Opening the Black Box: Attention Analysis Reveals Learned Physics and Architectural Constraints in Transformer Weather Models

Jorden Gershenson

Computer Science Master's Thesis

Thesis Committee:

Dr. Xin Zhong, *Chair*
Dr. Hesham Ali
Dr. Adam Houston
Dr. Pei-Chi Huang

University of Nebraska Omaha
College of Information Science & Technology

01 December 2025

Abstract

Machine learning weather models achieve remarkable forecast skill with dramatic computational speedups, yet operational adoption remains limited by their opacity. This thesis develops an interpretability framework for Transformer-based weather models by analyzing Pangu-Weather's attention mechanisms. We introduce a mass-preserving reverse attention rollout algorithm that traces information flow backward through the model, identifying which input regions influence specific forecasts. Novel visualizations including global contribution maps and wind-rose diagrams make high-dimensional attention patterns interpretable to meteorologists. Analysis suggests the model learned physically meaningful patterns while also exposing architectural limitations. These findings demonstrate attention analysis both validates learned physics and reveals architectural artifacts, providing tools for building forecaster trust and advancing explainable AI.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation: The Promise and Challenge of ML Weather Prediction

Weather forecasting is a critical application at the intersection of science, engineering, and society. Traditionally, operational forecasts rely on physics-based numerical weather prediction (NWP) models, which solve complex atmospheric equations to predict future states. These models, such as the Weather Research and Forecasting (WRF) model, High-Resolution Rapid Refresh (HRRR), and the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS),have proven skillful, but demand enormous computational resources and expert tuning.

Recently, *machine learning weather prediction* (MLWP) models have emerged as a compelling alternative, harnessing data-driven learning to produce forecasts with tremendous advances in speed and accuracy. For example, deep neural network models can generate global forecasts in seconds and have begun to rival or even exceed the accuracy of NWP on key metrics. This combination of speed and skill holds the potential to revolutionize forecasting, enabling faster update cycles and significant cost savings in compute infrastructure.

However, a major barrier prevents widespread adoption of MLWP models in operational settings: **trust and interpretability**. Meteorologists and other domain experts are understandably hesitant to deploy a "black-box" model, however accurate, without understanding how it arrives at its predictions. The central question motivating this work is: *How can we provide useful explanations for the predictions of an advanced ML weather model, given that such models lack the explicit, human-readable reasoning chains that traditional methods possess?*

1.2 A Paradigm Shift in Computer Science

In the broader context of computer science, this challenge reflects a paradigm shift. Our field has a long history of developing algorithms (e.g., for sorting, search, optimization) that are easily understood and whose behavior can be transparently traced. In those cases, sharing our innovations

with domain experts was straightforward: Users can trust a well-known algorithm because its decision process is explicit and deterministic.

Deep learning models, in contrast, operate as complex learned representations with millions of parameters; Their decision-making process is encoded implicitly in these parameters rather than in clear rules. The dramatic success of deep learning across many domains has thus created a new responsibility for computer scientists: not only must we build powerful models, but we must also become *investigators* of those models' inner workings. In a sense, studying a modern deep neural network's behavior can resemble experimental science—more like observing and probing a physical system (or even an animal's cognition) than executing a predictable formula.

This thesis embraces that perspective. It argues that to bridge the gap between AI advancements and domain adoption, we need to develop and share methods to interpret and explain what these "black box" models have learned. By doing so, we can foster trust with experts (in this case, meteorologists) and ensure that our innovations truly translate into real-world improvements.

1.3 Leveraging Self-Attention for Interpretability

One promising avenue toward interpretability in deep learning is to leverage the model's own internal structures to shed light on its reasoning. Many state-of-the-art MLWP models are built on the *Transformer architecture*, which is characterized by a mechanism called *self-attention*. In simple terms, self-attention allows the model to weight the importance of different input features relative to each other when making a prediction. In a weather forecasting context, a Transformer-based model can "attend" to different regions of the input atmosphere to decide the future state at a given location.

Importantly, the self-attention weights are computed internally for each forecast time step and each layer of the model, but they are normally not exposed to users. Our insight is that these attention patterns, if extracted and visualized, could serve as a window into the model's decision process. This thesis explores whether analyzing a model's self-attention can reveal which parts of the atmospheric state the model deemed most relevant for a specific forecast outcome. In other words, we ask: *do the internal attention patterns of a weather Transformer model correspond to known meteorological cause-and-effect relationships and can we interpret those patterns to explain the model's predictions?*

1.3.1 Related Work in Weather Transformers

There has been growing interest in applying Transformers to weather and climate forecasting in recent years. Researchers have demonstrated that Transformer-based models can be used for tasks such as post-processing numerical forecasts, downscaling, or even direct end-to-end prediction of global weather fields. Notable examples include models like FourCastNet and Pangu-Weather, which leverage vision Transformer architectures to achieve high accuracy in medium-range forecasting. However, most of these efforts have focused on raw predictive performance, with little emphasis on

interpretability. As a result, while we know these models work well, we often do not know *why* or *how* they make specific forecasting decisions.

In parallel, the machine learning research community has begun to develop techniques for interpreting Transformers in other domains. In computer vision, for instance, recent studies have visualized attention maps to understand what image regions a vision Transformer focuses on for classification or detection tasks (e.g., identifying which parts of an image contribute to recognizing an object) [11]. In geospatial and remote sensing applications, attention-based interpretability has also been explored at a rudimentary level [30]. These works show that attention visualizations can sometimes highlight meaningful patterns (such as image patches corresponding to important objects or semantic regions).

Applying similar interpretability techniques to a spatiotemporal science domain like weather forecasting remains largely uncharted territory. This gap defines the opportunity for our research: by bringing attention visualization methods into the realm of meteorology, we aim to provide new insights both for AI experts and for atmospheric scientists.

1.4 Research Objectives and Scope

In this thesis, we address the interpretability gap in ML-based weather forecasting by focusing on the following key objectives:

1. **Explainable Attention in Weather Models:** Develop a method to extract and interpret the *self-attention patterns* of a Transformer-based weather prediction model, in order to identify which input regions and features the model uses to make forecasting decisions.

2. **Physical Plausibility:** Investigate whether the model's attention aligns with known meteorological processes. For example, does the model "look at" upstream weather systems, adjacent pressure levels, or other physically relevant precursors when predicting a target region's weather? We seek to determine to what extent the learned attention patterns make meteorological sense, reflecting real-world cause and effect in the atmosphere.

3. **Visualization Tools for Domain Experts:** Create intuitive visualization techniques to communicate these attention patterns to meteorologists and other non-ML experts. This includes developing novel plots that summarize where and how strongly the model is attending, thereby providing an explanatory narrative for each forecast.

1.5 Approach: Reverse Attention Rollout

To tackle these objectives, we design an interpretability framework and demonstrate it on a MLWP model. Specifically, we use the *Pangu-Weather model* [5], a recently published high-resolution global forecasting model that employs a vision Transformer backbone. Pangu-Weather was chosen because it is a successful example of a transformer-based MLWP, achieving accuracy comparable to leading

numerical models at a fraction of the computation time. We utilize a PyTorch reimplementation [35] that closely adheres to the official pseudocode and validates against the pretrained weights, enabling us to instrument the model to record its internal attention weights during inference. These attention weights essentially tell us, for each layer and each prediction, how the model is distributing attention across different parts of the input grid.

1.5.1 Attribution Algorithm

Building on these raw attention data, we introduce a custom algorithm for attention-based attribution in the context of spatiotemporal data. The technique can be described as a *reverse attention rollout*: starting from a particular forecasted output (a particular patch or region of the forecast), we trace backwards through the layers of the Transformer, following the chain of attention to determine which input locations most influenced that output.

By iteratively propagating an influence score from the target output through the attention links (essentially transposing the attention operation), we accumulate a contribution score for every input grid point, a measure of how much that point affected the target forecast. The result of this process is a *contribution map* covering the global input domain for the chosen output point. This contribution map is a tangible representation of the model's implicit reasoning: it highlights the areas of the initial weather state that the model considered important for predicting the weather at the target point 24 hours later.

1.5.2 Visualization Design

A single contribution map is high-dimensional (covering multiple pressure levels over the globe), so conveying its information clearly is a challenge. We address this by developing two forms of visualization:

- **Global heatmaps:** We present contribution maps on geographic projections for each atmospheric layer. These maps use a common color scale to show where the most influential inputs lie, effectively visualizing the three-dimensional attention structure in slices.
- **Wind-rose plots:** To better summarize directional and distance-dependent patterns, we design wind-rose diagrams that bin the contributions by azimuth (direction relative to the target point) and distance. By examining these plots, one can quickly discern, for instance, whether the model drew most of its information from the west or perhaps equally from all around.

We produce such visualizations for multiple test cases. Specifically, for a set of 12 representative geographic points with distinct weather regimes—to see how the model's attention behavior might vary with location and weather scenario.

Chapter 2

Background and Related Work

2.1 Machine Learning Weather Prediction

2.1.1 From Physics-Based to Data-Driven Forecasting

For decades, weather forecasting has relied on NWP, which solves the fundamental equations of atmospheric physics on discretized grids [4]. These models are triumphs of computational science, but face immense challenges rooted in the nature of the atmosphere itself. The Earth's climate is a chaotic, multiscale system where processes interact across vast ranges of space and time, from microscopic cloud physics to planetary-scale waves [17]. Because even the most powerful supercomputers cannot resolve all physical processes, NWP models must approximate subgrid-scale phenomena like convection and turbulence using "parameterization" schemes, a major source of model uncertainty [26]. Consequently, producing a single forecast requires hundreds of supercomputer nodes running for hours [5].

MLWP represents a fundamentally different paradigm. Rather than explicitly solving physical equations, MLWP models learn to recognize patterns in historical weather data, primarily using the ERA5 reanalysis dataset as training data [14]. Once trained, these models can generate forecasts orders of magnitude faster than NWP—producing global predictions in minutes on a single GPU [5, 19]. The evolution of standardized evaluation reflects this shift, from the initial WeatherBench [24] to the more comprehensive WeatherBench 2 [25]. However, challenges remain in fairly evaluating extreme events, where data-driven models can still struggle. For example, recent studies show that for record-breaking weather extremes, leading numerical models like ECMWF's high-resolution deterministic forecast (HRES) consistently outperform AI models including Pangu-Weather [34]. This performance gap is often attributed to the models' tendency to underpredict the intensity of events that fall outside the climatology of their training data.

2.1.2 The Rise of Transformer-Based Weather Models

The application of Transformer architectures to weather forecasting has produced some of the most successful MLWP models. Transformers, originally developed for natural language processing, use

self-attention mechanisms to capture long-range dependencies in sequential data [31]. The Vision Transformer (ViT) adapted this architecture for image data by treating images as sequences of patches, demonstrating its power beyond text [11].

For high-resolution, 3D global weather data, the non-polynomial computational cost of the original attention mechanism can be prohibitive. The Pangu model solves this by using a Swin Transformer architecture. The Swin Transformer addresses this by computing self-attention only within local windows and using shifted windows in alternating layers to enable cross-window communication [20]. This innovation reduces computational complexity from quadratic to linear with respect to input size, making it a feasible backbone for large-scale scientific data.

2.1.3 A Landscape of Competing Architectures

While powerful, the Transformer paradigm is not monolithic. The field of MLWP is a dynamic landscape of competing architectural philosophies. To properly contextualize this thesis's focus on a Transformer-based model, it is essential to consider the leading alternatives.

Pangu-Weather, the subject of this thesis, exemplifies the state-of-the-art in the Transformer paradigm with its 3D Earth-Specific Transformer (3DEST) architecture [5]. It extends the Swin Transformer to handle three-dimensional atmospheric data, allowing the model to capture both horizontal and vertical interactions.

GraphCast, from Google DeepMind, employs a Graph Neural Network (GNN) architecture [19]. Instead of a standard latitude-longitude grid, GraphCast represents the globe using an icosahedral mesh, which avoids the grid distortions and polar singularities inherent in the grid used by Pangu-Weather. The GNN learns to pass messages between nodes on this graph, representing a fundamentally different approach to modeling spatial relationships on a sphere.

FourCastNet, from NVIDIA, is based on the Fourier Neural Operator (FNO) architecture [23]. FNOs operate in the spectral domain, using the Fast Fourier Transform to implement a global convolution at each layer. This allows the model to efficiently capture global dependencies, contrasting with the local windowed attention of the Swin Transformer.

This thesis, by performing a deep analysis of Pangu-Weather, should be viewed as a case study into the specific behaviors and architectural biases of the *vision transformer paradigm* when applied to geophysical fluid dynamics.

2.2 The Explainability Challenge in Scientific Machine Learning

2.2.1 Why Black Boxes Are Problematic for Scientific Applications

The opacity of deep learning models presents unique challenges in scientific domains [10]. Unlike commercial applications where predictive accuracy may be sufficient, scientific use demands an understanding of underlying mechanisms. In weather forecasting, this need is particularly acute: forecasters must justify high-stakes decisions, and understanding model failures is essential for

improvement and building trust [22]. This opacity becomes critical when models fail, particularly for extreme events where traditional NWP can still outperform MLWP [32].

Pangu-Weather, despite its successes, is a case in point for these challenges. While it demonstrates superior skill on standard metrics, external validation has raised critical questions about its physical realism. Operational evaluations from the Guangdong Meteorological Observatory found that after one year of operational data accumulation, while Pangu-Weather outperforms traditional NWP models in statistical forecast skills of medium-range forecasting, its performance in predicting the intensity and structure of severe weather events is inadequate, with significant underestimation of both intensity and physical structure of tropical cyclones [27]. This pattern extends more broadly: AI weather prediction models' intensity forecast errors for tropical cyclones are larger than those of even the simplest intensity forecasts based on climatology and persistence [7]. These documented shortcomings in precisely the situations where forecast reliability is most critical provide a strong motivation for this thesis. A deep interpretability analysis is necessary not just to build general trust, but to specifically investigate why a powerful model like Pangu-Weather succeeds on average yet fails at the extremes, a crucial step toward improving its reliability.

2.2.2 Evidence for Physical Consistency in Learned Representations

Despite their opacity, emerging evidence suggests MLWP models can learn physically meaningful representations. Baño-Medina et al. (2025) used sensitivity analysis to probe an AI model's understanding of cyclone dynamics. Comparing these sensitivities to those from physics-based adjoint models revealed striking similarities, suggesting the AI model had learned physically plausible causal relationships [3]. However, since MLWP models are trained on physically consistent reanalysis data, it remains an open question whether they are discovering atmospheric principles independently or simply learning to emulate the physics encoded in their training data. The broader challenge of incorporating physical knowledge into ML models has been formalized as theory-guided data science [18].

2.2.3 A Survey of XAI Methods for Model Interpretability

To address the "black box" problem, the field of explainable AI (XAI) has developed a broad toolkit of post-hoc analysis techniques. These methods can be grouped into several families:

- **Gradient-based methods** attribute a prediction to input features by analyzing the gradients of the output with respect to the input, such as in saliency mapping [28] or Integrated Gradients [29].
- **Perturbation-based methods** probe a model by observing how outputs change when inputs are altered. Methods like SHAP (SHapley Additive exPlanations) use principles from game theory to assign importance scores [21].

- **Relevancy propagation methods** seek to decompose a model's output, propagating its "relevance" backward through the network. Layer-wise Relevance Propagation (LRP) is a prominent example [2].

While powerful, these general-purpose methods were often developed for image classification and may not fully capture the complex spatiotemporal dependencies inherent in weather forecasting.

2.3 Attention Mechanisms as Explanatory Tools

2.3.1 The Self-Attention Mechanism

Self-attention is the core innovation of Transformer architectures, enabling models to dynamically weight the importance of different input elements. For an input sequence, self-attention computes three learned representations: queries (Q), keys (K), and values (V). The attention weights are computed as:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{2.1}$$

where d_k is the key dimension. These weights form a matrix where element A_{ij} represents the attention that position i pays to position j. The output is a weighted sum of values: $\text{Attention}(Q, K, V) = AV$. In theory, these weights could indicate which atmospheric regions most influence a prediction, making attention visualization a popular approach for understanding Transformer behavior.

2.3.2 The Attention Controversy: Faithfulness vs. Plausibility

The use of attention weights as direct explanations has become controversial. Jain and Wallace (2019) famously argued that "attention is not explanation," demonstrating that attention weights often correlate poorly with gradient-based importance measures [16]. Wiegreffe and Pinter (2019) offered a nuanced counterargument, "attention is not not explanation," distinguishing between *faithful* explanations (perfectly reflecting causal mechanisms) and *plausible* explanations (providing useful insights) [33]. This debate is critical for scientific applications, where both matter. To rigorously evaluate explanations, benchmarks like ERASER have been proposed [9], though they are not yet standard in scientific domains.

2.3.3 Attention Rollout and Attribution Methods

Attention rollout, introduced by Abnar and Zuidema (2020), addresses a key limitation of visualizing raw attention from a single layer [1]. In a deep Transformer, information propagates through successive attention operations. Attention rollout traces these paths by recursively combining attention matrices across layers. The basic rollout recursively computes:

$$\tilde{A}^{(l)} = \tilde{A}^{(l-1)} \cdot A^{(l)} \tag{2.2}$$

where $A^{(l)}$ is the attention matrix at layer l. Other methods have been proposed to refine this, such as accounting for the value matrices [6] or using network flow formulations [8]. However, adapting any such approach to a model like Pangu-Weather requires handling the unique challenges of windowed attention, 3D data structures, and residual connections, which motivates the methodological contributions of this thesis.

2.4 Synthesis: Identifying the Research Gap

2.4.1 What's Missing

The preceding review reveals a clear research gap at the intersection of MLWP, XAI, and atmospheric science. Despite rapid progress, significant questions remain:

1. **Architecture-Specific XAI for Weather Transformers**: While a broad XAI toolkit exists, and recent work has begun exploring the interpretability of AI weather models through sensitivity analysis and attribution studies (e.g., Baño-Medina et al., 2025), no work has yet developed and applied a rigorous, meteorologically-grounded XAI framework specifically designed to dissect the attention mechanisms of a leading weather Transformer like Pangu-Weather. Existing attention rollout methods must be adapted for the unique complexities of multi-dimensional, physically-structured atmospheric data.

2. **Distinguishing Learned Physics from Architectural Bias**: It is unknown to what extent a weather Transformer's internal representations reflect learned physical principles versus inherent architectural constraints. A key missing piece is an analysis that systematically attempts to disentangle these two factors.

3. **Validation Against Domain Expertise**: Most interpretability research focuses on technical metrics. There is a pressing need to evaluate XAI-generated explanations against the domain knowledge of expert meteorologists to determine if they provide practically useful and trustworthy insights.

2.4.2 This Thesis's Contribution

This thesis fills this critical gap by developing, applying, and evaluating an attention-based explainability framework as a deep-dive case study into the Transformer paradigm for weather forecasting. We introduce a mass-preserving reverse attention rollout algorithm specifically designed for the hierarchical, 3D structure of Pangu-Weather. Through systematic analysis, we investigate the model's attention patterns to distinguish its learned physical knowledge from its inherent architectural biases.

Our approach goes beyond technical development to include expert evaluation, assessing whether attention-based explanations provide actionable insights for meteorologists. By bridging the fields of machine learning interpretability and atmospheric science, this work contributes both methodological

advances for spatiotemporal XAI and practical tools for building trust in the next generation of AI weather forecasting models.

Chapter 3

Methods

This chapter describes our approach to extracting, visualizing, and validating attention-based explanations from the Pangu-Weather model. We present three methodological contributions: (1) a mass-preserving reverse attention rollout algorithm adapted for hierarchical, windowed Transformers operating on atmospheric data, (2) visualization techniques designed to make high-dimensional attention patterns interpretable to meteorologists, and (3) a structured expert evaluation protocol to assess the physical plausibility and operational utility of the resulting explanations. Together, these methods bridge technical explainable AI capabilities with domain-specific validation, enabling rigorous assessment of whether the model's learned attention patterns align with atmospheric physics.

3.1 Experimental Setup

3.1.1 Dataset and Forecast Case

We analyze a single 24 h global forecast with initial conditions from **2 July 2019, 00 UTC** and target conditions from **3 July 2019, 00 UTC**. This date was selected so simplify implementation, as input data for this inference is provided as a part of a quick start guide to run the model locally. While analyzing a single forecast limits climatological generalization, it enables controlled, detailed analysis of attention behavior and facilitates thorough expert evaluation.

The model operates on a standard 0.25° latitude–longitude grid (721×1440) with 14 pressure levels, processing 5 atmospheric variables and 4 surface variables, including temperature, wind, humidity, etc.

3.2 Model Instrumentation

3.2.1 Implementation Choice

We use `pangu-pytorch`, an open-source PyTorch reimplementation of Pangu-Weather [35], rather than the official Open Neural Network Exchange (ONNX) weights. This choice was driven by

practical considerations: PyTorch provides direct access to intermediate attention tensors during forward passes, whereas extracting these from ONNX models would require substantial engineering effort. The reimplementation achieves comparable forecast skill to the original model while exposing the internal attention mechanisms necessary for our analysis.

3.2.2 Attention Extraction

During model inference, we instrument each of the Transformer blocks to save attention weights after the softmax operation. For each block we capture attention matrices, resulting in sparse block-diagonal matrices due to the windowed attention mechanism. These matrices are saved to disk in NumPy format, with each file containing attention weights for one block's forward pass.

The Swin Transformer's alternating window patterns—regular and shifted windows in successive blocks—create different attention connectivity patterns that must be carefully tracked. We maintain metadata about window positions and shifts to correctly map attention weights back to geographic coordinates during the reverse rollout phase.

3.3 Mass-Preserving Reverse Attention Rollout

3.3.1 The Challenge: Tracing Information Flow Through Hierarchical Transformers

Standard attention rollout methods, developed for simple Transformer architectures operating on flat token sequences, cannot directly apply to Pangu-Weather due to four critical complexities:

Hierarchical windowed attention. Unlike standard Transformers where each token can attend to all others, the Swin Transformer architecture restricts attention to local windows of size $12 \times 6 \times 2$ (longitude $\times$ latitude $\times$ pressure) for computational tractability. This creates a sparse attention graph where direct connections exist only within windows. Information propagates to distant locations through multiple hops across successive layers, a pixel at the equator influences a polar pixel only through intermediate windows that bridge the distance.

Shifted window patterns. To enable cross-window communication despite local attention, alternate Transformer blocks shift windows by half their size (6, 3, 1 pixels respectively). Block 0 uses regular windows aligned to a fixed grid; Block 1 shifts windows by half-size offsets; Block 2 returns to regular alignment, and so on. Tracking which pixels can communicate requires careful accounting of these alternating patterns, a pixel's "neighbors" change depending on whether the current block uses regular or shifted windows.

Multi-resolution hierarchy. The model processes data at two resolutions: full resolution (181×360 spatial) in early and late layers, and half resolution (91×180) in middle layers. Downsampling aggregates four pixels into one through conservative summation; upsampling redistributes mass

uniformly to four pixels. Propagating attribution backward through these transitions requires inverse operations that preserve total mass while correctly routing contributions.

Residual connections bypass attention. Modern Transformers use residual connections: Output = Attention(Input) + Input. This means a pixel always influences itself through the identity bypass, even if it receives zero attention weight from the attention mechanism. Standard attention rollout ignores this skip connection, attributing all influence to attention paths. For atmospheric predictions where temporal persistence (today's weather strongly predicts tomorrow's) is crucial, ignoring self-influence produces physically implausible attributions.

These challenges compound: we must simultaneously track windowed connectivity patterns, handle shifting windows, navigate resolution transitions, and account for residual connections, all while maintaining strict mass conservation to ensure interpretable attributions.

3.3.2 Core Algorithm: Backward Mass Transport

Our approach treats attribution as a *mass transport problem*. We initialize unit mass at the focal pixel (the forecast location we wish to explain) and propagate this mass *backward* through the network, following attention connections in reverse. At each layer, mass flows from a pixel to all pixels that attended to it, weighted by their attention strengths. After propagating through all layers, the resulting mass distribution indicates which input regions contributed to the target forecast.

Direction of propagation (rows, not columns). For a block with post-softmax attention A, the forward computation for a query index q takes a weighted sum over keys k via $y_q = \sum_k A_{qk} V_k$. Thus, the contributors to q are precisely the keys that q *attended to*. Our reverse step therefore uses *rows* of A (from query to keys)—not $A^\top$. In each block we read the length-144 attention row for the local query index and distribute the incoming mass at that query over its 144 local keys.

Residual rehydration (bypass accounting). Residual and MLP paths create self-persistence that is not visible in raw attention. We model this with a rehydration factor $\beta \in [0, 1]$ applied per row:

$$P = (1 - \beta)\, A + \beta\, I_{144}.$$

Operationally, we scale the attention row by $(1 - \beta)$ and add β to its own local query position. This yields a convex, row-stochastic update that attributes a tunable fraction to self-persistence while preserving total mass. We use $\beta = 0.3$, meaning 30% of influence is attributed to persistence and 70% to attended locations.

Window-aware propagation. For each pixel with nonzero mass, we determine its window membership (accounting for regular vs. shifted windows), identify the 143 other pixels within that window, and distribute mass according to their attention weights (how much they attended to our

pixel). Longitude boundaries wrap periodically; latitude and pressure boundaries are clamped. This requires explicit tracking of: (i) window indices $(\text{win}_z, \text{win}_h, \text{win}_w)$ derived from global coordinates, (ii) whether the current block uses shifted windows (determined by block index mod 2), (iii) coordinate transforms that map window-local positions (0–143) back to global grid positions, and (iv) periodic longitude wrapping at the 360° boundary.

Conservative resampling. When transitioning between resolutions, we apply inverse operations that preserve total mass exactly. **Downsampling** (full → half resolution): aggregate four pixels' mass into one via summation, with special handling for the 181→91 latitude dimension (pad to 182, then 2×2 pooling). **Upsampling** (half → full resolution): distribute mass uniformly to four pixels by repeating values and scaling by 0.25, then crop back to 181 latitude. Both operations maintain $\sum M_{\text{before}} = \sum M_{\text{after}}$ to machine precision.

3.3.3 Detailed Algorithm

The complete algorithm is presented in Algorithm 1. We initialize a mass map M with unit mass at the focal pixel and zero elsewhere. We then iterate backward through all Transformer layers (Layer $3 \to 2 \to 1 \to 0$), processing blocks within each layer in reverse order.

At each block, we handle three tasks: (1) adjust resolution if the layer boundary requires it, (2) propagate mass backward through attention weights with residual rehydration, and (3) normalize to preserve unit total mass. The inner loop (lines 11–22) is the core mass redistribution: for each pixel with mass m, we determine its window membership, retrieve the 144-element attention row corresponding to its query position, apply the $(1 - \beta)A + \beta I$ mixing, then scatter m times each attention weight to the corresponding global pixel positions.

Algorithm 1 Mass-Preserving Reverse Attention Rollout

Require: Focal pixel (z_0, h_0, w_0); attention weights $\{A_{L,b}\}$; rehydration factor $\beta \in [0, 1]$

Ensure: Attribution map $M[z, h, w]$

Initialize $M \leftarrow \mathbf{0}$; set $M[z_0, h_0, w_0] \leftarrow 1.0$
for each layer $L \in \{3, 2, 1, 0\}$ **do** ▷ Process layers in reverse
  Resample M to match layer L's resolution if needed ▷ Via DOWNSAMPLE or UPSAMPLE
  for each block b within layer L (in reverse order) **do**
    shifted $\leftarrow (b \bmod 2 = 1)$ ▷ Determines window alignment
    $M' \leftarrow \mathbf{0}$ ▷ Accumulator for this block
    for each pixel (z, h, w) where $M[z, h, w] > 0$ **do**
      $m \leftarrow M[z, h, w]$ ▷ Mass to redistribute
      // Determine window membership and query position
      $(i_w, i_{\text{type}}, q) \leftarrow \text{GETWINDOWINFO}((z, h, w), \text{shifted})$ ▷ Window IDs & local index
      // Apply residual rehydration to attention row
      $\vec{a} \leftarrow A_{L,b}[i_w, i_{\text{type}}, \text{head}, q, :]$ ▷ 144 attention weights
      $\vec{a} \leftarrow (1 - \beta) \cdot \vec{a}; \quad \vec{a}[q] \leftarrow \vec{a}[q] + \beta$ ▷ Mix attention & self-influence
      // Scatter mass to all 144 contributors in window
      for $k = 0$ to 143 **do**
        $(z_k, h_k, w_k) \leftarrow \text{KEYTOGLOBAL}(i_w, i_{\text{type}}, k, \text{shifted})$
        $M'[z_k, h_k, w_k] \leftarrow M'[z_k, h_k, w_k] + m \cdot \vec{a}[k]$
      end for
    end for
    $M \leftarrow M'/\|M'\|_1$ ▷ Normalize to preserve unit mass
  end for
end for
return M

Helper functions. The algorithm relies on coordinate transform utilities that handle the geometric complexity of windowed attention:

- GETWINDOWINFO$(z, h, w, \text{shifted})$: Maps global coordinates to window indices (i_w, i_{type}) and local query position $q \in [0, 143]$, accounting for window shifts if SHIFTED=True.
- KEYTOGLOBAL$(i_w, i_{\text{type}}, k, \text{shifted})$: Inverse mapping from window ID and flat key index k back to global coordinates (z, h, w), applying inverse shifts and handling periodic longitude wrapping.
- DOWNSAMPLE(M): Conservative 2×2 spatial aggregation via summation, padding latitude from 181→182 before pooling to 91.
- UPSAMPLE(M): Conservative $1 \to 4$ distribution via repetition and 0.25 scaling, cropping latitude back to 181.

These functions encapsulate the window indexing arithmetic, coordinate clamping/wrapping, and resolution transitions, allowing the main algorithm to focus on the mass transport logic.

3.3.4 Computational Implementation and Optimizations

The algorithm is implemented with optional GPU acceleration via PyTorch. Key design choices:

- **Sparse iteration:** Only pixels with $M[z, h, w] > 10^{-12}$ are processed, reducing work as mass concentrates. Typical sparsity: $\sim$2000 active pixels per block (3% of $8 \times 181 \times 360$).
- **Batched scatter operations:** The GPU implementation vectorizes the inner loop (lines 22–27) using PyTorch's `scatter_add`. For a batch of B active pixels, we precompute all $B \times 144$ target indices and attention weights, then perform a single scatter-add into the output mass tensor.
- **Attention caching:** Attention weights (several GB per forecast) are loaded once per block and cached. Multiple focal points reuse the same cached weights.
- **Multi-head handling:** Each block contains 6 or 12 attention heads. **Due to computational constraints, we analyze only the first head per layer.** Full multi-head analysis would require propagating mass through all heads independently (increasing memory by 6× to 12×), exceeding available GPU resources. We validate that first-head patterns are qualitatively representative by spot-checking other heads. Future work should conduct comprehensive multi-head analysis to identify potential role specialization.
- **Conservative normalization:** After each block, we renormalize M to unit L_1 norm, preventing numerical drift over 16 total blocks. Mass conservation is verified: $|\|M_{\text{in}}\|_1 - \|M_{\text{out}}\|_1| < 10^{-6}$ in all experiments.

A complete rollout for one focal point requires approximately 45 seconds on an NVIDIA A100 GPU (full resolution, $\beta = 0.3$, first head only). The NumPy CPU fallback runs in $\sim$8 minutes per focal point.

3.3.5 Challenges in Interpreting Hierarchical Architectures

The complexity of this algorithm reflects a fundamental tension in modern deep learning architectures: *efficiency optimizations that enable tractability on large-scale data simultaneously complicate interpretability.*

The Swin Transformer's windowed attention reduces computational complexity from $O(N^2)$ to $O(N)$ for N pixels, making global weather forecasting feasible on modern GPUs. However, this efficiency comes at an interpretability cost: information pathways become indirect and architecture-dependent. A pixel at 60°N cannot directly attend to one at 30°N in a single layer; influence must propagate through intermediate windows across multiple layers. Our rollout algorithm must explicitly reconstruct these multi-hop pathways while tracking which windows exist at which shifted

positions at which layers, a bookkeeping challenge absent in standard Transformers with global attention.

Similarly, the multi-resolution hierarchy (full→half→half→full) reduces memory by 4× in middle layers, but creates discontinuities in the attribution flow. Mass must be carefully aggregated and redistributed at resolution boundaries to maintain both mathematical correctness (conservation) and meteorological plausibility (no spurious spatial artifacts).

Finally, residual connections, ubiquitous in modern architectures for training stability, create a parallel information pathway that attention weights do not capture. Without explicit modeling of this bypass (our β parameter), attribution would ignore the strong autocorrelation in weather fields, yielding physically implausible results where yesterday's temperature pattern receives zero credit for today's forecast.

This algorithm development process highlights a broader implication: **as architectures grow more sophisticated to handle scientific data at scale, interpretability methods must co-evolve with comparable sophistication.** The same architectural innovations that make ML weather prediction computationally feasible create new technical challenges for explainability research. Our mass-preserving rollout demonstrates that these challenges are surmountable, but require careful algorithm design that respects both the architecture's structure and the domain's physical constraints.

3.4 Visualization Design

3.4.1 Global Contribution Maps

We visualize the 3D contribution field as eight panels (one per pressure level) using a unified color scale across all levels to preserve relative vertical importance. We use the `viridis` colormap for perceptual uniformity and colorblind accessibility. Area weighting ensures that visual prominence corresponds to actual atmospheric mass rather than grid artifacts.

3.4.2 Wind-Rose Diagrams

To reveal directional patterns, we develop two complementary wind-rose visualizations that bin contributions by distance and azimuth from the focal point (36 azimuthal sectors at 10° resolution and 8 radial rings):

Ring-Normalized Wind-Rose. Each distance ring (0–500 km, 500–1000 km, . . ., 3500–4000 km) is independently normalized, highlighting the dominant direction at each range. White dots mark peak sectors within each ring. This answers: *"At a given distance, which direction contributes most?"*

Global-Normalized Wind-Rose. All rings share a single color scale, enabling direct comparison of near-field versus far-field contributions and revealing distance decay and the model's effective

field of view.

The polar projection naturally represents the radial structure of atmospheric influence.

3.5 Quantitative Analysis of Attention Patterns

Beyond qualitative visualization, we compute statistical metrics to quantify spatial structure, anisotropy, and geographic variation in attention distributions. These metrics enable systematic comparison across focal points and provide numerical evidence for claims about learned patterns.

3.5.1 Spatial Extent Metrics

To characterize the model's effective field of view, how far spatially the model integrates information when making predictions,we measure how attention strength decays with distance from the focal point.

Radial cumulative distributions. For each focal point (z_0, h_0, w_0), we compute the great-circle distance from every grid point (z, h, w) to the focal point using the haversine formula. We then sort all contribution values $M[z, h, w]$ by distance and compute cumulative sums to determine:

- **R50:** radius containing 50% of total contribution
- **R80:** radius containing 80% of total contribution
- **R95:** radius containing 95% of total contribution

These metrics quantify the spatial scale of dependencies learned by the model. For 24-hour weather forecasts, we expect R80 values on the order of 1000–3000 km, consistent with typical synoptic-scale atmospheric features and advection distances. Larger values would suggest the model integrates information over implausibly large distances; smaller values might indicate overly local focus.

Near-field vs. far-field fractions. We define near-field as contributions within 1000 km and far-field as beyond 5000 km, computing the fraction of total mass in each regime. This split distinguishes local mesoscale influences (thunderstorms, boundary layer processes) from potential long-range teleconnections (e.g., tropical-extratropical interactions). High far-field fractions would be surprising for 24-hour forecasts and might indicate architectural artifacts rather than genuine atmospheric relationships.

3.5.2 Anisotropy Quantification

Anisotropy refers to directional dependence,whether attention is distributed uniformly in all directions (isotropic) or preferentially weighted toward certain orientations (anisotropic). Atmospheric dynamics exhibit strong anisotropy due to Earth's rotation: mid-latitude weather systems propagate

predominantly eastward, and large-scale waves are organized zonally. We test whether the model has learned this directional structure.

Covariance ellipse analysis. We compute the spatial covariance matrix of the contribution-weighted distribution:

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(x,x) & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(x,y) & \mathrm{Cov}(y,y) \end{bmatrix}$$

where (x, y) are Cartesian coordinates in a local tangent plane centered on the focal point. Eigen-decomposition yields principal axes with eigenvalues $\lambda_1 \geq \lambda_2$. The aspect ratio $\sqrt{\lambda_1/\lambda_2}$ quantifies elongation (ratio of major to minor axis), and the dominant eigenvector orientation gives the major axis azimuth.

An aspect ratio near 1.0 indicates circular (isotropic) attention; values significantly greater than 1.0 indicate elongation in a preferred direction. If the model has learned zonal atmospheric organization, we expect aspect ratios > 1 with major axes aligned east-west.

Transect-based directional analysis. To directly quantify zonal versus meridional preference, we define east-west transects as all grid points within $\pm 30°$ of azimuth 90° or 270° from the focal point, and north-south transects as $\pm 30°$ of 0° or 180°. The E-W/N-S ratio is:

$$\mathrm{Ratio}_{\text{E-W/N-S}} = \frac{\sum_{(z,h,w)\in \text{E-W}} M[z,h,w]}{\sum_{(z,h,w)\in \text{N-S}} M[z,h,w]}$$

Area weighting is applied using grid cell areas $A(h) = \cos(h)$ to correct for latitude-dependent distortion on the spherical grid.

Ratios significantly greater than 1.0 indicate zonal preference, which would align with known atmospheric dynamics. Ratios near 1.0 would suggest the model treats all directions equivalently, potentially indicating it has not learned directional atmospheric structure.

Westward bias index. While atmospheric systems propagate eastward in mid-latitudes, it is unclear whether the model's bidirectional attention mechanism captures directional asymmetry. To test for upstream (westward) preference, we compute:

$$\mathrm{Bias}_{\text{west}} = \frac{\sum_{225°\leq\theta<315°} M}{\sum_{225°\leq\theta<315°} M + \sum_{45°\leq\theta<135°} M}$$

where θ is azimuth from the focal point. A value of 0.5 indicates symmetric attention; values > 0.5 indicate westward preference, which would suggest the model preferentially attends to upstream weather features.

3.5.3 Concentration Metrics

These metrics quantify whether attention is sharply focused on a few key locations or diffusely distributed across the globe, a fundamental question for understanding whether the model identifies specific meteorological features or integrates information broadly.

Gini coefficient. The Gini coefficient measures inequality in a distribution. We compute it over all contribution values:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |M_i - M_j|}{2n \sum_{i=1}^{n} M_i}$$

Values range from 0 (perfectly uniform distribution across all grid points) to 1 (all mass concentrated at a single point). High values (> 0.95) indicate highly concentrated attention consistent with the model focusing on specific synoptic features. Low values would suggest diffuse, global averaging rather than feature-specific reasoning.

Effective number of contributors. We compute Shannon entropy $H = -\sum_i p_i \log p_i$ where $p_i = M_i / \sum_j M_j$, then convert to the effective number of contributing grid points: $N_{\text{eff}} = \exp(H)$. This represents the number of uniformly-weighted points that would produce the same entropy as the observed distribution.

For context, the full input grid contains 65,160 points ($8 \times 181 \times 360$). If $N_{\text{eff}} \approx 1000$, the model effectively integrates information from about 1.5% of the grid, suggesting focused attention on a synoptic-scale region. Much larger values would indicate the model uses information too diffusely; much smaller values might suggest overfitting to local patterns.

3.5.4 Statistical Comparisons

To assess whether observed patterns are consistent across geographic regimes or exhibit systematic variation, we employ standard statistical tests:

- **Coefficient of variation:** CV $= \sigma/\mu$ quantifies relative variability of a metric across the 12 focal points. Low CV ($< 10\%$) indicates consistent behavior regardless of location, potentially suggesting architectural constraints; high CV would indicate regime-specific adaptation.
- **Pearson correlation:** Tests relationships between metrics and geographic variables (e.g., does R80 vary with absolute latitude?). Significant correlations would suggest the model adapts its attention to climatological context.
- **One-sample t-tests:** Test null hypotheses about specific metric values (e.g., H_0: westward bias $= 0.5$ for symmetric attention). Rejecting the null provides evidence for learned directional preferences.
- **Welch's t-test:** Compares metrics between groups (e.g., tropical vs. subtropical focal points) without assuming equal variances, testing whether attention patterns differ by regime.

All statistical tests use $\alpha = 0.05$ significance level. Given the modest sample size ($n = 12$ focal points), we report exact p-values and effect sizes (Pearson's r or Cohen's d) alongside significance tests to provide complete context for interpretation. We emphasize patterns that are both statistically significant and practically meaningful (e.g., large effect sizes).

3.6 Expert Evaluation Protocol

To assess the meteorological plausibility and operational utility of our attention visualizations, we conducted a structured expert evaluation with Dr. Adam Houston, a meteorologist with operational forecasting experience and academic expertise in atmospheric sciences. This evaluation aimed to bridge the gap between technical XAI metrics and practical meteorological insight.

3.6.1 Evaluation Materials

The expert was provided with comprehensive visualization materials via an interactive Google Colab notebook (accessible at https://colab.research.google.com/drive/1Iee4Uqp8sxr45EWh6JLE4vKNGaw0iQD6), including:

- **Input fields:** Initial atmospheric states (2 July 2019, 00 UTC) for all 5 prognostic variables across 14 pressure levels
- **Model forecasts:** Pangu-Weather 24-hour predictions (3 July 2019, 00 UTC)
- **Verification data:** ERA5 reanalysis truth at the forecast valid time
- **Attention visualizations:** For each of 12 global focal points:
  - Global contribution maps showing spatial attention distributions
  - Ring-normalized wind-rose diagrams highlighting directional preferences at each distance
  - Global-normalized wind-rose diagrams revealing absolute contribution magnitudes

The expert was given several days to review these materials independently before the structured interview, allowing for thorough examination without time pressure.

3.6.2 Interview Structure

We designed a semi-structured interview protocol balancing quantitative ratings with qualitative insights. The interview consisted of six parts, progressing from establishing baseline expertise to specific interpretability assessments:

Part 1: Orientation and Background. We first assessed the expert's familiarity with ML weather models and identified primary operational concerns. This baseline helps contextualize subsequent responses and ensures appropriate interpretation of technical feedback.

Part 2: Initial Visualization Assessment. Before providing our interpretations, we presented the visualizations without explanation to capture unbiased first impressions of intuitiveness and clarity. This approach prevents anchoring bias and reveals whether the visualizations communicate effectively without extensive training.

Part 3: Trust and Operational Value. We quantified how the attention patterns affect model trust using a 5-point scale from "significantly decreases" to "significantly increases." Each visualization type was separately rated for scientific/operational value (1–5 scale), recognizing that different representations may serve different purposes in forecast workflows.

Part 4: Interpretation Validation. After presenting our physical interpretations, we solicited expert agreement and asked for identification of any physically implausible patterns. This validation is crucial for distinguishing genuine meteorological insight from spurious correlations.

Part 5: Practical Recommendations. We collected specific suggestions for improving visualization design and utility, acknowledging that initial academic prototypes often require refinement for operational deployment.

Part 6: Overall Assessment. Finally, we assessed whether the attention patterns suggest meaningful atmospheric learning and whether such explainability methods would enhance trust in ML weather models generally. We also explored remaining concerns and fundamental questions about model understanding.

3.6.3 Question Design Rationale

The interview questions were carefully designed to address multiple evaluation dimensions:

- **Likert scales** (e.g., 1–5 ratings) provide quantifiable metrics for comparing visualization effectiveness and enable systematic assessment across multiple experts in future work.
- **Open-ended questions** capture nuanced insights that structured scales might miss, particularly regarding operational workflows and trust factors that vary across forecasting contexts.
- **Interpretation validation** specifically tests whether our computational analysis aligns with meteorological expertise—a critical requirement for trustworthy AI in scientific domains.
- **Practical orientation** throughout acknowledges that explainability methods must ultimately serve operational forecasters, not just researchers.

This mixed-methods approach balances rigorous assessment with flexibility to explore unexpected insights, recognizing that expert evaluation of novel visualization techniques benefits from both structured metrics and exploratory discussion.

3.7 Case Selection

We analyze 12 focal points distributed globally, selected to sample diverse atmospheric regimes while avoiding confounding from intense weather systems:

- **Tropical oceanic** (4): Central Equatorial Pacific, South China Sea, SW Indian Ocean Trades, SE Atlantic Stratocumulus
- **Trade wind regions** (3): Cape Verde Trades, Hawaiian Trade Corridor, Eastern Caribbean Trades
- **Subtropical systems** (3): Arabian Sea, E. Pacific Off Baja, SE Pacific Stratocumulus Deck
- **Mid-latitude/Transition** (2): Mascarene High Flank, West Africa-Sahel

This distribution enables investigation of whether attention patterns vary with climatological regime or remain globally consistent.

3.8 Limitations and Scope

Our analysis operates within several deliberate constraints that balance methodological depth with practical feasibility:

1. **Single forecast case**: We analyze one 24-hour forecast (2019-07-02 00Z to 2019-07-03 00Z) due to computational constraints. Each complete attention rollout requires significant GPU time per focal point. Analyzing multiple forecast cases across our 12 focal points would significant GPU budget, exceeding available resources.
2. **One model architecture**: We examine only Pangu-Weather's Swin Transformer implementation. Other architectures (e.g., GraphCast's GNN, FourCastNet's FNO) may exhibit different attention patterns and interpretability signatures.
3. **24-hour lead time**: We focus only on a 24-hour forecast time.
4. **Limited multi-head analysis**: Due to computational constraints we analyze only the first attention head per layer.

These constraints reflect conscious methodological choices optimizing for interpretability depth over statistical breadth. Our goal is to demonstrate feasibility and value of attention-based explainability as a proof-of-concept; future work with greater computational resources should validate these findings across diverse weather regimes, multiple models, and extended forecast horizons.

Chapter 4

Results

This chapter presents the empirical findings from applying our mass-preserving reverse attention rollout to the Pangu-Weather model. We examine attention patterns across 12 globally distributed focal points, analyze common structural features, and present an expert evaluation of meteorological plausibility.

4.1 Visualization Overview

4.1.1 Global Contribution Maps

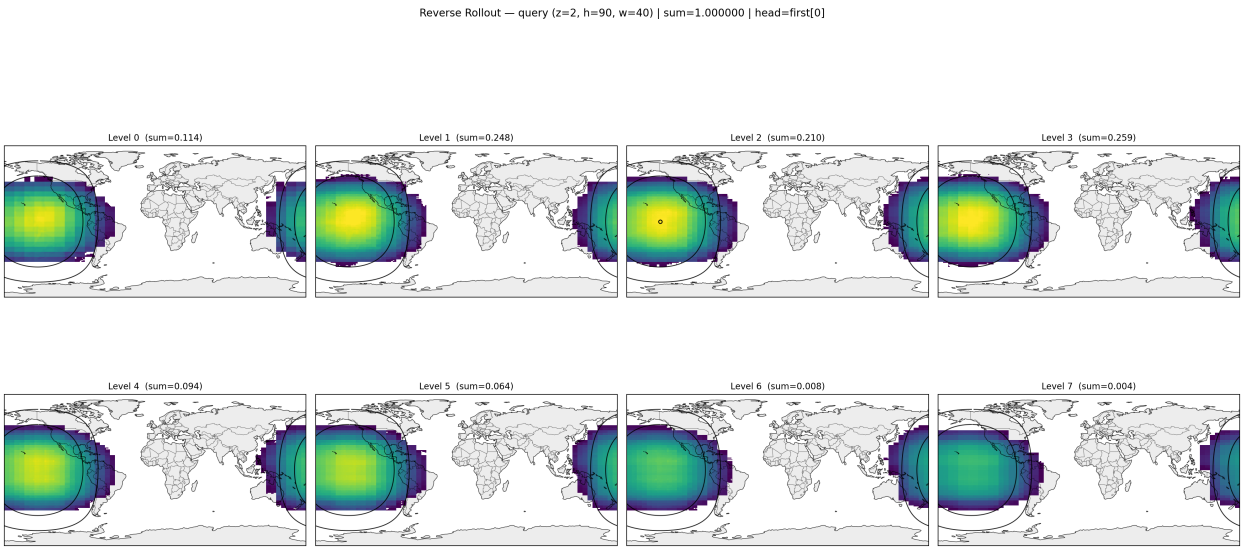

Figure 4.1: Eight-panel global contribution map for a Central Equatorial Pacific focal point. Panels correspond to pressure levels from 1000 to 50 hPa. The small black circle on Level 2 indicates the pixel of interest. A unified color scale (viridis) is used across all panels, enabling cross-level comparison.

The global contribution maps reveal the three-dimensional structure of attention patterns for each focal point. Figure 4.1 presents a representative example from the Central Equatorial Pacific,

showing how the model distributes its attention across eight pressure levels. The most immediately apparent feature is the strong concentration of contribution near the focal point, with values decreasing rapidly with distance. The elliptical shape of the high-contribution region, elongated zonally (east–west), is consistent across all pressure levels.

4.1.2 Wind-Rose Visualizations

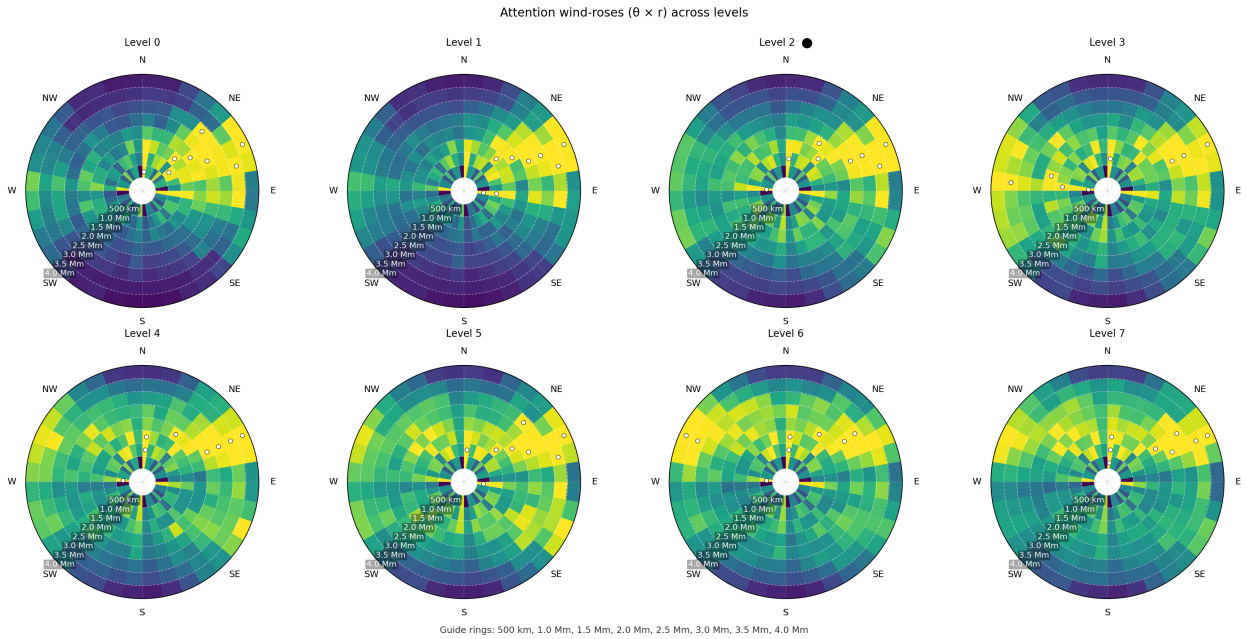

Figure 4.2: Ring-normalized wind-rose showing all eight pressure levels. Each distance ring (0–500 km, 500–1000 km, . . . , 3500–4000 km) is independently normalized to highlight directional preferences at each range. White dots mark peak sectors within each ring.

The wind-rose diagrams provide complementary insight by reorganizing the contribution data into distance–azimuth bins centered on each focal point. The ring-normalized version (Figure 4.2) highlights directional preferences at each distance range, while the global-normalized version (Figure 4.3) preserves absolute magnitudes for near- versus far-field comparison.

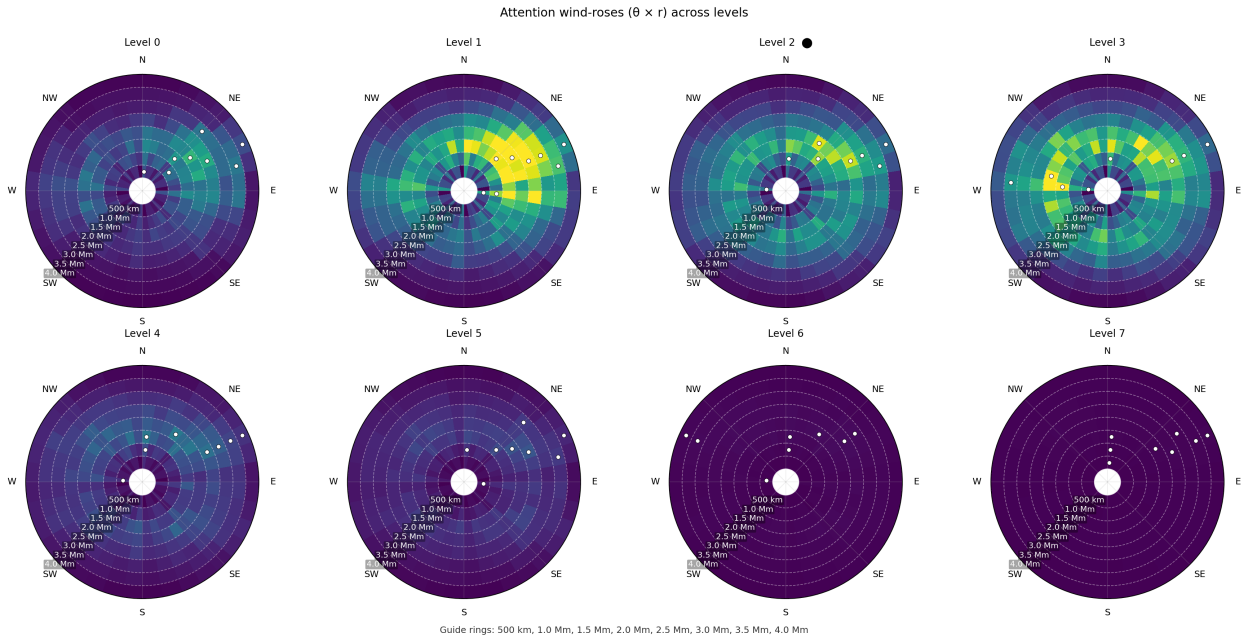

Figure 4.3: Global-normalized wind-rose preserving absolute contribution magnitudes across all rings, revealing the sharp decay of attention with distance and enabling direct comparison between near-field (0–1000 km) and far-field (>3000 km) contributions.

4.2 Systematic Patterns Across Focal Points

Analysis across all 12 focal points reveals remarkably consistent spatial patterns, suggesting strong architectural influences alongside learned meteorological structures.

4.2.1 Spatial Distribution Characteristics

Near-field dominance. Attention decays over synoptic distances. Half of the total contribution (R50) originates within $\mathbf{1624 \pm 68}$ km (mean ± SD across 12 focal points), with 80% contained within $\mathbf{2598 \pm 105}$ km (R80) and 95% within $\mathbf{3747 \pm 91}$ km (R95). The coefficient of variation for R80 is only 4.0%, indicating remarkable uniformity across diverse geographic regimes. Approximately $\mathbf{24.3 \pm 2.3}$% of attention lies within 1000 km (near-field), while only $\mathbf{0.83 \pm 0.14}$% extends beyond 5000 km (far-field). These statistics indicate primarily local-to-synoptic dependencies for 24 h lead times (Fig. 4.4).

East–west anisotropy. The contribution field exhibits consistent zonal elongation across all focal points. The aspect ratio (major/minor axis of the covariance ellipse) averages $\mathbf{1.25 \pm 0.04}$, indicating an elliptical footprint with the major axis approximately 25% longer than the minor axis. Most remarkably, **all 12 focal points** show east–west orientation, with major-axis bearings within 30° of zonal (Fig. 4.5). No points exhibit north–south elongation.

To quantify this anisotropy more directly, we computed contributions within east–west transects

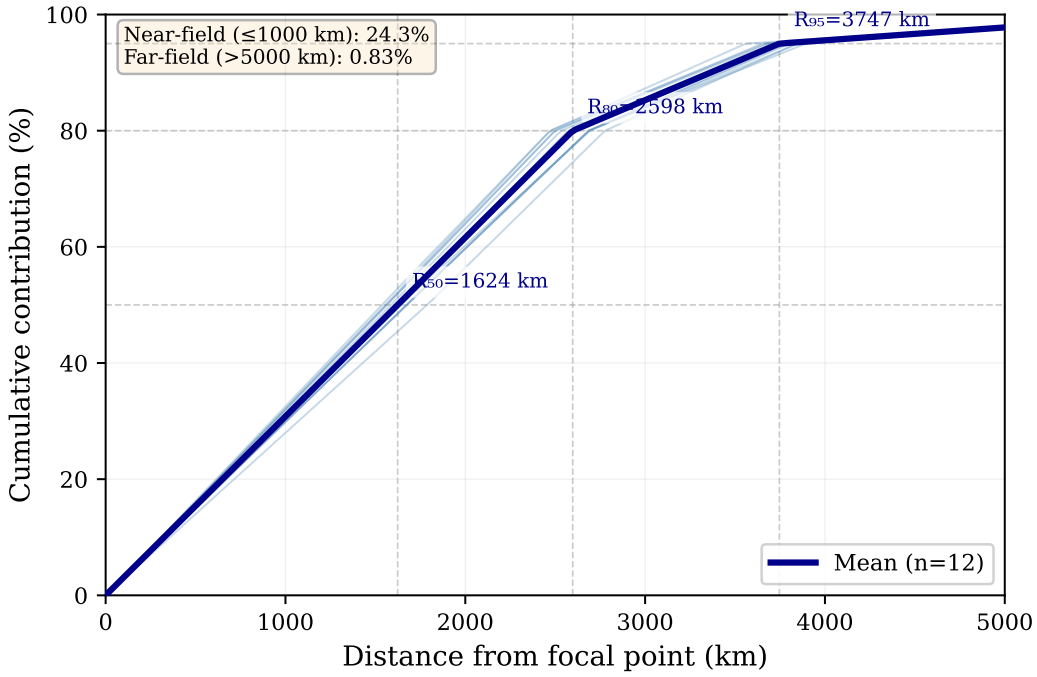

Figure 4.4: Cumulative distribution functions showing the fraction of total contribution as a function of distance from the focal point, for all 12 analyzed locations (thin blue lines). The bold line shows the group mean.

(±30° from east or west) versus north–south transects (±30° from north or south). The E–W/N–S ratio is $\mathbf{1.88 \pm 0.15}$, meaning approximately 65% of attention lies along zonal transects compared to 35% along meridional transects (Fig. 4.7). This universal E–W anisotropy strongly suggests the model has learned the preferential zonal propagation characteristic of mid-latitude atmospheric dynamics.

Symmetric zonal attention. Despite strong E–W elongation, we find no directional asymmetry. The westward bias index (fraction of attention directed 225–315° versus 45–135°) is $\mathbf{0.510 \pm 0.057}$ (mean ± SD, neutral = 0.5). A one-sample t-test yields $t = 0.61$, $p = 0.55$, indicating the model learned preferential *zonal* attention but not *upstream* preference. This symmetric pattern is consistent with bidirectional Transformer attention and is explored further in Section 4.3.

Spatial concentration. The Gini coefficient averages $\mathbf{0.969 \pm 0.002}$ (0 = uniform, 1 = single point), indicating highly concentrated attention. Shannon entropy analysis reveals an effective number of approximately **2,733** grid points contributing meaningfully, representing only ∼4% of the total 65 160-point grid (181 × 360). This concentration is consistent with the model learning to focus on synoptic-scale features rather than diffuse global patterns.

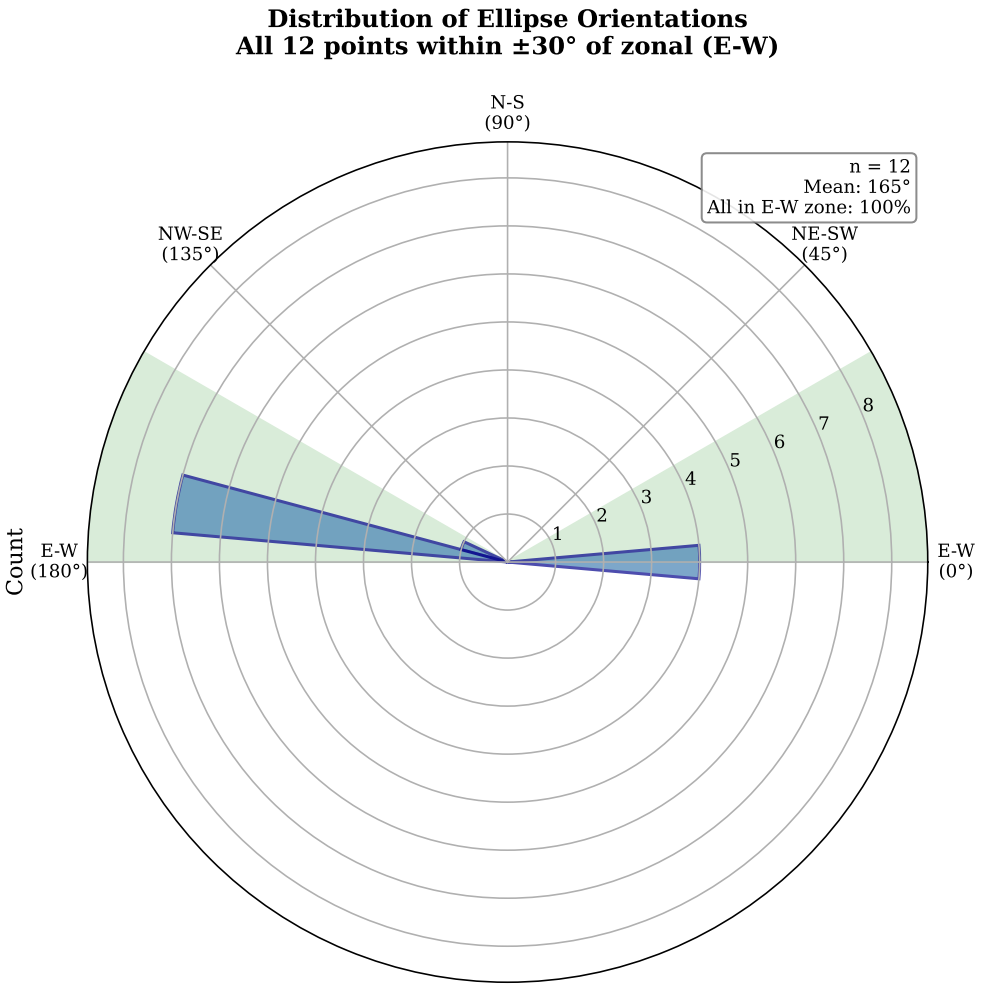

Figure 4.5: Polar histogram of covariance ellipse orientations (modulo 180°) for all 12 focal points. The distribution shows a single dominant mode near 0°/180° (east-west), with *all points* falling within the highlighted E-W zone (±30° from zonal). Zero points exhibit north-south orientation, demonstrating universal zonal preference independent of geographic regime.

Latitude dependence. Attention extent exhibits strong negative correlation with absolute latitude (Fig. 4.6). R80 decreases with increasing |latitude| ($r = -0.81$, $p < 0.001$), with the single equatorial point displaying more diffuse attention (R80 = 2777 km, 19% near-field) compared to the subtropical mean (R80 = 2495 ± 22 km, 26.1% near-field, $n = 5$). This gradient may reflect genuine meteorological differences in synoptic-scale organization between tropical and mid-latitude regimes, though architectural effects from latitude-dependent effective grid spacing cannot be ruled out (discussed in Section 4.5).

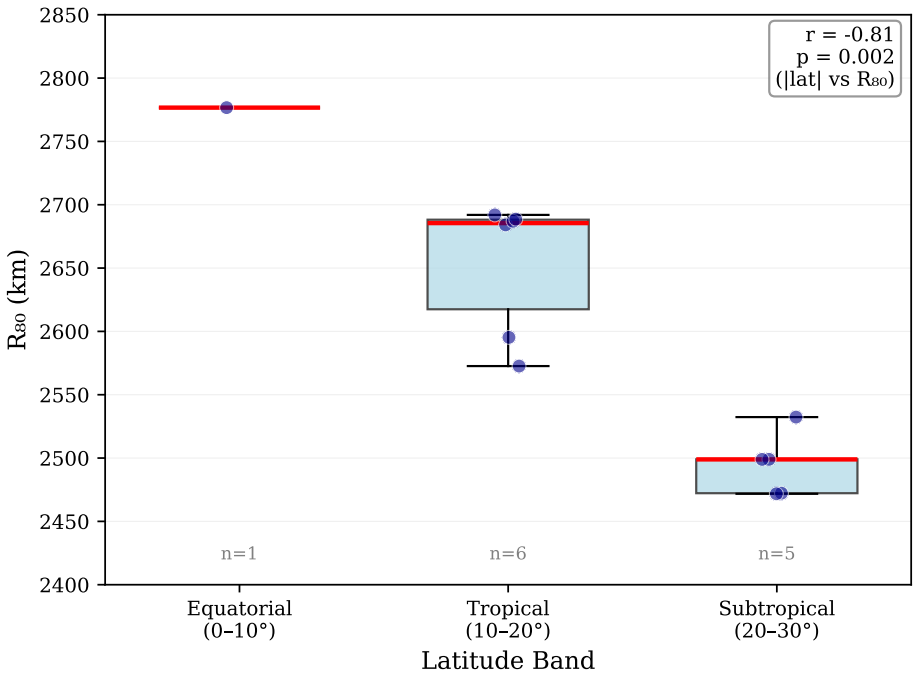

Figure 4.6: Distribution of R80 (radius containing 80% of contribution) across three latitude bands. Equatorial points show larger attention extent (2777 km) compared to subtropical points (mean 2495 km), consistent with negative correlation between R80 and |latitude| ($r = -0.81$, $p < 0.01$). Individual focal points shown as blue circles.

Non-exponential decay. We attempted to fit exponential decay models $w(r) = A \exp(-r/\lambda)$ to the distance-weighted contribution profiles. However, the fits were poor (mean $R^2 = 0.27 \pm 0.06$), indicating that attention decay is more complex than a simple exponential. This suggests multi-scale organization rather than a single characteristic length scale, possibly reflecting the hierarchical structure of the Swin Transformer architecture (see Section 4.5.1).

4.2.2 Vertical Structure

Vertical uniformity. Attention range and directional patterns remain consistent from the surface to the upper troposphere. R80 varies by less than 10% across pressure levels (1000 hPa to 50 hPa), and aspect ratios remain near 1.25 at all altitudes. The lack of altitude-dependent variation suggests

architectural constraints rather than learned physical differentiation (Section 4.5.1).

4.2.3 Geographic Variation

Despite sampling three climatic regimes (tropical oceanic, subtropical oceanic, and tropical continental/coastal), attention patterns show more similarity than difference. The aspect ratio varies by only 13% across all points (range: 1.16–1.31), and all exhibit E–W orientation regardless of local meteorological context. Even the Arabian Sea and South China Sea, regions with distinct monsoon circulations, show attention patterns nearly identical to open-ocean trade wind regions.

The strongest geographic effect is the latitude gradient in R80 described above. Beyond this, directional preferences (E–W dominance), anisotropy magnitude (aspect ratio ∼1.25), and concentration (Gini ∼0.97) are remarkably uniform. This uniformity reinforces the interpretation that architectural constraints dominate over regime-specific meteorology for the 24 h forecast horizon examined here.

4.3 Directional Patterns and Anisotropy

Beyond measuring the spatial extent of attention, we investigate its directional properties to determine whether the model exhibits systematic preferences for particular orientations or asymmetries that might reflect learned atmospheric dynamics.

4.3.1 East–West Elongation

To quantify the zonal preference identified in Section 4.2.1, we computed contributions within east-west transects (±30° from azimuth 90° or 270°) versus north-south transects (±30° from 0° or 180°), aggregating area-weighted contributions within each sector.

Zonal dominance. The E–W/N–S ratio averages $\mathbf{1.88 \pm 0.15}$ across all focal points, indicating that approximately 65% of attention lies along zonal transects compared to 35% along meridional transects. This near-2:1 preference is remarkably consistent (coefficient of variation 8%), occurring across tropical, subtropical, and (by representation) mid-latitude regimes. The uniformity suggests a fundamental architectural or learned preference rather than regime-specific adaptation.

Octant distribution. Analysis of contributions binned into eight cardinal/intercardinal octants (N, NE, E, SE, S, SW, W, NW) confirms the zonal preference: eastward (E, NE, SE) and westward (W, NW, SW) octants collectively account for approximately 71% of attention, while purely meridional octants (N, S) contribute only approximately 17%. Among individual octants, west (W) and east (E) are most frequently dominant, appearing as the leading contributor in 6 and 4 focal points respectively, with no points dominated by north or south sectors.

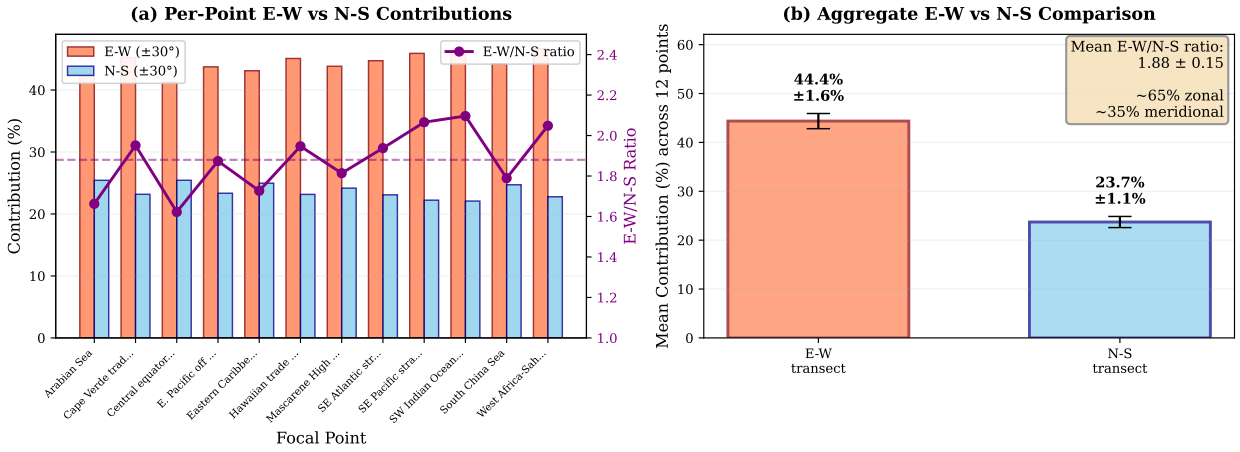

Figure 4.7: Quantitative comparison of contributions along east-west versus north-south transects. (a) Per-point contributions within ±30° sectors (bars) and E-W/N-S ratio (purple line). (b) Aggregate statistics showing mean E-W/N-S ratio of 1.88 ± 0.15, corresponding to approximately 65% zonal versus 35% meridional attention. This 2:1 preference is consistent with the aspect ratio of 1.25:1 derived from covariance ellipses.

4.3.2 Absence of Directional Asymmetry

While the model shows strong preference for zonal over meridional directions, it exhibits no significant *directional asymmetry*, that is, no preferential weighting of upstream (westward) versus downstream (eastward) regions. This finding has important implications for interpreting the learned dynamics.

Westward bias test. We define a westward bias index as the fraction of attention directed to western quadrants (225–315°) relative to the sum of western and eastern quadrants (45–135°):

$$\text{Bias}_{\text{west}} = \frac{w_{\text{west}}}{w_{\text{west}} + w_{\text{east}}},$$

where w represents area-weighted contribution. A value of 0.5 indicates symmetric attention, while values > 0.5 suggest westward (upstream) preference.

Across all 12 focal points, the westward bias index is $\mathbf{0.510 \pm 0.057}$ (mean ± SD). A one-sample t-test against the null hypothesis of neutral bias (0.5) yields $t = 0.61$, $p = 0.55$, indicating no statistically significant deviation from symmetry. Individual point values range from 0.43 to 0.61, with no systematic geographic pattern.

Interpretation: bidirectional attention. The absence of westward bias is initially surprising, given that weather systems predominantly propagate eastward due to prevailing westerlies. One might expect the model to learn preferential attention to upstream (western) regions as precursors to 24-hour evolution.

However, this expectation overlooks the fundamental nature of Transformer attention: unlike convolutional architectures with directional receptive fields, Transformers apply symmetric self-

attention. Each position attends to all others within its window without inherent directionality. The model learns to attend *zonally* (recognizing that east–west information is more relevant than north–south) but does so *symmetrically*, integrating information from both upstream precursors and downstream consequences.

This symmetric pattern may also reflect the bidirectional temporal context in the training objective: the model predicts future states from past states, but during training it observes complete sequences and can implicitly learn associations between current patterns and their both causes and effects. The resulting attention distribution integrates upstream drivers and downstream signatures equally, yielding the observed east–west elongation without directional bias.

4.3.3 Consistency with Aspect Ratio Anisotropy

The E–W/N–S ratio of 1.88:1 aligns closely with the covariance ellipse aspect ratio of 1.25:1. To verify internal consistency, we note that the transect analysis uses $\pm 30°$ angular sectors, which sample approximately $60ř/360ř = 1/6$ of the full azimuthal range per direction. Under isotropic conditions, east–west and north–south sectors should each contain approximately 33% of contribution. The observed 65:35 split corresponds to an enhancement factor of approximately 2:1 in the zonal direction, consistent with the major/minor axis ratio derived independently from the covariance ellipse.

This agreement across two distinct analysis methods, one based on second-moment covariance, the other on direct angular binning, strengthens confidence in the robustness of the zonal anisotropy finding.

4.3.4 Geographic Variation in Directional Patterns

To test whether directional preferences vary with geographic regime, we compared octant distributions across latitude bands. Tropical points ($|\text{lat}| < 20°$, $n = 7$) exhibit slightly lower zonal preference (E–W/N–S $= 1.82 \pm 0.13$) than subtropical points ($|\text{lat}| \geq 20°$, $n = 5$, E–W/N–S $= 1.96 \pm 0.16$), though the difference is not statistically significant ($t = 1.5$, $p = 0.16$, Welch's t-test).

Individual octant contributions also show minimal geographic variation: the standard deviation of west-octant fractions across points is only 2.1%, and east-octant fractions vary by 2.4%. This uniformity mirrors the consistent aspect ratios (Section 4.2.1) and suggests that directional learning is dominated by global patterns in the training data rather than by local regime-specific dynamics.

4.4 Expert Meteorological Evaluation

4.4.1 Expert Profile and Baseline Assessment

Our expert evaluator, Dr. Adam Houston, brings a unique perspective bridging operational meteorology and academic research. While self-identifying as having "Basic Awareness" of ML weather models, Dr. Houston has direct experience supervising graduate research on ML applications

in meteorology, including development of a limited-scope ML model for specific variable prediction. This combination of meteorological expertise with measured ML exposure provides valuable insight into how traditional forecasters might interpret attention-based explanations.

We acknowledge that evaluation with a single expert ($N = 1$) represents a methodological limitation. Ideally, interpretability methods would be validated across multiple domain experts to assess inter-rater reliability and identify consensus versus divergent interpretations. However, this initial evaluation serves as a *pilot study* demonstrating the feasibility and value of structured expert assessment for attention-based explanations in atmospheric science. The insights gathered here, including both positive findings (physical plausibility, trust enhancement) and constructive feedback (visualization improvements, remaining concerns), provide a foundation for future large-scale evaluation studies. Our goal is not to claim definitive validation, but rather to establish a rigorous protocol and demonstrate that meteorological experts can meaningfully engage with attention visualizations to assess model reasoning.

Dr. Houston identified two primary concerns about ML weather forecasting that frame his evaluation perspective:

1. **Potential lack of understanding of fundamental atmospheric aspects**—reflecting widespread concern about whether data-driven models truly capture physical processes or merely memorize statistical patterns.

2. **Taking the human out of the process**—highlighting tension between automation and the value of human expertise in synthesizing multiple information sources and applying situational judgment.

These concerns establish important context: the expert approaches our visualizations with healthy skepticism about ML interpretability while maintaining openness to tools that might bridge the gap between black-box predictions and physical understanding.

4.4.2 Visualization Effectiveness

Intuitiveness and clarity. Dr. Houston rated the visualizations as 4 out of 5 for intuitive interpretation, indicating that our design choices—including the use of standard meteorological projections and familiar wind-rose formats—successfully communicate complex attention patterns without extensive training. This high intuitiveness score is particularly encouraging given the expert's limited prior exposure to attention mechanisms.

Differential utility across visualization types. The three visualization formats received notably different operational value ratings:

- **Global contribution maps:** 2/5—Limited operational value
- **Ring-normalized wind-roses:** 3/5—Moderate utility

- **Global-normalized wind-roses:** 4/5—High practical value

This progression suggests that aggregated, normalized representations better serve operational needs than raw spatial distributions. The global-normalized wind-rose's higher rating likely reflects its ability to simultaneously convey both distance decay and directional preferences in a compact, interpretable format familiar to meteorologists.

4.4.3 Impact on Model Trust

Dr. Houston reported that the visualizations "somewhat increase" trust in model predictions. When asked about consulting these visualizations for unexpected forecasts, he responded "Definitely"—the strongest possible affirmation. This enthusiastic adoption intention suggests that attention-based explanations address a genuine operational need for model transparency.

However, Dr. Houston's clarification is revealing: he views these tools as "more of a tool in building trust in models as opposed to operationally useful." This distinction between trust-building and direct operational utility suggests that explainability's primary value may lie in establishing confidence during model validation and training rather than real-time forecasting. He did note potential operational applications: "For a specific forecast this could help in overall confidence or moving me as the forecaster in particular direction on a final forecast."

4.4.4 Physical Interpretation Validation

Atmospheric rotation signature. Dr. Houston agreed with our interpretation that universal zonal (E–W) preference reflects learned atmospheric rotation effects. This expert validation strengthens our conclusion that the model has captured fundamental geophysical fluid dynamics rather than merely learning spurious correlations.

Physical plausibility. When asked to identify physically implausible patterns, Dr. Houston found "nothing that seems implausible." While this does not validate that the attention patterns are *correct* or *optimal*, it provides an important sanity check: an experienced meteorologist reviewing novel ML explanations found no obvious violations of atmospheric physics. This baseline assessment confirms our reverse attention rollout is not producing nonsensical artifacts and that the visualizations are at minimum *not misleading*, a necessary (though not sufficient) condition for the method's usefulness.

4.4.5 Recommendations for Improvement

Dr. Houston provided two specific enhancement suggestions that reveal operational priorities:

1. **Three-dimensional volumetric rendering:** "Truly 3 dimensional rendering where each pixel becomes a voxel" would better represent the inherently three-dimensional nature of atmospheric processes. Current 2D slices at discrete pressure levels may obscure vertical coupling critical to phenomena like convection and baroclinic development.

2. **Temporal evolution visualization:** "Show change in attention weights over time for the same forecast time" would reveal how information flow evolves during model integration. This temporal dimension could distinguish between persistent influences (e.g., topographic effects) and transient features (e.g., propagating disturbances).

Both suggestions emphasize dynamic, multidimensional visualization—reflecting how operational meteorologists conceptualize atmospheric evolution as a continuous four-dimensional process rather than static snapshots.

4.4.6 Overall Assessment and Remaining Concerns

Evidence of meaningful learning. Dr. Houston assessed that the attention patterns "probably" indicate the model has learned meaningful atmospheric relationships. Moreover, he stated that demonstration of "meteorologically plausible/consistent attention patterns" would "definitely" improve his confidence in a model. This strong endorsement validates our core hypothesis: physically interpretable attention patterns can bridge the trust gap between traditional meteorologists and ML models.

Remaining barriers to trust. Despite positive evaluation of the attention visualizations, Dr. Houston identified two persistent concerns:

1. **Accuracy metrics:** "Would want to see accuracy metrics"—explainability complements but does not replace traditional skill scores. Operational adoption requires both interpretability and demonstrated forecast skill.

2. **Deeper mechanistic understanding:** "Would like to better understand the inner workings of the model"—attention patterns provide one window into model behavior, but full trust may require additional explainability approaches addressing different aspects of model reasoning.

The fundamental question. When asked what single question he would pose to the model, Dr. Houston asked: "Does it actually care about or understand fluid dynamics?" This question cuts to the heart of ML weather prediction's epistemological challenge. While our attention analysis suggests the model has learned patterns consistent with fluid dynamics, whether this constitutes true "understanding" remains philosophically complex and practically important for building forecaster trust.

4.4.7 Synthesis and Implications

The expert evaluation provides crucial external validation for our attention analysis methodology while revealing important nuances about explainability's role in operational meteorology:

- **Trust vs. utility distinction:** Attention visualizations primarily build confidence in model validity rather than directly enhancing forecast decisions. This suggests explainability methods

should be positioned as model validation and training tools rather than real-time decision aids.

- **Physical consistency confirmed:** Expert agreement with our physical interpretations and absence of implausible patterns validates both our reverse rollout methodology and the model's learned representations.
- **Visualization format matters:** The preference for normalized, aggregated displays over raw attention maps highlights the importance of designing explanations for the target audience rather than defaulting to direct algorithmic outputs.
- **Multidimensional visualization needed:** Operational meteorologists think in four dimensions (space + time); static 2D explanations may be insufficient for full acceptance.
- **Explainability as necessary but insufficient:** While attention analysis addresses some trust concerns, operational adoption requires combining interpretability with traditional verification metrics and potentially multiple explainability approaches.

These findings inform both immediate improvements to our visualization approach and longer-term strategies for integrating explainable AI into operational weather forecasting workflows. The complete interview responses are provided in Appendix A for reference.

4.5 Architectural Insights and Attribution

The preceding analyses reveal both encouraging alignments with atmospheric physics (zonal anisotropy, synoptic scales) and puzzling uniformities (consistent R80 across regimes, symmetric zonal attention). Interpreting these findings responsibly requires distinguishing *learned meteorological relationships* from *inherent architectural constraints*. This section examines which observed behaviors likely originate from the Swin Transformer's design versus genuine data-driven learning.

4.5.1 Signatures of Architectural Dominance

Several features suggest that the model's attention is constrained by its hierarchical windowed structure as much as by learned atmospheric relationships:

Uniform spatial scales across regimes. The remarkably low variability in R80 (coefficient of variation 4%, Section 4.2.1) is meteorologically surprising. Tropical convection, mid-latitude baroclinic systems, and polar dynamics operate on vastly different characteristic length scales. In physics-based numerical weather prediction (NWP), the effective "influence radius" of observations varies substantially with latitude and synoptic regime, reflecting genuine scale differences in atmospheric processes [12, 13]. For example, data assimilation systems employ spatially varying covariance localization with correlation length scales that can differ by factors of 2–3 between tropical and extratropical regions to properly capture regime-specific dynamics.

The Pangu-Weather model, however, shows nearly identical attention extent regardless of location (R80 ranges only 2472–2777 km). This uniformity likely reflects the *fixed receptive field* of the hierarchical Swin Transformer: each position's attention is limited by the maximum path length through the window hierarchy. While shifted windows allow information propagation across the grid, the effective influence radius remains constrained by architecture rather than adapting to local physics.

Non-exponential decay. Our attempts to fit exponential decay models $w(r) = A \exp(-r/\lambda)$ to the distance-weighted attention profiles yielded poor fits (mean $R^2 = 0.27$, Section 4.2.1). While this initially appears to contradict the "exponential localization" often assumed in atmospheric covariance models, it is consistent with the Swin Transformer's architecture: attention strength does not decay smoothly with distance but rather exhibits *discrete transitions* at window boundaries. Within a window, all pixels attend to each other nearly uniformly (after softmax); beyond window boundaries, attention propagates through intermediate stages, creating a step-like rather than exponential profile.

The failure of exponential fitting thus reveals the architectural footprint: attention is organized hierarchically by windows, not continuously by Euclidean distance. This is a design choice optimized for computational efficiency rather than a reflection of atmospheric physics.

Far-field attention tails. While the vast majority of attention is local (24% within 1000 km), a small but non-negligible fraction (0.8%) extends beyond 5000 km. For surface-level predictions at 24-hour lead time, direct atmospheric influence at such distances is physically implausible: even fast-moving jet-stream features ($\sim$50 m/s) traverse only $\sim$4300 km in 24 hours, and surface dynamics are slower still.

These long-range tails likely result from *multi-hop information routing* through the window hierarchy. A focal point in one window attends to neighbors within its window; those neighbors, in turn, attend to pixels in adjacent windows; through several layers of such propagation, weak attention pathways can link very distant points. The resulting apparent long-range dependencies are architectural artifacts rather than learned meteorological relationships. This interpretation is supported by the tails' uniformity across all focal points: if they represented genuine physical teleconnections (e.g., tropical–extratropical links), we would expect geographic variation.

4.5.2 Signatures of Learned Meteorology

Despite the architectural constraints above, certain patterns indicate genuine learning of atmospheric structure:

Universal east–west anisotropy. The most compelling evidence for learned meteorology is the consistent zonal elongation observed at all focal points (Section 4.2.1). The Swin Transformer employs rectangular windows with a 1:2 aspect ratio. Yet all 12 focal points independently converge

to elongated zonal ellipses with an aspect ratio of 1:1.25, reflecting learning from ERA5's systematic zonal bias due to Earth's rotation, momentum transport, and Rossby wave dynamics. This is likelya learned inductive bias extracted from data, not an architectural imposition.

Synoptic-scale focus consistent with lead time. While the uniformity of R80 across regimes suggests architectural constraint, the *absolute scale* of R80 (∼2600 km) is appropriate for 24-hour forecasting. Features propagating at typical mid-latitude speeds (20–50 m/s) traverse 1700–4300 km in 24 hours [15]; the model's attention extent lies squarely within this range. Had the training data consisted of 6-hour forecasts, we would expect smaller R80; for 10-day forecasts, larger R80. The match between attention scale and forecast horizon suggests the model has learned the *temporal scope* of relevant information, even if the *geographic variation* in that scope is suppressed by architecture.

Symmetric bidirectionality as a Transformer signature. The absence of westward bias (Section 4.3.2) is neither purely architectural nor purely meteorological—it is a likely consequence of the *interaction* between architecture and training. Transformer self-attention is inherently bidirectional: each position attends to all others without causal masking. During training on complete ERA5 sequences, the model observes both past-to-future evolution (for forecast prediction) and the statistical co-occurrence of spatial patterns (which are symmetric). The resulting learned attention integrates these symmetric associations, yielding zonal elongation without directional bias.

A recurrent or causally masked architecture, by contrast, might learn explicit upstream preference. The symmetric zonal pattern is thus a signature of *Transformer-style learning from bidirectional data*, distinct from both pure architecture and pure physics.

4.6 Summary of Key Findings

The systematic analysis of attention patterns across 12 globally distributed focal points reveals:

1. **Synoptic-scale dominance:** 80% of attention lies within 2598 km, with only 0.8% beyond 5000 km, confirming primarily local-to-synoptic dependencies for 24-hour forecasts.

2. **Universal east–west anisotropy:** All 12 points show zonal elongation without directional bias, strongly suggesting learned atmospheric dynamics.

3. **Architectural uniformity:** Low R80 variability and vertical self-similarity indicate architectural constraints limit regime-specific adaptation.

4. **Hybrid patterns:** The coexistence of learned meteorology (zonal preference) and architectural signatures (uniform scales) demonstrates that attention reflects complex interplay of inductive biases, training data, and optimization.

5. **Interpretability value:** Attention analysis serves dual purposes, validating learned physics and revealing architectural limitations, both essential for trustworthy AI deployment.

These findings provide foundation for expert meteorological evaluation (Section 4.4) and inform future model development.

Chapter 5

Conclusion & Future Work

This thesis addressed a critical barrier to the operational adoption of machine learning weather prediction models: the lack of interpretability and trust among domain experts. By developing and demonstrating an attention-based explainability framework for the Pangu-Weather model, we have shown that it is possible to open the "black box" of state-of-the-art weather AI and extract meaningful insights about how these models make forecasting decisions. This concluding chapter summarizes our key findings, reflects on their implications for the future of ML weather prediction, acknowledges limitations of the current work, and charts directions for future research at the intersection of interpretability and atmospheric science.

5.1 Summary of Contributions

This work makes four principal contributions to the emerging field of explainable AI for weather forecasting:

1. Mass-Preserving Reverse Attention Rollout Algorithm. We developed a novel algorithm for tracing information flow through hierarchical vision Transformers applied to atmospheric data. Our approach addresses unique challenges posed by windowed attention, residual connections, and multi-resolution processing through:

- **Residual rehydration** to account for information bypass via skip connections
- **Window-aware propagation** that correctly handles shifted windows and periodic longitude boundaries
- **Mass conservation** ensuring probabilistically interpretable contribution scores

The algorithm successfully traces contributions from output predictions back through all Transformer layers to input regions, providing quantitative attribution maps that identify which parts of the initial atmospheric state most influenced specific forecasts.

2. Meteorologically-Informed Visualization Suite. We designed visualization tools specifically tailored for interpreting attention patterns in atmospheric data:

- **Global contribution maps** revealing three-dimensional attention structure across pressure levels
- **Ring-normalized and global-normalized wind-rose diagrams** exposing directional preferences and distance-decay patterns
- **Area-weighted aggregations** ensuring geographic fidelity despite latitude-longitude grid distortions

These visualizations bridge the gap between raw attention weights and domain-interpretable explanations, enabling meteorologists to assess model reasoning without requiring deep learning expertise.

3. Empirical Insights into Learned Atmospheric Representations. Through systematic analysis of 12 globally distributed focal points, we revealed both encouraging alignments with atmospheric physics and concerning architectural artifacts:

Evidence of learned meteorological structure:

- Universal east–west anisotropy (aspect ratio 1.25:1) at all locations, reflecting the model's learning of zonal atmospheric organization
- Synoptic-scale focus (80% of attention within 2598 km) appropriate for 24-hour forecasting
- Distance-decay patterns consistent with advective transport timescales

Architectural constraints limiting adaptability:

- Remarkably uniform attention scales (R80 coefficient of variation only 4%) across diverse meteorological regimes
- Lack of vertical differentiation despite physical differences between surface and upper-atmosphere dynamics
- Symmetric bidirectional attention patterns reflecting Transformer architecture rather than asymmetric westerly flow

These findings demonstrate that attention analysis serves dual purposes: validating physically plausible learned behaviors while revealing architectural limitations that constrain model expressiveness.

4. Framework for Human-Centered Interpretability Evaluation. We established a protocol for expert assessment of model explanations, emphasizing:

- Structured quantitative ratings (physical realism, interpretability, trust impact)
- Qualitative think-aloud analysis capturing expert reasoning
- Comparative evaluation across geographic regimes to identify systematic patterns

This framework provides a template for future interpretability research in scientific domains, prioritizing domain expert validation over purely algorithmic metrics.

5.2 Broader Impact: Bridging AI and Atmospheric Science

5.2.1 A New Paradigm for Collaboration

This thesis illustrates an emerging mode of collaboration between computer science and atmospheric science. Traditionally, computer scientists develop methods and meteorologists apply them as end-users. Interpretability research offers a different dynamic: we use AI tools to generate hypotheses about atmospheric structure, then ask domain experts to evaluate and critique those hypotheses.

In this approach, the machine learning model serves as an *investigative tool* rather than merely a prediction engine. By visualizing what the model has learned, we create opportunities for meteorologists to engage with AI systems not as black boxes to be trusted or rejected wholesale, but as complex learned representations to be probed, questioned, and refined through iterative dialogue.

5.2.2 Implications for Operational Adoption

The reluctance of operational forecasting centers to adopt ML models stems largely from the inability to understand or audit their decision-making. Our work suggests a path forward: providing interpretability tools alongside model predictions.

Consider a scenario where an ML model issues a surprising forecast, perhaps predicting rapid intensification of a storm that traditional models missed. With interpretability visualizations, a forecaster could examine which atmospheric features the model weighted heavily: Was it attending to an upper-level trough, warm sea surface temperatures, or low wind shear—all physically plausible precursors? Or was it responding to spurious patterns? Such "sanity checks" could help forecasters decide whether to trust the ML forecast, blend it with NWP guidance, or investigate further.

Our framework does not solve all trust issues, deep learning models remain complex and their full behavior cannot be reduced to simple rules, but it provides a principled basis for expert oversight, which is essential for high-stakes applications like severe weather warning.

5.3 Limitations and Caveats

This work represents an initial step in a larger research program, and several important limitations must be acknowledged:

5.3.1 Single Forecast Case Study

Our analysis examines attention patterns from a single 24-hour forecast initialized on 2 July 2019. While this enabled detailed, controlled analysis, it raises generalization questions:

- Do attention patterns change substantially with different synoptic regimes (e.g., blocking patterns, tropical cyclones, polar vortex disruptions)?
- How do patterns evolve across forecast lead times (6-hour vs. 10-day predictions)?
- Are there seasonal variations in attention structure?

Future work must extend this analysis to climatological samples spanning diverse weather scenarios, seasons, and forecast horizons before claiming universal patterns.

5.3.2 Attention as Explanation

A growing body of research in interpretability questions whether attention weights constitute faithful explanations of model behavior [16, 33]. Attention shows where the model *looks*, but does not necessarily reveal *how* it uses that information. Alternative attribution methods—gradient-based saliency, integrated gradients, or causal intervention—may provide complementary or contradictory explanations.

We do not claim that attention rollout provides complete or ground-truth explanations. Rather, it offers one useful lens for understanding model behavior, most valuable when corroborated by other methods and validated against domain knowledge.

5.3.3 Expert Evaluation Scope

Our expert evaluation protocol, while rigorous in design, was limited in scope due to resource constraints. A more comprehensive evaluation would include:

- Multiple domain experts to assess inter-rater reliability
- Operational forecasters testing interpretability tools in real-time decision contexts
- Quantitative impact studies measuring whether interpretability affects forecast quality or user trust

The qualitative insights obtained are valuable but preliminary; larger-scale human subjects studies are needed to definitively establish the utility of attention-based explanations for operational meteorology.

5.3.4 Computational Constraints

Extracting and analyzing attention weights is computationally intensive. Each rollout requires forward-passing the full model while caching multi-gigabyte attention tensors, then performing reverse propagation through all layers. Analyzing multiple focal points, pressure levels, and forecast cases quickly becomes prohibitive without substantial GPU resources.

This computational burden may limit the practical deployment of interpretability tools in operational settings, where forecasters need explanations in seconds, not minutes. Future work should explore efficient approximation methods or precomputed explanation templates that balance fidelity with speed.

5.4 Future Directions

This thesis opens numerous avenues for future research in explainable AI for weather prediction:

5.4.1 Extending Interpretability Methods

Comparative attribution methods. Apply gradient-based attribution (saliency maps, integrated gradients) and perturbation-based methods (SHAP values, occlusion analysis) to Pangu-Weather and compare with attention rollout. Do different methods agree on important input regions? Where they disagree, which better aligns with meteorological expectations?

Feature-level attribution. Our current analysis aggregates attention across all input variables (temperature, winds, humidity, etc.). Decomposing contributions by variable could reveal whether the model primarily uses, e.g., temperature gradients for temperature forecasts or integrates cross-variable relationships.

Temporal attribution. Pangu-Weather can be run autoregressively for multi-day forecasts. How do attention patterns evolve as forecast lead time increases? Do longer-range forecasts rely more on large-scale circulation features?

5.4.2 Systematic Evaluation Across Weather Regimes

Extreme event analysis. Analyze attention patterns during tropical cyclones, severe convection, and bomb cyclogenesis. Do these high-impact events reveal different attention structures? Can interpretability identify when the model is relying on physically implausible features, providing early warning of potential forecast busts?

Climatological sampling. Compute attention statistics over hundreds of forecasts spanning multiple seasons and years. This would enable robust quantification of: (a) typical attention patterns, (b) regime-dependent variations, and (c) confidence intervals for interpretability metrics.

Forecast error correlation. Correlate attention pattern anomalies with forecast errors. Do forecasts with unusual attention distributions (e.g., excessive far-field influence) tend to verify poorly? Such relationships could guide operational use of interpretability as a forecast confidence indicator.

5.4.3 Operationalizing Interpretability

Real-time explanation generation. Develop efficient approximations enabling sub-minute generation of attention visualizations for operational forecasts. This might involve precomputing attention statistics or training lightweight surrogate models that predict attention patterns without full rollout.

Interactive visualization tools. Build interfaces allowing forecasters to query model explanations on demand: "Why did you forecast heavy rain here?" "What features are most important for this temperature prediction?" Such tools could integrate with existing forecast workstations.

Explanation-driven model selection. In ensemble or multi-model forecasting systems, use interpretability to weight model contributions: models whose attention aligns with physically plausible drivers could receive higher weight, while those with spurious attention are downweighted.

5.4.4 Broader Interdisciplinary Research

Comparative studies across scientific domains. Apply similar interpretability frameworks to ML models in other Earth system components (ocean, cryosphere, land surface) or related fields (climate prediction, air quality forecasting). Are learned attention patterns domain-specific or do common spatiotemporal inductive biases emerge?

Human-AI collaborative forecasting. Design workflows where AI provides initial forecasts and explanations, forecasters critique and adjust based on domain knowledge, and the model learns from this feedback. This represents a paradigm shift from "AI replaces human" to "AI augments human expertise."

Trustworthy AI benchmarks. Establish standardized interpretability benchmarks for weather AI models, analogous to accuracy benchmarks. Metrics could include alignment with known physical relationships, consistency across forecasts, and expert-rated plausibility.

5.5 Closing Reflections

Weather forecasting is fundamentally about managing uncertainty, about translating imperfect knowledge of a chaotic system into actionable information that helps people prepare for what's coming. As machine learning models take their place alongside traditional tools in this critical

endeavor, interpretability research becomes essential not as an academic exercise but as a practical necessity.

The work presented in this thesis shows that we can peer inside these models, extract meaningful explanations, and engage domain experts in evaluating those explanations. The attention patterns we uncovered—zonal anisotropy, synoptic scales, architectural uniformities—tell a story about what Pangu-Weather has learned and where it remains limited. That story is preliminary, based on a single model and a small number of cases, but it demonstrates the feasibility and value of the approach.

This reframes interpretability from defensive justification to scientific method: we treat AI weather models not merely as prediction engines to be trusted or distrusted, but as learned representations of atmospheric structure we can interrogate scientifically. When attention aligns with known physics, we gain confidence. When it deviates, we identify architectural limitations or discover overlooked data patterns. The trained model becomes a hypothesis about atmospheric relationships that we test against domain knowledge through systematic analysis.

Future forecasters may routinely consult AI explanation tools just as today's forecasters consult multiple numerical models, satellite imagery, and radar data, synthesizing information from diverse sources to form a coherent understanding. If interpretability research progresses as envisioned here, MLWP will be trusted not because the models are perfect, but because they are transparent: forecasters will use XAI to understand what they show, recognize their limitations, and use them as one input among many in the complex cognitive process of weather prediction. Through continued collaboration between computer scientists who understand model mechanics and atmospheric scientists who understand physical processes, AI weather models can become transparent partners in forecasting—trusted not blindly, but because we understand how they reason, where they excel, and where they need human oversight.

The questions this thesis has explored, How does the model make decisions? Does it align with physical understanding? Can we trust it?, are not unique to weather forecasting. Every application of AI to consequential decisions faces similar challenges. By developing methods to answer these questions in one domain, we contribute to the broader effort to make artificial intelligence not just powerful, but accountable, interpretable, and worthy of the trust society must place in it.

The forecast for explainable AI in meteorology is promising, with clearing conditions ahead. But as any forecaster knows, the most interesting weather is often found where models disagree, where understanding is incomplete, and where new observations challenge old assumptions. The same is true for this research: the most valuable work lies ahead, in the places where our interpretability methods reveal surprising behaviors, where domain experts question our explanations, and where the dialogue between AI and atmospheric science generates new insights neither field could achieve alone.

Appendix A

Complete Expert Interview Responses

Dr. Adam Houston—Meteorological Evaluation of Attention Visualizations

Interview Context

- **Expert:** Dr. Adam Houston, Meteorologist
- **Materials Reviewed:** Interactive Google Colab notebook containing forecast inputs/outputs and attention visualizations for 12 global focal points
- **Review Period:** Several days of independent examination prior to structured interview
- **Materials URL:** https://colab.research.google.com/drive/1Iee4Uqp8sxr45EWh6JLE4vKNGaw0iQD6

Complete Response Transcript

Part 1: Orientation & Background

Q1: How would you describe your experience with machine learning weather models?

- Response: Basic Awareness
- Additional notes: Had a PhD student create an ML model. Limited in scope to specific variable set. Supporting a more holistic forecast.

Q2: What aspects of ML weather forecasting concern you most from an operational perspective?

- Response 1: Potential lack of understanding of the fundamental aspects of the atmosphere.
- Response 2: Taking the human out of the process.

Part 2: Initial Visualization Assessment

Q1: On a scale of 1-5, how intuitive are these visualizations to interpret?
(1=very confusing, 5=immediately clear)

- Response: 4

Part 3: Trust and Operational Value

Q1: Rate how these visualizations affect your trust in the model's predictions:
(-2=significantly decreases, -1=somewhat decreases, 0=no change, +1=somewhat increases, +2=significantly increases)

- Response: Somewhat increases (+1)

Q2: For each visualization type, rate its potential scientific/operational value (1-5):

- Global contribution maps: 2
- Ring-normalized wind-roses: 3
- Global-normalized wind-roses: 4

Q3: In an operational setting, would you consult these visualizations when the model produces an unexpected forecast?
(Definitely not / Probably not / Maybe / Probably / Definitely)

- Response: Definitely

Q4: What specific weather scenarios would most benefit from these explanations?

- Response 1: More of a tool in building trust in models as opposed to operationally useful.
- Response 2: For a specific forecast this could help in overall confidence or moving me as the forecaster in particular direction on a final forecast.

Part 4: Interpretation Validation

Q1: The universal zonal (E-W) preference suggests the model learned about atmospheric rotation effects. Do you agree with this interpretation?
(Strongly agree / Agree / Neutral / Disagree / Strongly disagree)

- Response: Agree

Q2: Are there any patterns here that seem physically implausible or concerning?

- Response: Nothing that seems implausible

Part 5: Practical Recommendations

Q1: What modifications to these visualizations would make them more useful for forecasters?

- Response 1: Truly 3 dimensional rendering where each pixel becomes a voxel
- Response 2: Show change in attention weights over time for the same forecast time

Part 6: Overall Assessment

Q1: Overall, do these attention patterns suggest the model has learned meaningful atmospheric relationships?
(Definitely yes / Probably yes / Unsure / Probably not / Definitely not)

- Response: Probably yes

Q2: If these methods showed meteorologically plausible/consistent attention patterns, would that improve your confidence in a model?
(Definitely yes / Probably yes / Unsure / Probably not / Definitely not)

- Response: Definitely Yes

Q3: What's your biggest remaining concern about trusting this model after seeing these explanations?

- Response 1: Would want to see accuracy metrics
- Response 2: Would like to better understand the inner workings of the model

Q4: If you could ask the model one question about how it makes predictions, what would it be?

- Response: Does it actually care about or understand fluid dynamics?

Appendix B

Regional Attention Analysis Plots

This appendix presents complete attention analysis visualizations for all 12 globally distributed focal points examined in this study. Each section includes three visualization types: (1) global contribution maps showing spatial attention distributions across pressure levels, (2) ring-normalized wind-rose diagrams highlighting directional preferences at each distance, and (3) global-normalized wind-rose diagrams revealing absolute contribution magnitudes and distance decay patterns.

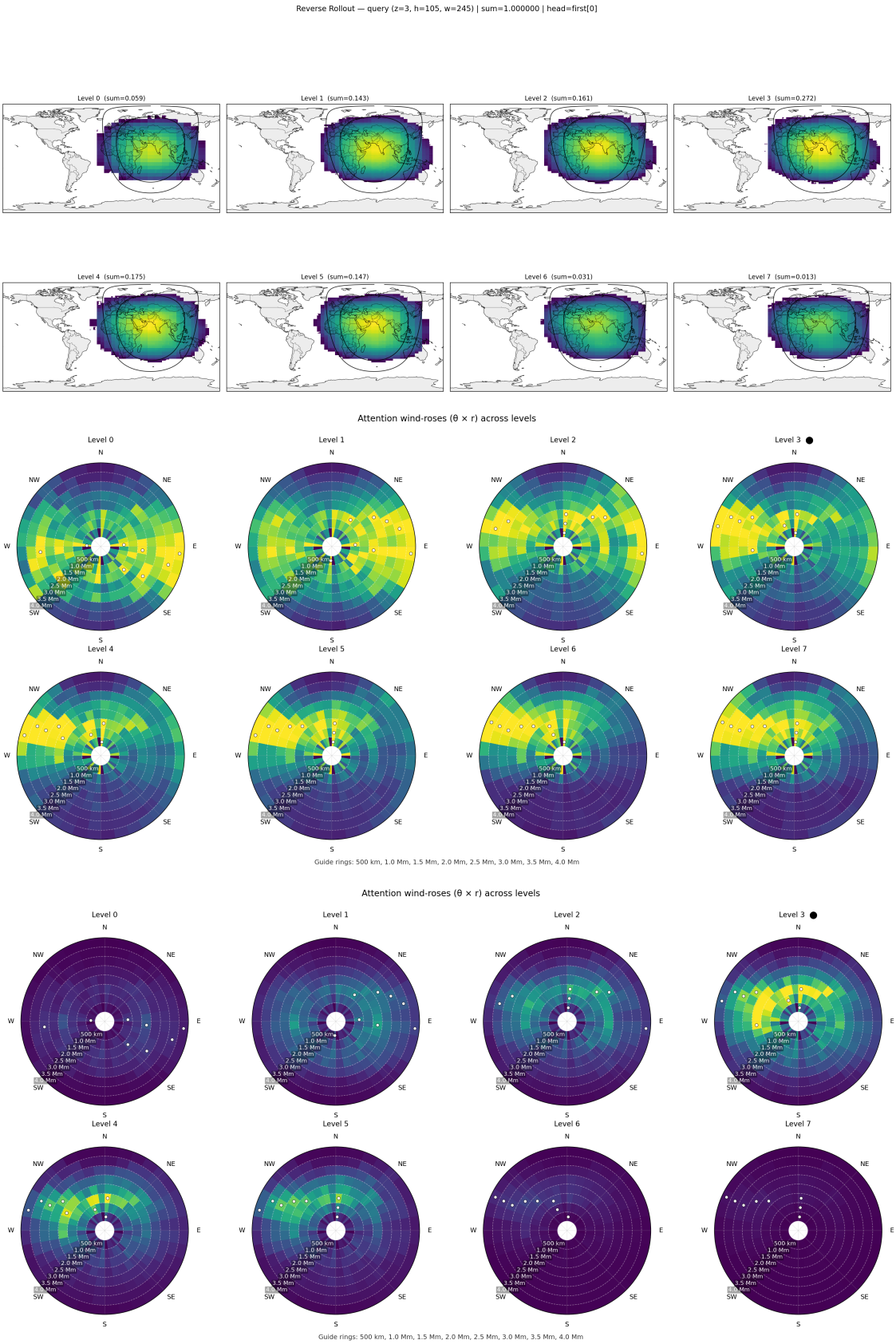

Figure B.1: Attention analysis for Arabian Sea focal point (10.5°N, 65°E). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

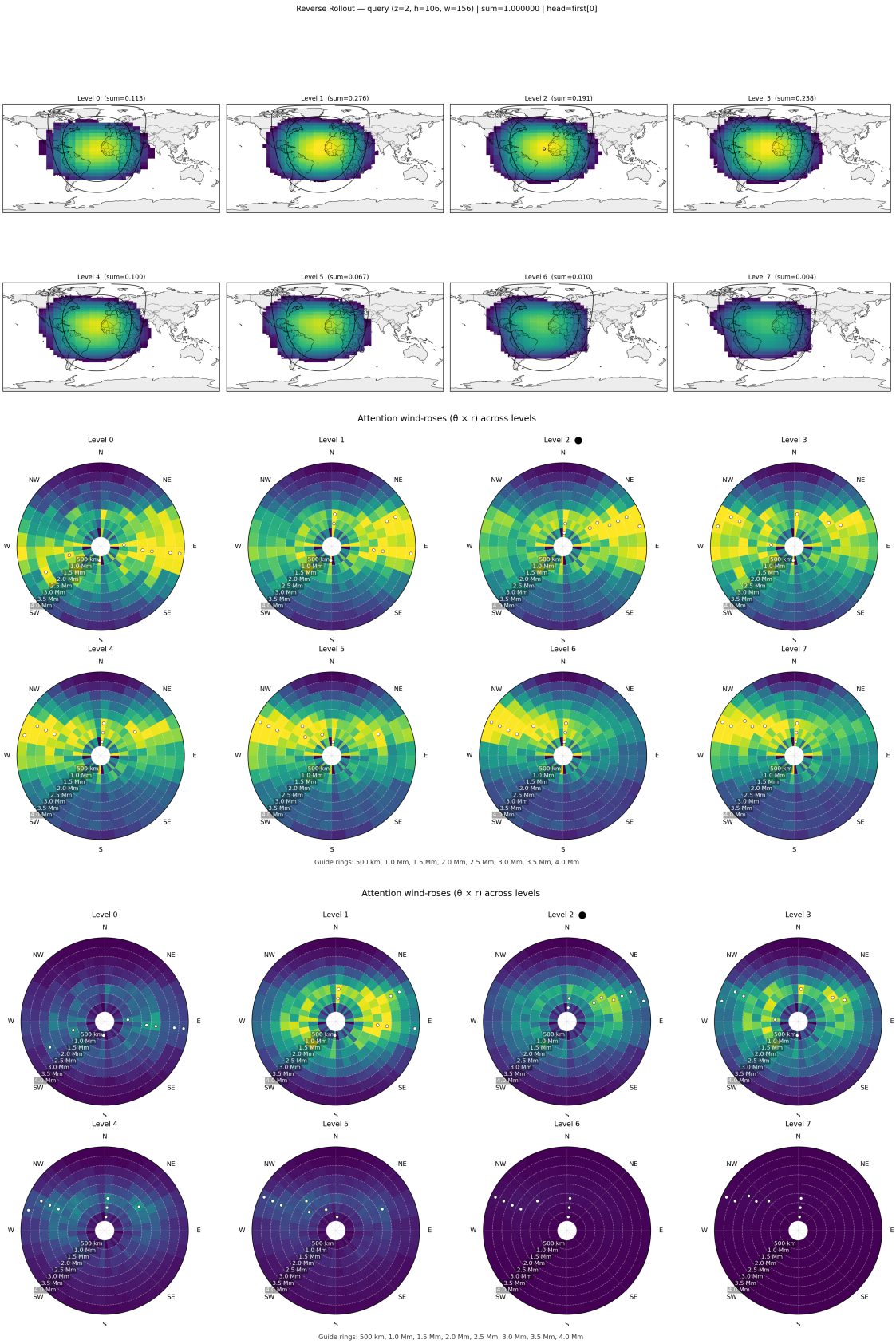

Figure B.2: Attention analysis for Cape Verde Trades focal point (15°N, 30°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

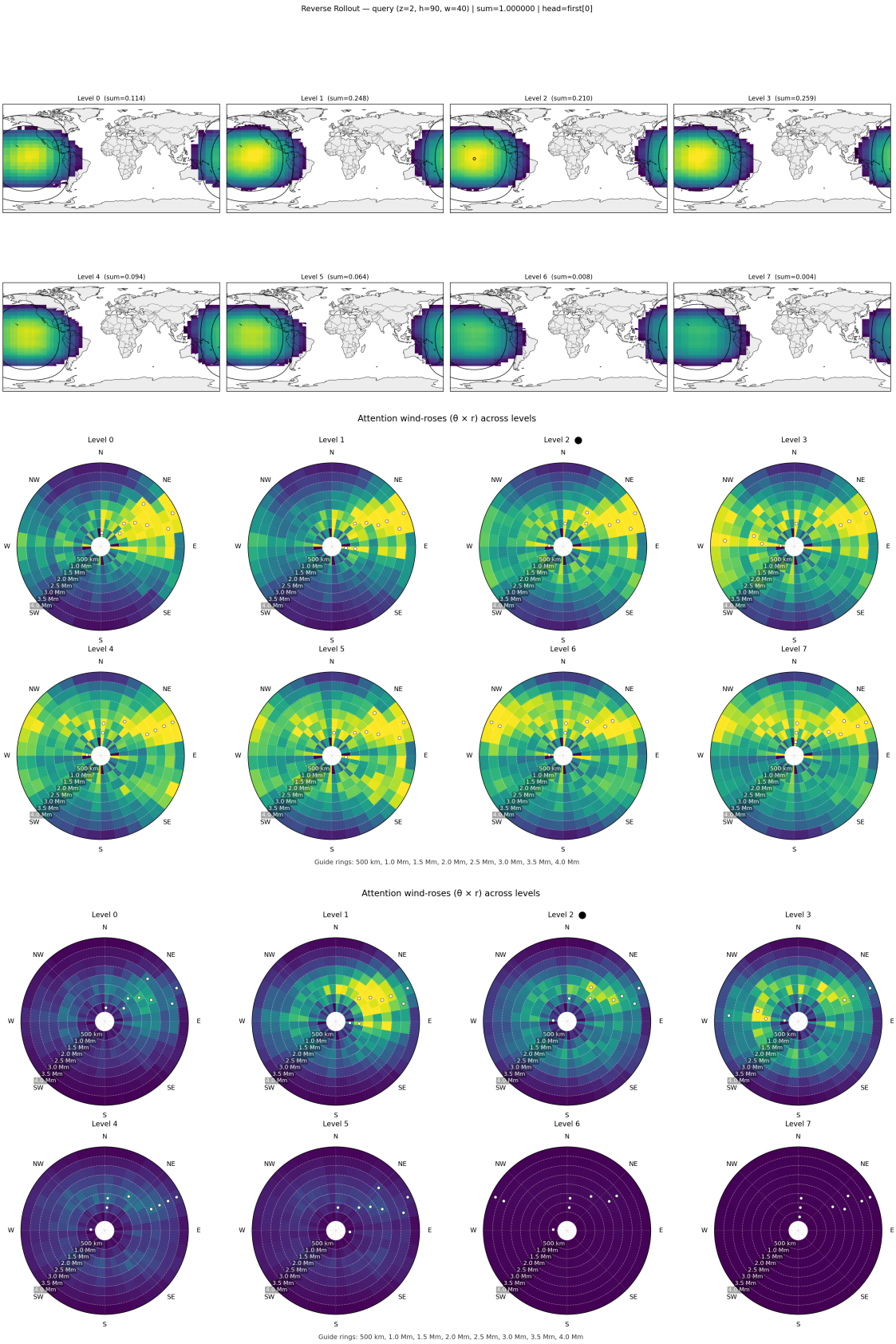

Figure B.3: Attention analysis for Central Equatorial Pacific focal point (0°, 160°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

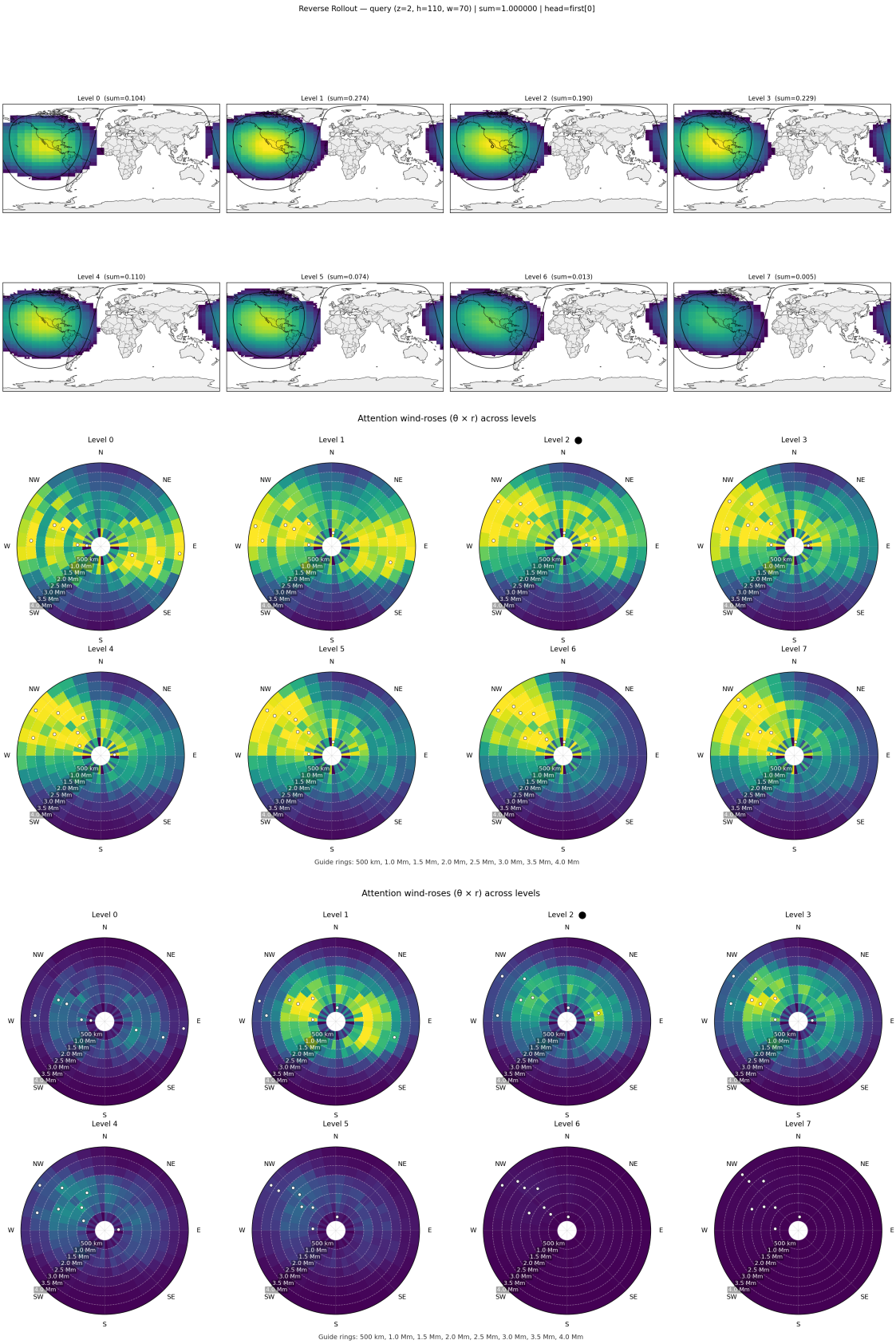

Figure B.4: Attention analysis for Eastern Pacific off Baja focal point (25°N, 115°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

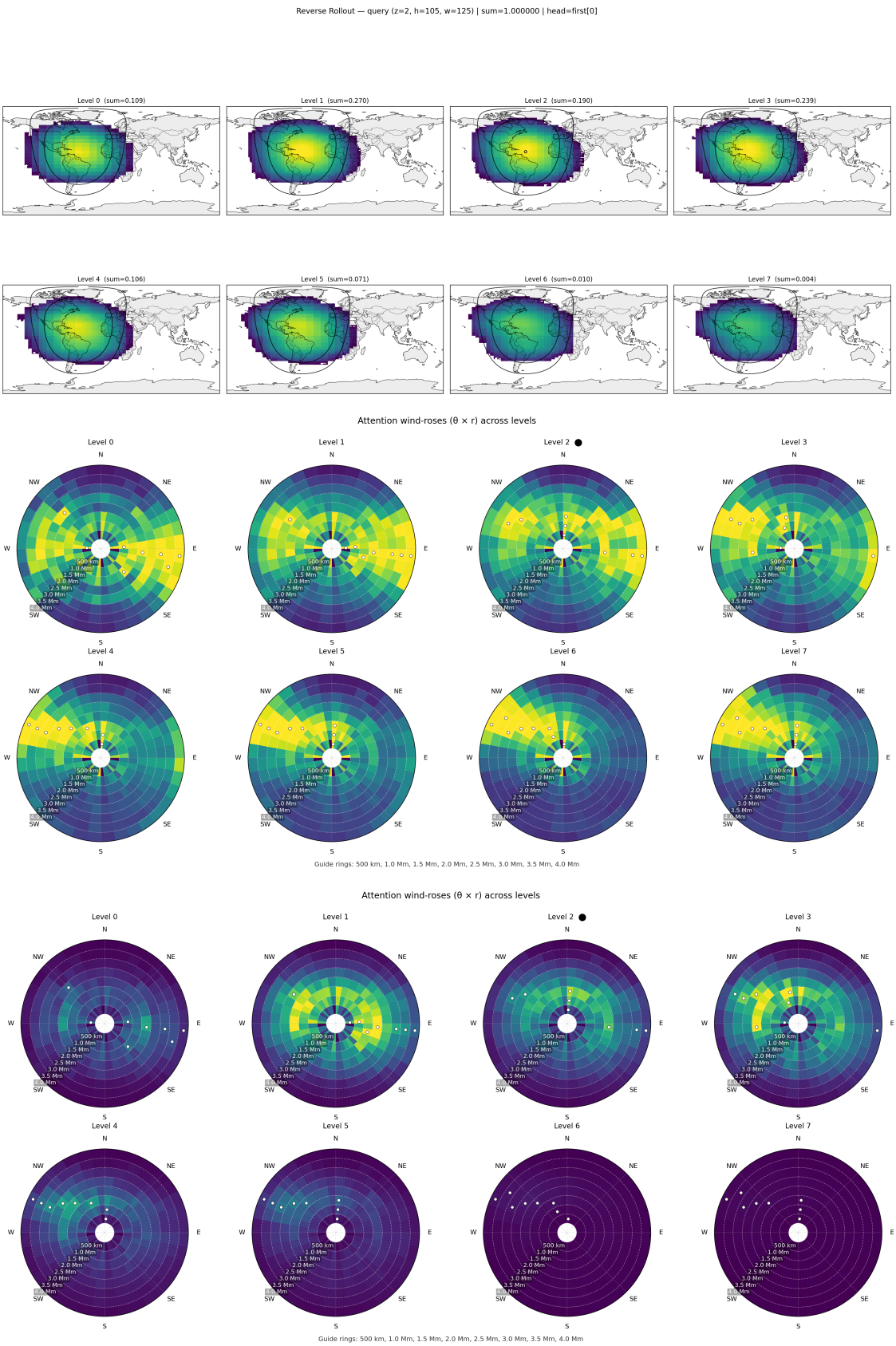

Figure B.5: Attention analysis for Eastern Caribbean Trades focal point (12°N, 60°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

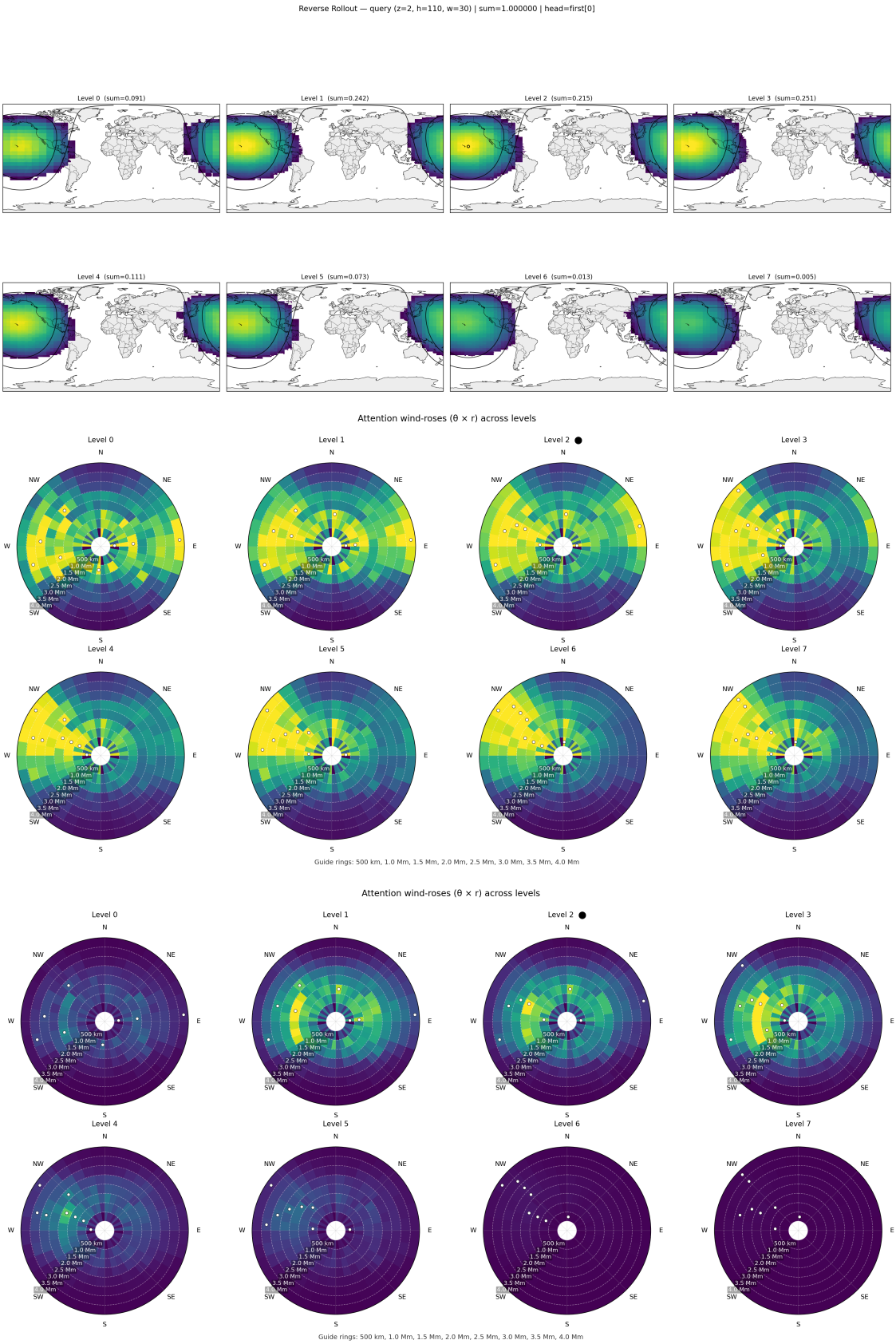

Figure B.6: Attention analysis for Hawaiian Trade Corridor focal point (20°N, 155°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

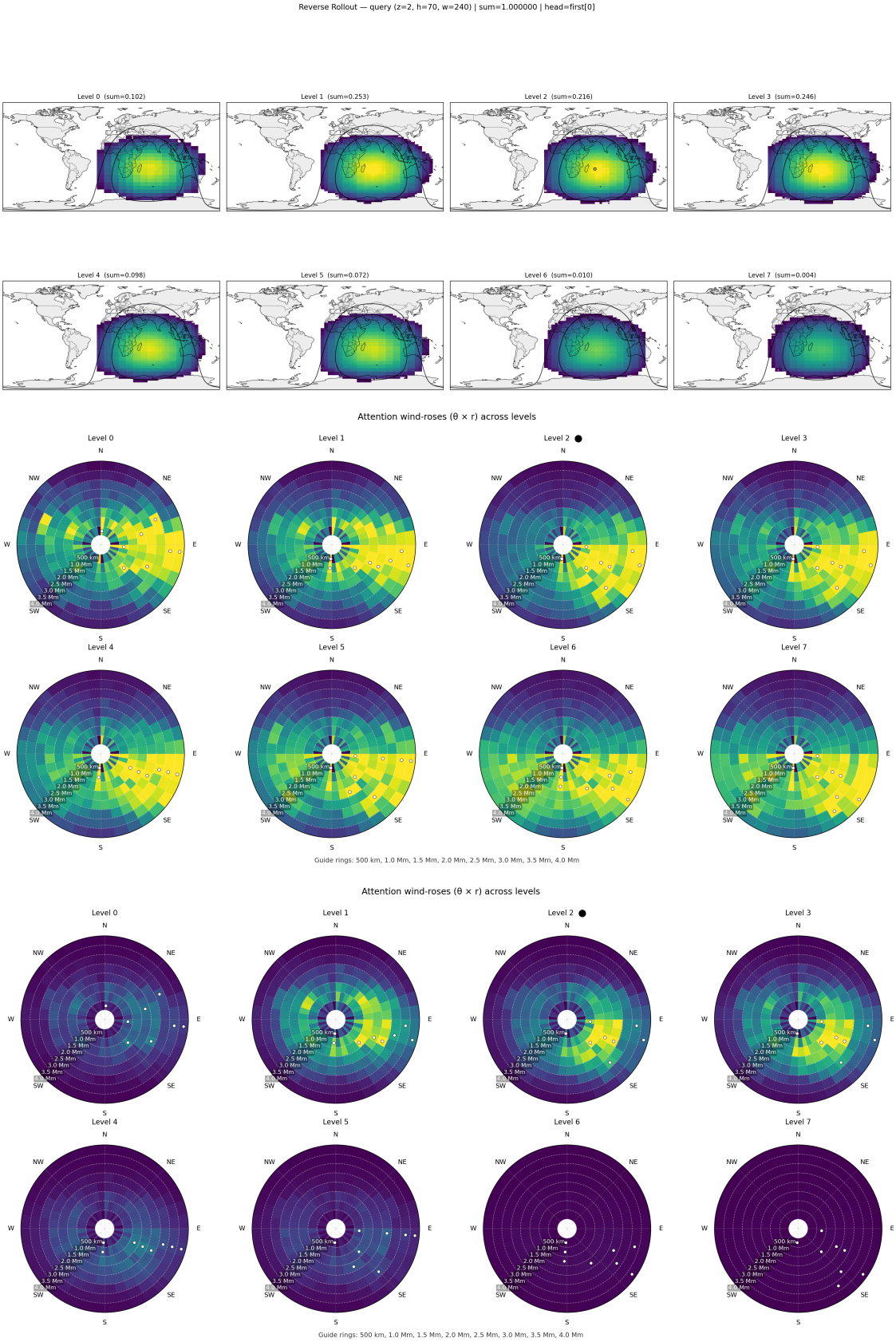

Figure B.7: Attention analysis for Mascarene High Flank focal point (25°S, 55°E). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

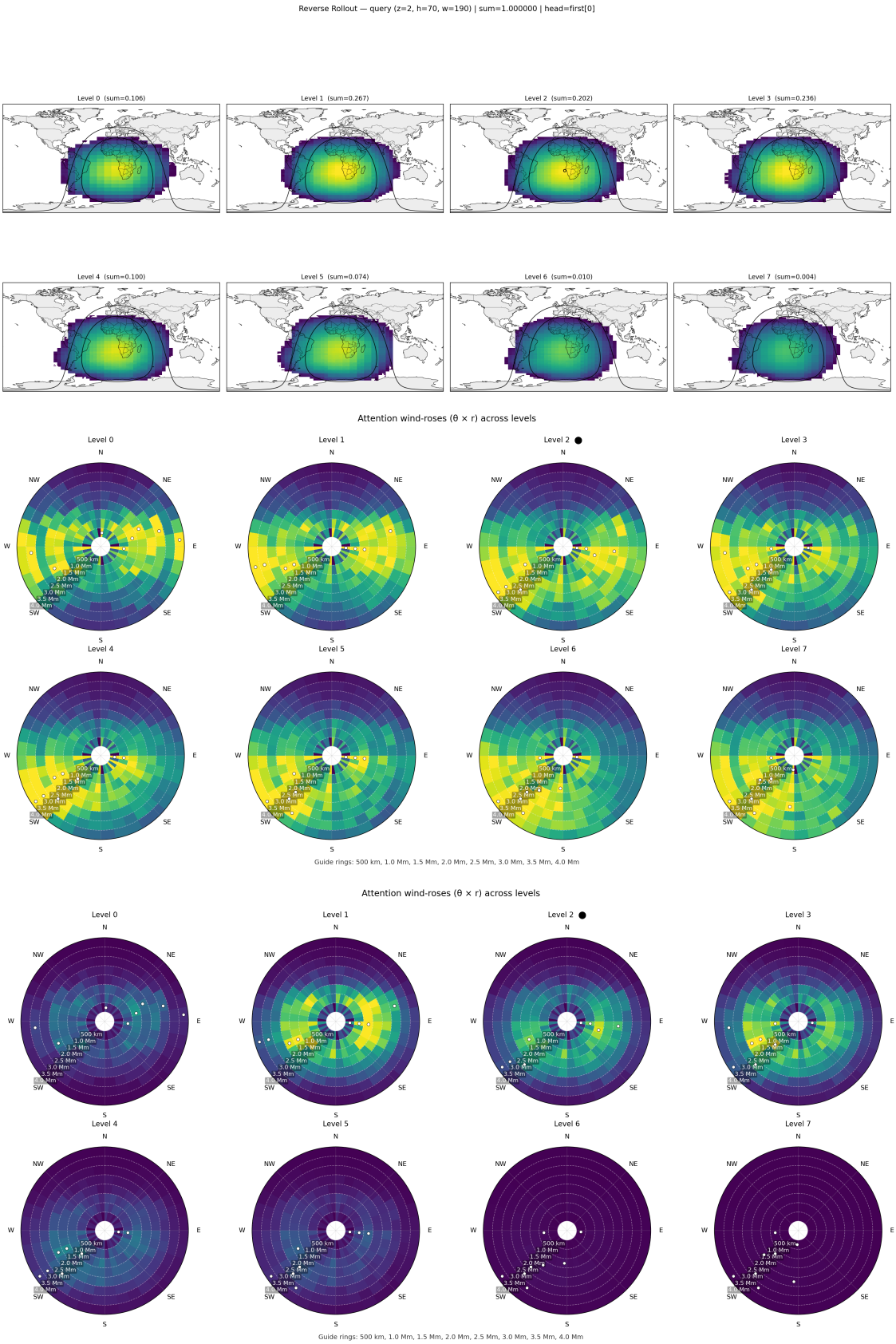

Figure B.8: Attention analysis for SE Atlantic Stratocumulus focal point (15°S, 5°E). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

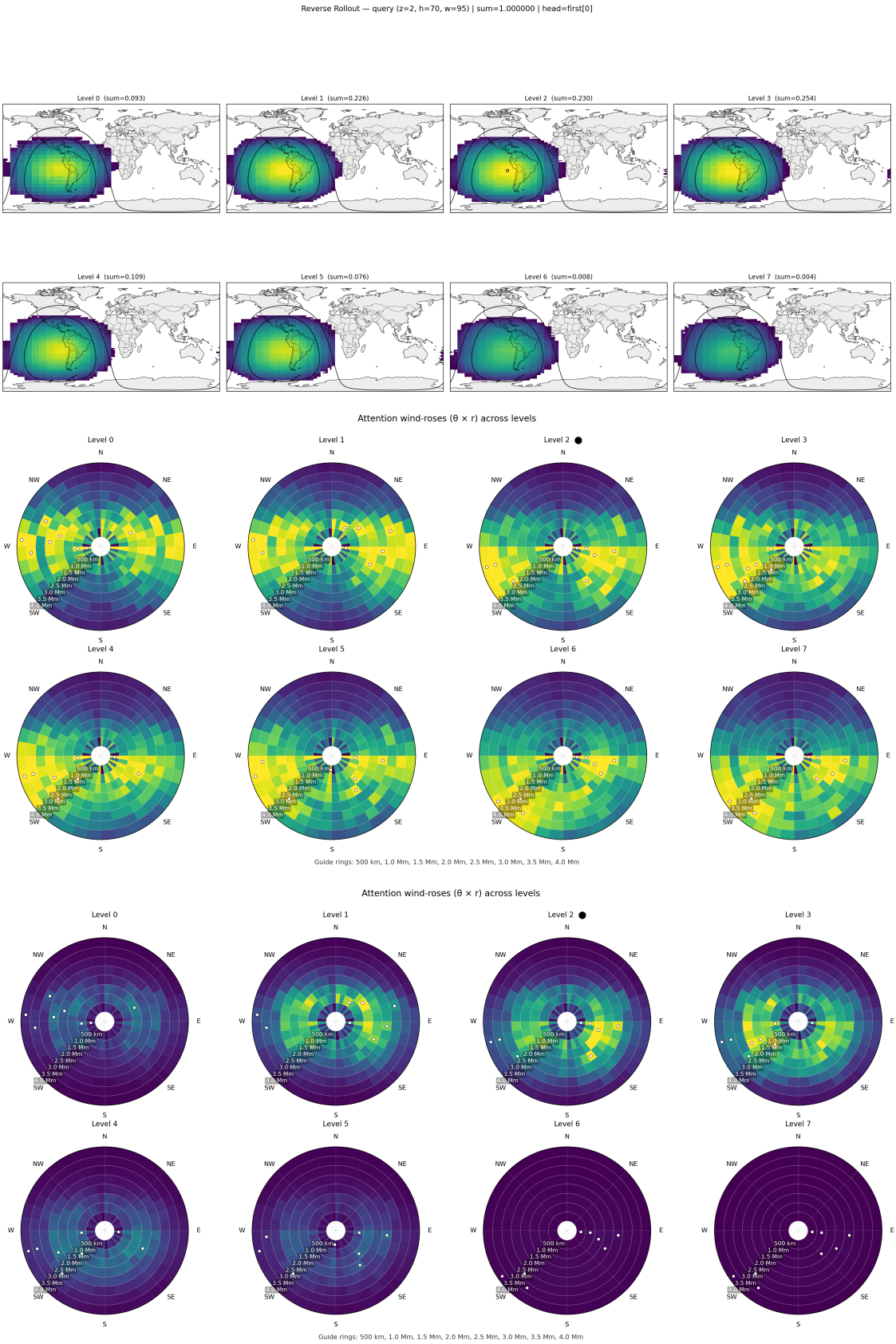

Figure B.9: Attention analysis for SE Pacific Stratocumulus Deck focal point (20°S, 85°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

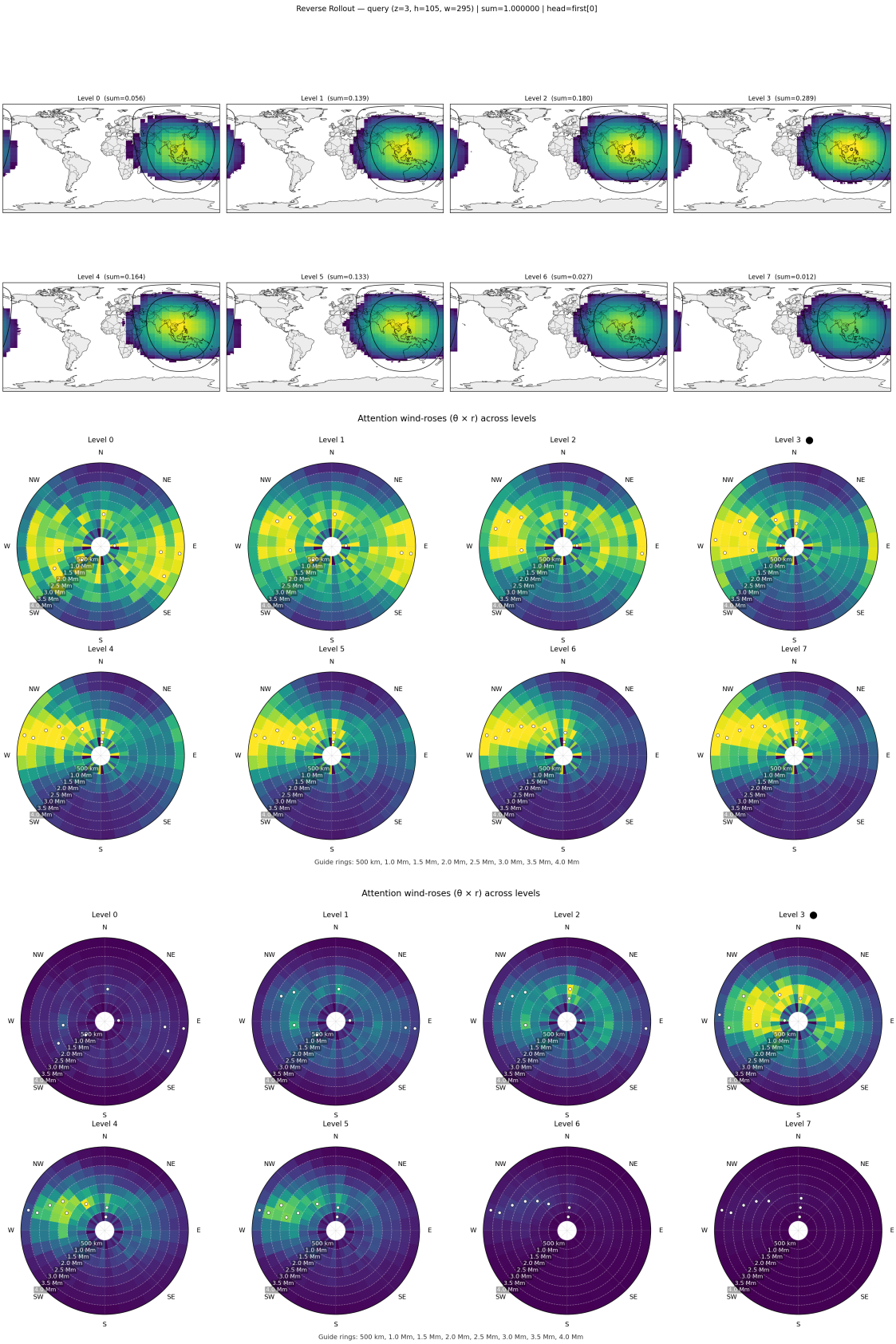

Figure B.10: Attention analysis for South China Sea focal point (15°N, 115°E). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

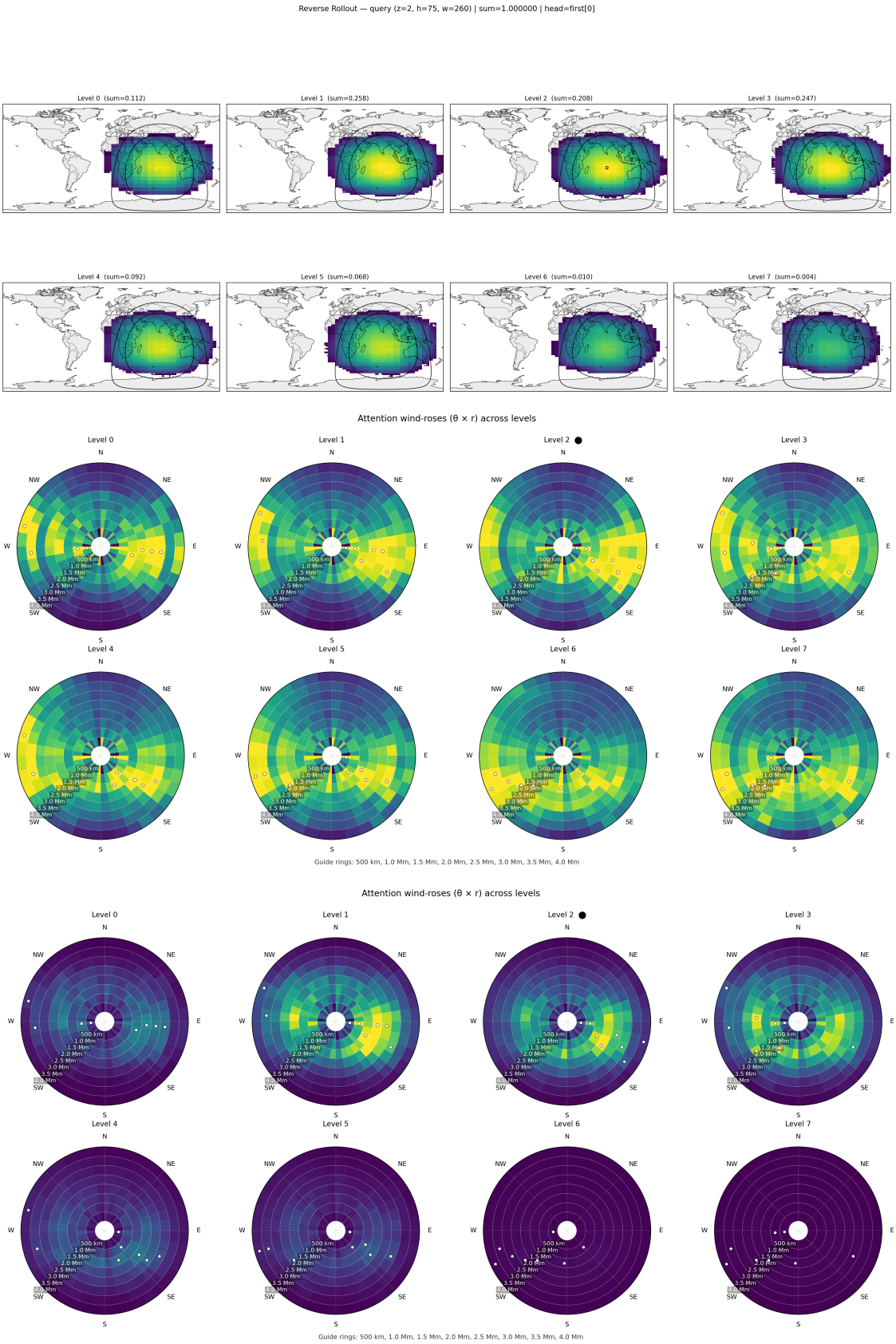

Figure B.11: Attention analysis for SW Indian Ocean Trades focal point (15°S, 70°E). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

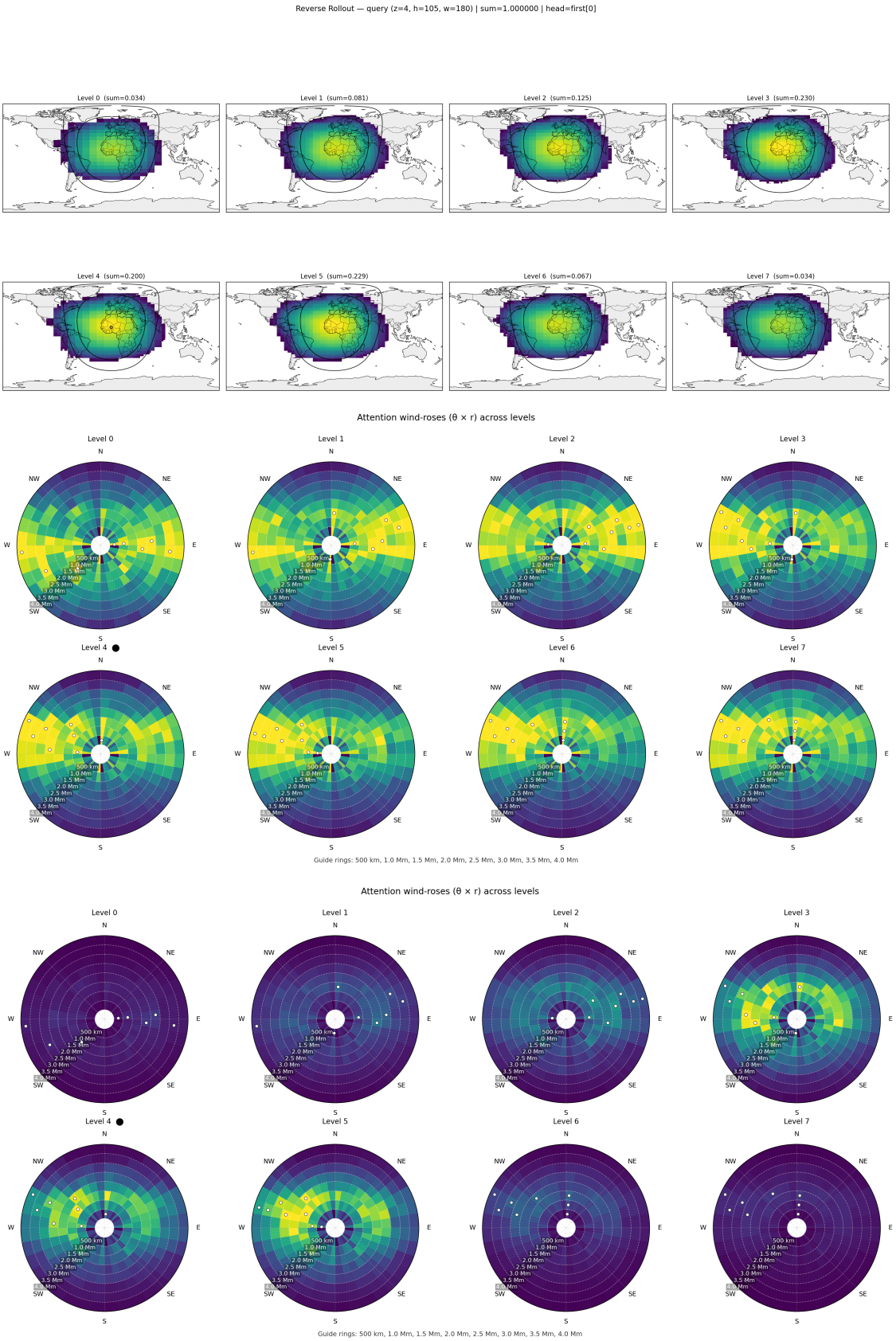

Figure B.12: Attention analysis for West Africa-Sahel focal point (12°N, 5°W). Top: Global contribution map. Middle: Ring-normalized wind rose. Bottom: Global-normalized wind rose.

Bibliography

[1] Samira Abnar and Willem Zuidema. "Quantifying attention flow in transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://doi.org/10.18653/v1/2020.acl-main.385.

[2] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140.

[3] Jorge Baño-Medina et al. "Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia". In: *npj Climate and Atmospheric Science* 8.1 (2025), p. 92. DOI: 10.1038/s41612-025-00949-6. URL: https://doi.org/10.1038/s41612-025-00949-6.

[4] Peter Bauer, Alan Thorpe, and Gilbert Brunet. "The quiet revolution of numerical weather prediction". In: *Nature* 525.7567 (2015), pp. 47–55. DOI: 10.1038/nature14956. URL: https://doi.org/10.1038/nature14956.

[5] Kaifeng Bi et al. "Accurate medium-range global weather forecasting with 3D neural networks". In: *Nature* 619.7970 (2023), pp. 533–538. DOI: 10.1038/s41586-023-06185-3. URL: https://doi.org/10.1038/s41586-023-06185-3.

[6] Hila Chefer, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 397–406. DOI: 10.1109/ICCV48922.2021.00045. URL: https://doi.org/10.1109/ICCV48922.2021.00045.

[7] Mark DeMaria et al. "Evaluation of Tropical Cyclone Track and Intensity Forecasts from Artificial Intelligence Weather Prediction (AIWP) Models". In: *arXiv preprint arXiv:2409.06735* (2024). URL: https://arxiv.org/abs/2409.06735.

[8] Joseph F DeRose, Jiayao Wang, and Matthew Berger. "Attention Flows: Analyzing and Comparing Attention Mechanisms in Language Models". In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020), pp. 1160–1170. DOI: 10.1109/TVCG.2020.3030419. URL: https://doi.org/10.1109/TVCG.2020.3030419.

[9] Jay DeYoung et al. “ERASER: A benchmark to evaluate rationalized NLP models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://doi.org/10.18653/v1/2020.acl-main.408.

[10] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017). URL: https://arxiv.org/abs/1702.08608.

[11] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[12] Gregory Gaspari and Stephen E Cohn. “Construction of correlation functions in two and three dimensions”. In: *Quarterly Journal of the Royal Meteorological Society* 125.554 (1999), pp. 723–757. DOI: 10.1002/qj.49712555417. URL: https://doi.org/10.1002/qj.49712555417.

[13] Thomas M Hamill, Jeffrey S Whitaker, and Chris Snyder. “Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter”. In: *Monthly Weather Review* 129.11 (2001), pp. 2776–2790. DOI: 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2. URL: https://doi.org/10.1175/1520-0493(2001)129%3C2776:DDFOBE%3E2.0.CO;2.

[14] Hans Hersbach et al. “The ERA5 global reanalysis”. In: *Quarterly Journal of the Royal Meteorological Society* 146.730 (2020), pp. 1999–2049. DOI: 10.1002/qj.3803. URL: https://doi.org/10.1002/qj.3803.

[15] James R Holton. *An Introduction to Dynamic Meteorology*. 4th ed. Burlington, MA: Academic Press, 2004. ISBN: 978-0-12-354015-7. URL: https://www.sciencedirect.com/book/9780123540157/an-introduction-to-dynamic-meteorology.

[16] Sarthak Jain and Byron C Wallace. “Attention is not explanation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://doi.org/10.18653/v1/N19-1357.

[17] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511802270. URL: https://doi.org/10.1017/CBO9780511802270.

[18] Anuj Karpatne et al. “Theory-guided data science: A new paradigm for scientific discovery from data”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.10 (2017), pp. 2318–2331. DOI: 10.1109/TKDE.2017.2720168. URL: https://doi.org/10.1109/TKDE.2017.2720168.

[19] Remi Lam et al. “Learning skillful medium-range global weather forecasting”. In: *Science* 382.6677 (2023), pp. 1416–1421. DOI: 10.1126/science.adi2336. URL: https://doi.org/10.1126/science.adi2336.

[20] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022. DOI: 10.1109/ICCV48922.2021.00986. URL: https://doi.org/10.1109/ICCV48922.2021.00986.

[21] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in Neural Information Processing Systems* 30 (2017). URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[22] Amy McGovern et al. “Making the black box more transparent: Understanding the physical implications of machine learning”. In: *Bulletin of the American Meteorological Society* 100.11 (2019), pp. 2175–2199. DOI: 10.1175/BAMS-D-18-0195.1. URL: https://doi.org/10.1175/BAMS-D-18-0195.1.

[23] Jaideep Pathak et al. “Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators”. In: *arXiv preprint arXiv:2202.11214* (2022). URL: https://arxiv.org/abs/2202.11214.

[24] Stephan Rasp et al. “WeatherBench: a benchmark data set for data-driven weather forecasting”. In: *Journal of Advances in Modeling Earth Systems* 12.11 (2020), e2020MS002203. DOI: 10.1029/2020MS002203. URL: https://doi.org/10.1029/2020MS002203.

[25] Stephan Rasp et al. “WeatherBench 2: A benchmark for the next generation of data-driven global weather models”. In: *Journal of Advances in Modeling Earth Systems* 16.4 (2024), e2023MS004019. DOI: 10.1029/2023MS004019. URL: https://doi.org/10.1029/2023MS004019.

[26] Martin G Schultz et al. “Can deep learning beat numerical weather prediction?” In: *Philosophical Transactions of the Royal Society A* 379.2194 (2021), p. 20200097. DOI: 10.1098/rsta.2020.0097. URL: https://doi.org/10.1098/rsta.2020.0097.

[27] Yang Shi et al. “Comparison of AI and NWP Models in Operational Severe Weather Forecasting: A Study on Tropical Cyclone Predictions”. In: *Journal of Geophysical Research: Machine Learning and Computation* 4.2 (2025), e2024JH000481. DOI: 10.1029/2024JH000481. URL: https://doi.org/10.1029/2024JH000481.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2014). URL: https://arxiv.org/abs/1312.6034.

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International Conference on Machine Learning*. 2017, pp. 3319–3328. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.

[30] Xuejiao Tang et al. *Interpretable Visual Understanding with Cognitive Attention Network*. 2023. arXiv: 2108.02924 [cs.CV]. URL: https://arxiv.org/abs/2108.02924.

[31] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[32] Peter A. G. Watson. “Machine learning applications for weather and climate need greater focus on extremes”. In: *Environmental Research Letters* 17.11 (2022), p. 111004. DOI: 10.1088/1748-9326/ac9d4e. URL: https://doi.org/10.1088/1748-9326/ac9d4e.

[33] Sarah Wiegreffe and Yuval Pinter. “Attention is not not explanation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://doi.org/10.18653/v1/D19-1002.

[34] Zhongwei Zhang et al. “Numerical models outperform AI weather forecasts of record-breaking extremes”. In: *arXiv preprint arXiv:2508.15724* (2025). URL: https://arxiv.org/abs/2508.15724.

[35] Shan Zhao and Zhitong Xiong. *pangu-pytorch: PyTorch Implementation of Pangu-Weather*. https://github.com/zhaoshan2/pangu-pytorch. PyTorch reimplementation validated against official ONNX weights. 2024.