

Spatially-Aware Temperature Forecasting using Graph Attention and Temporal Convolution Networks

Roshan Khanal

*Dept. of Computer Engineering
Nepal College of Information Technology
Pokhara University, Kathmandu, Nepal
ujyaloroshan369@gmail.com*

Ashim Khadka*

*Dept. of Computer Engineering
Nepal College of Information Technology
Pokhara University, Kathmandu, Nepal
ashim.khadka@ncit.edu.np*

Bibek Gautam

*Dept. of Electronics and Computer Engg.
Pulchowk Campus, Institute of Engineering
Tribhuvan University, Kathmandu, Nepal
bibekgautam2052@gmail.com*

Saroj Panthi

*Dept. of BCA
Padmashree College
Kathmandu, Nepal
sarojpanthi2062@gmail.com*

*Corresponding author

Abstract—Accurate local temperature forecasting is difficult in topographically complex regions such as Nepal, where atmospheric behavior is governed by strong spatiotemporal dependencies that conventional statistical and deep learning models fail to fully capture. In this study, a new Graph Attention Network-Temporal Convolutional Network (GAT-TCN) model is proposed for multi-step temperature forecasting over Bagmati Province, Nepal. A 3×3 grid of ERA5 reanalysis data is encoded as a spatial graph, where a GAT learns heterogeneous dependencies among neighboring grid points while a TCN captures temporal dynamics from a 24-hour historical window using causal, dilated convolutions. Spatial and temporal representations are combined through a late-fusion strategy, and 14-day forecasts are generated autoregressively. Experimental results show that the proposed late-fusion GAT-TCN model outperforms GAT, TCN, GCN-LSTM, GAT-LSTM, and an early-fusion GAT-TCN variant, achieving an R^2 of 0.944 and MAE of 0.149°C on Day 1, degrading gracefully to R^2 of 0.768 at Day 14. Transfer learning experiments using a progressive layer-unfreezing strategy further show that the pretrained model adapts effectively to mountain and Terai regions, confirming the generalizability of the proposed spatiotemporal framework.

Index Terms—temperature forecasting, graph attention network, temporal convolutional network, spatiotemporal deep learning, ERA5, transfer learning

I. INTRODUCTION

Temperature prediction is important in agriculture, water management, disaster preparedness, and energy planning. In geographically complicated areas like Nepal, temperature variations have a strong dependence on circulation and processes related to the terrain and hence are hard to predict at high resolution. NWP models provide accurate physical representation of the atmosphere but require heavy computation, lose accuracy at fine scales, and linear models like ARIMA and SARIMA consider the process to be linear and stationary in nature [1].

ECMWF's ERA5 reanalysis data set [2] makes use of deep sequence based models like LSTM for predicting weather [3], which take into account temporal dependency but fail to consider spatial dependency among the nearby locations as it treats each location individually [4]. Graph Neural Networks (GNNs), particularly Graph Attention Networks (GATs) [5], address this gap. It can learn adaptive, non-uniform spatial dependencies among neighboring locations. Hybrid graph-recurrent models such as GAT-LSTM [6] combine both dimensions, but they rely on recurrence. It makes training slow and unstable over long forecast horizons. Temporal Convolutional Networks (TCNs) [7] offer a parallelizable alternative. It models long-range temporal dependency through dilated causal convolutions without the instability associated with recurrence.

Still, the main issue here is the fact that spatiotemporal learning in modern hybrid models e.g., GCN-LSTM has problems with uniform spatial information aggregation and training instability because of recurrence over long forecast ranges. That's why current spatiotemporal architectures cannot make use of adaptive attention weighted spatial dependencies and stable parallelizable temporal learning, thus lacking the ability to perform accurate forecasting of multi-day localized temperatures in complex topographies, such as Nepal.

The objective of this study is therefore to design and evaluate a spatiotemporal forecasting model that couples graph attention-based spatial encoding with temporal convolution-based sequence modeling, so as to jointly capture heterogeneous spatial dependencies among neighboring grid points and stable long-range temporal dynamics, thereby improving multi-day temperature forecasting accuracy over conventional and hybrid graph-recurrent baselines.

This work proposes a GAT-TCN framework that models spatial dependency through graph attention and temporal de-

pendency through dilated causal convolution, fusing the two representations to forecast temperature 14 days ahead for Bagmati Province, Nepal. The main contributions of this study are: (1) a spatial graph formulation of a 3×3 ERA5 grid with an attention-based spatial encoder; (2) a TCN-based temporal encoder with a 24-hour receptive input window; (3) a late-fusion architecture that outperforms early-fusion and baseline spatiotemporal models; and (4) a transfer learning evaluation demonstrating adaptability to mountain and Terai regions.

II. RELATED WORK

Temperature prediction using statistical techniques, such as SARIMA, works well when used on stationary series, however, cannot deal with the nonlinearity and nonstationarity in the atmosphere [1], [8]. The research indicates that the comparative analysis of such models reveals their weakness on nonlinear, high-frequency, multi-step prediction tasks despite the fact that such models work efficiently computationally [9]. That leads us to consider data-driven approaches [10].

Classic machine learning approaches: support vector regressions, random forests, and Naive Bayes classifiers, were among the first data-driven approaches. These models have shown the possibility of predicting the weather directly from the historical data [11]. These approaches are independent feature-wise and have no built-in temporal model [12], the ensembles of such models outperform the single ones on short-term prediction tasks [13]. These models continue to treat the forecasting task as a time-independent regression problem, rather than a true sequence modeling problem [14].

Deep sequence models, such as LSTMs and their variants, produced high improvements in learning of temporal patterns for temperature sequences [3], [15], [16]. LSTM variant with attention mechanism, SFA-LSTM [4] and BiLSTM-attention architectures [17], [18] provided improvements in accuracy by weighting relevant temporal parts. But spatial interaction between locations is not modeled in these methods. Convolution-based spatio-temporal approaches such as ConvLSTM contain fixed grid-based spatial context [16], [19], [20], but cannot represent irregular or heterogeneous spatial dependencies.

Graph-based approaches address this limitation directly. GNNs model locations as graph nodes and have shown strong performance transferring information across geographically related points [2], [21]. GATs [5] extend this by learning attention-weighted and non-uniform neighbor contributions. Hybrid spatio-temporal graph models like GCN-LSTM [22] and GAT-LSTM [6] combine graph-based spatial encoders with recurrent temporal encoders. This improve performance compare to purely temporal or purely spatial baselines. But their sequential recurrent structure limits training stability and parallelism over long horizons. TCNs [7], [23] have been shown to outperform RNN-based architectures for long-horizon, multi-step time-series forecasting. TCNs have stable gradient propagation and controllable receptive field growth.

The literature reviewed above reveals three recurring shortcomings: (i) purely temporal models ignore spatial interaction

among neighboring locations [3], [4], [15], [16]; (ii) fixed-grid convolutional models cannot represent heterogeneous, non-uniform spatial dependence [16], [19], [20]; and (iii) hybrid graph-recurrent models such as GCN-LSTM and GAT-LSTM, while combining both dimensions, remain constrained by uniform spatial aggregation and by the sequential, error-accumulating nature of recurrence over long forecast horizons [6], [22]. GAT-TCN is proposed specifically to resolve these three gaps jointly: the GAT component replaces uniform or fixed-grid spatial aggregation with attention-weighted, adaptive neighbor contributions, allowing the model to learn which surrounding grid points matter most at a given time rather than assuming equal or purely distance-based influence; the TCN component replaces recurrence with causal, dilated convolutions, giving the temporal encoder a wide, explicitly controllable receptive field, stable gradient propagation, and parallel computation, all of which are particularly valuable for the long, autoregressive 14-day rollout used in this study. Combining these two components through late fusion, rather than a recurrent hybrid, is therefore expected to yield more accurate and more stable multi-day forecasts than existing spatio-temporal baselines, which is examined empirically in Section IV.

III. METHODOLOGY

A. System Overview

The proposed system, illustrated in Fig. 1, consists of a data preprocessing stage followed by two parallel feature-extraction branches: a spatial branch based on GAT and a temporal branch based on TCN. The two learned representations are combined through late fusion and passed to a fully connected layer that outputs the predicted temperature.

B. Data

Twenty years (2006–2025) of hourly ERA5 reanalysis data were collected over a 3×3 spatial grid ($0.25^\circ \times 0.25^\circ$ resolution) centered on the Kathmandu Valley, Bagmati Province, Nepal (Fig. 2), yielding nine spatial nodes and 175,200 hourly samples per node. Seven meteorological variables were used per node: 2m air temperature, surface pressure, 2m dew point temperature, mean sea-level pressure, total precipitation, and the 10m zonal and meridional wind components. These physically fundamental variables were chosen over derived quantities because they jointly govern surface temperature through horizontal advection, thermodynamics, and synoptic-scale atmospheric processes [24]. The central grid node was used as the forecasting target, with the remaining eight nodes providing spatial context. Exploratory analysis confirmed a strong diurnal cycle, clear seasonal variability, and high spatial correlation between neighboring nodes (Pearson $r > 0.97$), justifying the use of both temporal and graph-based spatial modeling.

Data was chronologically split into training (2006–2019), validation (2020–2021), and test (2022–2025) sets to prevent information leakage, and all features were standardized using Z-score normalization computed from the training set only.

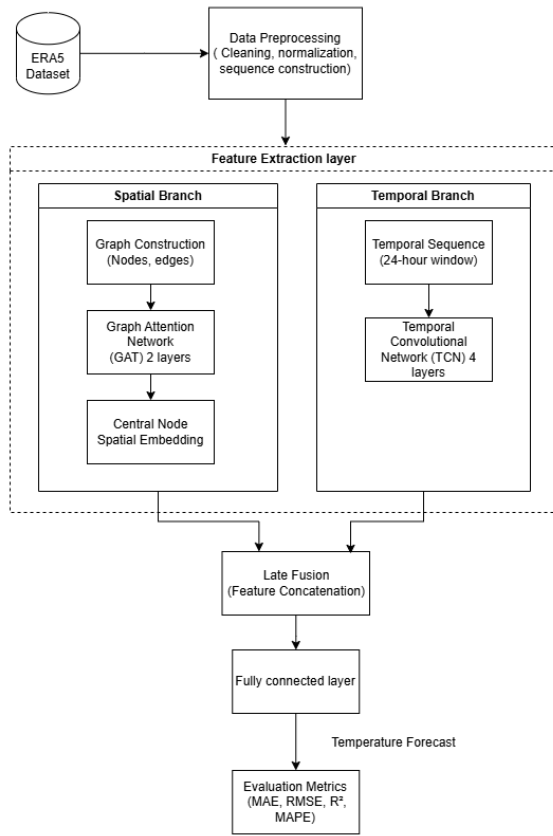


Fig. 1. System architecture of the proposed GAT-TCN model.



Fig. 2. Study region and 3×3 spatial grid configuration.

Sequences were constructed using a sliding 24-hour input window.

C. Spatial Graph Construction

The study area was represented as an undirected graph $G = (V, E)$ with nine nodes corresponding to the 3×3 ERA5 grid points (Fig. 3). Every node has connection with its horizontal, vertical, and diagonal neighbors, forming an 8-neighborhood static graph that remains fixed during training and inference.

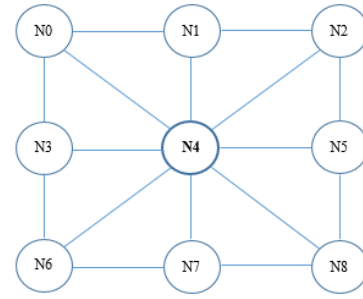


Fig. 3. Spatial graph representation of the 3×3 grid; N4 is the central target node.

D. Graph Attention Network

For each node i with feature vector $h_i \in \mathbb{R}^7$, a shared linear transformation is applied:

$$z_i = Wh_i \quad (1)$$

The attention coefficient between node i and neighbor $j \in N(i)$ is computed as

$$e_{ij} = \text{LeakyReLU}(a^T[z_i \| z_j]) \quad (2)$$

and normalized via softmax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})} \quad (3)$$

The updated node representation aggregates neighbor information using K attention heads, averaged rather than concatenated to control dimensionality:

$$h'_i = \frac{1}{K} \sum_{k=1}^K \sigma \left(\sum_{j \in N(i)} \alpha_{ij}^k z_j \right) \quad (4)$$

The spatial encoder uses two stacked GAT layers, each with two attention heads producing 16-dimensional embeddings. Although all nine node embeddings are computed, only the central node embedding $s_t \in \mathbb{R}^{16}$ is forwarded to the fusion stage.

E. Temporal Convolutional Network

The temporal branch receives a 24-hour sequence from the central node, $X_{temp}(t) \in \mathbb{R}^{24 \times 7}$. Causal convolution ensures predictions depend only on present and past inputs:

$$y(t) = \sum_{k=0}^{K-1} w(k) x(t-k) \quad (5)$$

Dilated convolution extends the receptive field without added depth:

$$y(t) = \sum_{k=0}^{K-1} w(k) x(t-d \cdot k) \quad (6)$$

The TCN consists of four residual blocks (kernel size 3, dilations 1, 2, 4, 8; 128 channels; dropout 0.5), illustrated in

Fig. 4. Each block applies two dilated causal convolutions with weight normalization and a residual connection:

$$z_l = \text{ReLU}(\text{Conv1D}_{l,2}(\text{ReLU}(\text{Conv1D}_{l,1}(x_l)))) \quad (7)$$

$$x_{l+1} = \text{ReLU}(x_l + z_l) \quad (8)$$

With kernel size $K = 3$ and four dilation levels, the effective receptive field is 61 hours, exceeding the 24-hour input window and allowing the network to draw on the full diurnal cycle. The final temporal representation is taken from the last time step, $h_{temp} \in \mathbb{R}^{128}$.

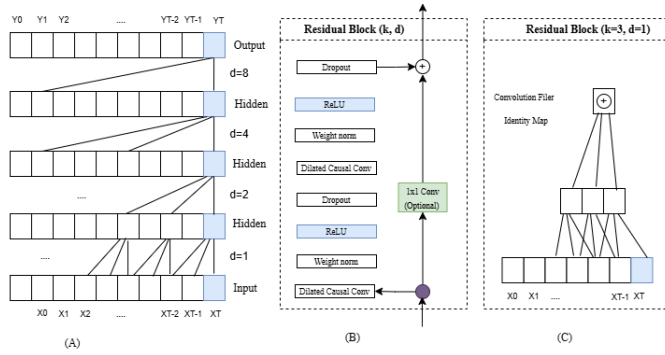


Fig. 4. TCN architecture with dilated causal residual blocks.

F. Fusion and Autoregressive Forecasting

The spatial and temporal embeddings are concatenated,

$$z_t = [s_t || h_{temp}] \in \mathbb{R}^{144} \quad (9)$$

and passed through a fully connected layer to produce the one-step prediction:

$$\hat{y}_{t+1} = W_o z_t + b \quad (10)$$

Fourteen-day (336-hour) forecasts are generated autoregressively, where each predicted value is appended to the temporal input window to predict the subsequent hour, without access to future ground-truth observations. An early-fusion variant, in which the GAT embedding is concatenated with raw features at every time step before temporal modeling, was also implemented for comparison.

G. Training

Both branches were trained jointly using the Adam optimizer (learning rate 0.001, batch size 64) to minimize MSE loss, with gradient-norm clipping (threshold 1.0), a ReduceLROnPlateau scheduler, and early stopping (patience 12 epochs, max 150 epochs). Hyperparameters (GAT hidden size and head count, TCN width and depth, dropout, batch size, learning rate) were tuned via structured random search.

IV. RESULTS AND DISCUSSION

A. Forecasting Performance

Table I compares the day-wise R^2 of the proposed late-fusion GAT-TCN against an early-fusion GAT-TCN variant and two hybrid graph-recurrent baselines, GCN-LSTM and GAT-LSTM. The late-fusion GAT-TCN achieved the best performance across all 14 forecast days, with $R^2=0.944$, $MAE=0.149^\circ\text{C}$, and $MAPE=0.94\%$ on Day 1, degrading gracefully to $R^2=0.768$ by Day 14. Purely temporal (TCN) and purely spatial (GAT) models performed markedly worse at longer horizons, confirming that both spatial and temporal information are needed. Fig. 5 shows that the late-fusion model closely tracks the diurnal temperature cycle throughout the 14-day horizon, with deviations growing modestly as the forecast horizon increases, as expected under autoregressive rollout.

TABLE I
DAY-WISE R^2 COMPARISON OF FORECASTING MODELS

Day	GAT-LSTM	GCN-LSTM	GAT-TCN (Early)	GAT-TCN (Late)
1	0.748	0.821	0.913	0.944
3	0.623	0.750	0.801	0.886
7	0.566	0.705	0.743	0.835
10	0.538	0.672	0.713	0.802
14	0.496	0.636	0.676	0.768

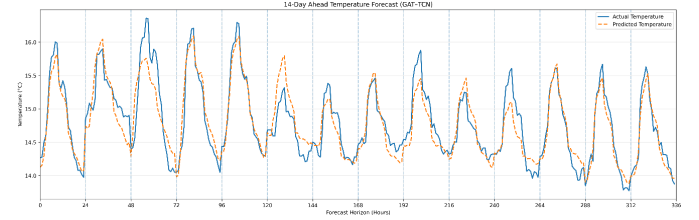


Fig. 5. 14-day-ahead temperature forecast produced by the late-fusion GAT-TCN model against ground-truth observations.

Late fusion architecture where each branch specialized in representation before merging them was superior to early fusion demonstrating that too early combination of spatial and temporal signals led to low representational quality. In addition, 48-hour input window was also analyzed but resulted in poor accuracy compared to 24-hour window (Day 1 R^2 dropped from 0.944 to 0.828), which may be because the shorter window fit well with dominant diurnal periodicity without introducing noise from other history.

B. Sensitivity Analysis

A sensitivity analysis of the model on dropout rate, TCN depth/width, GAT hidden size, and number of attention heads revealed that the dropout rate of 0.5 gave the best trade-off between overfitting and underfitting; the TCN width and depth led to higher accuracy until 128 channels and 4 layers after which the return was not significant; and two attention heads with 16-dimensional GAT embeddings outperformed both fewer (under-fit) and more (over-parameterized) ones,

showing that the small spatial grid was an appropriate choice for small spatial encoders. Fig. 6 illustrates the effect of dropout rate on forecasting error across the first seven forecast days: rates below 0.4 overfit and lose stability at longer horizons, while rates above 0.6 underfit, with 0.5 consistently achieving the lowest MAE.

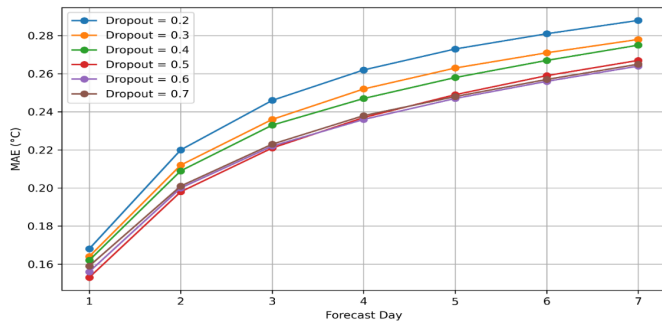


Fig. 6. Effect of dropout rate on MAE across forecast days 1–7.

C. Transfer Learning

To assess generalizability, the pretrained Bagmati model was fine-tuned on five years of ERA5 data over two held-out regions – a mountain region and a Terai (lowland) region – using a three-phase progressive unfreezing strategy (output layer only → partial layers → full model with discriminative learning rates). The adapted model achieved Day 1 R^2 of 0.925 (mountain) and 0.935 (Terai), decreasing to 0.741 and 0.569 respectively by Day 14. The larger reduction in Terai likely due to its flatter terrain and different synoptic drivers compared to the Kathmandu Valley domain, but overall, it confirms that the learned spatiotemporal representations transfer well between topographically different regions of Nepal. Fig. 7 shows the 14-day transfer-learning forecast for the Terai region, where the model continues to track the diurnal cycle closely despite being adapted from a climatically distinct training domain.

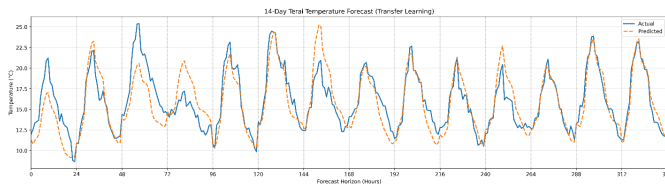


Fig. 7. 14-day-ahead temperature forecast for the Terai region using the transfer-learned GAT-TCN model.

V. CONCLUSION

This paper presented a GAT-TCN framework for spatially-aware, multi-step temperature forecasting that models heterogeneous spatial dependencies through graph attention and long-range temporal dependencies through dilated causal convolution. Evaluated on 20 years of ERA5 data over Bagmati Province, Nepal, the proposed late-fusion architecture consistently outperformed purely spatial, purely temporal, and hybrid graph-recurrent baselines across a 14-day forecast horizon,

and transfer learning experiments confirmed its adaptability to mountain and Terai regions. Next, we plan to take into account additional atmospheric variables like radiation, soil moisture, and cloud cover, as well as explore dynamically changing graphs based on atmospheric conditions and simultaneous multi-location prediction.

REFERENCES

- [1] A. M. Elshewey et al., “A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset,” *Sustainability*, vol. 15, no. 1, p. 757, Dec. 2022.
- [2] R. Keisler, “Forecasting Global Weather with Graph Neural Networks,” arXiv:2202.07575, Feb. 2022.
- [3] I. M. Nizar, D. Adytia, and A. W. Ramadhan, “Forecasting of Temperature by using LSTM and Bidirectional LSTM approach: Case Study in Semarang, Indonesia,” in *Proc. ICoDSA, Bandung, Indonesia*, Oct. 2021, pp. 146–150.
- [4] M. A. R. Suleman and S. Shridevi, “Short-Term Weather Forecasting Using Spatial Feature Attention Based LSTM Model,” *IEEE Access*, vol. 10, pp. 82456–82468, 2022.
- [5] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” arXiv:1710.10903, Feb. 2018.
- [6] U. Orji, Ç. Güven, and D. Stowell, “Enhanced Load Forecasting with GAT-LSTM: Leveraging Grid and Temporal Features,” arXiv:2502.08376, Feb. 2025.
- [7] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” arXiv:1803.01271, Apr. 2018.
- [8] P. Aghelpour, B. Mohammadi, and S. M. Biazar, “Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA,” *Theor. Appl. Climatol.*, vol. 138, no. 3–4, pp. 1471–1480, Nov. 2019.
- [9] J. Cifuentes, G. Marulanda, A. Bello, and J. Reneses, “Air Temperature Forecasting Using Machine Learning Techniques: A Review,” *Energies*, vol. 13, no. 16, p. 4215, Aug. 2020.
- [10] M. K. Alomar et al., “Data-driven models for atmospheric air temperature forecasting at a continental climate region,” *PLOS ONE*, vol. 17, no. 11, p. e0277079, Nov. 2022.
- [11] M. Biswas, T. Dhoom, and S. Barua, “Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data,” *Int. J. Comput. Appl.*, vol. 182, no. 34, pp. 20–24, Dec. 2018.
- [12] M. S. Rahman, F. A. Tumpa, M. S. Islam, A. A. Arabi, M. S. B. Hossain, and M. S. U. Haque, “Comparative Evaluation of Weather Forecasting using Machine Learning Models,” arXiv:2402.01206, Feb. 2024.
- [13] Nikunj Kumar B. Nayak, “Comparative Analysis of Weather Forecasting Using LSTM, BiLSTM, and CLSTM Networks,” *J. Electr. Syst.*, vol. 20, no. 3, pp. 1511–1517, May 2024.
- [14] H. Zhang, Y. Liu, C. Zhang, and N. Li, “Machine Learning Methods for Weather Forecasting: A Survey,” *Atmosphere*, vol. 16, no. 1, p. 82, Jan. 2025.
- [15] E. A. Nketiah, L. Chenlong, J. Yingchuan, and S. A. Aram, “Recurrent neural network modeling of multivariate time series and its application in temperature forecasting,” *PLOS ONE*, vol. 18, no. 5, p. e0285713, May 2023.
- [16] E. Haque, S. Tabassum, and E. Hossain, “A Comparative Analysis of Deep Neural Networks for Hourly Temperature Forecasting,” *IEEE Access*, vol. 9, pp. 160646–160660, 2021.
- [17] X. Hao, Y. Liu, L. Pei, W. Li, and Y. Du, “Atmospheric Temperature Prediction Based on a BiLSTM-Attention Model,” *Symmetry*, vol. 14, no. 11, p. 2470, Nov. 2022.
- [18] N. Zrira, A. Kamal-Idrissi, R. Farssi, and H. A. Khan, “Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism,” *J. Sea Res.*, vol. 198, p. 102472, Apr. 2024.
- [19] S. F. Tekin, A. Fazla, and S. S. Kozat, “Numerical Weather Forecasting using Convolutional-LSTM with Attention and Context Matcher Mechanisms,” arXiv:2102.00696, Oct. 2023.
- [20] Z. Zhang and Y. Dong, “Temperature Forecasting via Convolutional Recurrent Neural Networks Based on Time-Series Data,” *Complexity*, vol. 2020, pp. 1–8, Mar. 2020.
- [21] Y. Varshney, V. Kumar, D. K. Dubey, and S. Sharma, “Forecasting Precision: The Role of Graph Neural Networks and Dynamic GNNs in Weather Prediction,” vol. 3, no. 1.

- [22] L. G.-D. Sáenz, J. Cifuentes, and G. Marulanda, “Short-Term Spatio-Temporal Forecasting of Air Temperatures using Deep Graph Convolutional Neural Networks.”
- [23] P. Hewage et al., “Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station,” *Soft Comput.*, vol. 24, no. 21, pp. 16453–16482, Nov. 2020.
- [24] J. M. Wallace and P. V. Hobbs, *Atmospheric Science: An Introductory Survey*, 2nd ed. Amsterdam: Academic Press, 2008.