A Graph Neural Network Approach for Localized and High-Resolution Temperature Forecasting

Joud El-Shawa[1,2] **Elham Bagheri**[1,2] **Sedef Akinli Kocak**[1] **Yalda Mohsenzadeh**[1,2*]
[1]Vector Institute for Artificial Intelligence, Toronto, Canada
[2]Western University, London, Canada
[*]ymohsenz@uwo.ca

Abstract

Heatwaves are intensifying worldwide and are among the deadliest weather disasters. The burden falls disproportionately on marginalized populations and the Global South, where under-resourced health systems, exposure to urban heat islands, and the lack of adaptive infrastructure amplify risks. Yet current numerical weather prediction models often fail to capture micro-scale extremes, leaving the most vulnerable excluded from timely early warnings. We present a Graph Neural Network framework for localized, high-resolution temperature forecasting. By leveraging spatial learning and efficient computation, our approach generates forecasts at multiple horizons, up to 48 hours. For Southwestern Ontario, Canada, the model captures temperature patterns with a mean MAE of 1.93°C across 1–48h forecasts and MAE@48h of 2.93°C, evaluated using 24h input windows on the largest region. While demonstrated here in a data-rich context, this work lays the foundation for transfer learning approaches that could enable localized, equitable forecasts in data-limited regions of the Global South.

1 Introduction

A global study estimated that from 2000-2019 approximately 489,000 heat-related deaths occurred each year [1], a figure also highlighted by the World Health Organization [2], and one that has likely risen as the climate crisis accelerates. Heatwaves threaten health, ecosystems, and economies worldwide. Their impacts are uneven: low-income, racialized, and Global South communities are most exposed, despite minimal contribution to emissions [3, 4]. Factors such as poor housing insulation, limited cooling infrastructure, and underfunded public health amplify vulnerability [5, 6]. At the national scale, low- and middle-income countries face severe risks due to lower adaptive capacity and constrained resources [7].

Unfortunately, current operational weather forecasts often lack the granularity and lead time needed to adequately warn and protect local communities. Existing forecast systems typically operate at 10-30 km scales, smoothing over urban heat islands or neighborhood "hot spots." This leads to systematic underestimation of heat risk in precisely those marginalized areas most in need of targeted interventions [8]. Addressing this inequity requires forecasting systems that are high-resolution, adaptive, and accessible across diverse contexts, from urban heat islands in megacities to rural, resource-limited regions.

Traditional numerical weather prediction (NWP) models are computationally costly, limited in resolution, and rely on parameterizations that often miss land-atmosphere feedbacks critical for heat extremes [9, 10]. Recent machine learning advances demonstrate alternatives: GraphCast achieves skillful global forecasts at 28 km resolution [11], FourCastNet uses Fourier operators for global predictions at a similar resolution [12], and neural models improve extreme heat anomaly forecasts [13]. Most approaches, however, emphasize global scales. Localized prediction remains

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

(a) Region A (b) Region B (c) Region C

Figure 1: Bounding boxes around Regions A–C.

underdeveloped, despite its importance for frontline adaptation. Li et al. [14] showed that Graph Neural Networks (GNNs) can classify regional heatwave events with high accuracy, underscoring their potential for fine-scale, actionable forecasting. Our work extends this logic: using GNNs to produce continuous, high-resolution temperature forecasts that can be post-processed to identify heatwaves. This provides a step toward developing practical, fine-scale forecasting frameworks that can eventually be adapted to data-limited or vulnerable contexts in the Global South.

2 Methods

Data. Our study leverages the National Oceanic and Atmospheric Administration (NOAA) Unrestricted Mesoscale Analysis (URMA) dataset as the primary source of training and evaluation data [15]. URMA provides high-resolution (2.5 km, hourly), gridded analyses of surface meteorological variables, such as temperature, winds, pressure, and elevation, which are well-suited for our task of fine-scale forecasting. While the end goal is operational heatwave forecasting, we forecast 2-meter air temperature as the primary target because heatwave definitions vary across jurisdictions (for example, consecutive hot days, exceedance of absolute thresholds, or percentile-based criteria). Predicting temperature provides a low-level signal that can be post-processed to align with region- or agency-specific heatwave definitions, without retraining the model.

Regions. Following discussions with stakeholders from climate NGOs in the Global South, we decided to focus on Southwestern Ontario in Canada as our case study region since it incorporates various land types, such as urban, farmland, forest, and water bodies. Specifically, we have 3 bounding boxes in this region, as depicted in Figure 1: **Region A** is from 42.78°N, 81.45°W to 43.18°N, 81.05°W ($\approx$ 44 km by 33 km); **Region B** is from 42.50°N, 81.50°W to 43.50°N, 79.50°W ($\approx$111 km by 163 km); **Region C** is from 42.00°N, 81.00°W to 45.00°N, 78.00°W ($\approx$333 km by 243 km). This setup allows us to run computationally intensive experiments on the smallest window (A), then demonstrate scalability and proof of concept on larger domains with consistent data and preprocessing. All three regions are covered by the URMA analysis grids for the variables of interest.

Preprocessing. We obtained hourly data for all variables spanning multiple years, with exact ranges specified in the *Appendix*, for each region. Missing or invalid entries were filled using spatial interpolation (i.e., mean of nearby grid points). Each input feature was standardized using the mean and standard deviation of the training set. The target variable (temperature) was also normalized during model training, though final errors are reported in physical units (°C) for clarity.

Proposed model. We developed and trained a hybrid Graph Convolutional Network (GCN) with a Gated Recurrent Unit (GRU) for each region. Each grid point in the region is represented as a graph node with meteorological features (temperature, winds, pressure, etc.), connected by edges to capture spatial interactions. Graph convolution layers model neighborhood effects [16], while GRUs capture temporal dependencies [17]. Train/validation/test splits varied by pipeline and region (specified in the *Appendix*). The objective was to predict temperature 1-48 hours ahead from the current time (at 1, 6, 12, 18, 24, 36, and 48h). The models were optimized using mean squared error (MSE).

Table 1: Per-region performance. Mean across horizons; 48h at the farthest forecast horizon. (MAE = Mean Absolute Error, RMSE = Root Mean Squared Error)

Region	Mean MAE (°C)	MAE@48h (°C)	RMSE@48h (°C)
A	2.55	3.78	4.84
B	2.48	3.73	4.84
C	1.93	2.93	3.90

Table 2: Region A performance comparisons. Mean across horizons; 48h at the farthest horizon.

Model	Mean MAE (°C)	MAE@48h (°C)	RMSE@48h (°C)
Baseline (tabular)	2.55	3.78	4.84
Embeddings (ClimateBERT+PCA)	3.34	4.34	5.54
Control (random weights)	9.11	8.89	10.49

Embeddings. Data from resource-limited regions is often sparse, inconsistent, and difficult to align with information-rich datasets. In a separate setup, to harmonize heterogeneous inputs, we also explore language-model embeddings as an intermediate representation. We convert each Region A observation (per timestamp and location) into a short paragraph, for example: *"temperature is 291.6 K, dew point is 283.7 K, u wind component is 4.0 m/s, v wind component is -2.1 m/s, surface pressure is 99209 Pa, ... , elevation is 172.0 meters."* These descriptions are then encoded using a ClimateBERT model [18], yielding a 768-dimensional vector. To control dimensionality and reduce noise, we apply Principal Component Analysis (PCA) [19], fitting on train and transforming train, validation, and test. The reduced embeddings are then used as node features within the GCN–GRU forecasting pipeline.

3 Results

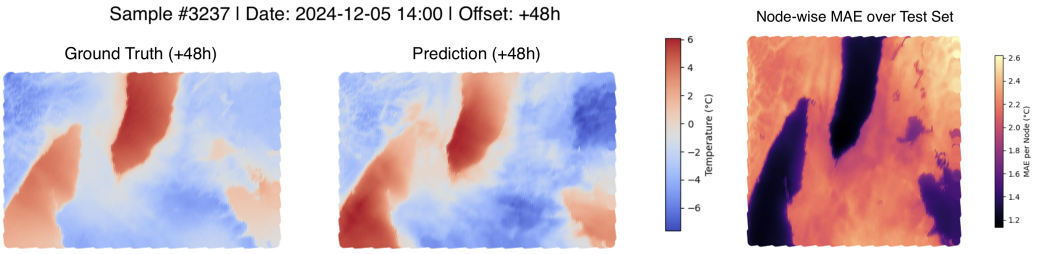
As summarized in Table 1, the GNNs achieved strong performance across all three regions. Performance improved as the spatial window expanded (Region C > Region B > Region A), which is consistent with larger graphs capturing richer neighborhood interactions and mesoscale context. These results indicate that graph-based models can provide accurate, high-resolution forecasts suitable for downstream early-warning pipelines.

On Region A, as seen in Table 2, performance with embeddings shows a modest decrease in mean MAE relative to the tabular baseline, while a control model initialized with random weights performs noticeably worse, indicating that the embedding features carry meaningful signals. This approach provides a standardized input format that can accommodate missing or unstandardized variables, which is valuable when extending to data-limited regions. We expect comparable performance with minimal fine-tuning as additional local data becomes available.

4 Discussion

Table 1 shows a clear trend: MAE and RMSE decrease as region size increases, suggesting that larger graphs capture richer spatial context. Training on Region C strains memory, so we also sampled every 6 h. The 6 h model reached mean MAE 2.39, MAE@48h 3.15, and RMSE@48h 4.16, versus the hourly model's 1.93, 2.93, and 3.90, respectively—i.e., modest degradations (+0.46, +0.22, +0.26) for substantially lower compute. This indicates that coarser temporal sampling can cut compute demands significantly while retaining most of the forecasting skill. Additionally, the use of embeddings provides a way to standardize non-standard data for modelling.

Although the results are promising, our evaluation is limited to held-out test performance on URMA data. Most benchmarked state-of-the-art systems run at much coarser scales (about 10–50 km, 3–6h), whereas our model operates at 2.5 km and hourly. This resolution mismatch makes fair comparisons difficult and highlights the need for future matched-resolution baselines. Future work should also test transferability to regions with sparser data and different climate regimes.



(a) Randomly sampled test timestamp showing ground truth and model predictions 48 hours ahead in Region C.

(b) Average node-wise MAE across the test set in Region C.

Figure 2: Example results from Region C.

Localized climate forecasting is not just a technical problem, but an equity issue. Marginalized groups, already at heightened risk due to structural inequalities [5], are further disadvantaged when coarse models fail to detect their specific vulnerabilities [8, 20]. Our approach offers a transferrable solution: models trained in data-rich regions can be adapted for under-monitored contexts, strengthening forecasting capacity and promoting climate resilience where it is most needed. For Global South governments with limited resources, such forecasts could support early-warning systems that prioritize vulnerable neighborhoods, enabling more efficient allocation of scarce resources (e.g., cooling centers, medical services, energy). By producing forecasts at community scale, GNN-based models can support municipal planning and public health interventions in ways coarse global models cannot.

This work contributes to the dialogue on climate AI [21] by showing that small, region-specific GNNs can provide outsized impact compared to large-scale models requiring vast compute. Lightweight, locally deployable systems may offer the most practical path toward equitable climate resilience.

5 Conclusion

We introduced a GCN–GRU framework for community-scale 2-meter temperature forecasting at 2.5 km using NOAA URMA across three regions in Southwestern Ontario. The models achieved strong performance, with MAE and RMSE improving as spatial context increased. A lighter 6-hour sampling preserved most of the hourly model's skill, and a language-model embedding pathway provided a practical route to standardize inputs while maintaining useful performance.

Although the long-term goal is heatwave early warning, forecasting temperature offers a flexible, low-level signal that can be post-processed to match specific heatwave definitions. Future directions include expanding predicted targets beyond temperature to include humidity, wind, and related variables, enabling a composite index and more reliable event detection. To address limitations such as resolution mismatch with coarser baselines and evaluation only on URMA data, we plan matched-resolution baselines for fairer comparisons and broader geographic coverage. Beyond this, integrating with operational dashboards will help the system support timely and equitable early warnings at low computational and environmental cost. The same framework could also extend to other extremes, such as wildfires, floods, or droughts.

References

[1] Qi Zhao et al. Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019. *The Lancet Planetary Health*, 5(7):e415–e425, 2021.

[2] World Health Organization. Heatwaves. `https://www.who.int/health-topics/heatwaves/#tab=tab_1`, 2019. Accessed: 2025-07-20.

[3] Ella S Parsons, Ashley Jowell, Erika Veidis, Michele Barry, and Sonoo T Israni. Climate change and inequality. *Pediatric Research*, 2024.

[4] Thilagawathi A Deivanayagam, Sonora English, Jason Hickel, Jon Bonifacio, Renzo R Guinto, Kyle X Hill, Mita Hug, and Rita Issa. Envisioning environmental equity: climate change, health, and racial justice. *The Lancet*, 2023.

[5] Genee S Smith, E Anjum, C Francis, L Deanes, and C Acey. Climate change, environmental disasters, and health inequities: The underlying role of structural inequalities. *Current Environmental Health Reports*, 9(3):284–293, 2022.

[6] Alana Hansen, Linda Bi, Arthur Saniotis, and Monika Nitschke. Vulnerability to extreme heat and climate change: is ethnicity a factor? *Global Health Action*, 6(1), 2013.

[7] Carol Ziegler, Vincent Morelli, and Omotayo Fawibe. Climate change and underserved communities. *Primary Care: Clinics in Office Practice*, 44(1):171–181, 2017.

[8] Tebello Putsoane, Johannes Bhanye, and Abraham Matamanda. Extreme weather events and health inequalities: Exploring vulnerability and resilience in marginalized communities. In *Climate Change and the Built Environment*. Springer, 2024.

[9] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[10] Sonia I Seneviratne et al. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3–4):125–161, 2010.

[11] Remi Lam, Alvaro Sanchez-Gonzalez, et al. Graphcast: Learning skillful medium-range global weather forecasting. *Science*, 381(6653):1101–1106, 2023.

[12] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022.

[13] Ignacio Lopez-Gomez, Amy McGovern, Shreya Agrawal, and Jason Hickey. Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, 2(1), January 2023.

[14] Peiyuan Li, Yin Yu, Daning Huang, {Zhi Hua} Wang, and Ashish Sharma. Regional heatwave prediction using graph neural network and weather station data. *Geophysical Research Letters*, 50(7), April 2023.

[15] National Centers for Environmental Prediction (NCEP). Noaa real-time mesoscale analysis (rtma) / unrestricted mesoscale analysis (urma). `https://registry.opendata.aws/noaa-rtma/`. Accessed: 2025-06-20.

[16] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[17] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[18] Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*, 2022.

[19] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society*, 2016.

[20] Sonal Jessel, Samantha Sawyer, and Diana Hernandez. Energy, poverty, and health in climate change: a comprehensive review of an emerging literature. *Front Public Health*, 7:357, 2019.

[21] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2), February 2022.

Appendix

Table 3: Description of selected NOAA URMA variables.

Abbreviation	Definition
t2m	2-meter air temperature (K)
d2m	2-meter dewpoint temperature (K)
u10	10-meter u-component of wind (m/s)
v10	10-meter v-component of wind (m/s)
sp	Surface pressure (Pa)
orog	Orography (m)

Table 4: Baseline experimental setup: data frequency and splits.

Region	Frequency	Start Date	End Date	Split	Train;Val;Test Periods/Timestamps
A	1hr	01/2022	06/2025	Manual	2022 ; 2023 ; 2024-end
B	1hr	01/2022	06/2025	Manual	2022 ; 2023 ; 2024-end
C	1hr	01/2022	12/2024	Ratio	18,396 ; 3,942 ; 3,943
C (6hr)	6hr	06/2017	06/2025	Ratio	8,170 ; 1,751 ; 1,751

Additional information: Features included: t2m, d2m, u10, v10, sp, orog. Splits were chosen based on the best-performing models identified via grid search. For regions with data extending to June 2025 (A and B), we included these months in the test set rather than shifting timestamp boundaries, to avoid bias from introducing additional seasonal data. Ratio-based splits followed a 70%/15%/15% scheme. For Region C, due to resource limitations, we excluded the additional six months available in Regions A and B that were extra for testing.

Table 5: Baseline experimental setup: hyperparameters and forecast horizons.

Region	Hyperparameters	Forecast Horizons (hrs)
A	LR=0.0001, Win=48, BS=16, HD=32, Dist=8	1, 6, 12, 18, 24, 36, 48
B	LR=0.001, Win=24, BS=16, HD=32, Dist=8	1, 6, 12, 18, 24, 36, 48
C	LR=0.001, Win=24, BS=16, HD=32, Dist=4	1, 6, 12, 18, 24, 36, 48
C (6hr)	LR=0.001, Win=24, BS=16, HD=32, Dist=4	6, 12, 18, 24, 36, 48

Abbreviations: LR = learning rate; Win = input window (hrs), BS = batch size; HD = hidden dimensions; Dist = distance (threshold for graph connectivity (km)). Final hyperparameters were selected based on the best-performing models identified via grid search.

Table 6: Training setup: GPU hardware, number of GPUs, and runtime per region.

Region	GPU Model	#GPUs	Runtime
A	NVIDIA L40S (48GB)	2	35 min
B	NVIDIA L40S (48GB)	2	42 min
C	NVIDIA H100 SXM (80GB)	2	5h 40 min
C (6hr)	NVIDIA L40S (48GB)	2	5h 45 min

Cluster specs: L40S nodes - Dell 750xa, 2×Intel Xeon Gold 6338, 512 GB RAM, 4×NVIDIA L40S 48GB. H100 nodes - Dell XE9680, 2×Intel Xeon Gold 6442Y, 2048 GB RAM, 8×NVIDIA H100 SXM 80GB.