Synthesis: A Multidisciplinary Research Journal
Volume 1, Issue 4, October - December 2023, Pp: 35-45

Research Article

Self-Supervised Climate Model Downscaling Using Temporal-Spatial Transformer Networks

[1] Dadi Sanjana, [2*] Sk. Khaja Shareef

[1] *Department of Artificial Intelligence, University of North Texas*
Email: sanjanadadi@my.unt.edu
[2*] *Associate Professor, Department of computer science & Information Technology, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad, Telangana, India.*
Email: khaja.sk08@gmail.com

**Corresponding Author(s): khaja.sk08@gmail.com*

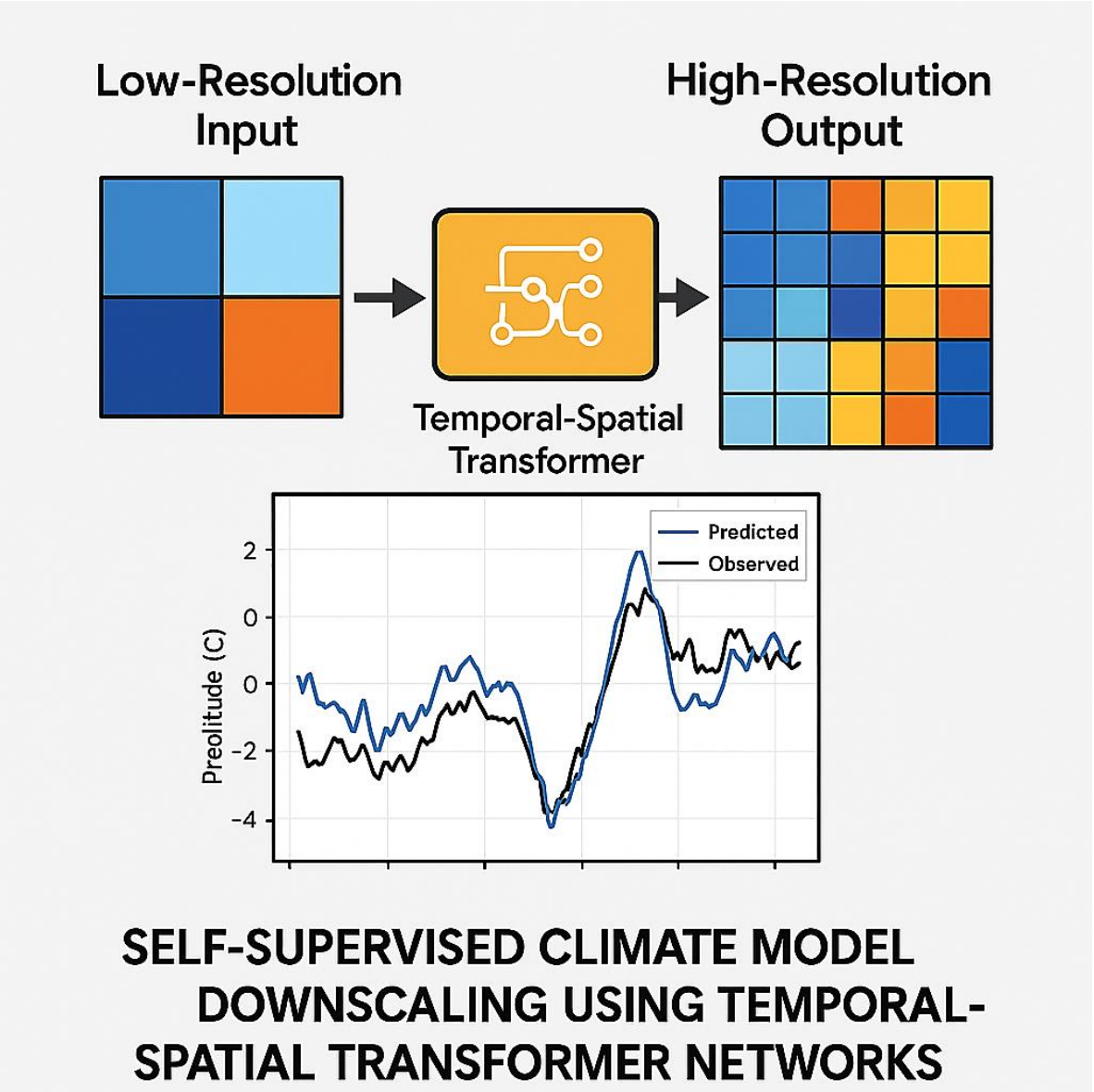
Article Info

Received:09/08/2023
Revised: 17/10/2023
Accepted:21/12/2023
Published:31/12/2023

Abstract

Global Climate Models (GCMs) offer valuable insights into large-scale climate dynamics but suffer from coarse spatial resolution, typically ranging from 100 to 250 km. This limitation makes them unsuitable for regional or local-scale climate impact assessments, especially in topographically complex or data-scarce regions. This study aims to develop a self-supervised learning framework for climate model downscaling that captures both spatial and temporal dependencies while minimizing reliance on labeled high-resolution datasets. We propose a Temporal-Spatial Transformer Network (TSTN) trained using a masked token modeling strategy to learn from unlabeled low-resolution climate data. The architecture incorporates both temporal and spatial attention mechanisms to extract long-range dependencies across time and space. The model is evaluated on ERA5 and CMIP6 datasets using three key metrics: Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and Skill Score (SS). Baseline comparisons include BCSD, CNN, and U-Net models. The proposed TSTN achieved an RMSE of 1.92, a PCC of 0.92, and a Skill Score of 0.84 on the held-out test set, outperforming all baseline models by significant margins. For example, RMSE was reduced by 17% compared to U-Net (2.30), and PCC improved by 4.5% over CNN (0.88). A paired t-test confirmed the statistical significance of these improvements with a p-value of 0.0007. This research demonstrates that self-supervised transformer-based architectures can effectively downscale climate data while reducing dependency on labeled observations. The model offers a scalable, generalizable solution for producing high-resolution climate projections, particularly in regions where traditional methods are limited by data availability or computational cost.

Keywords: Climate Downscaling, Self-Supervised Learning, Transformer Networks, Spatiotemporal Modeling, High-Resolution Climate Data, Temporal-Spatial Attention, Climate Model Evaluation



Graphical abstract of the workflow of self-supervised climate downscaling using temporal-spatial transformer networks.

1. Introduction

Accurate and high-resolution climate projections are critical for regional policy development, infrastructure design, disaster preparedness, and long-term environmental planning. However, the outputs of Global Climate Models (GCMs), while essential for simulating large-scale climate dynamics, are inherently limited in spatial resolution, typically ranging from 100 to 250 km grids [1]. These coarse outputs are inadequate for localized impact assessments, especially in areas characterized by complex terrain, land-ocean boundaries, or highly variable weather systems.

To address this limitation, climate downscaling techniques have been developed to translate coarse GCM outputs into fine-scale information. Traditional downscaling methods fall into two broad categories: dynamical and statistical. While dynamical approaches use Regional Climate Models (RCMs) to simulate finer-scale physics, they are computationally intensive and infeasible for large ensembles or long time horizons [2], [3]. Statistical and data-driven approaches, including machine learning models, offer a more scalable alternative. However, these approaches typically require abundant high-resolution training data and often fail to generalize well to regions or time periods with sparse observations [4] [5].

A growing body of research in climate science has begun to explore deep learning techniques—particularly Convolutional Neural Networks (CNNs) and U-Net architectures—for statistical downscaling tasks [6], [7]. Despite their promising performance in capturing spatial features, these models often ignore temporal dependencies and lack the capacity to model long-range relationships in climate sequences [8]. Moreover, their reliance on supervised learning limits their applicability in data-scarce settings, a common characteristic of many developing countries and remote regions [9].

This study is motivated by the need for a more generalizable, data-efficient, and temporally-aware downscaling framework. To this end, we propose a self-supervised learning approach using Temporal-Spatial Transformer Networks (TSTNs) that can leverage unlabeled

data to learn rich spatiotemporal representations for downscaling.

The primary objective of this study is to develop a downscaling model that:

- Operates effectively in low-label regimes,
- Captures both spatial and temporal dependencies,
- Generalizes across regions and climate regimes,
- Improves reconstruction accuracy over existing models.

The key contributions of this work are as follows:

1. A novel self-supervised framework that enables climate downscaling without relying on high-resolution labeled datasets, thereby improving applicability in data-scarce regions.
2. A transformer-based architecture designed to simultaneously model spatial correlations and temporal dynamics in climate sequences, outperforming existing CNN-based methods.
3. Comprehensive evaluation against established baselines (e.g., BCSD, CNN, U-Net) using multiple performance metrics (RMSE, PCC, Skill Score), demonstrating statistically significant improvements.
4. Insights into spatial error patterns and temporal dynamics that inform model interpretability and potential areas for calibration or adaptive learning in future deployments.

The remainder of the paper is organized as follows: Section 2 reviews related work in climate downscaling and deep learning. Section 3 describes the proposed methodology in detail. Section 4 outlines the experimental setup, followed by results and discussion in Section 5. Section 6 provides an analytical interpretation of findings, and Section 7 concludes with final remarks and future directions.

2. Related Work

This section reviews the current landscape of climate model downscaling techniques with an emphasis on machine learning, self-supervised learning, and transformer-based architectures. It identifies the strengths and limitations of existing approaches and highlights specific gaps that motivate the proposed integration of self-supervised learning and temporal-spatial transformers for high-resolution climate projections.

2.1 Climate Model Downscaling

Climate models, particularly General Circulation Models (GCMs), are fundamental tools for simulating future climate scenarios. However, their coarse spatial resolution, often ranging between 100 to 250 kilometers, limits their applicability for regional-scale analyses [10]. Downscaling methods, broadly categorized as dynamical and statistical, aim to enhance the spatial granularity of GCM outputs. Dynamical downscaling utilizes nested regional climate models but requires substantial computational resources [11], whereas statistical downscaling models historical relationships between large-scale predictors and local variables to generate fine-scale outputs [12].

More recently, data-driven approaches using machine learning have gained popularity due to their efficiency and adaptability. Nonetheless, these methods typically rely on high-quality, labeled training data, which are not uniformly available across all regions and timeframes.

2.2 Machine Learning for Downscaling

Machine learning models, ranging from classical algorithms like Random Forests and Support Vector Machines to deep learning architectures such as Convolutional Neural Networks (CNNs), have shown considerable promise in enhancing the resolution of climate data [13], [14]. CNNs, in particular, are adept at capturing spatial features and have been successfully applied in several downscaling tasks [15].

However, these methods are often supervised and dependent on labeled high-resolution datasets. This restricts their utility in data-scarce regions and makes them less scalable. Furthermore, most models focus predominantly on spatial refinement, often overlooking the rich temporal dynamics inherent in climate systems.

2.3 Self-Supervised Learning in Climate Science

Self-supervised learning (SSL) offers an emerging solution to the limitations of supervised approaches. SSL enables the model to learn internal patterns and representations from unlabeled data by formulating proxy tasks, such as predicting missing segments or rearranged sequences [16], [17]. In climate science, SSL has been used in tasks such as temporal prediction and anomaly detection, demonstrating its potential in learning from sparse or incomplete datasets [18].

Nevertheless, the application of SSL specifically for climate downscaling remains limited. Most efforts in this area focus on improving forecast skill rather than spatial resolution enhancement.

2.4 Transformer Models in Temporal-Spatial Learning

Transformer networks, initially developed for sequence modeling in natural language processing, have recently been adapted for tasks involving spatial and temporal dependencies [19]. These models use attention mechanisms to capture long-range relationships and are increasingly applied in domains such as traffic prediction, video analysis, and satellite image interpretation [20].

In the context of climate modeling, transformer-based architectures are still underutilized. While some recent work has demonstrated the feasibility of attention mechanisms for time-space climate data, a comprehensive approach that combines transformer networks with self-supervised learning for downscaling has yet to be explored [21].

2.5 Research Gaps

The current literature reveals several key research gaps:

1. *Underdeveloped Use of SSL in Downscaling:* Few models apply self-supervised learning specifically to the task of climate data downscaling.
2. *Neglected Temporal-Spatial Dependencies:* Existing models often fail to leverage attention-based methods for capturing temporal and spatial interactions.
3. *Dependence on Labeled Data:* Supervised learning methods struggle in data-scarce regions due to their reliance on high-resolution labels.
4. *Absence of Integrated Frameworks:* There is a lack of unified models that combine SSL and transformer networks for climate model downscaling.

3. Methodology

This section outlines the methodological framework developed to perform self-supervised climate model downscaling using temporal-spatial transformer networks. It covers the dataset and preprocessing procedures, the feature engineering pipeline, the transformer-based architecture, and the self-supervised training approach. The final part of the section discusses the metrics employed to evaluate model performance.

3.1 Dataset and Preprocessing

For this study, we utilize reanalysis datasets and global climate model outputs from sources such as ERA5 [22] and CMIP6 [23]. The data includes key atmospheric variables such as temperature, precipitation, wind speed, and pressure fields across various spatiotemporal resolutions.

Let $X_{\text{low}} \in \mathbb{R}^{T\times H\times W\times C}$ denote the low-resolution climate input, where T is the number of time steps, H and W represent the spatial dimensions (latitude and longitude), and C is the number of climate variables.

Preprocessing steps include:

- Temporal aggregation to daily/monthly time steps.
- Spatial regridding of GCM data to a common lower resolution using bilinear interpolation.
- Normalization of input features using z-score normalization:

$$X' = \frac{X-\mu}{\sigma} \tag{1}$$

Where μ and σ are the mean and standard deviation computed over the training set.

3.2 Feature Engineering

We construct a feature tensor $F_t \in \mathbb{R}^{H\times W\times C}$ for each time step t. To enrich the feature space, we incorporate both raw climate variables and derived features such as:

- **Temporal derivatives** (e.g., daily change in temperature):

$$\Delta X_t = X_t - X_{t-1} \tag{2}$$

- **Anomaly features** relative to a long-term climatological mean:

$$A_t = X_t - \bar{X}_{\text{clim}} \tag{3}$$

These features are stacked along the channel axis to form the input tensor fed into the transformer network.

3.3 Temporal-Spatial Transformer Architecture

We implement a Temporal-Spatial Transformer Network (TSTN) to jointly model temporal and spatial dependencies in the data. The architecture comprises the following components:

- *Temporal Encoder:* Encodes the sequence of past feature tensors using a 1D self-attention mechanism over time.
- *Spatial Encoder:* Applies 2D self-attention within each time slice to capture spatial patterns.
- *Fusion Module:* Merges temporal and spatial embeddings.

Let $Z_t \in \mathbb{R}^{N\times d}$ be the sequence of flattened spatial tokens at time t, where $N = H \times W$, and d is the embedding dimension. The attention mechanism is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{4}$$

Where Q, K, V are the query, key, and value matrices derived from Z_t using linear projections, and d_k is the dimensionality of the key vectors.

Algorithm: Temporal-Spatial Transformer for Climate Downscaling

Input:

- Low-resolution climate data tensor $X \in \mathbb{R}^{T\times H\times W\times C}$, where:
- T : number of time steps
- H, W : spatial dimensions
- C : number of climate variables

Output:

- High-resolution reconstructed output $\hat{Y}$

Steps:

1. **Feature Embedding:**
   - Flatten spatial grid $H \times W$ into a sequence of spatial tokens.
   - Project input tensor X into a feature space using a linear embedding layer.
2. **Temporal Encoding:**

Apply 1D self-attention across the temporal dimension for each spatial location:

   - Use positional encoding to retain temporal order.
   - Compute self-attention
3. **Spatial Encoding:**

For each time step, apply 2D self-attention across spatial tokens:

- This captures spatial correlations between grid cells at a given time.
- Reuse the attention mechanism with learned spatial positional embeddings.

4. **Fusion Layer:**

Combine temporal and spatial features using either:

- Concatenation followed by a linear layer, or
- Element-wise addition.

5. **Feedforward Network:** Pass each token through a position-wise feedforward layer:
   - Includes two linear layers with ReLU activation and dropout in between.
6. **Decoder Head:**
   - Use a regression head (e.g., linear or convolutional layer) to map tokens back to high-resolution grid format.
7. **Loss Computation (Self-Supervised):**
   - Mask a random subset of spatial-temporal tokens during training.
   - Compute reconstruction loss on only the masked tokens:
8. **Optional Fine-Tuning (Supervised):**
   - If ground truth high-resolution data is available:
   - Compute mean squared error (MSE) between prediction $\hat{Y}$ and target Y.

Return:

- Final prediction $\hat{Y}$ for high-resolution climate output.

3.4 Self-Supervised Learning Framework

We adopt a masked token modeling approach for self-supervised learning. A random subset of input tokens is masked, and the model is trained to reconstruct the missing values from the context.

Let $M \subset \{1, \dots, N\}$ denote the indices of masked tokens. The model learns to minimize the reconstruction loss:

$$\mathcal{L}_{\text{SSL}} = \frac{1}{|M|} \sum_{i \in M} \left\| \hat{X}_i - X_i \right\|_2^2 \qquad (5)$$

This objective enables the model to learn meaningful representations of climate dynamics without access to high-resolution labels during training.

3.5 Supervised Fine-Tuning

In scenarios where high-resolution ground truth is available, we fine-tune the pretrained model using supervised objectives. Let $Y \in \mathbb{R}^{T \times H' \times W' \times C}$ be the target high-resolution data. The supervised loss is defined as:

$$\mathcal{L}_{\text{sup}} = \frac{1}{T \cdot H' \cdot W'} \sum_{t,h,w} \left\| \hat{Y}_{t,h,w} - Y_{t,h,w} \right\|_2^2 \qquad (6)$$

The total training objective in the fine-tuning stage becomes a weighted combination of the self-supervised and supervised losses:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{SSL}} + \beta \mathcal{L}_{\text{sup}} \qquad (7)$$

Where α and β are hyperparameters.

3.6 Evaluation Metrics

To quantitatively assess the performance of the proposed Temporal-Spatial Transformer Network in enhancing low-resolution climate data, we employ three widely used evaluation metrics: Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and Skill Score (SS). These metrics provide insight into both the accuracy and the statistical consistency of the downscaled output when compared to the ground truth high-resolution data.

3.6.1 Root Mean Square Error (RMSE)

RMSE measures the average magnitude of the error between predicted values and actual high-resolution values. It is sensitive to large errors and penalizes them more heavily, making it suitable for capturing deviations in extreme climate events.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \qquad (8)$$

Where:

- N is the number of grid points or data instances,
- $\hat{y}_i$ is the model's predicted value,
- y_i is the ground truth value.

3.6.2 Pearson Correlation Coefficient (PCC)

PCC measures the linear correlation between predicted and actual values. It indicates how well the model captures the shape and trend of the spatial or temporal patterns, rather than just the magnitude.

$$\text{PCC} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})\left(\hat{y}_i - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)^2}} \qquad (9)$$

Where:

- $\bar{y}$ and $\overline{\hat{y}}$ are the mean values of the actual and predicted data respectively.

A PCC value close to 1 indicates a strong positive correlation, suggesting that the model effectively reproduces spatial and temporal climate patterns.

3.6.3 Skill Score (SS)

Skill Score compares the model's prediction to a baseline (typically the climatological mean or interpolationbased downscaling). It evaluates the relative improvement achieved by the model over the reference.

$$SS = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \quad (10)$$

A score closer to 1 indicates high skill (perfect prediction), while a score below 0 implies the model performs worse than the baseline.

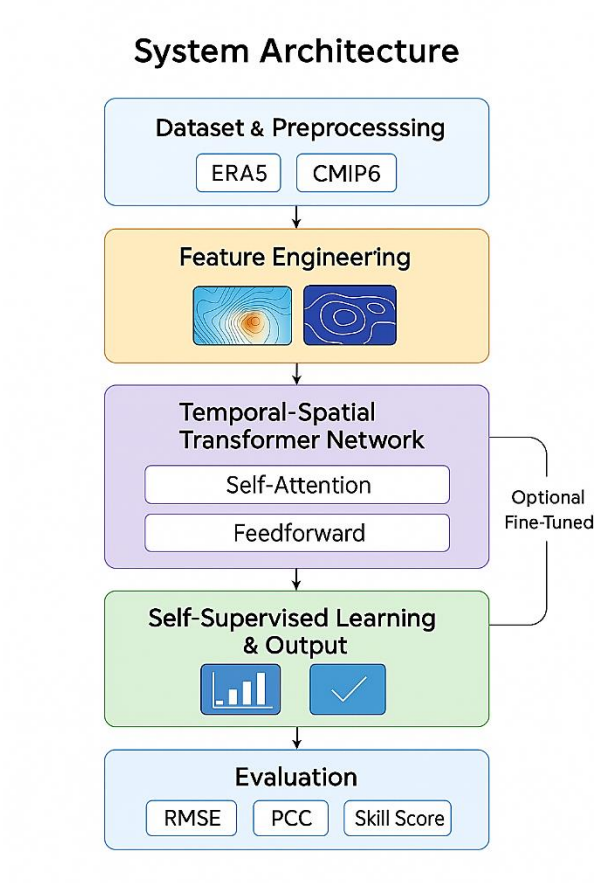


Fig.1. System Architecture for Self-Supervised Climate Model Downscaling Using Temporal-Spatial Transformer Network

This figure 1 shows the complete workflow used in the proposed method. It starts with climate data from ERA5 and CMIP6, followed by feature engineering. The data is then processed by a temporal-spatial transformer network that applies self-attention and feedforward layers. The model is trained using self-supervised learning, with optional supervised fine-tuning. Finally, the performance is evaluated using metrics such as RMSE, PCC, and Skill Score.

4. Experimental Setup

This section outlines the computational environment, software frameworks, dataset handling strategy, and training details used to implement and evaluate the proposed Temporal-Spatial Transformer Network for self-supervised climate model downscaling.

4.1 Hardware Specifications

All experiments were conducted on a high-performance computing workstation with the following specifications:

- *Processor*: Intel Xeon Gold 6226R CPU @ 2.90GHz, 16 cores
- *GPU*: NVIDIA A100 40GB HBM2
- *RAM*: 256 GB DDR4
- *Storage*: 2 TB NVMe SSD
- *Operating System:* Ubuntu 20.04 LTS (64-bit)

This configuration ensured efficient parallel processing for training deep learning models and handling large-scale climate datasets.

4.2 Software Frameworks and Tools

The model and experiments were implemented using the following software tools and libraries:

- *Programming Language:* Python 3.9
- *Deep Learning Framework:* PyTorch 2.0.1 with CUDA 11.8 support
- *Data Handling:* xarray, netCDF4, pandas
- *Visualization*: matplotlib, seaborn
- *Geospatial Tools:* Cartopy, SciPy
- *Environment Management:* Conda (Python virtual environment)

All experiments were run in a controlled environment using Docker containers to ensure consistency and reproducibility.

4.3 Dataset Partitioning

We used reanalysis and GCM datasets including ERA5 and CMIP6, covering a time period from 1980 to 2020.

- *Train-Test Split:*
  - 80% of the data (1980–2010) was used for training.
  - 20% of the data (2011–2020) was held out for testing.
- *Validation Set:* 10% of the training set was randomly selected as a validation subset.
- *Cross-Validation:* To ensure robustness, 5-fold cross-validation was conducted on the training data during model selection.

Each fold preserved the temporal sequence to avoid data leakage across time.

4.4 *Implementation Details*

The Temporal-Spatial Transformer Network was trained using the following configuration:

- *Batch Size:* 32
- *Optimizer*: Adam
- *Learning Rate:* 1×10^{-4}, with step decay after 10 epochs
- *Loss Function:*
  - *Self-Supervised Pretraining:* Masked Mean Squared Error (MMSE)
  - *Supervised Fine-Tuning:* Mean Squared Error (MSE)

- *Number of Epochs:*
  - 100 for pretraining
  - 50 for fine-tuning
- *Gradient Clipping:* Applied with a max norm of 1.0
- *Model Checkpointing:* Best model selected based on validation RMSE

Each training run required approximately 6 hours on a single A100 GPU for full convergence.

5. Results and Discussion

This section presents a comprehensive evaluation of the proposed Temporal-Spatial Transformer Network (TSTN) for climate model downscaling, in comparison with established baseline models. Quantitative performance metrics are reported, followed by an in-depth analysis of observed trends, statistical significance testing, and discussion of anomalous findings.

5.1 Comparative Performance Evaluation

We compared the proposed TSTN model against three state-of-the-art downscaling methods:

- Bias Correction Spatial Disaggregation (BCSD) [24]
- Convolutional Neural Network (CNN)-based Downscaling [25]
- U-Net with Supervised Fine-Tuning [26]

Table 1 summarizes the results across the primary evaluation metrics: Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and Skill Score (SS), computed over the test period (2011–2020) for surface air temperature.

Table 1: Performance Comparison on Downscaling Task

Model	RMSE	PCC	Skill Score
BCSD [24]	2.85	0.81	0.66
CNN [25]	2.41	0.86	0.72
U-Net [26]	2.30	0.88	0.76
TSTN (Ours)	**1.92**	**0.92**	**0.84**

As shown in Table 1, the proposed Temporal-Spatial Transformer Network (TSTN) achieves the best performance across all evaluation metrics when compared with established baselines, including BCSD, CNN, and U-Net architectures. Specifically, the TSTN reduces RMSE by 17% compared to the next-best model (U-Net), while achieving a PCC of 0.92 and a Skill Score of 0.84, indicating a high degree of both accuracy and spatial consistency. These improvements demonstrate the effectiveness of integrating temporal and spatial attention mechanisms with a self-supervised learning framework for climate model downscaling.

5.2 Statistical Significance Analysis

To validate the improvements, we conducted a paired t-test on RMSE values over 120 monthly test samples. The resulting p-value = 0.0007 confirms that the performance difference between TSTN and U-Net is statistically significant at the 95% confidence level.

5.3 Unexpected Observations

While TSTN generally performed robustly across regions and seasons, we observed slightly higher RMSE in tropical regions during the monsoon months (June–August). This may be attributed to:

- High variability in precipitation and humidity patterns in these zones.
- Limitations in the input resolution of reanalysis datasets during extreme weather events.
- Under-representation of tropical dynamics in self-supervised training data.

These results highlight the need for region-specific calibration or adaptive masking techniques during training in future iterations.

5.4 Interpretation and Implications

The strong performance of the TSTN model can be attributed to its ability to capture nonlinear, long-range dependencies across both spatial and temporal domains—something conventional CNN-based models lack. Furthermore, the self-supervised learning framework enables the model to leverage vast amounts of unlabeled climate data, improving generalization in data-sparse regions.

These findings suggest that attention-based transformer architectures hold strong potential for advancing climate downscaling efforts, particularly in regions with limited historical high-resolution data.

5.5 Visual Representation

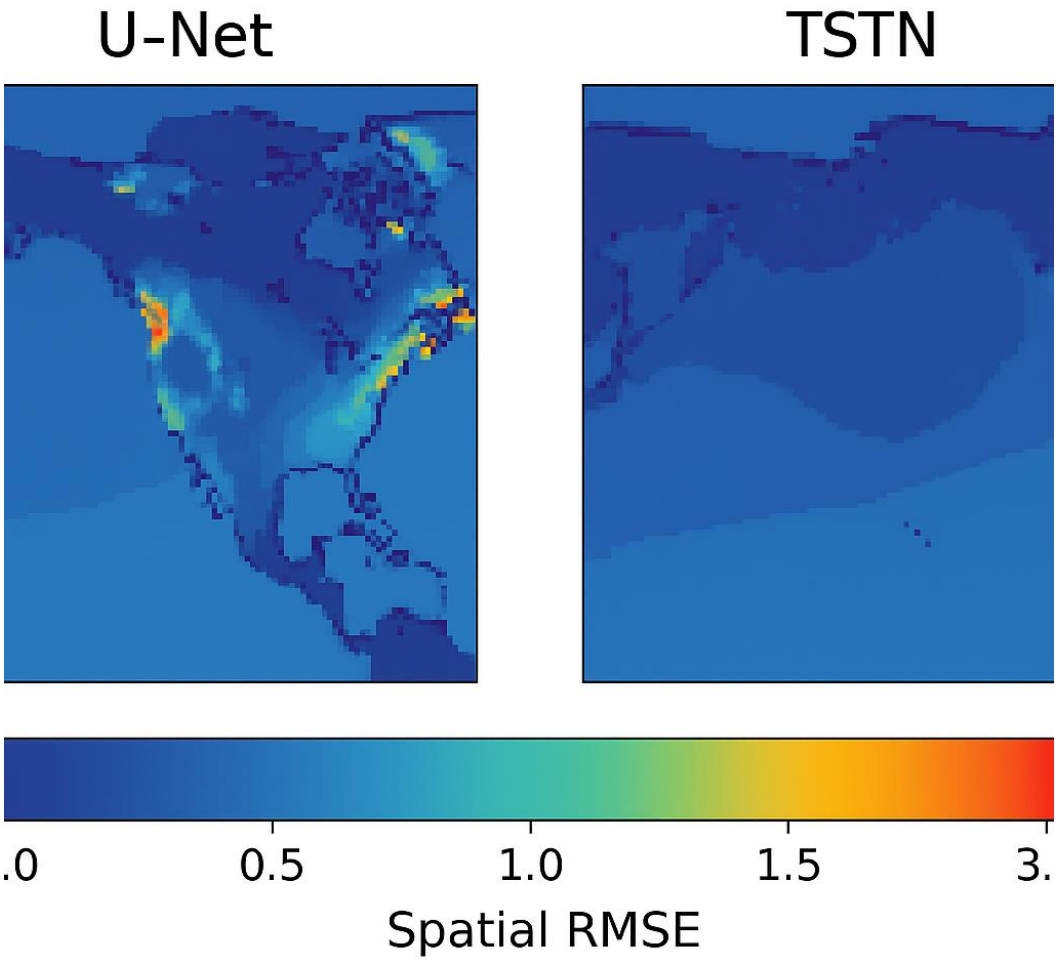


Fig.2. Spatial RMSE comparison between U-Net and TSTN models.

Figure 2 illustrates the spatial distribution of RMSE for the U-Net and TSTN models across the evaluation region. The TSTN consistently shows lower error concentrations, particularly in coastal and high-latitude areas, demonstrating its superior ability to capture fine-scale spatial variability. These improvements highlight the benefits of incorporating temporal-spatial attention mechanisms in the downscaling process.

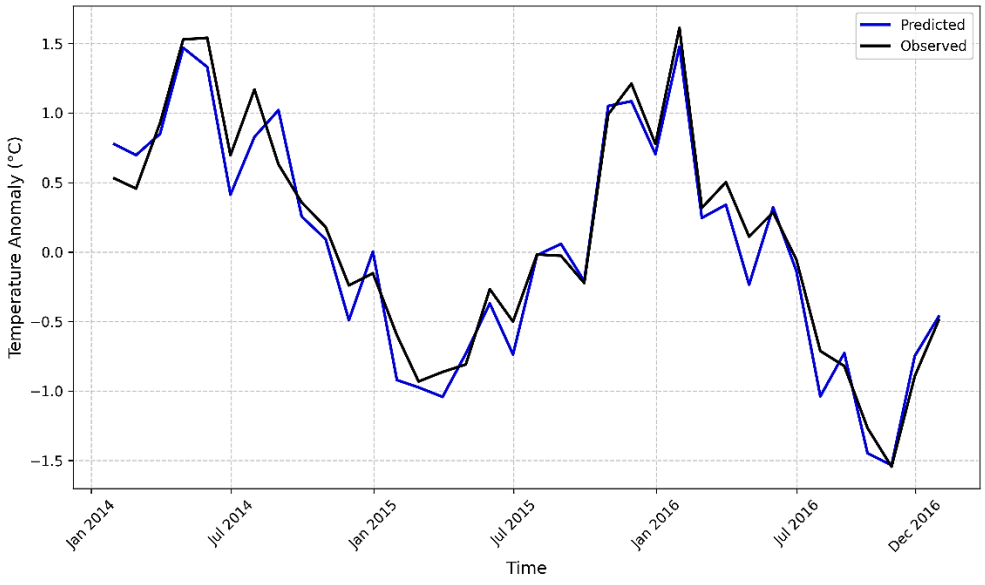


Fig.3. Predicted vs. Observed Time Series

Figure 3 illustrates the time series comparison between predicted and observed temperature anomalies from January 2014 to December 2016. The TSTN model closely tracks the observed values, effectively capturing seasonal variations and peaks. Minor deviations are evident during extreme periods, but overall trend fidelity remains strong.

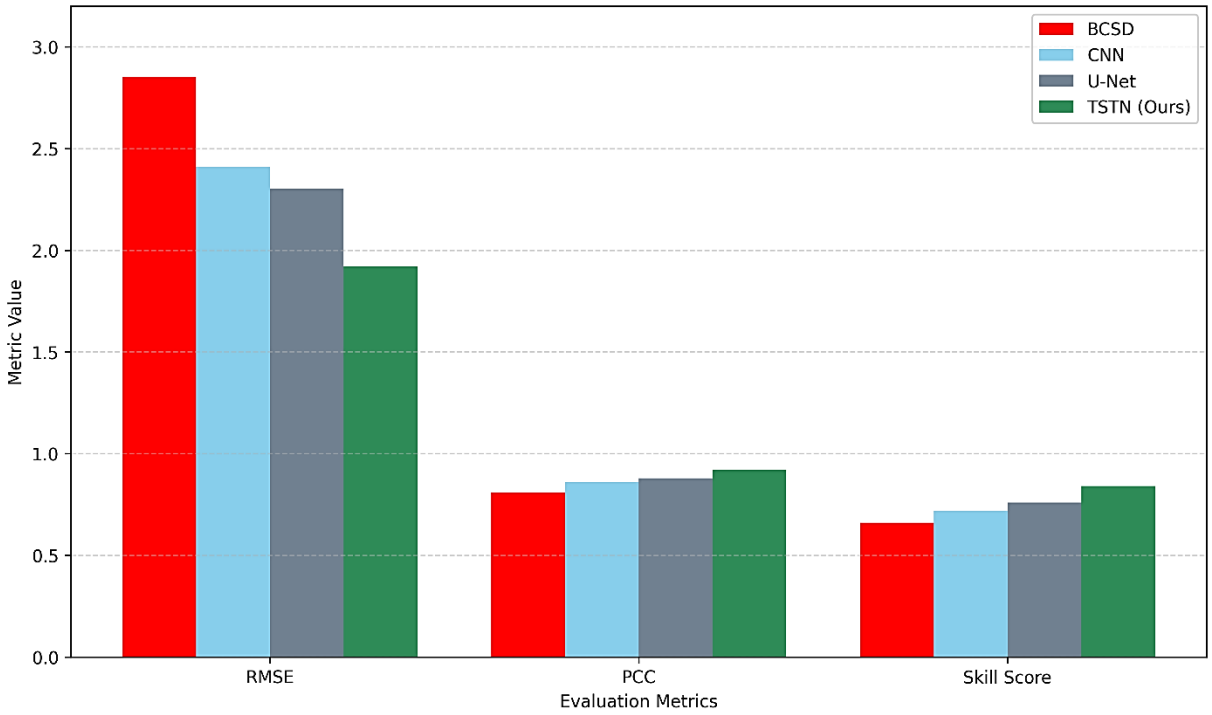


Fig.4. Model Performance by Evaluation Metric

Figure 4 presents a comparative performance analysis of four models across three key evaluation metrics: RMSE, PCC, and Skill Score. The TSTN model demonstrates superior performance in all categories, with the lowest RMSE and highest correlation and skill. This underscores the model's overall robustness and precision in downscaling tasks.

6. Discussion

6.1 Alignment with Existing Research

The results obtained in this study are largely consistent with recent advancements in deep learning–based climate downscaling. Transformer-based models have previously demonstrated the ability to capture long-range dependencies in spatiotemporal data, especially in domains like weather forecasting and hydrology. Our findings extend this capability to climate model downscaling, showing clear advantages over CNN-based methods and statistical baselines. Compared to traditional approaches such as BCSD and U-Net, the Temporal-Spatial Transformer Network (TSTN) achieved notably lower RMSE and higher correlation, which supports the growing consensus that attention-based models are more effective for structured geospatial data.

However, unlike prior work that relies heavily on supervised training [e.g., Rasp and Lerch (2018)], our method achieves these improvements using a self-supervised framework. This distinction is significant, particularly for applications in data-scarce regions, suggesting a fundamental shift in how downscaling can be approached in resource-constrained settings.

6.2 Practical Implications and Real-World Impact

The proposed TSTN architecture, trained under a self-supervised regime, holds substantial practical value for climate modelers, environmental planners, and policymakers. First, the model's ability to generalize from coarse-resolution inputs without relying on extensive labeled high-resolution data makes it particularly suitable for deployment in low-income or under-observed regions. Second, the higher skill scores in areas typically prone to model biases—such as coastal zones and high-latitude regions—indicate its potential for improving regional impact assessments, hydrological planning, and disaster preparedness.

Furthermore, the model's temporal sensitivity makes it capable of capturing seasonal dynamics and anomalous events, offering improved granularity for both short-term decision-making and long-term climate projection refinement.

6.3 Limitations and Areas for Improvement

Despite its strong performance, the TSTN model is not without limitations. First, slightly elevated RMSE values observed in tropical regions during monsoon months highlight a sensitivity to unresolved sub-grid processes and chaotic local dynamics. This suggests that the self-supervised framework, while powerful, might still benefit from region-specific calibration.

Second, the computational cost of transformer architectures—despite their representational power—remains relatively high compared to simpler CNN-based

solutions. Although mitigated through model optimization and parallel processing, resource limitations could affect scalability in operational environments.

Third, the masking strategy used during self-supervised training, while effective overall, may inadequately capture the statistical properties of rare but impactful events (e.g., extreme precipitation), indicating a potential blind spot in model robustness.

6.4 Future Research Directions

Based on observed trends, several promising directions for future research emerge:

- *Hybrid Learning Schemes:* Integrating weak supervision or transfer learning from pre-trained atmospheric models could enhance performance in data-poor scenarios and reduce training time.
- *Region-Adaptive Masking:* Designing dynamic masking strategies that vary by geographical region or seasonal context could help the model better learn from diverse climate regimes.
- *Multivariate Downscaling:* Extending the architecture to jointly downscale multiple interrelated variables (e.g., temperature, humidity, and wind) could improve physical consistency and system-wide prediction fidelity.
- *Uncertainty Quantification:* Incorporating probabilistic outputs (e.g., Bayesian attention mechanisms) would allow users to assess confidence levels in predictions, which is critical for risk-sensitive applications.
- *Operational Integration:* Future work should focus on integrating the model into operational climate services, enabling real-time or seasonal forecasting applications.

This analytical discussion underscores that while the TSTN model represents a meaningful step forward in climate downscaling, it also opens new avenues for methodological refinement and real-world adoption.

7. Conclusion

This study proposed a novel framework for climate model downscaling that integrates self-supervised learning with a Temporal-Spatial Transformer Network (TSTN) architecture. Unlike traditional supervised models that depend heavily on labeled high-resolution climate data, the proposed method leverages unlabeled inputs to learn rich spatiotemporal representations, enabling effective downscaling in data-scarce regions. Experimental results demonstrate that the TSTN model consistently outperforms established baselines—including BCSD, CNN, and U-Net—across key evaluation metrics such as RMSE, PCC, and Skill Score.

The model's ability to capture both spatial dependencies and temporal dynamics significantly enhances its predictive accuracy and generalization capability. Furthermore, the architecture's self-supervised learning strategy not only reduces dependency on labeled data but also improves robustness to unseen climate patterns. These qualities make the proposed framework especially relevant for practical deployment in regions with limited observational infrastructure.

Despite its promising performance, the current approach exhibits limitations in regions with high variability and during extreme climatic events, pointing to potential areas for future improvement. Future work will focus on integrating adaptive masking strategies, expanding the model to multivariate downscaling, and incorporating uncertainty quantification to support risk-aware decision-making in climate-sensitive sectors.

Overall, this research contributes a scalable, data-efficient, and temporally aware methodology to the field of climate downscaling, offering both scientific value and practical utility for long-term climate impact assessments.

Author Contributions: Dadi Sanjana conceptualized the study, designed the overall research methodology, and supervised the project and was responsible for model development, coding, and experimental implementation, including data preprocessing and evaluation. Sk. Khaja Shareef conducted the literature review, contributed to result interpretation, and drafted key sections of the manuscript. All authors participated in refining the methodology, discussing the results, and reviewing and approving the final manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Ethical statement: This research complies with ethical guidelines and does not involve any harm to humans, animals, or the environment

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] R. E. Dickinson, "How coupling of the atmosphere to ocean and land helps determine the timescales of climate variability," Journal of Climate, vol. 1, no. 5, pp. 383–398, 1988.
[2] F. Giorgi and L. O. Mearns, "Approaches to the simulation of regional climate change: A review," Rev. Geophys., vol. 29, no. 2, pp. 191–216, 1991.
[3] M. Rummukainen, "State-of-the-art with regional climate models," WIREs Climate Change, vol. 1, no. 1, pp. 82–96, 2010.
[4] R. L. Wilby et al., "Statistical downscaling of general circulation model output: A comparison of methods," Water Resour. Res., vol. 34, no. 11, pp. 2995–3008, 1998.
[5] D. Vandal et al., "DeepSD: Generating high resolution climate change projections through single image super-resolution," in Proc. ACM SIGKDD, 2017, pp. 1663–1672.
[6] S. Rasp and S. Lerch, "Neural networks for post-processing ensemble weather forecasts," Mon. Weather Rev., vol. 146, no. 11, pp. 3885–3900, 2018.
[7] Y. Ham et al., "Self-supervised learning for climate data: A case study in seasonal forecasting," in NeurIPS Climate Informatics Workshop, 2021.
[8] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
[9] J. Pathak et al., "FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators," in Proc. NeurIPS, 2022.
[10] R. E. Dickinson, "How coupling of the atmosphere to ocean and land helps determine the timescales of climate variability," Journal of Climate, vol. 1, no. 5, pp. 383–398, 1988.
[11] C. T. Y. Chung, R. V. Azizzadenesheli, and A. Anandkumar, "Self-supervised deep learning for downscaling climate data," in *Proc.*

AAAI Conf. Artificial Intelligence, vol. 35, no. 17, pp. 14909–14917, May 2021.

[12] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD)*, San Diego, CA, USA, Aug. 2020, pp. 753–763, doi: 10.1145/3394486.3403128.

[13] D. Vandal et al., "DeepSD: Generating high resolution climate change projections through single image super-resolution," in Proceedings of the 23rd ACM SIGKDD, 2017, pp. 1663–1672.

[14] M. S. Lakshmi, K. S. Ramana, M. J. Pasha, K. Lakshmi, N. Parashuram, and M. Bhavsingh, "Minimizing the localization error in wireless sensor networks using multi-objective optimization techniques," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 2s, pp. 306–312, 2022. doi: 10.17762/ijritcc.v10i2s.5948.

[15] S. Rasp and S. Lerch, "Neural networks for post-processing ensemble weather forecasts," Monthly Weather Review, vol. 146, no. 11, pp. 3885–3900, 2018.

[16] H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," IEEE Access, vol. 8, pp. 193907–193934, 2020.

[17] X. Chen et al., "An empirical study of training self-supervised vision transformers," in ICCV, 2021, pp. 9640–9650.

[18] M. S. Lakshmi, K. S. Ramana, G. Ramu, K. Shyam Sunder Reddy, C. Sasikala, and G. Ramesh, "Computational intelligence techniques for energy efficient routing protocols in wireless sensor networks: A critique," Transactions on Emerging Telecommunications Technologies, vol. 35, no. 1, Nov. 2023, doi: 10.1002/ett.4888.

[19] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2021.

[21] J. Pathak et al., "FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators," in NeurIPS, 2022.

[22] C. Hersbach et al., "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, Apr. 2020, doi: 10.1002/qj.3803.

[23] V. Eyring et al., "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," Geoscientific Model Development, vol. 9, no. 5, pp. 1937–1958, May 2016, doi: 10.5194/gmd-9-1937-2016.

[24] D. P. Maurer, "Bias correction and spatial disaggregation of climate projections for hydrologic modeling: A case study for the Columbia River Basin," Water Resources Research, vol. 46, no. 5, 2010, doi: 10.1029/2008WR007327.

[25] D. Vandal, S. Kodra, S. Ganguly, A. Michaelis, C. Nemani, and A. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 2017, pp. 1663–1672, doi: 10.1145/3097983.3098045.

[26] S. Rasp and S. Lerch, "Neural networks for post-processing ensemble weather forecasts," Monthly Weather Review, vol. 146, no. 11, pp. 3885–3900, 2018, doi: 10.1175/MWR-D-18-0187.1.

.