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

New strides in machine learning (ML) and artificial intelligence (AI) have significantly influenced the field of weather forecasting in recent years. In this paper, we describe the development of a new ML-driven architecture for weather forecasting across the globe. 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, multi-layer perceptron-based, and recurrent neural network-based models. We also formulated our architecture after consolidating recent insights from leading AI-driven forecasting systems, emphasizing the progression towards hybrid models that adeptly combine spatial and temporal data processing. Our model incorporates halo regions for enhanced contextual understanding and employs 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 advancements in machine learning have paved the way for models capable of forecasting weather on both local and global scales. Advancing beyond conventional methodologies, recent now incorporate hybrid ML-driven architectures for more intelligent climate forecasting on everything from temperature to more complex phenomena like fire radiative energy (FRE). A notable study utilized machine learning methods, combining weather, fuel data, and topographic information to forecast 1- and 2-day FRE with significant accuracy, using a random forest approach. This model explained 48% of the variance in 1-day forecasts and demonstrated skill in predicting daily increases and decreases in FRE, outperforming persistence models, particularly during severe fire conditions (Thapa et al., 2024).

Another significant milestone in the ML climate 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) encoder and decoder with a sliding window transformer processor, trained on the ECMWF's ERA5 re-analysis and operational NWP data. This model demonstrated strong accuracy in medium-range forecasts, surpassing traditional numerical models in upper-air and surface parameter predictions. The AIFS approach notably demonstrated promising results with techniques such as attention-based graph processing and extensive parallelism, enabling efficient training on high-resolution input data.

The shift towards leveraging transformers and GNNs for weather prediction has been part of a broader trend highlighted in recent foundational research. Models like *AIFS* represent a new generation of AI-based systems capable of balancing computational efficiency with high accuracy. These

.

systems operate by emulating NWP processes through learned representations, achieving substantial performance improvements in anomaly correlation and root mean square error (RMSE) scores. The AIFS model, with its modular design and ensemble training capabilities, has set a new benchmark for AI-driven weather forecasting.

Further reinforcing this trend, recent reviews on AI foundation models for weather and climate (Mukkavilli et al., 2023) illustrate how transformers and neural operators can be integrated into comprehensive systems capable of handling varied spatiotemporal scales and downstream tasks. The review emphasizes the versatility and computational advantages of using foundational AI models for predictive tasks such as downscaling, pattern recognition, and forecasting extreme weather events like tropical cyclones and atmospheric rivers. These models have demonstrated the potential for efficient, large-scale deployment — achieving notable breakthroughs in training time and inference speed.

This evolving landscape is exemplified by the latest advancements in models that showcase the application of transformer-based architectures tailored to atmospheric data. Beyond their success in the natural language paradigm, transformers can handle both short- and long-range dependencies and trends, making them a natural choice for forecasting in a weather context. Combining transformers with architectures leveraging the spatial locality and translational invariance inherent in global climate forecasting has been shown to be potentially powerful for weather-based recognition tasks (Chen et al., 2023). The use of hierarchical temporal aggregation and Earth-specific priors has allowed these models to achieve results competitive with leading operational systems. This body of literature forms the basis for our model developed herein.

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 formulation behind key methods employed.

2.1 Dataset Exploration and Partitioning

The dataset utilized for this study 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 was 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).

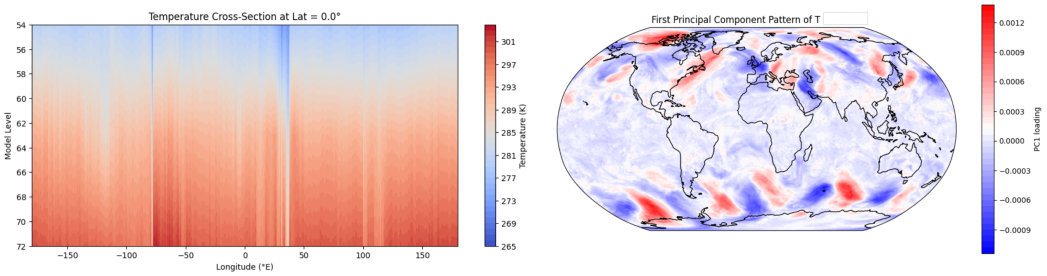

Figure 1: Visualizations collected during data exploration.

To prepare the dataset, we conducted an exploratory analysis that included the computation of statistical measures such as mean, variance, and covariance of each variable across spatial and vertical dimensions. This analysis informed our understanding of the data's distribution and highlighted areas with significant atmospheric interactions that would require enhanced modeling precision.

Given the relatively high dimensionality of the incoming data, 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) 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 the manifold structure of high-dimensional data.

Further exploration included Isomap, Locally Linear Embedding (LLE), and Multi-Dimensional Scaling (MDS), but these methods struggled with preserving the data's non-linear manifold properties.

Laplacian eigenmaps were ultimately adopted for their ability to construct a manifold learning-based embedding preserving 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. 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 enhancing key local spatial relationships between climate datapoints.

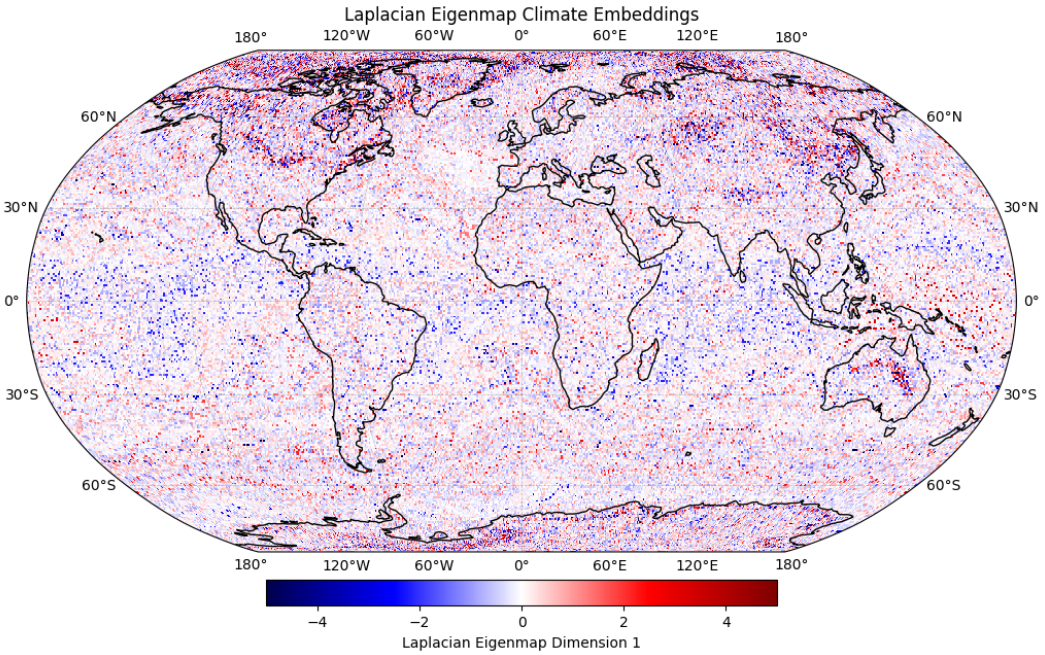

Figure 2: Rough 1D visualization of Laplacian eigenmap embeddings across different regions.

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 atmospheric conditions.

2.5 Transition to Neural Network Architectures

The inadequacies of statistical 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 model temporal dependencies, we transitioned to Recurrent Neural Networks (RNNs). 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 prowess in long-range dependency modeling, the computational cost of self-attention proved limiting for large-scale weather data.

2.6 CNN-Transformer Hybrid Architecture

The final architecture combined the 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.

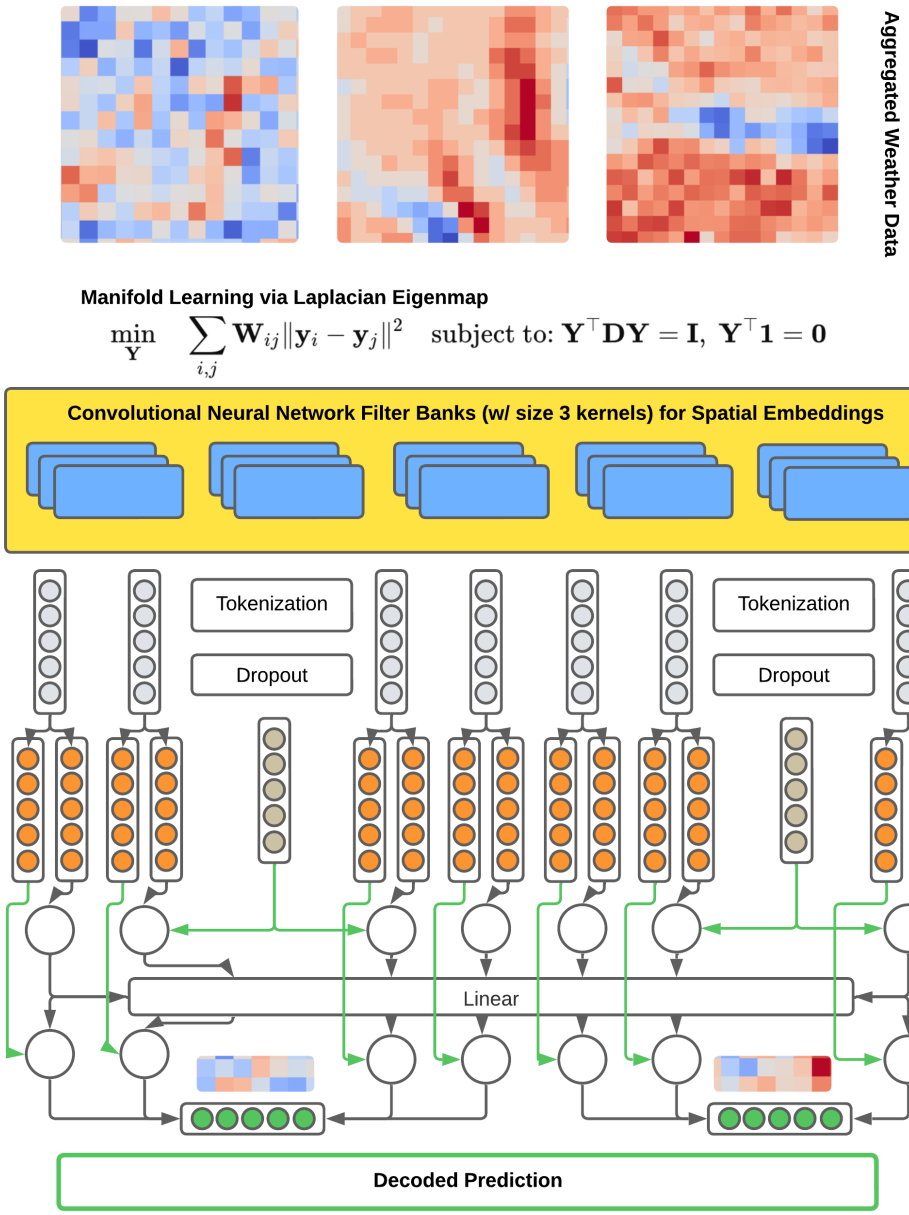

Figure 3: The CNN-Transformer model architecture.

3 Results

3.1 Model Evaluation

Figure 5 illustrates the performance comparison of the proposed CNN-Transformer hybrid against baseline methods. The CNN-Transformer outperformed other methods, 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 as evident in the gulf between pure transformer and CNN-transformer performance is clearly evident.

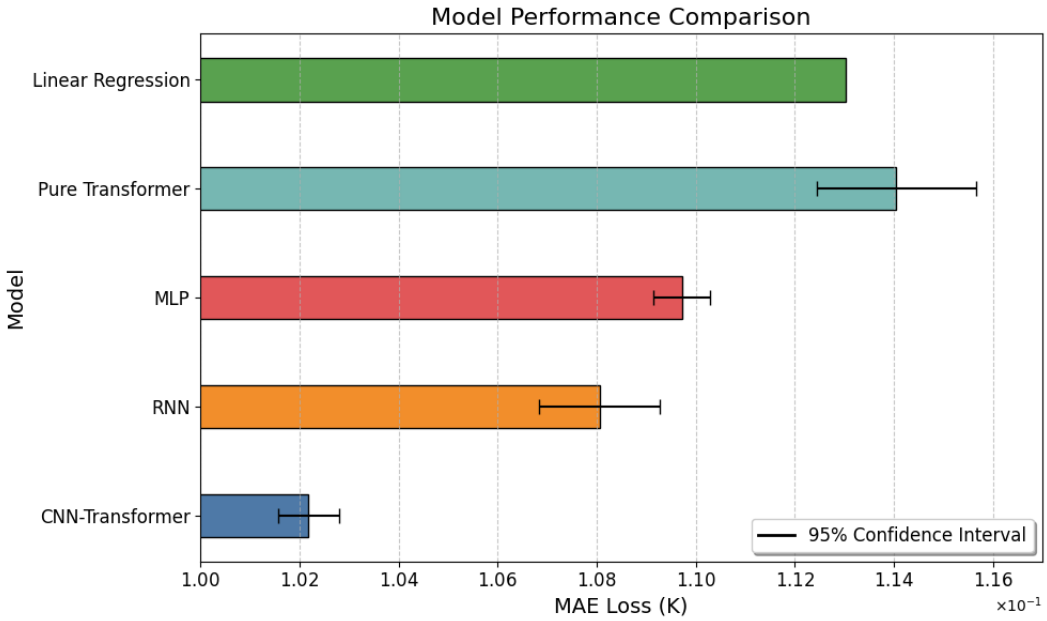

Figure 4: Model performance comparison.

We also conducted careful evaluation to determine the optimal halo size; the empirical optimal value of four aligns with the theoretical optimum given the spacing of frames in the data.

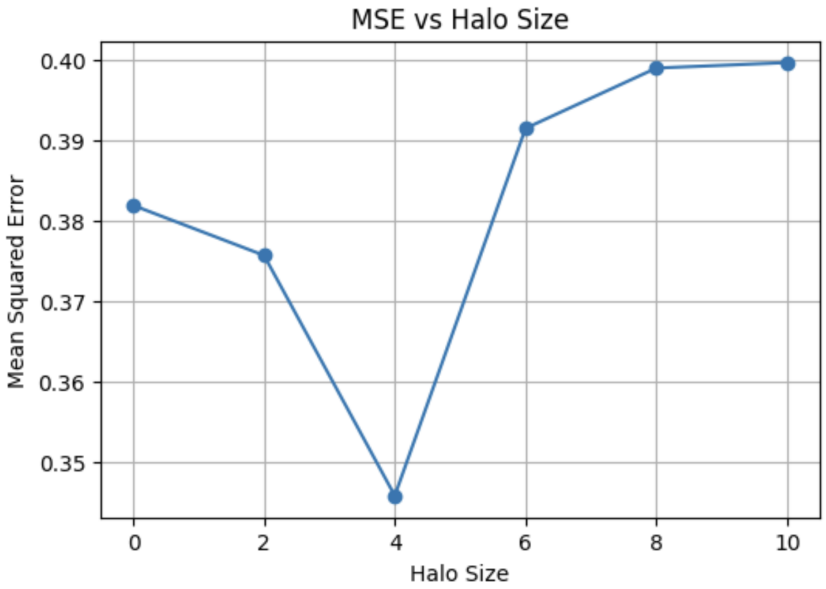

Figure 5: Optimal halo size.

3.2 Visualization

Figure 9 showcases the deviations of predicted weather patterns from ground truth data. The CNN-Transformer demonstrates relatively close agreement with observed data compared to baselines, accurately capturing spatiotemporal dynamics.

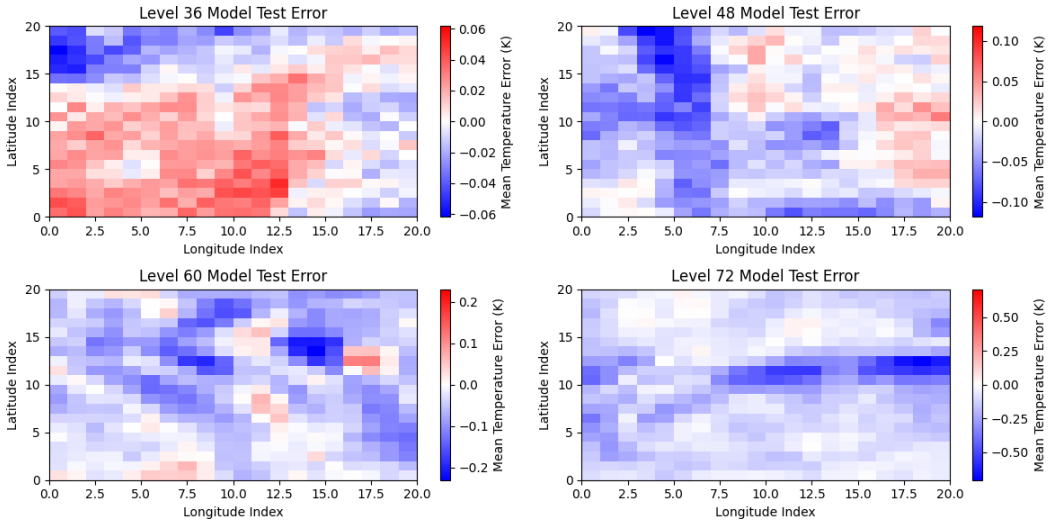

Figure 6: Predicted weather patterns compared with ground truth.

3.3 Final Results

The final trained CNN-Transformer model posted a total mean squared error performance on the test set of **0.041858** (Kelvin) 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—is depicted in Figure 8.

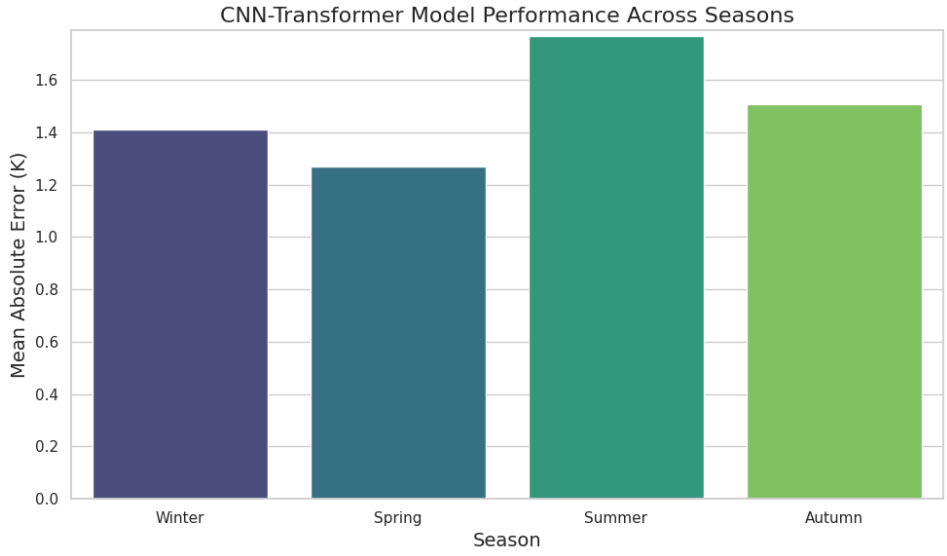

Figure 7: Seasonal model performance to assess anomalies.

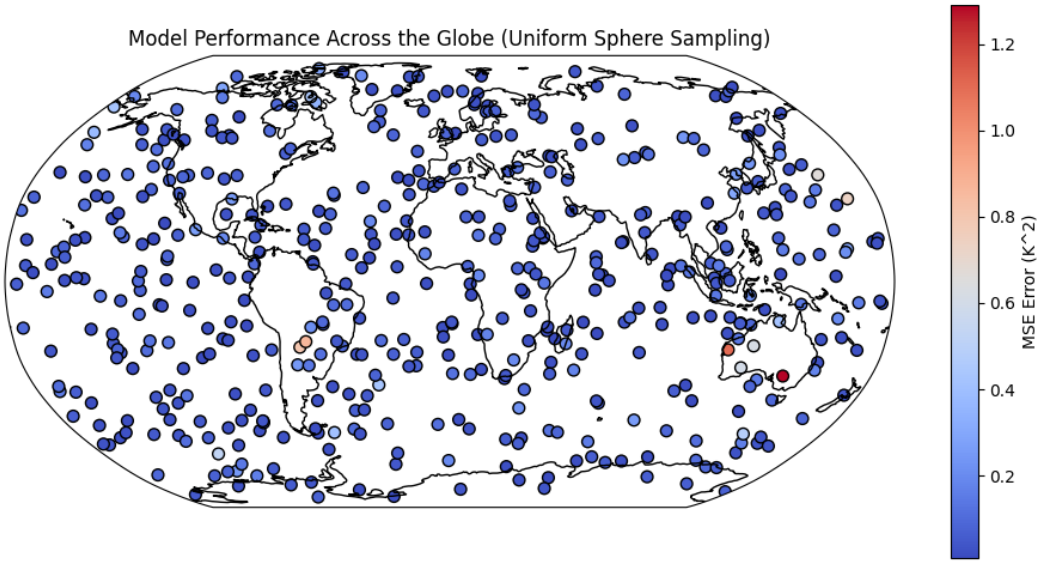

Figure 8: Model performance across global regions.

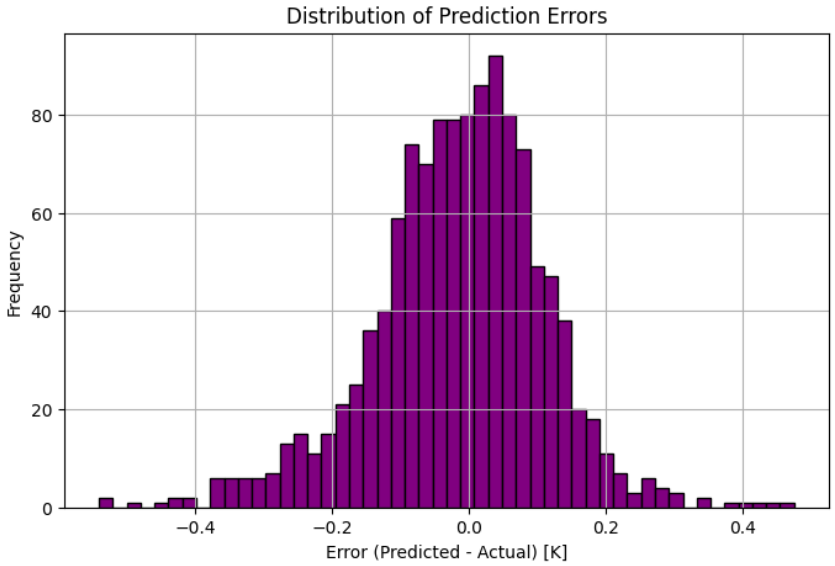

Figure 9: Model test performance distribution in tropical regions.

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 (observe 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 provides a more comprehensive representation of the atmospheric state — enabling the model to capture the multifaceted interactions that drive weather patterns. Each variable contributes unique information; for instance, wind components are essential for understanding air movement, while moisture content is critical for precipitation forecasting. By integrating these variables, the model gains a holistic understanding of the atmospheric conditions, thereby contributing to its 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 to previous results [10] and potentially to 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 incorporating 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.

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 performance distribution ($\sigma < 0.04$) across the tropics (Figure 9) is even tighter.

Future work may focus on several avenues to enhance the model's robustness and enable further generalizability. One promising direction is the incorporation of additional atmospheric variables, such as humidity profiles and cloud cover, which could provide further insights 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 its applicability on a global scale. Additionally, investigating the interpretability of the model's predictions through techniques such as attention visualization can provide valuable insights into the underlying mechanisms driving its forecasts, fostering greater trust and transparency in its 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] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, *Accurate medium-range global weather forecasting with 3D neural networks*, *Nature*, 2023.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

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

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

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