Supplement of Atmos. Chem. Phys., 26, 947–961, 2026
https://doi.org/10.5194/acp-26-947-2026-supplement

*Supplement of*

# Intra-city scale graph neural networks enhance short-term air temperature forecasting

**Han Wang et al.**

*Correspondence to:* Jiachuan Yang (cejcyang@ust.hk)

**Contents of this file:**

## Sect. S1: Architecture of two graph neural networks (GNNs).

We employ both directed and undirected learning mechanisms to aggregate spatial information from neighboring nodes for the purpose of comparison. The implementation details are described below:

GSAGE: A mean operator was utilized to aggregate surrounding information; the aggregation can be formulated as:

$$\mathbf{h}_{\mathcal{N}(i)}^{k-1} = \sum\nolimits_{j\in\mathcal{N}(i)} \mathbf{h}_j^{k-1}, \tag{Eq. S1}$$

where $\mathbf{h}_j^{k-1}$is the representation of the nodes in node $i$'s immediate neighborhood.

After aggregating neighboring feature vectors, GSAGE concatenates the node's current representation $\mathbf{h}_i^{k-1}$ with aggregated neighborhood vector $\mathbf{h}_{\mathcal{N}(i)}^{k-1}$:

$$\mathbf{h}_i^k = \sigma\left(\boldsymbol{W}^k \cdot \left[\mathbf{h}_i^{k-1} \parallel \mathbf{h}_{\mathcal{N}(i)}^{k-1}\right]\right), \tag{Eq. S2}$$

where $\boldsymbol{W}$ are learned, and ∥ denotes vector concatenation, and this concatenation can be understood as the simple form of "skip connection". The aggregation process is also illustrated in Figure 1c.

GAT: Contrary to GSAGE, GAT can adeptly assign the importance of their neighbors. We applied this to examine whether there are super nodes that can provide more important information than other nodes. An improved version, GATv2, was applied to avoid static attention problems (Brody et al., 2021) in our study. A scoring function $e$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ assigns the importance score $\alpha_{ij}$ for every edge $(j, i)$ , which indicates the importance of the features of the neighbor j to the node i:

$$\alpha_{ij} = softmax_j\left(e(\boldsymbol{h}_i, \boldsymbol{h}_j)\right) = \mathrm{softmax_j}\left(\boldsymbol{a}^\top \,\mathrm{LeakyReLU}\left(\boldsymbol{W}^k \cdot \left[\boldsymbol{h}_i \parallel \boldsymbol{h}_j\right]\right)\right), \tag{Eq. S3}$$

where $\boldsymbol{a}, \boldsymbol{W}$ are learned, and $\alpha_{ij}$ usually is unequal with $\alpha_{ji}$.

With edge weights, GAT computes the node representation as a weighted average over its neighbors:

$$\boldsymbol{h}_i^k = \sigma\left(\sum\nolimits_{j\in\mathcal{N}(i)} \alpha_{ij} \cdot \boldsymbol{W}\boldsymbol{h}_j^{k-1}\right). \tag{Eq. S4}$$

## Sect. S2: Model configuration analysis

Table S1 lists the top five model configurations (out of 100 trials) ranked by their validation performance. Although the search range for time lag was set up to 200, the optimal configurations tend to select relatively short lags. This suggests that while incorporating temporal sequences benefits the model, excessively long input windows (though theoretically containing more information) may introduce redundant or noisy signals that ultimately degrade performance. Similar observations have been reported in a purely time-series forecasting study (Wang et al., 2024).

The optimal number of GNN layers is generally two, indicating that moderate spatial aggregation effectively captures global spatial dependencies, whereas deeper GNNs may lead to over-smoothing across locations. Regarding the number of neighbors, the results show that models typically perform better when incorporating a larger number of spatial connections, implying that richer inter-station relationships enhance representational learning.

**Table S1. Model configurations with top five validation performances (the brackets [ ] indicate the search range for each parameter).**

| Validation<br>RMSE | Time lag<br>[1, 200] | Hidden dimension<br>[10, 200] | GNN layer<br>[1, 3] | Neighbour size<br>[1, 15] | learning rate<br>[5e-5, 1e-3] | parameter<br>number |
|---|---|---|---|---|---|---|
| 0.903 | 44 | 116 | 2 | 15 | $2.17\times10^{-4}$ | 164959 |
| 0.915 | 30 | 135 | 2 | 12 | $1.61\times10^{-4}$ | 222757 |
| 0.916 | 32 | 109 | 2 | 12 | $1.61\times10^{-4}$ | 145849 |
| 0.918 | 5 | 171 | 2 | 13 | $1.13\times10^{-4}$ | 356029 |
| 0.922 | 2 | 62 | 3 | 12 | $1.93\times10^{-4}$ | 55869 |

**Reference:**

Brody, S., Alon, U., and Yahav, E.: How Attentive are Graph Attention Networks?, in: International Conference on Learning Representations, 2021.

Wang, H., Zhang, J., and Yang, J.: Time series forecasting of pedestrian-level urban air temperature by LSTM: Guidance for practitioners, Urban Climate, 56, 102063, https://doi.org/10.1016/j.uclim.2024.102063, 2024.

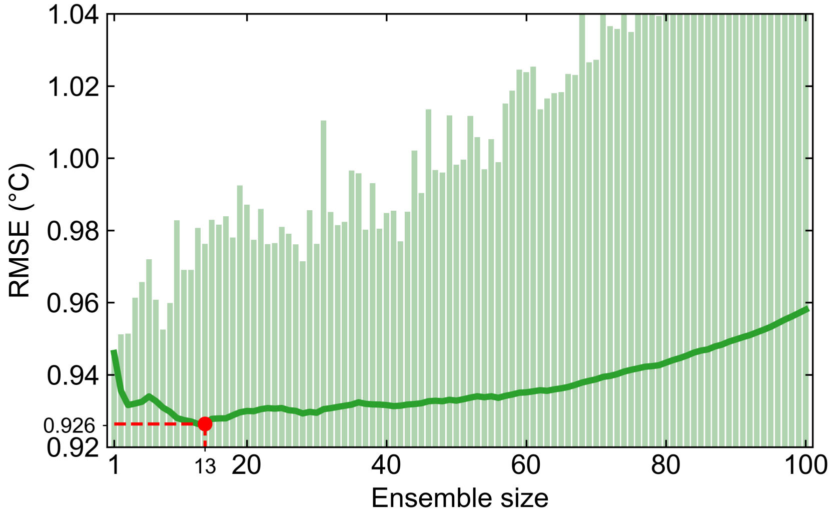

**Figure S1.** Variation in Hyper-GSAGE performance with different ensemble member sizes. Each column represents the performance of an individual GSAGE model, sorted by ascending validation error. The green line denotes the Hyper-GSAGE performance with a corresponding number of best models. It should be noted that the observed increase in RMSE with large ensemble sizes beyond 13 is primarily due to the inclusion of failure models. Conducting additional trials within the optimal hyperparameters range generally achieves a better performance, and this graph is only for illustrative purposes.

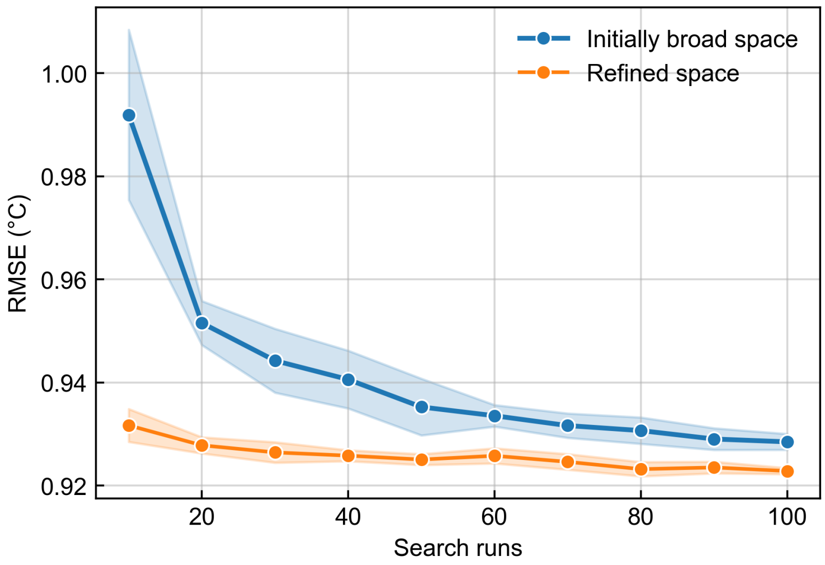

**Figure S2.** Variation in hyper-GSAGE performance (ensemble size of 10 members) across different hyperparameter search runs. The blue line represents searches initiated from a very broad search space (as defined in Table S1), while the orange line represents searches within a refined space based on initial search results. Performance stabilizes at fewer than 50 runs, with less than 1% variation beyond this point.

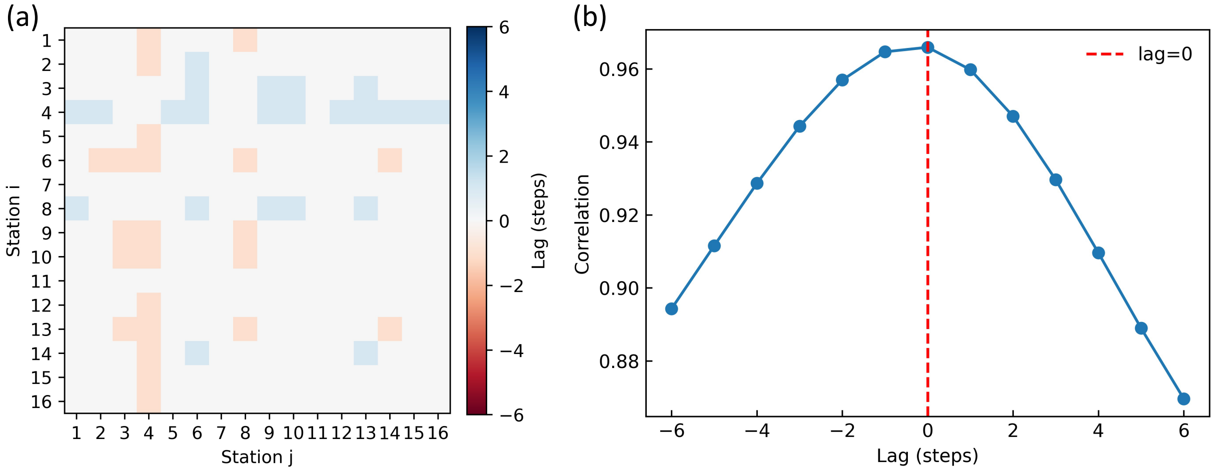

**Figure S3. (a)** Lag of maximum correlation between station pairs, evaluated over a range of −6 to +6 time steps. In our results, the lags vary between −1 and +1, with most pairs peaking at 0. A lag of 0 indicates that the $T_a$ time series at the two stations exhibit the highest correlation at the same (synchronized) time step. Positive lags indicate that temperature variations at station *i* lag behind those at station *j*, whereas negative lags indicate the opposite. **(b)** Lagged correlation between stations 2 and 3 (station IDs as shown in Fig. 2), chosen because they are the farthest apart.

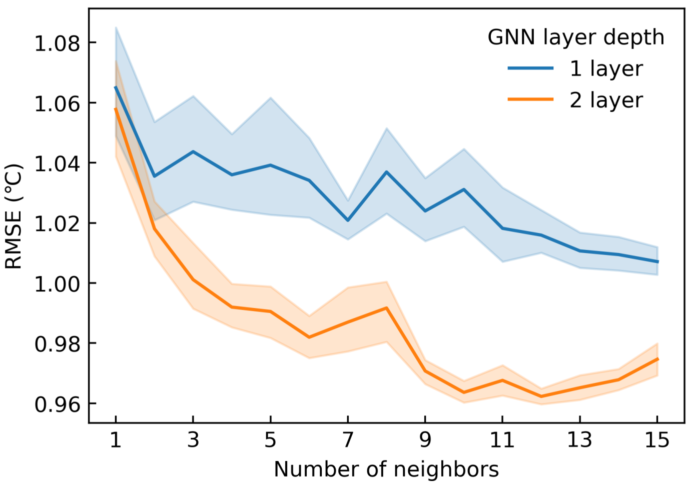

**Figure S4.** Variation of RMSE with the number of neighboring nodes used to form edge connections, classified by graph depth (one-layer and two-layer GNNs). The solid curve denotes the mean RMSE, and the shaded area represents the standard deviation across models trained with different hyperparameter settings.

70

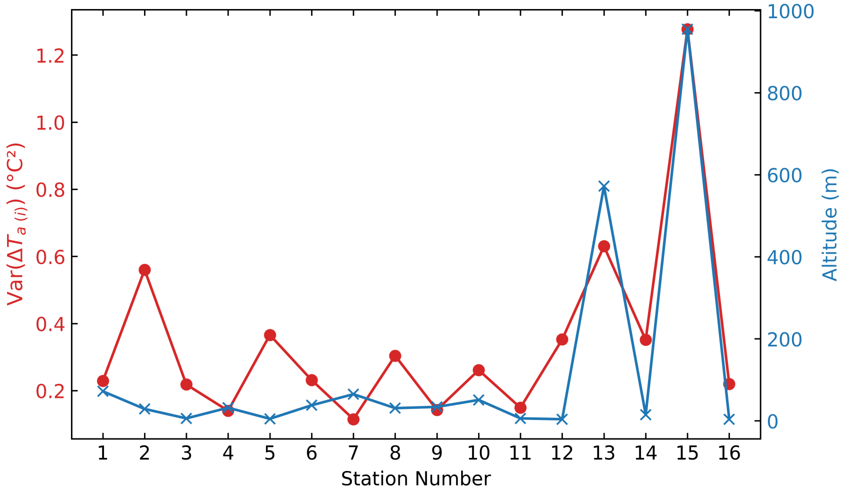

**Figure S5.** Variance of daily $T_a$ anomalies and altitude at 16 studied stations. The red line shows the variance of daily mean $T_a$ anomalies (left y-axis) for each station, calculated as the deviation of each station's Ta from the mean value across all stations (see equation (7) in Methods) while applied to daily scale. Lower variance indicates more synchronized Ta daily mean evolution between the local station and the global pattern. Blue crosses show the altitude of each station (right y-axis). Station numbers are shown on the x-axis.

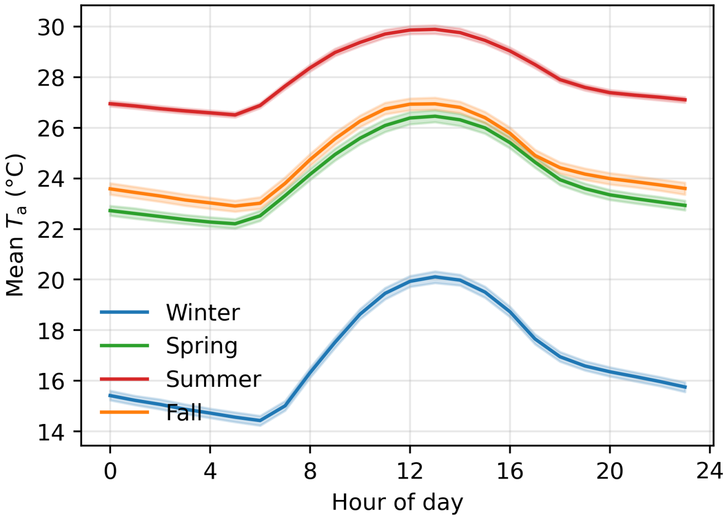

**Figure S6.** Diurnal variation of $T_a$ cross seasons. Solid lines represent the mean Ta, with shaded bands indicating 95% confidence intervals. 80

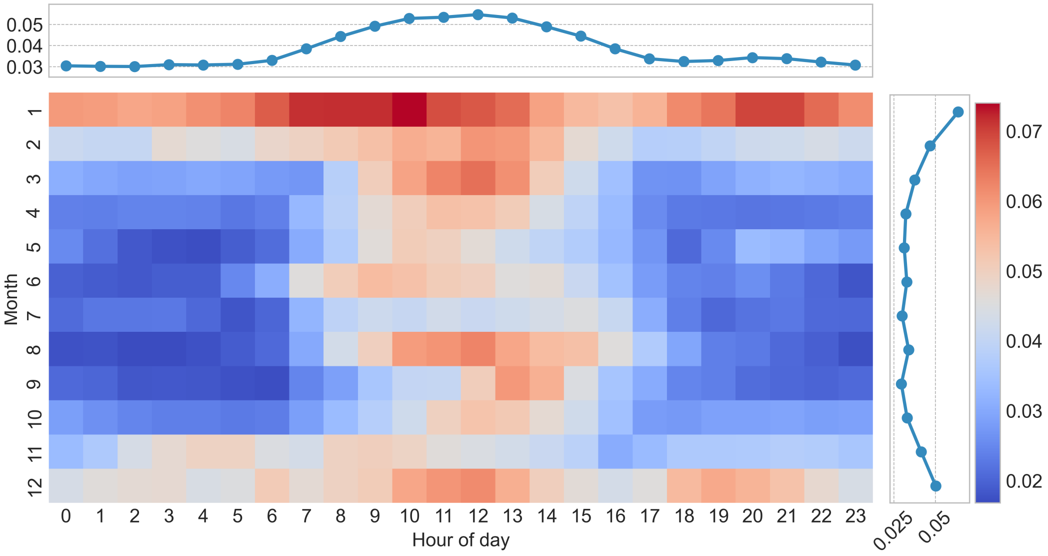

**Figure S7.** Same as Fig.3, but the RMSEs are normalized by mean values of the corresponding grids.

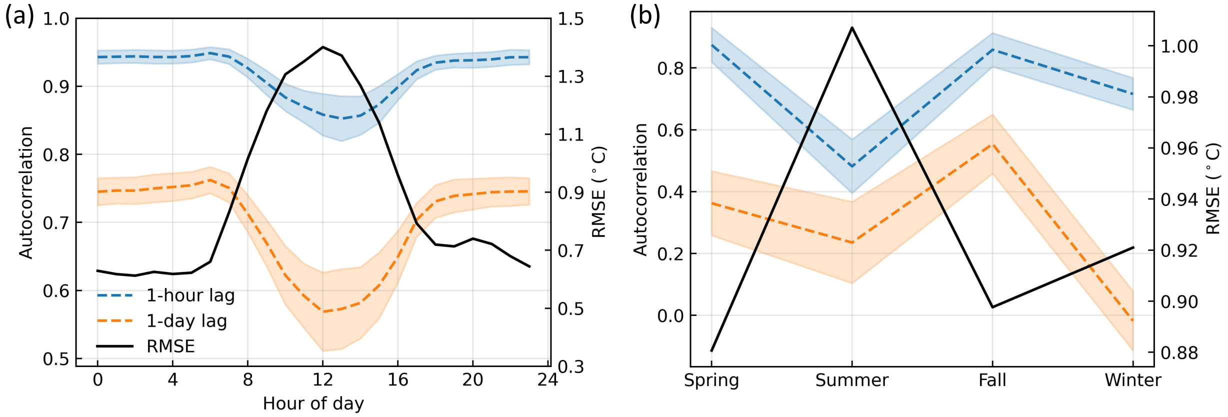

**Figure S8. (a)** Diurnal and **(b)** seasonal variations in the autocorrelation coefficients of observed $T_a$ at 1-hour and 24-hour time lags (dashed line), together with the corresponding RMSE of Hyper-GSAGE forecasts (solid line). Shaded areas denote one standard deviation across all stations. Both lag correlations exhibit a pronounced midday minimum, indicating diminished $T_a$ persistence during the daytime, and enhanced persistence at nighttime. RMSE varies inversely with autocorrelation coefficients, indicating greater forecast uncertainty during periods of lower persistence. 85

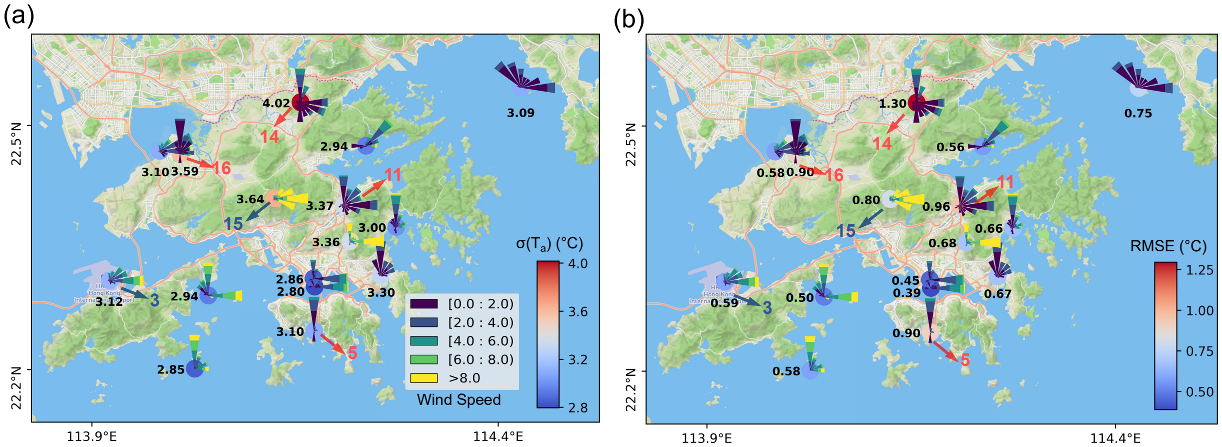

**Figure S9.** Local wind with corresponding **(a)** $T_a$ variability (i.e. time-series standard deviation) and **(b)** RMSE during winter nighttime (basemap © Mapbox). The wind pattern at each site is denoted by Windrose map, where the length denotes the frequency, and color denotes the velocity. Consistent wind distributions in mountains peak (station No.15) and plain airport (station No.3) demonstrate the easterly background wind, while locations No.5, 11, 14 and 16 show larger variances and forecast errors with northern wind.

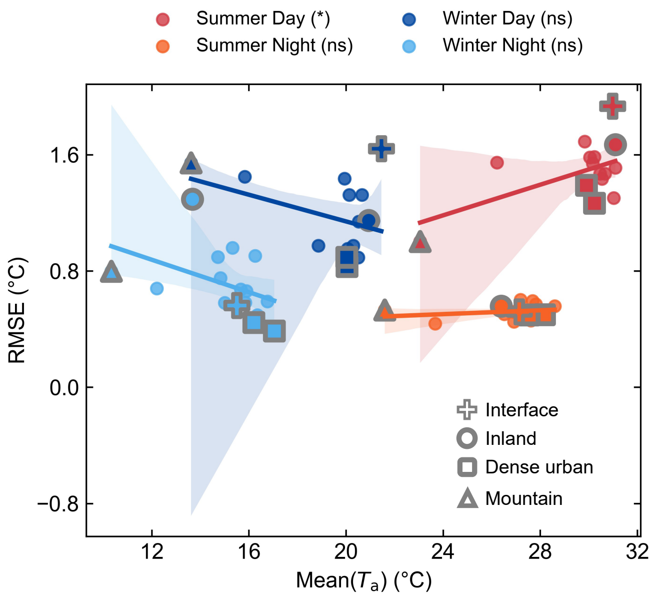

**Figure S10.** The relationship between local mean $T_a$ and forecast RMSEs by Hyper-GSAGE. Each point represents a location, and shaded areas indicate 95% confidence intervals derived from bootstrapping. Key locations discussed in the text are highlighted using the shapes indicated in the legend. A single asterisk (*) next to the period indicates a significant relationship ($p \leq 0.05$), while “ns” denotes non-significance.