CNN-Transformers with Manifold Learning-Based Embeddings for Global Weather Prediction

Ben Choi
NASA Goddard Space Flight Center
Harvard College
benchoi@college.harvard.edu

Abstract

Recent strides in machine learning (ML) and artificial intelligence (AI) have significantly influenced the field of data-driven weather forecasting. In this paper, we describe the development of a new ML-driven architecture for global climate forecasting. This architecture—taking the form of a conbined convolutional neural network (CNN) and transformer—was developed after careful literature review, ablation studies, and consideration of key alternatives, including regression-based, multilayer perceptron-based, and recurrent neural network-based models. We also formulated our architecture after consolidating recent insights from leading AI-driven forecasting systems; our work builds upon a recent lineage of hybrid models that adeptly combine spatial and temporal data processing. In particular, we incorporate halo regions for enhanced contextual understanding and employ Laplacian eigenmaps for manifold learning-based dimensionality reduction. Our model incorporates data for temperature (t), wind (u & v), moisture (Qv), and surface pressure (ps), and performs well against baseline methods in preliminary experimentation.

1 Introduction

Recent advances in machine learning (ML) have paved the way for models capable of forecasting weather on both local and global scales. Advancing beyond conventional methodologies, contemporary architectures now incorporate hybrid ML-driven structures for more intelligent climate forecasting. Recent ML-based forecasting models have demonstrated success on everything from predicting fundamental parameters like temperature (Anjali et al., 2019; Azari et al., 2022; Mukkavilli et al., 2023) to analyzing complex phenomena like tropical cyclones (Chan et al., 2021; Lang et al., 2024) and fire progression (Thapa et al., 2024).

In 2022, NVIDIA's FourCastNet (Pathak et al., 2022) demonstrated a successful large-scale deep learning model for state-of-the-art numerical weather prediction (NWP). Beyond leveraging advances in high-performance computing, NVIDIA's FourCastNet-based prediction system uses transformers to refine spatial embeddings of climate data and yield effective downstream predictions. While pure transformers have demonstrated subpar results on forecasting problems (Zeng et al., 2023), Pathak et al., 2022 presents a modified formulation with spatial (token) mixing suitable for large-scale climate data. Beyond their high impact in other areas of machine learning, transformer-based architectures evidently demonstrate promise in the climate forecasting space—if combined with effective pre-processing and embedding methods.

Another milestone in the ML weather forecasting domain has been the *Artificial Intelligence Forecasting System (AIFS)* (Lang et al., 2024) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). AIFS leverages a hybrid architecture combining a Graph Neural Network (GNN) with graph convolutions and a sliding window transformer processor. AIFS has

.

demonstrated strong accuracy for both upper-air variables as well as surface weather parameters, surpassing traditional numerical models on key tests; Lang et al., 2024 also demonstrated state-of-the-art results with AIFS in tropical cyclone prediction, underscoring the overall versatility of the system. Beyond representing another innovative use of transformers in the NWP space, the AIFS system notably also demonstrates the promise of graph-based processing methods and convolutions that encapsulate spatial relationships—concepts which we fold into our eventual proposed system.

Inspired by the promise of combining attention-based networks with effective spatial embeddings of climate data, we present a preliminary architecture for global weather prediction leveraging a CNN-Transformer hybrid structure. We aim to extend the existing lineage of ML-driven forecasting models by proposing a streamlined, versatile architecture that can function across all global regions. Our architecture was developed after careful progression through traditional alternatives; more precise formulations are included in the subsequent sections of this report.

2 Methodology

The development of the CNN-Transformer architecture was rooted in a systematic exploration of data and progressive modeling iterations—ultimately leading to a hybrid capable of capturing complex spatiotemporal dynamics in weather forecasting. This section details the dataset exploration, model development process, and mathematical formulations behind key methods employed.

2.1 Dataset Exploration and Partitioning

The dataset utilized for this study (from the NASA Global Modeling and Assimilation Office) encompassed a comprehensive collection of atmospheric variables, including temperature (T), zonal wind component (u), meridional wind component (v), moisture content (Qv), and surface pressure (ps. These variables were extracted from high-resolution weather datasets, focusing on the uppermost 36 levels in the atmospheric profile to capture essential features influencing temperature predictions. The data were partitioned into non-overlapping three-dimensional cubes of size $64 \times 64 \times 36$, where 64 represented the grid dimensions in latitude and longitude, and 36 corresponded to the selected vertical levels; subsequent dimensionality reduction was performed (as described in section 2.3).

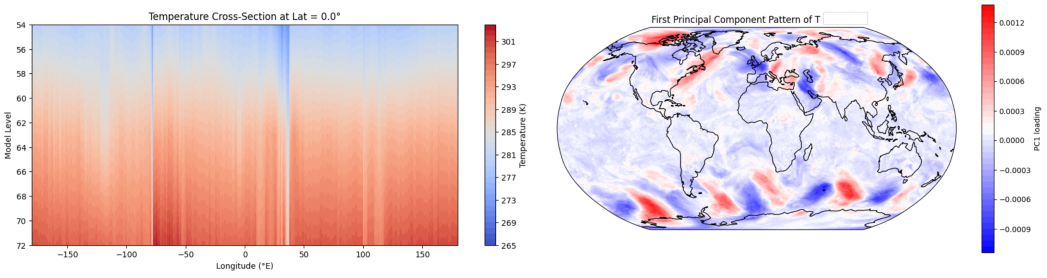

Figure 1: Visualizations collected during data exploration. A sample temperature cross section is depicted in the left panel; global temperature component fluctuations are depicted right.

To prepare the dataset, we conducted an exploratory analysis that included the computation of statistical measures (e.g., mean, variance, covariance) of each variable across spatial and vertical dimensions. This analysis informed our understanding of the data distribution and highlighted areas with significant atmospheric interactions that would require enhanced modeling precision. Given the relatively high dimensionality of the incoming data and potential for complex nonlinear interactions, we formulated a version of the manifold hypothesis—that is, that there exists exploitable low-dimensional manifold structure within the high-dimensional climate data—and performed subsequent nonlinear dimensionality reduction (see *2.3* below).

2.2 Halo Region Concept

To improve the model's capacity to understand boundary interactions within the data cubes, we introduced the concept of halo regions. These are extended margins around each $64 \times 64 \times 36$ cube, allowing the model to incorporate contextual data beyond the strict boundaries of the primary input. Let H denote the halo size, which varied from 0 to 10. The augmented input cube with halo can be defined as:

$$\text{Input}_{\text{halo}} = \text{Input}_{\text{core}} + H_{\text{margin}}, \tag{1}$$

where H_{margin} extends the grid on each side by H units, yielding an overall input size of $(64 + 2H) \times (64 + 2H) \times 36$. An optimal halo size $H = 4$ was empirically determined to provide a balance between added context and computational overhead (see *Results*).

2.3 Dimensionality Reduction Techniques

Initial experiments with dimensionality reduction included Principal Component Analysis (PCA), a linear technique defined by:

$$\arg\max_{\mathbf{w}} \frac{\mathbf{w}^T S \mathbf{w}}{\mathbf{w}^T \mathbf{w}}, \tag{2}$$

where S is the covariance matrix. PCA, while effective for variance preservation, failed to retain much of the manifold structure of high-dimensional data.

Subsequent unsupervised exploration progressed toward nonlinear methods such as Isomap, Locally Linear Embedding (LLE), and Multi-Dimensional Scaling (MDS). While partially effective, the preservation of local relationships and interactions—a critical feature of climate data—was suboptimal.

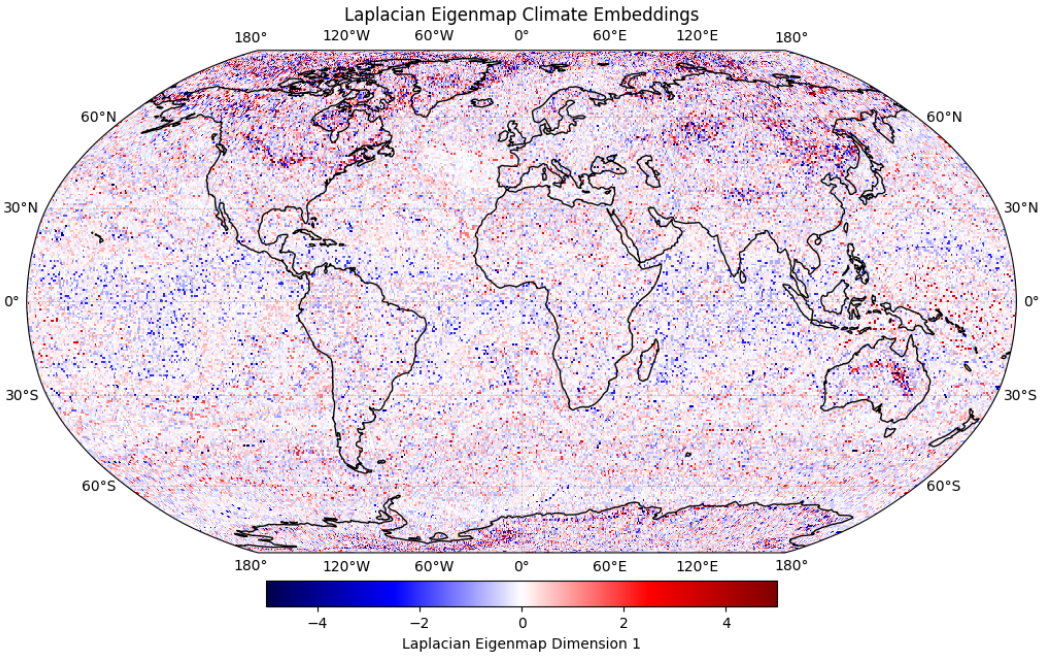

Figure 2: Approximate 1D visualization of Laplacian eigenmap embeddings across global regions.

Laplacian eigenmaps were ultimately adopted for their ability to construct a manifold learning-based embedding with an emphasis on the preservation of local relationships:

$$L = D - W, \tag{3}$$

where D is the degree matrix and W is the adjacency matrix. The eigenvalue problem:

$$L\mathbf{y} = \lambda \mathbf{y}, \tag{4}$$

was solved to obtain eigenvectors $\mathbf{y}$ that compressed the data into a lower-dimensional representation while preserving local structure. Unlike methods such as Isomap or Multi-Dimensional Scaling (MDS), which rely on global distance metrics and can struggle with complex manifold structures, Laplacian eigenmaps excel in maintaining local neighborhood distances via exploiting aforementioned eigenvalue problem encompassing spatial adjacency relationships. This makes the Laplacian eigenmap particularly effective for capturing potential nonlinear structures inherent in climate data. Total dimensionality reduction was on the order of roughly 20 to 1, with the projection effectively encapsulating key local spatial relationships between climate data points while mitigating problems associated with the curse of dimensionality.

2.4 Model Development and Progressive Complexity

Model training employed an L2 norm loss function, with careful variable selection implemented after much discussion and deliberation. While initial drafts incorporated only temperature, zonal, and meridional wind variables, the final CNN-Transformer also incorporates moisture in terms of specific humidity (Qv) and surface pressure (ps) to yield richer, more well-informed predictions.

Initial modeling efforts commenced with linear regression, defined mathematically as:

$$T_{\text{pred}} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i, \tag{5}$$

where T_{pred} is the predicted temperature, X_i are the input features, and β_i are the coefficients determined through least squares minimization. While linear regression offered insights into basic predictive capabilities, it failed to capture the non-linear dependencies inherent in weather data.

Subsequent trials employed autoregressive models (ARMA and ARIMA). The ARMA(p, q) model is represented by:

$$T_t = \sum_{i=1}^{p} \phi_i T_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{6}$$

where ϕ_i are autoregressive coefficients, θ_j are moving average coefficients, and ϵ_t is white noise. While ARMA and ARIMA models improved temporal prediction, their stationarity assumptions limited broader application across diverse conditions.

2.5 Transition to Neural Network Architectures

The inadequacies of traditional models prompted the exploration of multilayer perceptrons (MLPs), represented as:

$$\mathbf{h}^{(l)} = \sigma(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \tag{7}$$

where $\mathbf{h}^{(l)}$ denotes the activations at layer l, $W^{(l)}$ are weight matrices, $\mathbf{b}^{(l)}$ are biases, and σ is the activation function. Although MLPs modeled non-linearities, they struggled with spatial coherence and failed to capture long-term dependencies.

To better model temporal dependencies, we examined a Recurrent Neural Network (RNN)-based architecture. An RNN processes input sequentially:

$$\mathbf{h}_t = f(W_h\mathbf{h}_{t-1} + W_x\mathbf{x}_t + \mathbf{b}), \tag{8}$$

where $\mathbf{h}_t$ is the hidden state at time t, W_h and W_x are weight matrices, and f is the activation function. RNNs demonstrated enhanced performance in capturing temporal dynamics but struggled with long-term dependencies.

To overcome these challenges, we evaluated a pure Transformer model characterized by the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{9}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimensionality of the keys. Despite the model's effectiveness in long-range dependency modeling, the computational cost of self-attention and lack of effective spatial encapsulation proved limiting for large-scale weather data.

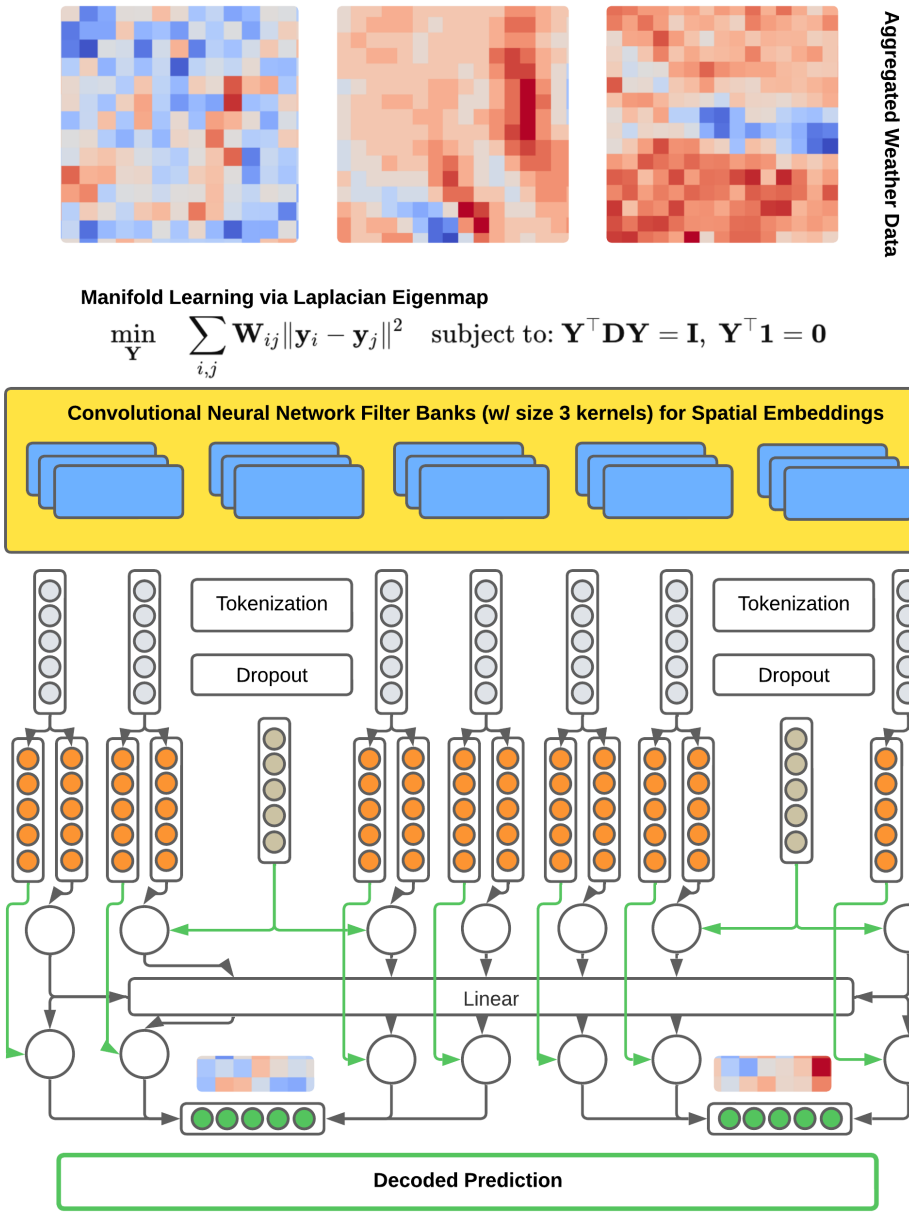

Figure 3: The CNN-Transformer model architecture. Aggregated weather data is routed through the manifold learning-based embedding process before being fed into a multilayer CNN. Subsequent spatial embeddings are leveraged by the transformer to produce decoded predictions.

2.6 CNN-Transformer Hybrid Architecture

The final architecture combined a convolutional neural network (CNN) module for spatial feature extraction with a transformer encoder for temporal modeling. The CNN component processed $64 \times 64 \times 36$ cubes, outputting feature maps represented by:

$$\text{Output}_{\text{CNN}} = \sigma(W_{\text{conv}} * X + b). \tag{10}$$

These feature maps were reshaped and fed into the Transformer encoder to capture temporal dependencies, modeled as:

$$\text{Transformer Output} = \text{Attention}(Q, K, V) \text{ followed by feedforward operations.} \tag{11}$$

As can be seen in the *Results* section covered below, the neural network baselines generally outperformed their regression counterpart, but the pure transformer did not perform well, likely due to the inability to leverage spatial properties inherent in weather data. By building in a CNN, however, to develop spatial encodings of the Laplacian eigenmap-reduced data, we were able to combine the best of both worlds in terms of temporal and spatial encoding. Moreover, the addition of CNN embeddings can somewhat mimic the structure of climate partial differential equations (PDEs) through its layered architecture, which applies local convolutional filters that approximate differential operators. This allows CNNs to capture local spatial dependencies in the data, akin to how PDEs model physical processes such as heat transfer and fluid dynamics by describing how variables like temperature or pressure change over space and time. The repeated application of these convolutional filters in the network's layers enables hierarchical learning of spatial features, providing a mechanism for modeling complex relationships similar to those governed by PDEs in climate systems.

3 Results

3.1 Model Evaluation

Figure 4 illustrates the performance comparison of the proposed CNN-Transformer hybrid against baseline methods. The CNN-Transformer outperformed other tested models, demonstrating robust generalization and accuracy across diverse atmospheric conditions. The effect of combining spatial embeddings with the CNN in conjunction with transformer-driven temporal handling is clearly evident in the gulf between pure transformer and CNN-transformer performance.

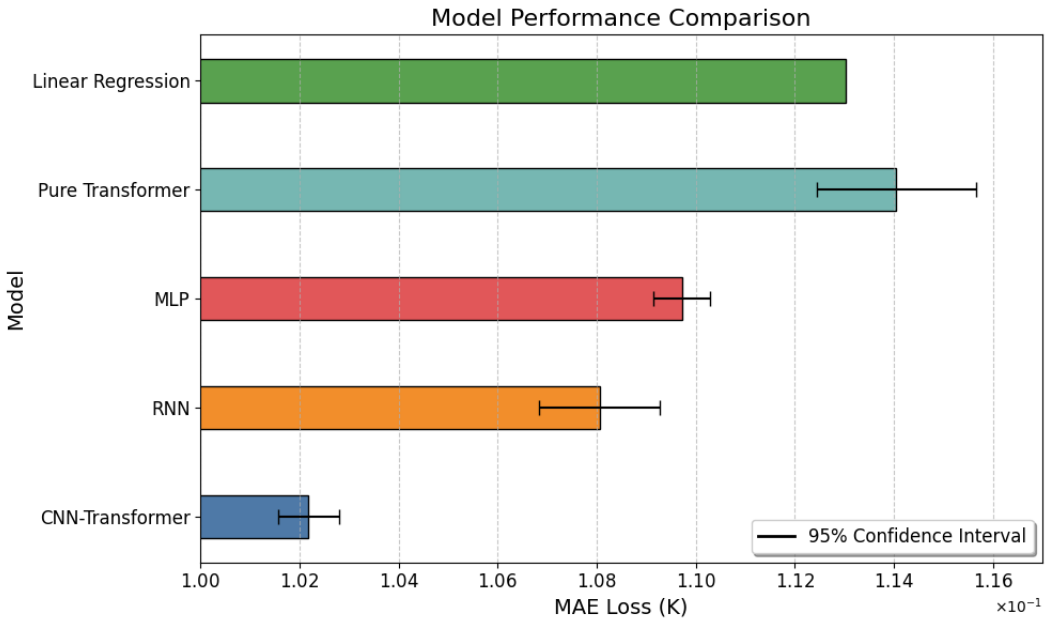

Figure 4: Model performance comparison with 95% confidence intervals for the linear regression, pure transformer, MLP, RNN, and CNN-Transformer models.

We also conducted careful evaluation to determine the optimal halo size (Figure 5); the empirical optimal value of four aligns with the theoretical optimum given the spacing of frames in the data. The subsequent increase in error for large halo sizes, however, was somewhat surprising, and is discussed further in Section 4 (Discussion).

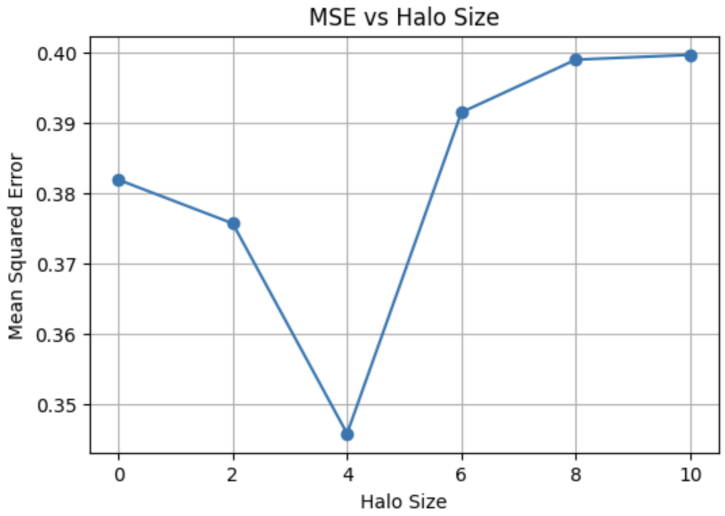

Figure 5: Empirical determination of optimal halo size.

3.2 Visualization

Figure 6 showcases the deviations of predicted weather patterns from ground truth data. The CNN-Transformer demonstrates relatively close agreement with observed data compared to baselines, providing evidence that the model is accurately able to capture spatiotemporal dynamics.

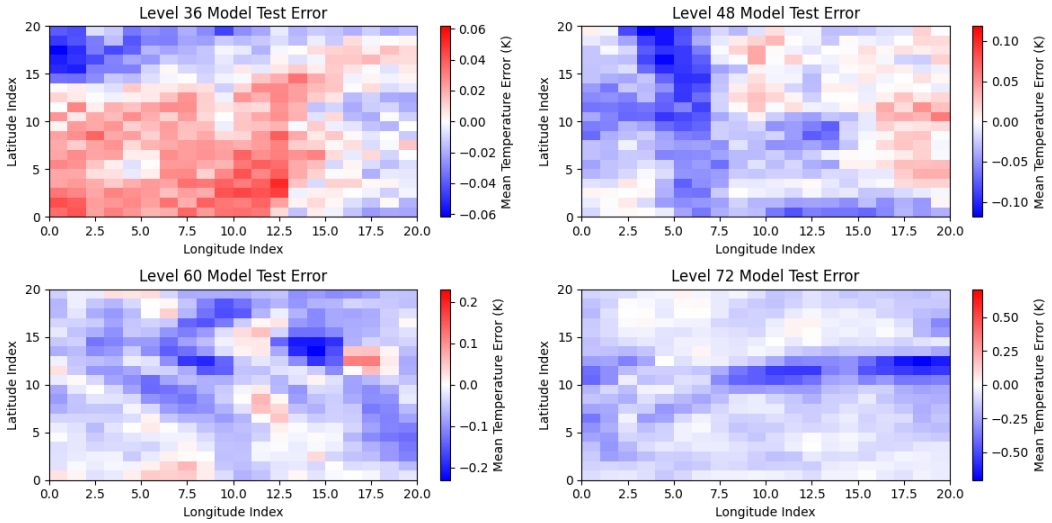

Figure 6: Predicted weather patterns compared with ground truth.

3.3 Final Results

The final trained CNN-Transformer model posted a total mean squared error performance on the test set of 0.0419 (Kelvin^2) after 10 training epochs, corresponding to an average 0.205° miss across the aforementioned test set. Further analyses were conducted to analyze performance across seasons (with latitudes >30° in the hemispheres); seasonal performance (in terms of summed MAE in Kelvin) is conveyed in Figure 7. As a key initial objective of the project was to train a single model capable of functioning across all global regions, we also conducted analyses of model performance on different test regions around the globe. Spherical coordinate-based sampling was implemented to prevent oversampling the poles. The results of this analysis—demonstrating the overall performance of the model across various test global regions—are depicted in Figure 8.

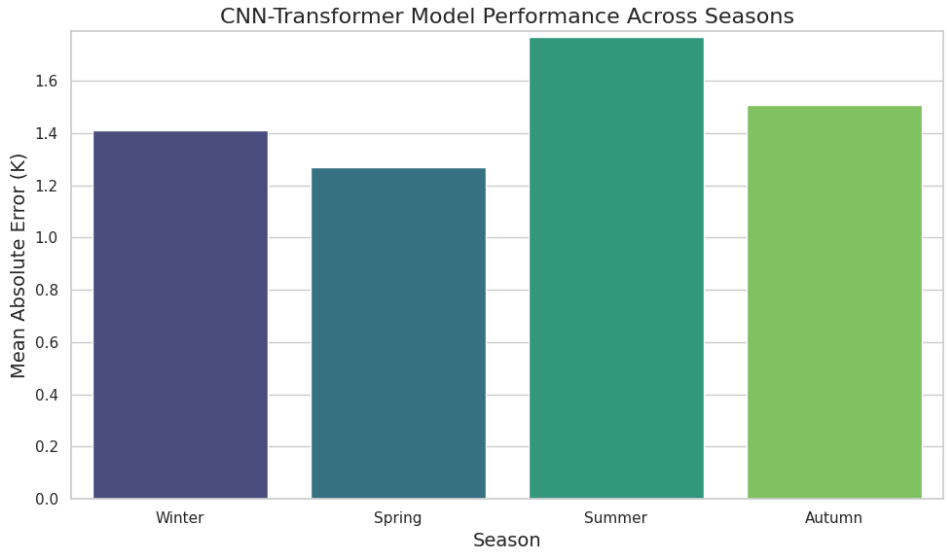

Figure 7: Seasonal model performance to assess anomalies.

4 Discussion

The integration of CNNs with time series transformer architectures in our proposed model demonstrates significant potential advantages for global weather prediction. The CNN component effectively captures spatial hierarchies and local patterns within the high-dimensional climate data, leveraging its ability to detect and encode spatial features through convolutional filters. This spatial encoding is crucial for weather prediction, where local atmospheric phenomena can have substantial impacts on broader climatic patterns. By incorporating CNNs, the model benefits from translational invariance and parameter sharing, which not only reduce computational complexity but also enhance the model's ability to generalize across different spatial regions; CNNs also confer possible advantages based on their close association with climate PDEs as mentioned in Section 2.

The time series transformer component complements the CNN by adeptly modeling temporal dependencies and long-range interactions within the weather data. Unlike traditional recurrent architectures, transformers use self-attention mechanisms that allow the model to weigh the significance of different time steps dynamically. This capability is particularly advantageous for capturing the intricate temporal dynamics inherent in weather systems, where events at disparate times can influence each other (evident in the performance delta between the RNN and CNN-Transformer in Figure 4). The hybrid CNN-Transformer architecture thus combines the strengths of both spatial feature extraction and temporal modeling, leading to improved predictive performance over models relying solely on either component.

Our model incorporates multiple atmospheric variables, including temperature (T), zonal wind component (u), meridional wind component (v), moisture content (Qv), and surface pressure (ps). The inclusion of these variables is designed to provide a more comprehensive representation of the atmospheric state—enabling the model to capture the multifaceted interactions driving weather patterns. By integrating these variables, we aim to construct a model with a holistic understanding of atmospheric conditions, thereby positively contributing to predictive accuracy.

As seen in Figure 7, CNN-Transformer model test performance was best in the winter and worst in the summer. These seasonal discrepancies in performance correspond with previous results (Ünal et al., 2023) and potentially with conventional intuition, which posits that the chaotic, small-scale, and convective nature of summer weather systems makes forecasting more challenging.

The implementation of halo regions with an optimal size of four significantly enhances the model's contextual understanding. Halo regions extend the input data beyond the primary $64 \times 64 \times 36$ cubes, providing additional spatial context that is crucial for accurate predictions at the boundaries. An empirical analysis determined that a halo size of four strikes an optimal balance between incor-

porating sufficient contextual information and managing computational overhead. This size aligns with the temporal spacing of the data, ensuring that the model captures relevant spatial dependencies without introducing excessive dimensionality. Interestingly, increasing the halo size beyond four did not yield performance improvements and even led to a slight increase in prediction error. This unexpected trend can be possibly attributed to both the inherent noise threshold present in sampling as well as the curse of dimensionality; adding more dimensions without corresponding data can dilute the model's ability to learn meaningful patterns. Additionally, larger halo regions may introduce noise and irrelevant information, which can negatively impact the model's performance by overwhelming the learning process with unnecessary complexity.

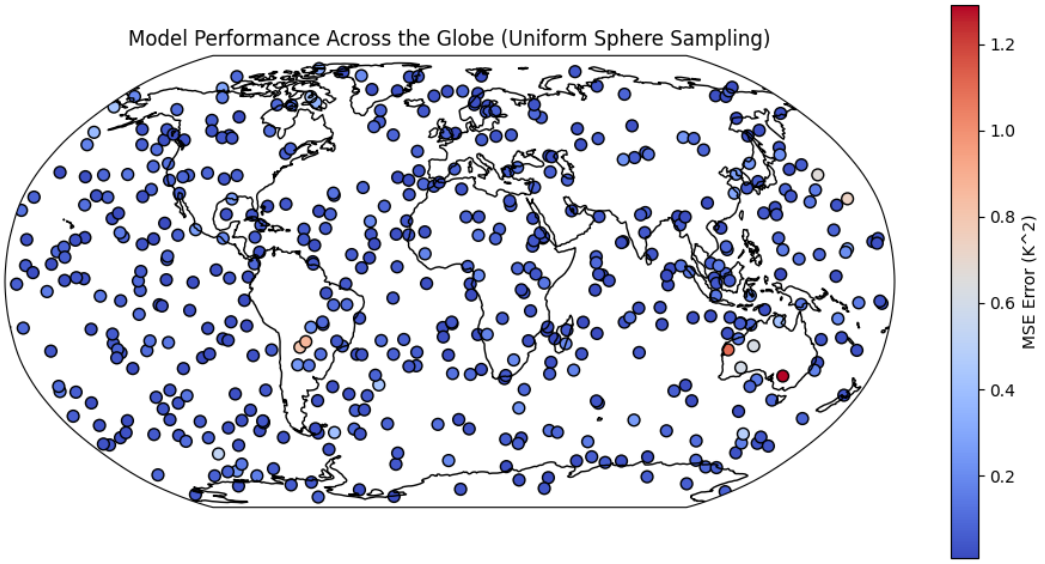

Figure 8: Model performance across global regions.

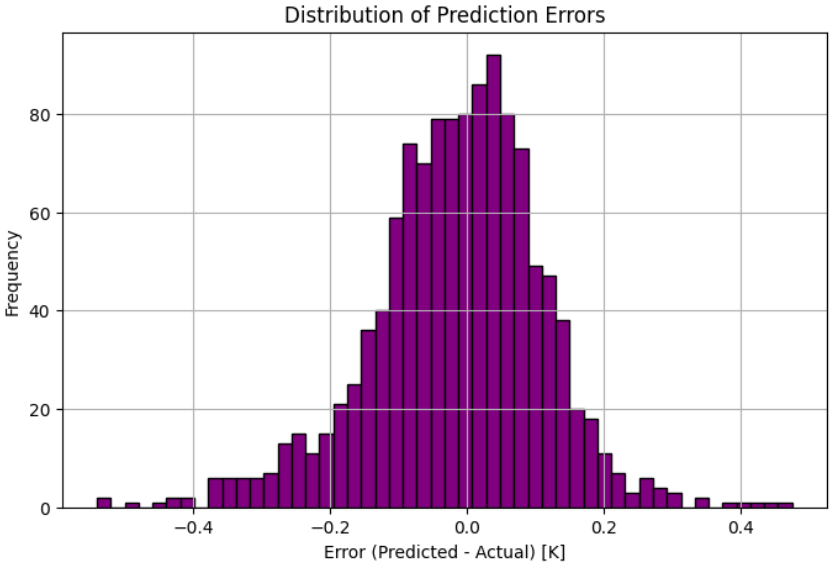

Figure 9: Model test performance distribution in tropical regions.

As mentioned previously, creating a single model capable of functioning across all global regions was a major goal of this project, and we observe fairly consistently strong performance in this regard

in Figure 8. Model performance is observed to be slightly worse over land as opposed to over water, with a few isolated incidents of poor performance in Oceania and South America. While perhaps further work needs to be done to ensure the model is capable of generalizing to more complex weather patterns over land, the model's overall performance can be characterized as consistent with minimal variance across hemispheres and vastly different climate regions; the empirical performance distribution across the tropics (Figure 9) is even tighter.

Future work may focus on several avenues to enhance the model's robustness and enable further generalization. One promising direction is the incorporation of additional atmospheric variables, such as humidity profiles and cloud cover, which could provide further insight into complex weather phenomena. Furthermore, expanding the temporal scope of the data to include longer sequences may enable the model to capture more extended temporal dependencies, potentially improving its forecasting capabilities. Another area of exploration involves the integration of ensemble learning techniques, where multiple models are trained and their predictions aggregated to reduce variance and improve overall accuracy. Validation of the model across diverse climatic regions and under varying weather conditions will also be essential to ensure applicability on a global scale. Additionally, investigating the interpretability of the model's predictions through techniques such as attention visualization may provide valuable insights into the underlying mechanisms driving model forecasts, fostering greater trust and transparency in future applications.

Acknowledgements

I would like to extend a sincere and heartfelt thanks to Arlindo da Silva and NASA GSFC for their invaluable sponsorship, accommodation, and support of this work.

References

[1] T. Anjali, K. Chandini, K. Anoop, and V. L. Lajish, *Temperature prediction using machine learning approaches*, In *2019 2nd International conference on intelligent computing, instrumentation and control technologies (ICICICT)*, vol. 1, pp. 1264-1268, IEEE, July 2019.

[2] B. Azari, K. Hassan, J. Pierce, and S. Ebrahimi, *Evaluation of machine learning methods application in temperature prediction*, *Environmental Engineering*, vol. 8, no. 1, pp. 1-12, 2022.

[3] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, *Accurate medium-range global weather forecasting with 3D neural networks*, *Nature*, 2023.

[4] M. H. K. Chan, W. K. Wong, and K. C. Au-Yeung, *Machine learning in calibrating tropical cyclone intensity forecast of ECMWF EPS*, *Meteorological Applications*, vol. 28, no. 6, pp. e2041, 2021.

[5] M. Chantry, H. Christensen, et al., *Opportunities and challenges for machine learning in weather and climate modelling: hard, medium, and soft AI*, *Philosophical Transactions of the Royal Society A*, 2021.

[6] S. Chen, T. Shu, H. Zhao, Y. Y. Tang, *MASK-CNN-Transformer for real-time multi-label weather recognition*, *Knowledge-Based Systems*, 2023.

[7] K. Chen, L. Bai, F. Ling, P. Ye, T. Chen, J. J. Luo, *Towards an end-to-end artificial intelligence driven global weather forecasting system*, *arXiv preprint arXiv:2306.01465*, 2023.

[8] S. Dewitte, J. P. Cornelis, R. Müller, A. Munteanu, *Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction*, *Remote Sensing*, 2021.

[9] S. Lang, M. Alexe, M. Chantry, J. Dramsch, et al., *AIFS—ECMWF's data-driven forecasting system*, *arXiv preprint arXiv:2406.01465*, 2024.

[10] S. K. Mukkavilli, D. S. Civitarese, J. Schmude, J. Jakubik, A. Jones, N. Nguyen, C. Phillips, S. Roy, S. Singh, C. Watson, R. Ganti, H. Hamann, U. Nair, R. Ramachandran, K. Weldemariam, *AI Foundation Models for Weather and Climate: Applications, Design, and Implementation*, *arXiv preprint arXiv:2309.10808*, 2023. https://doi.org/10.48550/arXiv.2309.10808.

[11] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, ... and A. Anandkumar, *FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators*, *arXiv preprint arXiv:2202.11214*, 2022.

[12] L. J. Slater, L. Arnal, M. A. Boucher, et al., *Hybrid forecasting: blending climate predictions with AI models*, *Hydrology and Earth System Sciences*, 2023.

[13] L. H. Thapa, P. E. Saide, J. Bortnik, M. T. Berman, A. da Silva, D. A. Peterson, et al., *Forecasting daily fire radiative energy using data driven methods and machine learning techniques*, *Journal of Geophysical Research: Atmospheres*, 129, e2023JD040514, 2024. https://doi.org/10.1029/2023JD040514.

[14] Ü. Ünal, A. Kaya, T. Altınsoy, et al., *Climate Model-Driven Seasonal Forecasting Approach with Deep Learning*, *Environmental Data Science*, 2023.

[15] A. Zeng, M. Chen, L. Zhang, and Q. Xu, *Are transformers effective for time series forecasting?*, In *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, pp. 11121-11128, June 2023.