Latent Representations of Land–Sea Boundaries and Extreme Temperature in Aurora's Encoder (Student Abstract)

Benjamin Richards[1], **Pushpa Kumar Balan**[2]
[1]IMT Atlantique, 4 Rue Alfred Kastler, 44300 Nantes, France
[2]University of Central Missouri, Warrensburg, MO, USA
benjamin.richards@imt-atlantique.net, pushpakumarbalan@gmail.com
https://richardsbenjamin.github.io/

Abstract

Deep learning models are emerging as strong alternatives to numerical weather prediction, yet their internal representations remain poorly understood. We analyze the latent space of Microsoft's Aurora model to test whether its embeddings align with known physical processes. First, we show that land–sea distinctions are strongly captured, with errors mainly at coastlines. Second, we examine extreme surface temperatures using percentile-based thresholds, finding that embeddings reveal a gradient from moderate to severe events, though recall degrades at the rarest percentiles. These results suggest that Aurora's encoder encodes physically consistent features but underestimates rare extremes. Our study combines deep learning forecasting, interpretable representation learning, and classical ML probing, illustrating how cross-disciplinary AI methods can yield insight into foundation models.

Code — https://github.com/richardsbenjamin/auroraencoderanalysis

Datasets — https://weatherbench2.readthedocs.io/en/latest/data-guide.html

Introduction

Numerical weather prediction (NWP) has been the cornerstone of operational forecasting for decades. These physics-based models solve governing equations on discretised grids and are essential for applications ranging from disaster risk management to agriculture (Bauer, Thorpe, and Brunet 2015). Recently, deep learning has emerged as a promising alternative, enabled by high-quality reanalysis datasets such as ERA5 and advances in large-scale training on GPU/TPU clusters. Aurora is a data-driven forecasting system with an encoder–processor–decoder architecture. At its core, Aurora is a deep learning model based on a transformer architecture trained for global weather forecasting, and we use its encoder to extract high-dimensional latent embeddings that summarise atmospheric states. Inputs include atmospheric variables (temperature, winds, humidity, geopotential) at multiple pressure levels, surface variables (2 m temperature, winds, sea-level pressure), and static variables (land–sea mask, soil type, surfacet geopotential), all on a 0.25° grid. The encoder produces latent embeddings by patching surface and atmospheric fields separately, yielding spatially localised representations

Data and Methodology

The input data used for our analysis consists of various time steps from the WeatherBench 2 ERA5 dataset (Rasp et al. 2023) The input is passed to the encoder to produce the latent embeddings for each time step. To interpret these encoder's embeddings, we rely on principal component analysis (PCA) and concept analysis. PCA, a popular dimensionality reduction technique, is commonly used to explore latent representations (López-González et al. 2024). Concept vector analysis, another common method for exploring latent spaces, typically consists of identifying positive and negative samples for a concept, then extracting activations from a network layer, and training a linear classifier to distinguish these samples.

A concept can be any abstraction that can be well-defined by the user (Schwalbe 2022). In this work we focus on the land–sea distinction and extreme temperature events as concepts, and logistic regression for the classification. A high classification accuracy implies that the concept has been learnt by the deep learning model. By combining various machine learning techniques, including methods from representation learning and classical machine learning, our framework unifies multiple AI methodologies to assess whether Aurora encodes physically meaningful distinctions in its latent representations.

Land-Sea Analysis

We investigate whether Aurora's encoder captures the physical distinction between land and ocean. Since a land–sea mask is provided as input during training, we hypothesise that this boundary should be reflected in the latent space. This distinction is crucial because land and ocean exhibit different thermodynamic responses to forcing. However, there is no guarantee that the latent space has learnt this representation. To test this, we train a logistic regression classifier on latent vectors to predict whether a patch corresponds to land or ocean. Patches between longitudes 120°–210° (covering East Asia and Australia) are held out for testing, yielding a 75/25 train–test split. The classifier

achieves 99.87% accuracy, indicating that the encoder has effectively internalised the land–sea boundary. Misclassifications occur primarily along coastlines, where the boundary is inherently ambiguous (Fig. 1). This result suggests that Aurora learns geographically meaningful representations aligned with real-world dynamics.

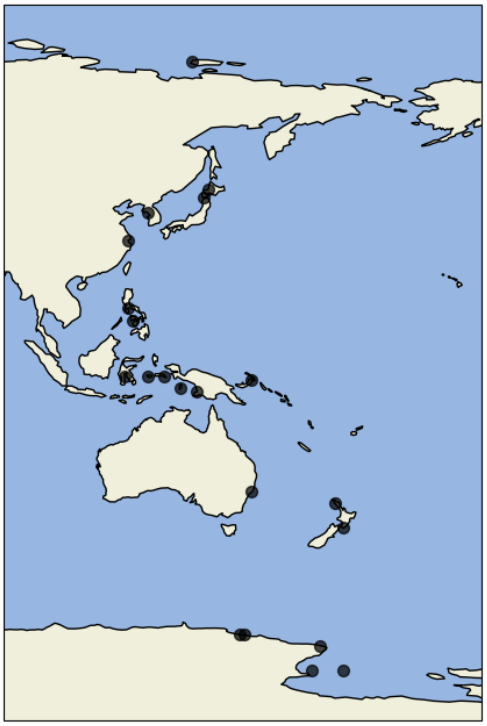

Figure 1: Location of land-sea classification errors.

Extreme Temperature Analysis

We next examine how the encoder represents extreme surface temperatures. Using ECMWF climate projection statistics, we define extremes based on the 75th, 90th, 95th, and 99th percentiles of 2-metre temperature across Europe. ERA5 values exceeding these thresholds are assigned binary labels, resampled to patch resolution, and used as inputs to logistic regression classifiers.

To visualise latent structure, we apply PCA on embeddings 2. The results show a clear gradient: moderate events scatter broadly, while increasingly severe extremes concentrate in a distinct cluster. This suggests that intensity itself is encoded along a coherent axis of variation within the latent space.

Classification performance varies with percentile thresholds (Table 1). At the 75th and 90th percentiles, the model achieves high accuracy with balanced precision and recall, reliably detecting moderate extremes. However, at the 95th and 99th percentiles, recall declines sharply while precision remains high. Thus, the model is conservative: it rarely overpredicts extremes but misses many of the rarest cases.

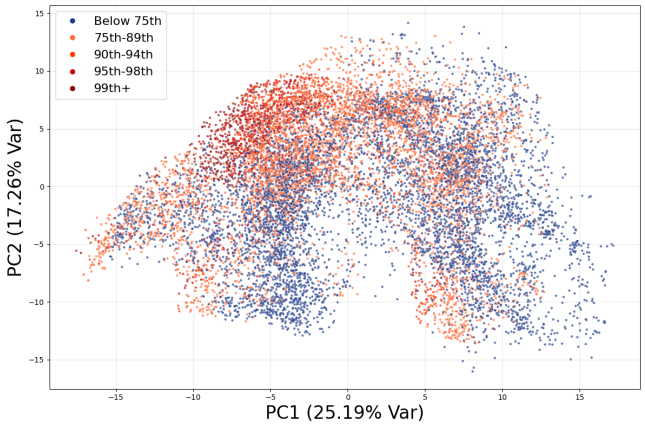

Figure 2: Temperature Extremes PCA.

Percentile	Accuracy	Precision	Recall
p75	0.930	0.925	0.939
p90	0.969	0.935	0.865
p95	0.976	0.897	0.813
p99	0.992	0.917	0.775

Table 1: Logistic regression classification performance.

Conclusion and Future Work

Our results indicate that Aurora's encoder learns physically consistent representations, including the land–sea boundary and structured gradients of extreme temperatures. These preliminary findings motivate further exploration of extreme events and temporal dynamics in the latent space. Current experiments have been limited to small-scale embeddings (10 samples); future work will expand to 200 samples (100 GB) and extend the concept analysis to phenomena such as tropical cyclones, atmospheric rivers, and equatorial processes. We also plan to examine whether sequential structure is embedded by predicting latent vectors autoregressively across time steps.

References

Bauer, P.; Thorpe, A.; and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature*, 525(7567): 47–55.

López-González, C. I.; Gómez-Silva, M. J.; Besada-Portas, E.; and Pajares, G. 2024. Analyzing and interpreting convolutional neural networks using latent space topology. *Neurocomputing*, 593: 127806.

Rasp, S.; Hoyer, S.; Merose, A.; Langmore, I.; Battaglia, P.; Russel, T.; Sanchez-Gonzalez, A.; Yang, V.; Carver, R.; Agrawal, S.; Chantry, M.; Bouallegue, Z. B.; Dueben, P.; Bromberg, C.; Sisk, J.; Barrington, L.; Bell, A.; and Sha, F. 2023. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. arXiv:2308.15560.

Schwalbe, G. 2022. Concept Embedding Analysis: A Review. arXiv:2203.13909.