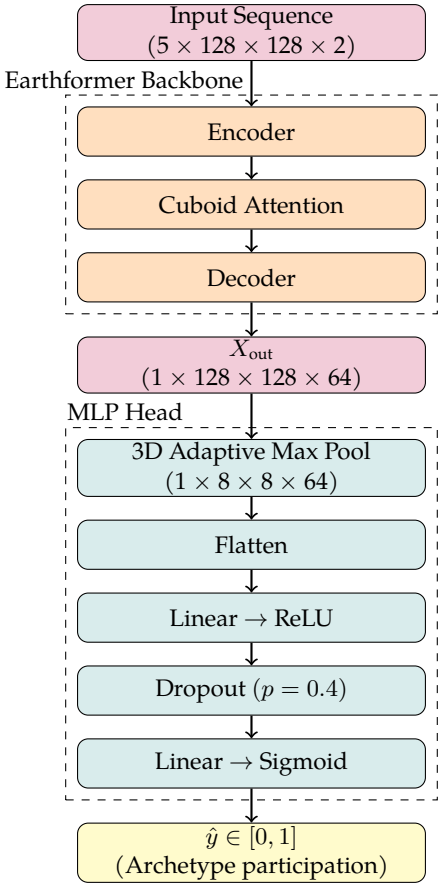


Fig. 2: Architecture of EarthformerPredictor, the Earthformer-based predictor with lightweight MLP head. For clarity, the internal hierarchical structure of the Earthformer backbone is abstracted and shown here as a single encoder–decoder module. See [9] for the full architectural details, including multiple stacked cuboid attention blocks, cuboid decomposition and reassembly, downsampling and upsampling stages, and the use of global vectors.

4 Experimental Design

Preprocessing Various preprocessing steps are followed to prepare the data in the most adequate format for each step of the pipeline. Common to each step is discarding all months except June, July, and August, and coarsening the spatial resolution from 0.7 to 2.1 degrees.

Stream function After the common preprocessing steps, the stream function is deseasonalized by subtracting the seasonal trend from the values. Additionally, this field is weighted with square-root cosine weighting, which compensates for the convergence of meridians toward the poles: grid cells at higher latitudes represent a smaller physical area than those near the equator, so the weighting ensures that each latitude band contributes proportionally to its actual surface area in subsequent analyses.

Because of the way the stream function is computed, point-wise magnitudes are arbitrary. Shifting the entire stream function field up or down by the same amount does not change the winds it produces, since the winds depend solely on the spatial gradients of the stream function, not on absolute values. We therefore follow the standard practice to subtract the spatial mean [23]. Doing so removes an uninformative offset.

For archetypal analysis, we use only the data between latitudes [30,60] to ensure that the computed archetypes reflect the mid-latitude circulation regimes most relevant to amplified Rossby-wave behavior. In contrast, predictive modeling leverages a broader range of values between latitudes [15,75].

Surface air temperature The composite analysis uses the TAS field between latitudes [15,75], deseasonalized after applying the common steps, yielding temperature anomalies.

Outgoing longwave radiation Following the common steps, the OLR field is deseasonalized and lagged by 14 days before it is used as a variable for prediction. We utilize the values between latitudes [-20,30] to capture tropical variability where deep convection and the associated heating anomalies are most pronounced.

4.1 Experimental Procedure

Base experiment The predictive experiment is carried out by training the EarthformerPredictor model on sequences of 250hPa stream function data concatenated with OLR values lagged 14 days behind the stream function. No model parameters are frozen during training. The optimization is performed using the AdamW optimizer [16], which incorporates decoupled weight decay as a regularization mechanism, together with a learning rate schedule that applies linear warm-up followed by cosine annealing [15]. The model is trained for 100 epochs with RMSE as the loss function.

The dataset consists of 9,200 samples, of which 20% are reserved for validation. Each sample contains a 5-day $(s = 5)$ input sequence of atmospheric fields and a corresponding target archetype coefficient at a 5-day lead time $(l = 5)$ starting from the last day of the input sequence. To emphasize persistence and reduce noise, the target series is smoothed with a 7-day rolling mean computed over the original target day and the six preceding days:

$$\tilde{y}_t = \frac{1}{7} \sum_{k=0}^{6} y_{t-k}$$

where $\tilde{y}_t$ is the smoothed target at time t. While this smoothed value is not a lead-time target in the strict, point-wise sense, it provides a more stable signal for learning while preserving the temporal interpretation. In what follows, references to "lead time" implicitly refer to prediction of these smoothed target values unless stated otherwise.

Table 1 summarizes the configuration of critical experimental parameters. We run our experiments on two separate archetypes to assess whether the model's performance generalizes across regimes.

Table 1: Summary of key training configuration choices.

Component	**Setting**
Target smoothing	7-day rolling mean
OLR lag	14 days
Loss function	RMSE
Optimizer	AdamW
Learning rate	5×10^{-5} (with warm-up + cosine annealing)
Warm-up schedule	Linear warm-up over first 20% of training
Learning rate decay	Cosine annealing to 0
Weight decay	1×10^{-4}
Batch size	16
Epochs	100

Target smoothing ablation Taking a 7-day rolling average of the original targets is beneficial for highlighting persistent conditions. We conduct an ablation study testing the model's performance without the smoothing step to assess the importance of smoothing for predictive performance.

Investigating the horizon of predictability We vary the lead time to test the model's capabilities with lead times both smaller and greater than 5 days, considering $l \in \{3, 5, 7, 10\}$. Doing so allows us to evaluate the model's behavior across different forecasting horizons.

4.2 Evaluation Metrics

Alongside the loss function, we quantify the model's accuracy as the fraction of predictions whose absolute error is below 0.05.

We additionally assess model performance using correlation-based metrics. Pearson's r is utilized to measure linear association, while Kendall's τ and Spearman's ρ evaluate rank-order between predictions and targets.

4.3 Ethical Considerations

As this study uses publicly available climate data and does not involve human subjects, there are no direct ethical concerns.

5 Results

5.1 Archetypal Analysis

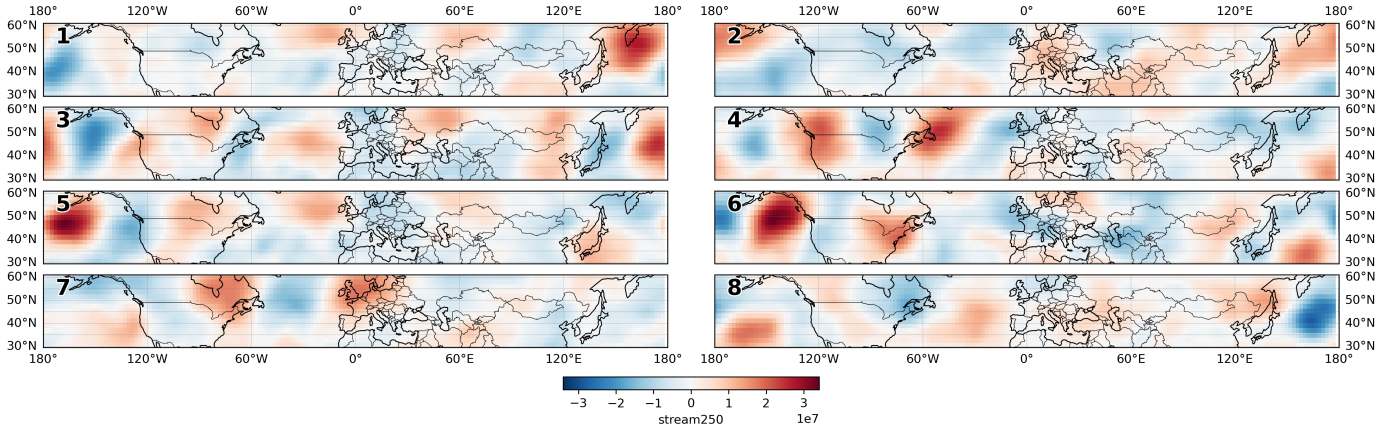


Fig. 3: Stream function Archetypes ($p = 8$) at 250 hPa; shading shows stream function anomalies.

Figure 3 shows the eight archetypal circulation patterns identified from the preprocessed 250hPa stream function field. These archetypes capture dominant Rossby-wave structures that characterize the mid-latitude upper-tropospheric flow. Red shading indicates positive ψ (anticyclonic ridges), whereas blue shading indicates negative ψ (cyclonic troughs), with larger amplitudes denoting stronger rotational flows and greater Rossby-wave activity. The figure illustrates that archetypes 4,5,6,7 exhibit the most severe traits, with high magnitudes and strong meridional flows over the US/EU regions. The patterns within these archetypes display the clearest evidence of QSW behaviour capable of sustaining persistent surface extremes:

- Archetype 4 features a pronounced trough–ridge–trough succession over the Pacific, western North America, and eastern North America, representing a positive phase of the Pacific–North American (PNA) pattern [5]. This tripole-like configuration of alternating anomalies from the Pacific to Eurasia reflects a weakened jet stream over the central Pacific and enhanced blocking in the North Pacific. The resulting flow supports warm temperatures over western North America and colder weather across eastern North America. Europe lies under the trough end of the wave train, bringing cold, unsettled conditions.
- Archetype 5 displays a negative PNA-like tripole, starting with a strong ridge over the Pacific, followed by a strong trough surrounding the U.S. West Coast, and a downstream ridge over eastern North America. This configuration leads to split weather in the U.S., with cooler temperatures in the West and a warmer regime in the East. Starting from the North Atlantic, a pronounced ridge extends northeastward, followed downstream by a deep trough centered over Europe. This configuration reflects a wave pattern that promotes cyclonic circulation over the continent. As a result, Europe experiences increased storm activity, cooler temperatures, and a higher likelihood of precipitation extremes.

- Archetype 6 shows an even more amplified positive PNA phase, characterized by a high intensity ridge over the Gulf of Alaska, a deep trough across eastern North America and a strong ridge over western North America. The meridional flow structure leads to a cold–warm split in the U.S between the east and the west. The anomalies across Europe are weak and mostly negative, indicating a broad but shallow cyclonic pattern yet no strong wave activity. Absence of a pronounced ridge–trough pairing suggests limited QSW activity and relatively zonal flow. Europe likely experiences near-normal conditions, with no strong signal for persistent extremes or blocking.
- Archetype 7 exhibits a well-organized, large-scale Rossby-wave train extending from the Pacific across North America into Europe and western Asia. Over the U.S., a broad trough dominates the eastern part of the country, favoring cooler and more unsettled conditions, while a weaker ridge lies along the western coast. Over Europe, a ridge centered over northwestern Europe and the British Isles transitions into a weak trough over southeastern Europe, consistent with downstream wave propagation. This configuration typically supports anticyclonic, dry, and warm conditions over western and northern Europe, while more unsettled and cooler weather tends to occur farther southeast under the downstream trough.

While Archetypes 4–7 capture the most severe Rossby-wave activity, Archetypes 2 and 3 also exhibit distinct but more moderate circulation features that influence the US and European regions:

- Archetype 2 shows a ridge centered over Europe and a broad trough across the North Atlantic, forming a classic Euro–Atlantic blocking structure [13]. This configuration favors warm, stable conditions over much of Europe and cooler, stormier weather over the western North Atlantic and eastern US.
- Archetype 3 displays a weaker, eastward-shifted ridge–trough couplet across the Pacific–North American sector, indicative of a transitional or low-amplitude PNA configuration. The ridge over western North America and a downstream trough over the eastern US support modest temperature contrasts across the continent. Over Europe, the pattern extends into a faint ridge–trough sequence.

The remaining archetypes display weaker signals, making them less significant overall:

- Archetype 1 exhibits a broad trough over eastern Europe and a weak ridge over East Asia.
- Archetype 8 shows weak ridging over the eastern Pacific and slight troughing over eastern North America, consistent with a low-amplitude, near-zonal flow. Over Europe, a faint ridge–trough pattern indicates minimal activity.

5.2 Composite Analysis

Figure 4 reflects the temperature composites conditioned on each of the archetypes seen in 3. These composites demonstrate the average surface temperature associated with each of the archetypes and provide a quantitative measure of their impact across American and European regions of interest, as defined by the shaded areas. z^+ denotes an aggregate score of positive deviations from the temporal mean, whereas z^- denotes negative deviations, and $|z|$ counts both positive and negative deviations.

Ranking the archetypes by their positive anomaly score, we obtain the following grouping:

- The strongest warm anomalies within the ROIs begin with Archetype 7 which stands out with a score of 1.18, followed by Archetype 5 with 0.93, and lastly Archetype 3 shows a score of 0.76.
- Intermediate yet notable positive deviations come from Archetype 2 and Archetype 6, with scores of 0.67 and 0.57 respectively.
- The weakest responses within the composite set consist of Archetype 1, 8, and 4, all scoring negligibly low respectively with 0.41, 0.37, 0.23.

Based on these results, **Archetype 7** is the most consequential archetype for Rossby-wave-related extreme events over American, European, and West Asian breadbasket regions during summer months. Therefore it is chosen as the main archetype to be predicted.

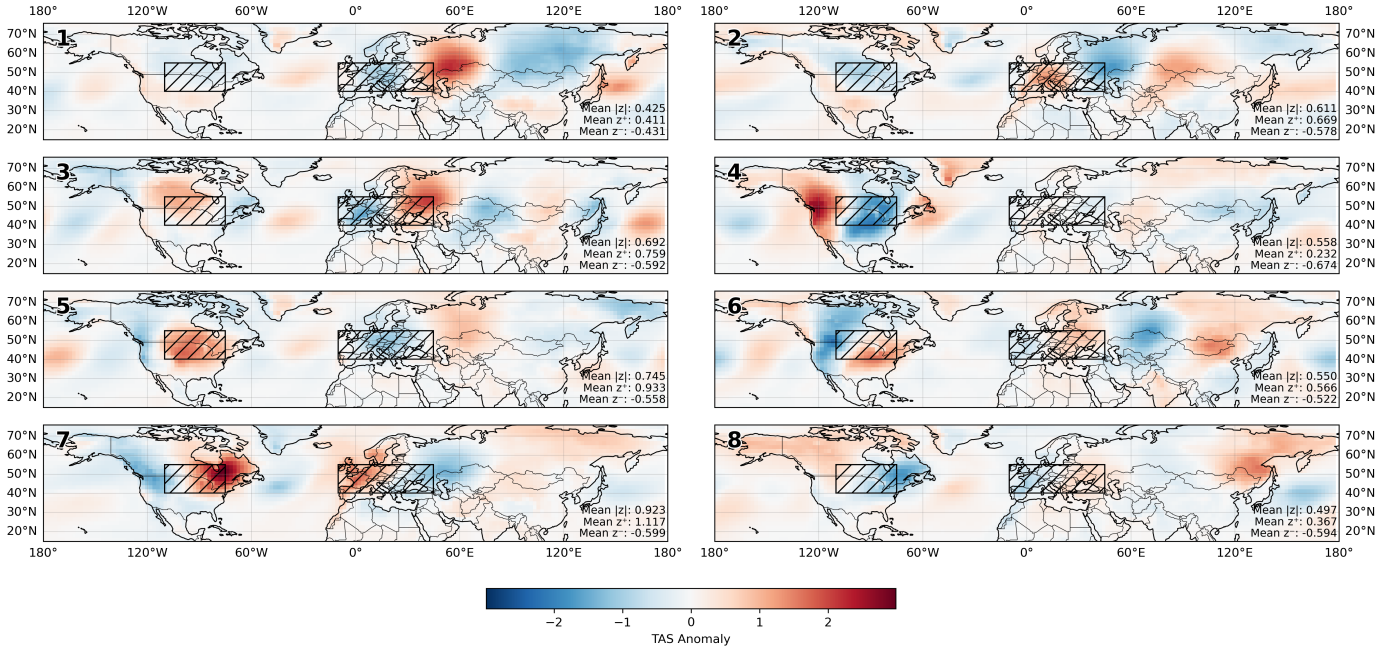


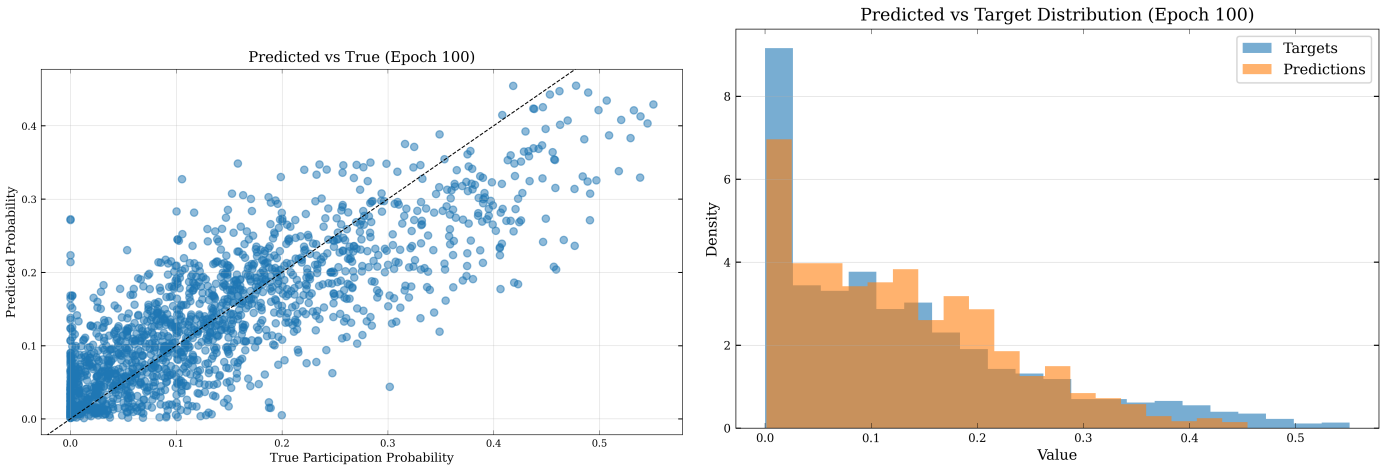
Fig. 4: TAS composites of each archetype; hatched boxes indicate breadbasket regions of interest.

5.3 Predictive Modeling

Predicting Archetype 7 In this part of the experiment, the EarthformerPredictor model was trained to forecast archetype 7 prevalence scores with a lead time of 5 days. The loss curve of the training alongside a plot of the learning rate is provided in Appendix A.

Figure 5a illustrates a scatter plot of the predicted values plotted against the true archetype coefficients on the validation set. The points are mostly clustered around the diagonal, displaying strong predictive skill. The dispersion around the diagonal suggests residual uncertainty, with the model showing signs of underestimation when predicting higher participation values.

Figure 5b compares the distributions of predicted and target archetype probabilities. Both distributions are strongly right-skewed, indicating that low participation probabilities dominate the dataset. The close overlap between the two histograms suggests that the model learned the overall probability density of the target variable reasonably well, though a slight overestimation in the mid-range values (around 0.1–0.3) can be observed.

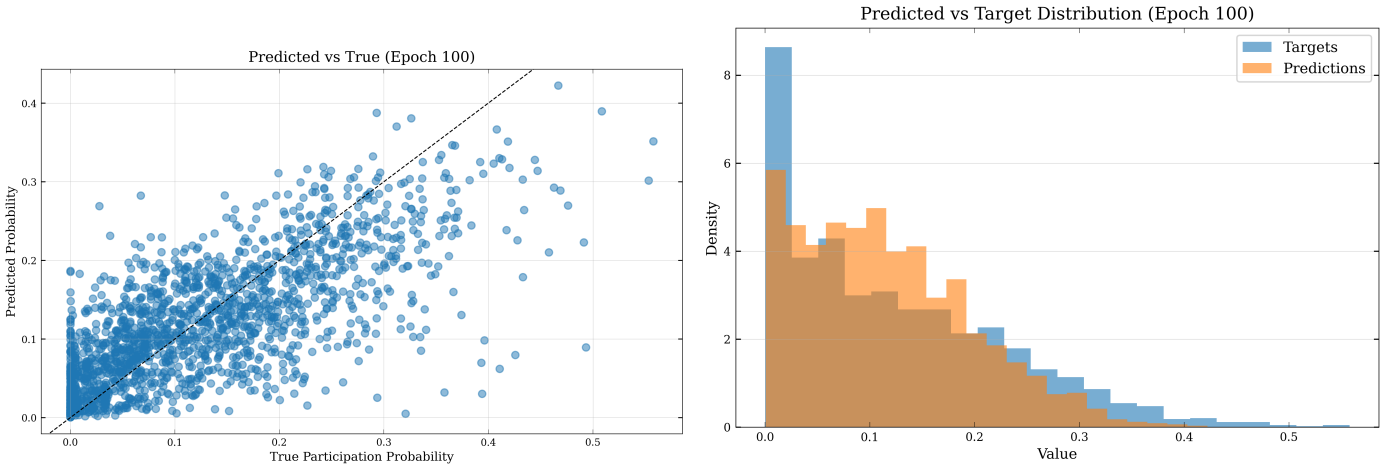


(a) Predicted vs. true labels, Archetype 7

(b) Predicted vs. true histograms, Archetype 7

Fig. 5: Prediction performance for Archetype 7. Panel (a) shows predicted versus true participation values, while panel (b) compares the two distributions.

Generalization to Other Archetypes To assess whether the predictive framework generalizes beyond a single circulation regime, we repeated the experiment using Archetype 5, which also plays an important role in summertime mid-latitude variability. The model achieved performance comparable to the Archetype 7 experiment, as shown in Figure 6, indicating that the method is not overly specialized to a particular archetype. Minor differences in performance are likely attributable to variability in the training data rather than structural limitations.



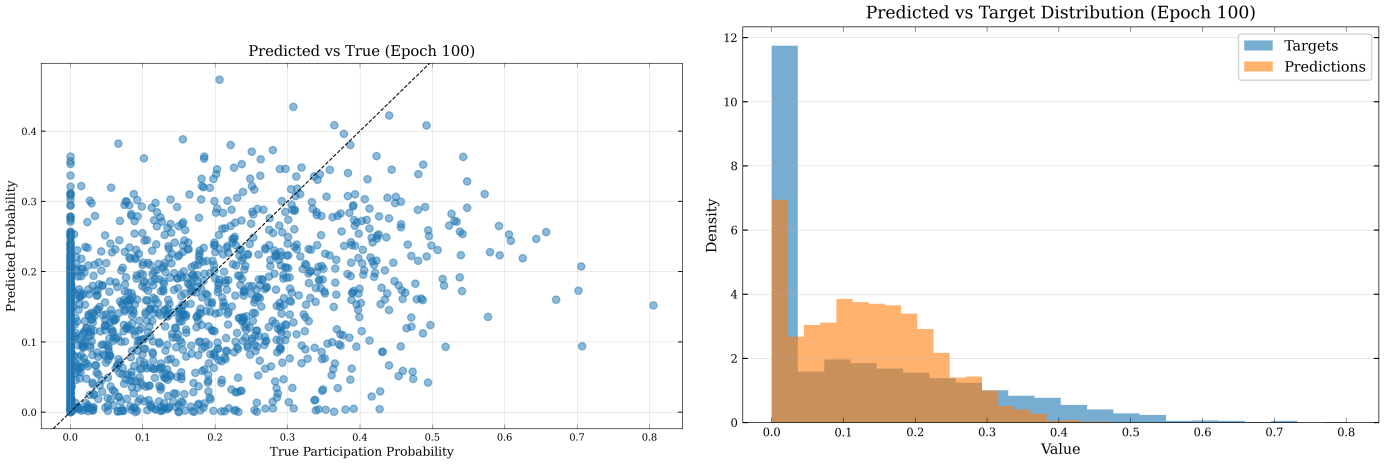
(a) Predicted vs. true labels, Archetype 5

(b) Predicted vs. true histograms, Archetype 5

Fig. 6: Prediction performance for Archetype 5. Model shows a minor performance reduction while retaining comparable predictive behavior across distinct archetypes.

5.4 Importance of Target Smoothing

We additionally tested the model using the original target series without rolling average-based smoothing to evaluate the effect of smoothing on predictability. As shown in Figure 7a, predictive skill deteriorates substantially without smoothing. Figure 7b demonstrates that compared to the smoothed targets, the unsmoothed series contains a much larger proportion of zero-valued entries, resulting in a sparser and more discontinuous target distribution. Smoothing improves the signal-to-noise ratio of the target series, allowing the model to learn a more meaningful mapping.



(a) Predicted vs. true labels

(b) Predicted vs. true histograms

Fig. 7: Prediction performance for Archetype 7 without target smoothing. Model displays reduced skill due to higher noise in target series.

5.5 Quantitative Evaluation Across Archetypes and Smoothing Configurations

We measured model performance on validation sets across Archetypes 5 and 7, and reported results for unsmoothed target series of Archetype 7. Table 2 lists Pearson r, Spearman ρ, Kendall τ, RMSE, and Accuracy for the three configurations. Two baselines are included for context: climatology (predicting the mean of the target series) and persistence (taking the target of the final input timestep). Our results show that for the smoothed target series the model performs better than the baselines for both archetypes.

Table 2: Comparison metrics across archetypes and smoothing configurations. Compared to persistence, the model achieves 45-60% improvements across correlation and error metrics for smoothed targets, whereas gains drop to ∼30% or less without smoothing, and accuracy no longer improves. This displays the role of target smoothing in enhancing predictability.

Experiment	Method	Pearson r	Spearman ρ	Kendall τ	RMSE	Accuracy
A7, smoothed	Climatology	–	–	–	0.13	0.27
	Persistence	0.56	0.55	0.39	0.12	0.40
	Model	**0.83**	**0.82**	**0.63**	**0.06**	**0.60**
A5, smoothed	Climatology	–	–	–	0.11	0.29
	Persistence	0.50	0.51	0.36	0.11	0.42
	Model	**0.74**	**0.74**	**0.56**	**0.06**	**0.59**
A7, unsmoothed	Climatology	–	–	–	0.16	0.18
	Persistence	0.30	0.28	0.21	0.19	**0.33**
	Model	**0.39**	**0.38**	**0.27**	**0.13**	0.32

5.6 Quantitative Evaluation Across Lead Times

To further characterize model performance, we evaluated the predictive skill of the model across multiple lead times on their respective validation sets. Table 3 reports Pearson r, Spearman ρ, Kendall τ, RMSE, and Accuracy for lead times of 3–10 days. Climatology and persistence baselines are provided for each configuration.

Table 3: Predictive performance for Archetype 7 across lead times, comparing climatology, persistence, and model forecasts. Across all lead times, the model consistently outperforms persistence, with relative improvements of 40-60% at 5 days and up to ∼95% for rank-based correlation metrics at 7 days. Even at a 10-day lead time, the model retains a 30-65% improvement over persistence, indicating predictive skill beyond simple temporal carryover.

Lead Time	Method	Pearson r	Spearman ρ	Kendall τ	RMSE	Accuracy
3 days	Climatology	–	–	–	0.13	0.27
	Persistence	0.79	0.78	0.59	0.09	0.54
	Model	**0.90**	**0.89**	**0.71**	**0.05**	**0.70**
5 days	Climatology	–	–	–	0.13	0.27
	Persistence	0.56	0.55	0.39	0.12	0.40
	Model	**0.83**	**0.82**	**0.62**	**0.06**	**0.60**
7 days	Climatology	–	–	–	0.13	0.27
	Persistence	0.35	0.34	0.23	0.15	0.33
	Model	**0.63**	**0.63**	**0.45**	**0.09**	**0.45**
10 days	Climatology	–	–	–	0.13	0.27
	Persistence	0.20	0.19	0.12	0.16	0.29
	Model	**0.27**	**0.29**	**0.20**	**0.11**	**0.33**

6 Discussion

The experimental results demonstrate that combining archetypal analysis with temperature composites and a transformer-based prediction model offers a coherent framework for identifying and forecasting circulation regimes associated with concurrent heat extremes. Archetypal analysis offers a compact representation of dominant upper-tropospheric wave patterns, several of which display characteristics consistent with QSWs.

The predictive modeling results indicate that Earthformer can learn meaningful relationships between recent upper-level flow patterns, lagged OLR input, and archetype participation. The model reproduces the overall distribution of the target variable and captures the dominant variability, although it tends to underestimate higher participation levels. Despite the novelty of the task and the absence of benchmark models, the results demonstrate that short-lead forecasts of archetype participation are feasible, even though predictability decreases progressively toward a 10-day lead time. This suggests that Rossby-wave-driven extreme-event precursors, when captured within an archetypal configuration, can be inferred directly from upper-level flow fields.

The configuration for our base experiment uses a 5-day lead time and a 7-day rolling-average target. Because of this setup, the earliest two days contributing to the target window overlap temporally with the most recent input days, meaning that the prediction target is not fully disjoint from the input period in a strict temporal sense. However, this overlap does not constitute direct information leakage, as the model does not observe the target values themselves and no input variable explicitly encodes future archetype participation. Instead, the model must infer the evolving circulation state from upper-level flow fields and learn how such states translate into sustained regime prevalence. In this context, the overlap reflects the intrinsic persistence of large-scale circulation rather than providing predictive cues, and may even facilitate learning by reinforcing physically meaningful continuity.

Furthermore, when the lead time is increased to 7 days – such that no part of the rolling-average target overlaps with the input window – the model still substantially outperforms a persistence baseline. In this fully non-overlapping configuration, forecast skill remains approximately 40% higher than persistence, demonstrating that the model captures predictive information beyond simple temporal carryover.

Beyond forecasting, the archetypal framework itself provides additional value as a diagnostic tool. The extracted circulation regimes present interpretable structures that climate scientists can monitor in model output or reanalysis data, and the composite analysis assesses their relevance within the context of temperature anomalies. In this sense, the methodology provides both a descriptive mechanism for understanding Rossby-wave configurations and a predictive component aimed at anticipating their emergence.

6.1 Limitations

Several limitations should be considered when interpreting the findings: First, archetypal analysis itself introduces methodological constraints. Performing SVD as a dimensionality-reduction step alters the representation of variability by emphasizing leading modes, and the re-projection of archetypes back into physical space can obscure small-scale or low-variance structures. In addition, archetypes reside on the convex hull of the reduced dataset, making the method sensitive to outliers and potentially limiting interpretability when patterns are highly nonlinear.

Second, the prediction task remains challenging because archetype participation is a target dominated by low values. This class imbalance reduces the model's sensitivity to high-participation events, precisely the cases that are of particular interest for concurrent extremes. The use of rolling-average smoothing alters the standard interpretation of predictability and lead time, as the model is trained to anticipate aggregated regime prevalence rather than daily snapshots of circulation states. As a result, the predicted target values are better understood as reflecting the probability of persistent circulation conditions, rather than the precise timing of individual extreme events.

Furthermore, the architecture inherits restrictions from Earthformer itself. Because the attention mechanism operates locally within cuboids, with only restricted communication across their boundaries, hemisperic-scale wave structures may be fragmented across blocks and not captured fully coherently. Varying input size or exploring alternative architectural choices may therefore be beneficial to better represent large-scale dynamical structures.

Additionally, the prediction model is currently trained to estimate the participation of a single archetype, limiting its context to one circulation regime. A multi-target formulation could leverage relationships between archetypes, including co-occurrence or transitions, which might improve forecast skill.

Moreover, the analysis relies on the LENTIS dataset, which inherits known structural biases from EC-Earth v3. Although the model captures large-scale dynamical processes realistically and retains the chaotic nature of the atmosphere, such biases may reduce the direct applicability of the results to real-world observations.

The study uses only the first 100 years of the available 1600-year dataset due to computational constraints. The developed pipeline is robust enough such that scaling the work to utilize the full dataset is relatively easy with sufficient resources. A longer training set could improve both archetype richness and predictive performance, particularly for rare regimes.

Finally, climate change also poses an inherent challenge: LENTIS represents conditions around 2000–2010, and future warming may produce circulation configurations that fall outside the archetypal structures learned from the dataset. This may restrict the model's ability to anticipate unprecedented extreme-event patterns.

6.2 Future Work

While the current pipeline is functional and provides clear evidence that amplified Rossby-wave circulation extremes can be identified and predicted, it is still possible to expand and improve upon this work in various directions. A straightforward next step would be to run the analyses on the full 1600-year LENTIS dataset, which would increase the frequency of rare circulation patterns and strengthen the statistical robustness of both the archetypes and the prediction model. Alternatively, applying the same methodology to observational datasets such as ERA5 would allow direct assessment of real-world skill and reveal potential discrepancies between model-derived and observed circulation regimes.

The archetypal analysis component could be extended by experimenting with different numbers of archetypes or by exploring alternative configurations that may capture other forms of extreme behavior. On the prediction side, Earthformer offers a strong backbone, but testing additional spatiotemporal architectures may yield improvements in skill or computational efficiency. Extending the set of predictor variables to include fields such as additional radiative or land–atmosphere indicators may also help increase lead time by capturing earlier precursors of QSW-like circulation development.

Seasonal and regional generalizations offer another valuable avenue for further research. Applying the framework to winter months and using negative anomalies from the composite analysis would allow the identification of circulation regimes linked to cold spells, which are equally important for crop productivity. Similarly, redefining the regions of interest would make it possible to study the effects of Rossby-wave-related circulation archetypes on other agriculturally or socially relevant zones, including Asia or northern Africa.

An additional area for improvement concerns the highly skewed distribution of archetype participation, in which large values associated with extreme circulation regimes appear comparatively rare. Future work could address this by adopting distribution-aware loss functions, such as a negative log-likelihood formulation, which would allow the model to explicitly represent uncertainty and asymmetry in the target distribution. A promising alternative would be a two-stage or multi-task prediction framework. In such an approach, the model would first predict whether a circulation regime becomes active – for instance, by exceeding a intensity or persistence threshold – and then conditionally predict the strength of the regime if it is active. Such approaches could improve sensitivity to rare, high-impact circulation states without altering the underlying physical interpretation of the prediction task.

7 Conclusion

This study set out to identify large-scale Rossby-wave circulation regimes associated with concurrent heat extremes and to assess whether these regimes can be systematically predicted from upper-tropospheric flow fields. Using archetypal analysis applied to the 250 hPa stream

function, we extracted a set of circulation structures that capture dominant amplified Rossby-wave configurations. Further, composite analysis showed that these archetypes differ substantially in their surface temperature impacts, with Archetype 7 and 5 producing the strongest warm anomalies across multiple American, European, and Asian, breadbasket regions. We then trained a transformer-based model to forecast the participation of these archetypes at a subseasonal lead times up to 10 days, with a 5-day lead time used as a primary proof of concept. We demonstrated that performance hinges on smoothing the target series by means of a rolling average. The model successfully reproduced the main characteristics of the target distribution and achieved meaningful predictive skill despite the absence of prior benchmarks for this task. As seen in the quantitative evaluation across lead times, the model displays a level of accuracy inversely proportional to the chosen lead time.

Ultimately, the obtained results indicate that combining pattern-extraction methods with spatiotemporal deep learning provides a viable pathway for anticipating heat extremes driven by amplified Rossby-wave circulation patterns. Our findings offer both interpretive value and predictive potential for applications related to agricultural risk assessment, particularly in a climate system where circulation-related extremes are expected to intensify. The code and data used in this research are openly available on GitHub [17].

8 Acknowledgements

We kindly acknowledge Tamara Hoppé for making the LENTIS stream function data available, and we acknowledge sponsoring by the VU HPC Council and the IT for Research (ITvO) ADA Linux computational cluster at the VU University Amsterdam.

References

1. Aslak, U., Thøgersen, J.C.: py_pcha: Fast python implementation of archetypal analysis using principle convex hull analysis (pcha). https://github.com/ulfaslak/py_pcha (2024)
2. Black, A.S., Monselesan, D.P., Risbey, J.S., Sloyan, B.M., Chapman, C.C., Hannachi, A., Richardson, D., Squire, D.T., Tozer, C.R., Trendafilov, N.: Archetypal analysis of geophysical data illustrated by sea surface temperature. Artificial Intelligence for the Earth Systems **1** (8 2022). https://doi.org/10.1175/aies-d-21-0007.1
3. Chapman, C.C., Monselesan, D.P., Risbey, J.S., Hannachi, A., Lucarini, V., Matear, R.: The Typicality of Regimes Associated with Northern Hemisphere Heatwaves (7 2025), https://figshare.le.ac.uk/articles/journal_contribution/The_Typicality_of_Regimes_Associated_with_Northern_Hemisphere_Heatwaves/30009697
4. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**(4), 338–347 (1994). https://doi.org/10.1080/00401706.1994.10485840, https://www.tandfonline.com/doi/abs/10.1080/00401706.1994.10485840
5. Dalman, L.: Climate variability: Pacific–north american pattern (2023), https://www.climate.gov/news-features/understanding-climate/climate-variability-pacific-north-american-pattern, accessed: 2025-01-26
6. Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F.J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M.P., Keskinen, J.P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P.A., Moreno-Chamarro, E., Nieradzik, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O.T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X., Zhang, Q.: The ec-earth3 earth system model for the coupled model intercomparison project 6. Geoscientific Model Development **15**(7), 2973–3020 (2022). https://doi.org/10.5194/gmd-15-2973-2022, https://gmd.copernicus.org/articles/15/2973/2022/
7. Fei, C., White, R.H.: Large-amplitude quasi-stationary rossby wave events in era5 and the cesm2: Features, precursors, and model biases in northern hemisphere winter. Journal of the Atmospheric Sciences **80**, 2075–2090 (8 2023). https://doi.org/10.1175/JAS-D-22-0042.1
8. Fitzpatrick, R.: Theoretical Fluid Mechanics. 2053-2563, IOP Publishing (2017). https://doi.org/10.1088/978-0-7503-1554-8, https://doi.org/10.1088/978-0-7503-1554-8
9. Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., Yeung, D.Y.: Earthformer: Exploring space-time transformers for earth system forecasting (7 2022), http://arxiv.org/abs/2207.05833
10. Ghosh, N., Santoni, D., Nawn, D., Ottaviani, E., Felici, G.: A comprehensive review of transformer-based language models for protein sequence analysis and design (2025), https://arxiv.org/abs/2507.13646
11. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010), https://proceedings.mlr.press/v9/glorot10a.html
12. Happé, T., Wijnands, J.S., Fernández-Torres, M., Scussolini, P., Muntjewerf, L., Coumou, D.: Detecting spatiotemporal dynamics of western european heatwaves using deep learning. Artificial Intelligence for the Earth Systems **3**(4), e230107 (2024). https://doi.org/10.1175/AIES-D-23-0107.1, https://journals.ametsoc.org/view/journals/aies/3/4/AIES-D-23-0107.1.xml
13. Kautz, L.A., Martius, O., Pfahl, S., Pinto, J.G., Ramos, A.M., Sousa, P.M., Woollings, T.: Atmospheric blocking and weather extremes over the euro-atlantic sector – a review. Weather and Climate Dynamics **3**(1), 305–336 (2022). https://doi.org/10.5194/wcd-3-305-2022, https://wcd.copernicus.org/articles/3/305/2022/
14. Kornhuber, K., Coumou, D., Vogel, E., Lesk, C., Donges, J.F., Lehmann, J., Horton, R.M.: Amplified rossby waves enhance risk of concurrent heatwaves in major breadbasket regions (1 2020). https://doi.org/10.1038/s41558-019-0637-z
15. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017), https://arxiv.org/abs/1608.03983
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
17. Mitrani, M.: concurrent-heatwave-prediction: Predicting high-risk atmospheric patterns linked to northern hemispheric crop failures with spatiotemporal transformers. https://github.com/markmitrani/concurrent-heatwave-prediction (2025)
18. Muntjewerf, L., Bintanja, R., Reerink, T., van der Wiel, K.: The knmi large ensemble time slice (knmi–lentis). Geoscientific Model Development **16**(15), 4581–4597 (2023). https://doi.org/10.5194/gmd-16-4581-2023, https://gmd.copernicus.org/articles/16/4581/2023/

19. Mørup, M., Hansen, L.K.: Archetypal analysis for machine learning and data mining. Neurocomputing **80**, 54–63 (2012). https://doi.org/https://doi.org/10.1016/j.neucom.2011.06.033, https://www.sciencedirect.com/science/article/pii/S0925231211006060, special Issue on Machine Learning for Signal Processing 2010
20. Ossó, A., Bladé, I., Karpechko, A., Li, C., Maraun, D., Romppainen-Martius, O., Shaffrey, L., Voigt, A., Woollings, T., Zappa, G.: Advancing our understanding of eddy-driven jet stream responses to climate change – a roadmap. Current Climate Change Reports **11**(1), 2 (Nov 2024). https://doi.org/10.1007/s40641-024-00199-3, https://doi.org/10.1007/s40641-024-00199-3
21. Pereira, G.A., Hussain, M.: A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships (2024), https://arxiv.org/abs/2408.15178
22. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat: Deep learning and process understanding for data-driven earth system science. Nature **566**(7743), 195–204 (Feb 2019). https://doi.org/10.1038/s41586-019-0912-1, https://doi.org/10.1038/s41586-019-0912-1
23. Röthlisberger, M., Bosart, L.F., Keyser, D., Martius, O.: Recurrent synoptic-scale rossby wave patterns and their effect on the persistence of cold and hot spells . https://doi.org/10.1175/JCLI-D-18, https://doi.org/10.1175/JCLI-D-18-
24. Schulzweida, U.: Cdo user guide (Oct 2020). https://doi.org/10.5281/zenodo.5614769, https://doi.org/10.5281/zenodo.5614769
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
26. White, R.H., Kornhuber, K., Martius, O., Wirth, V.: From atmospheric waves to heatwaves: A waveguide perspective for understanding and predicting concurrent, persistent, and extreme extratropical weather. Bulletin of the American Meteorological Society **103**, E923–E935 (3 2022). https://doi.org/10.1175/BAMS-D-21-0170.1
27. Zhang, K., Randel, W., Fu, R.: Relationships between outgoing longwave radiation and diabatic heating in reanalyses. Climate Dynamics **49** (10 2017). https://doi.org/10.1007/s00382-016-3501-0

A Additional Figures

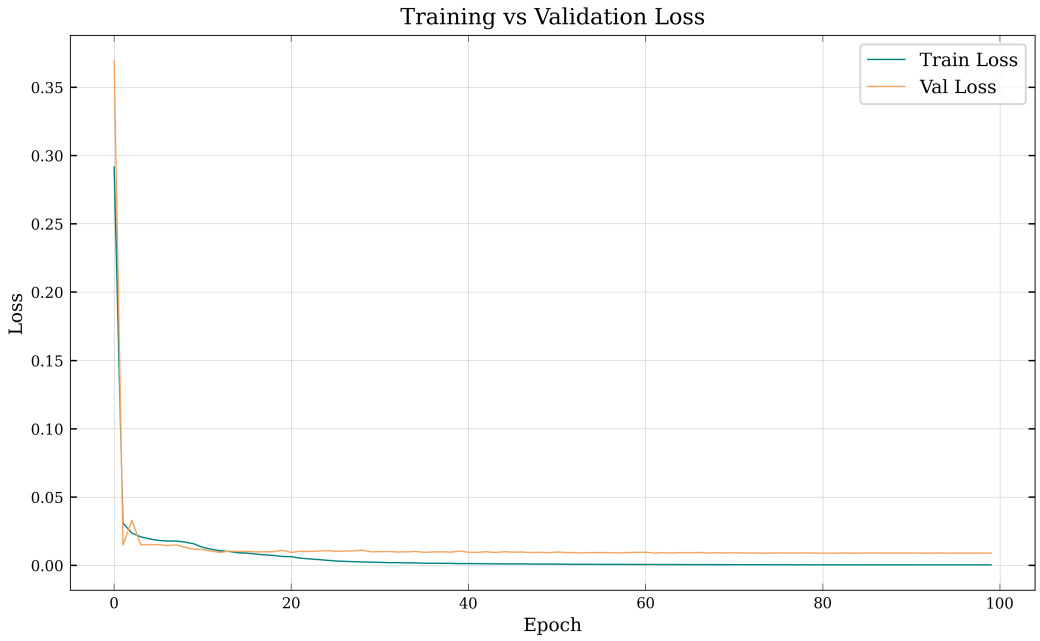


Fig. 8: Loss curves during training. Both losses decrease steeply in the initial epochs and then stabilize, without any signs of overfitting.

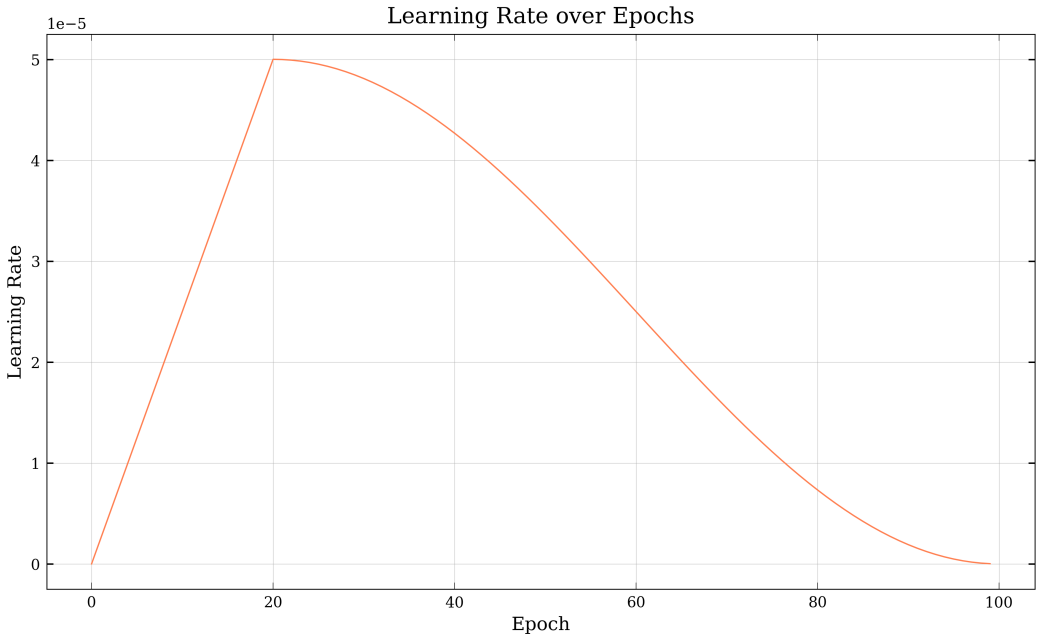


Fig. 9: Evolution of learning rate over epochs. For the initial 20% of epochs, a linear warm-up strategy is followed, then the learning rate is brought back to 0 with cosine annealing.