

Downscaling urban land surface temperature using AlphaEarth satellite GeoAI embeddings: A cross-city evaluation in the US

Bijoy Mitra^{*}, Guiming Zhang^{*}

Department of Geography & the Environment, College of Natural Sciences and Mathematics, University of Denver, CO 80208, USA

ARTICLE INFO

Keywords:

AlphaEarth foundations
Deep learning
SHAP
Spatial prediction
Land surface temperature
Urban thermal stress

ABSTRACT

Fine-scale land surface temperature (LST) data are critical for urban heat management, yet open-source LST datasets have long been limited to 30–500 m resolution. Recently, Google released its AlphaEarth Foundations (AEF), a 64-dimensional GeoAI embedding dataset covering the entire globe at a spatial resolution of 10 m. This study introduces a novel application and comprehensive evaluation of the AEF embedding dataset for downscaling 30-m Landsat LST to 10-m resolution in urban environments using deep learning. As case studies, LST downscaling was conducted across 26 US cities using an artificial neural network that takes the embeddings as input to predict LST values and is trained with spatially stratified samples. We achieved a mean root-mean-squared error (RMSE) of 1.57 °C and a mean R^2 of 0.733 on the test set. Spatial cross-validation reveals minimal deviations (standard deviation of RMSE $\sim$ 0.0012 on the log-transformed scale) across the folds. The prediction errors were not spatially clustered and had no significant hotspots for any of the cities. Further, SHAP analysis demonstrated that AEF layers 31, 18, and 1 had the highest feature importance for predicting LST across the cities. Finally, the model trained for each city was applied to predict LST in other cities to assess the model's cross-city geographic transferability. Model transferability succeeds across places with similar forest structure or aridity but is hindered by extreme temperatures and the distinct characteristics of tropical ecosystems. The downscaled high-resolution LST product can be used to monitor urban thermal stress and inform neighborhood-level climate adaptation planning.

1. Introduction

Land surface temperature (LST) is a fundamental Earth surface parameter and is one of 10 essential terrestrial climate variables identified by NASA, the US Climate Change Science Program, and the Global Climate Observing System (Hollmann et al., 2013). LST drives critical processes including surface energy balance, evapotranspiration, and regional climate dynamics (Peng et al., 2014). However, global mean temperatures are rising rapidly, with human-induced warming reaching 1 °C above pre-industrial levels in 2017, at a rate of 0.2 °C per decade (IPCC, 2022). Current Paris Agreement pledges are insufficient to prevent warming from exceeding 1.5 °C, thereby threatening food security, ecosystem stability, and urban habitability through intensified heat waves, floods, and sea-level rise (IPCC, 2022, 2023; Solomon, 2007).

Nevertheless, the local climate in cities and urban regions often differs from that of surrounding rural landscapes (Heinl et al., 2015; Tran et al., 2017). While the global population is increasingly concentrated in urban areas, these areas are also major economic hubs

^{*} Corresponding authors.

E-mail addresses: bijoy.mitra@du.edu (B. Mitra), guiming.zhang@du.edu (G. Zhang).

that attract more people to relocate. Currently, 55% of the world's population lives in cities, and by 2050, this figure will rise to 68% (van Dijk et al., 2021). Such intense urbanization has severely strained urban ecosystems and services, with far-reaching demographic and environmental effects. Urban areas have the highest levels of soil sealing and continuous built-up areas, thereby negatively impacting ecosystem services (Artmann, 2014; Laforteza et al., 2009). Moreover, urban materials such as concrete and asphalt have high thermal capacity and low albedo (Artmann, 2014). They absorb more incoming shortwave solar radiation and slow nighttime cooling, trapping heat in the urban environment and raising LST. Such heat island effects may worsen in urbanized regions due to urban expansion and anticipated global warming (Laforteza et al., 2009).

Several studies have incorporated an integrated LST approach to assess the health of urban areas worldwide. Mentaschi et al. (2022) found that 3-day surface urban heat island (SUHI) extremes were often more than twice the warm-season median SUHI in metropolitan areas globally between 2003 and 2020, with local exceedances reaching 10 Kelvin. In the US, the incidence of elevated temperatures is also evident. Three North American regions have quite unusual LST increases: the freezing northwest region ($0.12^{\circ}\text{C}/\text{y}$), the west coast from 20°N to 40°N ($0.07^{\circ}\text{C}/\text{y}$), and the tropics south of 20°N ($0.04^{\circ}\text{C}/\text{y}$) (Yan et al., 2020). Since 1970, 240 American cities and all 50 states have experienced warming, with Reno, Nev., recording the largest increase in LST (7.8°F), followed by Las Vegas, Nev. (5.9°F) and El Paso, Texas (5.7°F) (Xian, 2023). According to the USGS, the average trend in the 50 US cities' mean SUHI intensity is 0.32°C per decade (0.58°F per decade), with a maximum temperature of 2.88°C (Xian, 2023).

The macro-urban form, building morphology, reflective and thermal characteristics of constituent land-surface materials, and the intensity of energy-related urban activities all influence the spatial variation in LST (Li et al., 2023; Morabito et al., 2016). Metropolitan areas exhibit high spatial variability in temperature, but meteorological station networks only provide partial representations of air-temperature variations in heterogeneous urban/suburban environments (Li et al., 2023). Remotely sensed thermal infrared (TIR) data from satellite instruments can capture a balanced image of LST on different spatial and temporal scales. Since the 1960s, numerous TIR instruments have been deployed on various satellites. This includes the Chinese Fengyun (FY) satellite series, the National Oceanic and Atmospheric Administration (NOAA) satellite series, the Landsat satellite series, the Earth Observing System Terra and Aqua satellites, and the Meteoros Second Generation (MSG) satellite series (Li et al., 2023).

MODIS provides LST data at 1–8-day intervals, but it is not suitable for monitoring cities due to its coarse spatial resolution (about 1 km). Moreover, MODIS LST contains missing values due to non-overlapping satellite orbits and instrumental failures (Weng et al., 2014; Zhang et al., 2022a). Landsat, on the other hand, offers 30-m-resolution data (TIRs are natively collected at 100-m and resampled to 30-m by the USGS), although it is highly sensitive to cloud contamination and requires extensive atmospheric correction (Weng et al., 2014). In addition, several instances of impervious surface mapping have been acquired using high-resolution satellite imagery, e.g., IKONOS, ISPR, QuickBird, and RapidEye (Lu and Weng, 2009; Munaf et al., 2015; Zhang and Guindon, 2012). These, however, are only accessible for particular geographical regions.

As urban populations continue to grow and their economic potential increases, precise temperature monitoring in urban environments and understanding of their spatial dynamics are critical. Existing LST studies are limited to either spatiotemporal LST calculations from satellite instruments (Heinl et al., 2015; Morabito et al., 2016; Pande et al., 2024; Tran et al., 2017) or the development of machine learning (ML)-based predictive models (Jia et al., 2021; Pande et al., 2024; M. T. Rahman et al., 2017). Applications of ML to develop LST models fall into two major categories: developing models from land-cover indices and enhancing spatial or temporal resolution by combining multiple satellite outputs. For instance, Zhang et al., 2022a presented a seamless daily LST dataset (1-km resolution) interpolated using correlation weighting to fill gaps in the LST database. (Tanoori et al., 2024) tested 6 ML algorithms, e.g., Decision tree, RF, DNN, SVM, AdaBoosting, and XGB, to predict LST from landcover compositions including built-up, vegetation, and soil. Pande et al. (2024), however, tested LST prediction using climatic indices with XGBoost, Bagging-XGBoost, and AdaBoost, and found the Bagging ensemble to be among the best predictive models. The use of deep learning is also evident. Jia et al. (2021) presented a conv2D-LSTM architecture for extracting features from land cover and climatic variables to predict LST. Purely convolutional-based structures, e.g., U-Net (Han et al., 2024) and 3D-CNN (Fu et al., 2022) were also used to model LST.

Artificial Neural Network (ANN) is one of the prominent models for modeling and predicting LST. (Shwetha and Kumar, 2016) presented the prospect of using infrared and passive microwave to predict LST under cloudy sky conditions. (Karunakaran and Sudalaimuthu, 2025) incorporated geographically weighted regression (GWR) and ANN to develop LST prediction models from meteorological, LULC, and soil texture parameters. (Olafsson et al., 2025) also presented an ANN model that simulated day- and nighttime LST from MODIS data. ANN has also been combined with Cellular Automata-Markov Chain to simulate future LST (Kafy et al., 2021; Rahman and Esha, 2022).

The above works indicate that deep learning (DL) models are applicable to LST studies. However, these studies rely heavily on handpicked physical or thematic parameters to model LST. Moreover, the model's generalizability to other geographic regions (i.e., geographic transferability) is rarely examined. LST modeling and prediction across regions are often difficult due to missing data and challenges in generalizing models across space and time (Lute and Luce, 2017). Furthermore, climate variability often impedes the generalization of DL models due to physical barriers such as impacts, geographical or topographical configurations, and socioeconomic structures, even though DL models are more resilient in capturing spatial patterns (Priyatikanto et al., 2023).

Recently, Google's AlphaEarth Foundations (AEF) has published a 10-m high-resolution geospatially embedded vector dataset that combines major multispectral and hyperspectral imagery (Brown et al., 2025). AEF powers Google Earth Engine's Satellite Embedding dataset, which has approximately 1.4 trillion embedding footprints every year. Their model's learning efficiency, as measured by its average error rate, was 24% lower than that of the tested deep learning and classical geospatial models (Brown et al., 2025). AlphaEarth embeds data into a compact representation that highlights stable land features and reduces atmospheric/radiometric fluctuation, overcoming sensor-specific restrictions. It compresses hundreds of noisy bands into a compact, information-rich embedding that suppresses air noise while retaining the most important spectral properties of land features. In particular, the dataset is not raw satellite

image outputs but rather derived features from a massive vision transformer AI model (ViT); thus, it is free of cloud and atmospheric noise (Brown et al., 2025).

Handcrafted indices like NDVI or albedo have limitations for modeling LST. They reduce complex land surface characteristics that impact LST to single spectral indices, and they are susceptible and sensitive to atmospheric noise or sensor artifacts. Moreover, such manual feature selection heavily relies on domain knowledge and a set of features handpicked for one study area may not work well for other regions. AEF embeddings address these issues by condensing many spectral bands into 64 learned features that are already cloud- and noise-free, retaining the representation of complex land surface characteristics in an embedding space. The embeddings also provide a high-resolution, universal feature set for LST modeling for anywhere in the world. This leads to three research questions: (1) Can AEF embeddings be used to model and predict LST at 10 m accurately? (2) Do models trained in one study area (e.g., city) generalize well to other regions?

Consequently, this study presents a comprehensive evaluation of downscaling 30-m Landsat LST to 10-m in 26 US cities using AEF embeddings with an ANN model. This study contributes to the literature in three ways. First, it evaluates AEF embeddings as predictors for high-resolution LST downscaling across diverse urban environments. Second, it assesses cross-city model transferability to examine how well city-specific models generalize to other urban settings. Third, it employs SHAP-based AI explainability analysis to identify embedding bands critical for LST modeling and prediction. The models were spatially cross-validated within their respective cities, and the spatial pattern of prediction errors was examined. High-resolution LST at 10-m scale is critical for identifying intra-urban thermal gradients that drive UHI intensity. Coarser products like 500-m MODIS or 30-m Landsat LST cannot resolve neighborhood-scale heat patterns relevant to urban planning interventions. The downscaled LST produced in this study directly enables fine-scale UHI mapping, supporting targeted mitigation strategies such as the placement of green infrastructure and the identification of heat-vulnerable populations.

2. Methodology

2.1. Study area

With a population of approximately 347 million, the US ranks as the third-most populous and third-largest country in the world. The average urban settlement rate is roughly 80–85%. The US comprises almost 19,500 incorporated areas and 387 metropolitan statistical areas. Each area has a city with a population of more than 100,000 (Marshall, 2007; Nguyen et al., 2018; Seto et al., 2011). Most are significant commercial centers in the global economy. In recent years, Southern cities such as Austin and San Antonio, and Mountain West cities like Phoenix and Denver, have been among the fastest-growing. Some studies point to housing and job

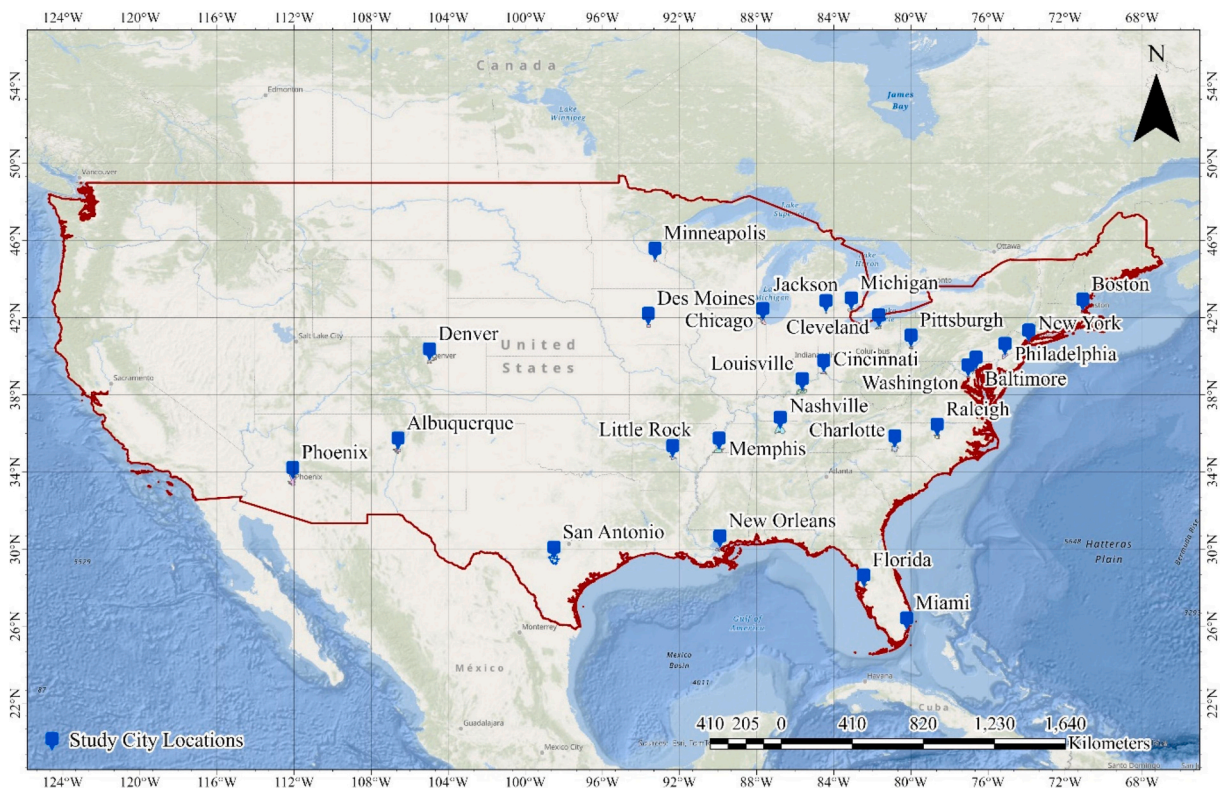


Fig. 1. 26 U.S. cities selected, representing a range of geographic regions, climate types, and urban development patterns.

opportunities as key drivers of urban growth (Marshall, 2007; Nguyen et al., 2018). Others note urban growth as a response to climate migration, for example, toward the South and West. Based on economic expansion, urban expansion, and potential effects of climate change, this study predicted LST for 26 major US cities experiencing significant heat stress. Fig. 1 presents the selected cities and their boundary extents.

2.2. Data

The AlphaEarth Foundations Satellite Embedding V1 is a robust geospatial embedding dataset at 10 m resolution. It contains geospatial embeddings learned from a variety of remote sensing imagery. While it is not conventional satellite imagery, it integrates optical, radar, LiDAR, and other data streams through an artificial intelligence (AI) model and operates like a virtual satellite (Brown et al., 2025; Tollefson, 2025). This dataset's 64-dimensional unified digital representation, termed as “embedding vector,” records the spatiotemporal trajectories of Earth surface conditions. The dataset is currently available for land surfaces, shallow seas, intertidal and reef zones, and interior and coastal waterways on an annual scale.

Here, embeddings are feature vectors that summarize connections across multi-source, multi-modal data in a less immediately interpretable but more powerful fashion than spectral imagery and indices, which correspond to physical measurements (Brown et al., 2025). These 64-byte summaries or “embeddings” use the combined encoded data from each source to reconstruct the temporal progression of variables (Brown et al., 2025). Primarily, the AEF embedding dataset integrates data from dozens of public sources, including optical satellite photos, radar, 3D laser mapping, climate models, and more. Then it makes the data usable by enhancing the model's ability to summarize each $10\text{ m} \times 10\text{ m}$ square concisely. These summaries significantly reduce the cost of planetary-scale research and need sixteen times less storage space than other algorithms we investigated.

Embedding representations overcome common issues such as clouds, scan lines, sensor artifacts, and missing data, providing seamless, analysis-ready features without major preprocessing or signal correction (Brown et al., 2025; Tollefson, 2025). The AEF embedding vectors for the year 2024 were used in this study to downscale 30-m-resolution Landsat LST data to 10-m-resolution LST. The 30-m LST dataset was derived from Landsat-8 Collection 2 Level-2 Surface Temperature imagery. This is the highest-quality, analysis-ready thermal product publicly available (Vermote et al., 2016). All scenes covering each study area for the full year 2024 were included systematically. No manual or subjective scene selection was applied. Landsat Collection 2 Level-2 products are radiometrically calibrated and atmospherically corrected by USGS before distribution (Masek et al., 2006). This eliminated any need for additional preprocessing by the authors. Thermal bands were rescaled using the official USGS scale factor (0.00341802) and offset (149.0) to generate analysis ready product within Google Earth Engine. Values were then converted from Kelvin to Celsius in accordance with standard USGS protocol. Cloud contamination was handled using a bitwise QA_PIXEL masking approach. Bits 3, 4, and 2 were used to flag and remove cloud, cloud shadow, and cirrus pixels, respectively (Foga et al., 2017). Only high-confidence clear-sky pixels passed this multi-condition mask. This conservative masking strategy ensured radiometrically clean observations entered

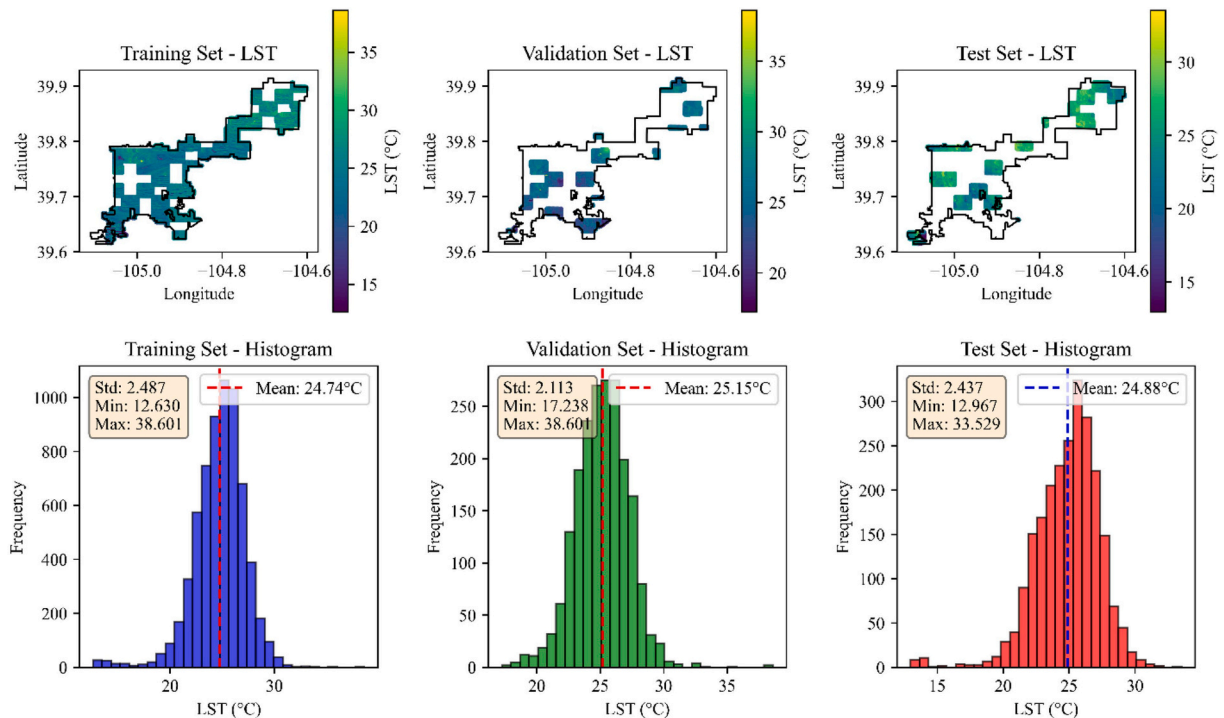


Fig. 2. A reference sample distribution of the training, validation, and testing sets for Denver.

the composite. Annual mean compositing across all valid clear-sky pixels reduced noise and removed single-date acquisition bias. The entire workflow was implemented in Google Earth Engine (Gorelick et al., 2017). This ensured full reproducibility and computational consistency across the study areas. The LST layer was exported at 30-m resolution in GeoTIFF format as the primary training target for the model. This systematic, standard-compliant pipeline produced spatially consistent, radiometrically reliable training data.

2.2.1. Data sampling technique

The training samples for each city were developed by extracting LST values and AEF embedding vector values from Landsat LST and Satellite Embedding V1 raster images at a set of sample locations. We adopted a fishnet strategy to select sample locations, sampling every 3 Landsat pixels to ensure sampling heterogeneity. For ML-based modeling, it is common to randomly split the dataset into 2 sets, e.g., training and testing, at an 80–20% ratio. However, this can introduce data leakage into the testing set and reduce model robustness. For our model development, spatial stratification was adopted to reduce data leakage, using a 64–16–20% split to obtain the training, validation, and testing sets. The approach first divides the dataset into spatial blocks, then uses an 80–20 ratio to obtain a random train-test split of the blocks, and finally leaves out 20% of the random blocks from the training set as a separate validation set (16% of the entire dataset). The validation set is used to monitor model performance during training and to determine when to stop training before the model overfits, without moving it to model testing. An illustration of the spatial stratification strategy applied to the city of Denver is presented in Fig. 2.

The fishnet sampling interval of one sample per three Landsat pixels was selected to avoid spatial redundancy among neighboring pixels while maintaining sufficient density to represent urban surface variability. Spatial blocks were defined by dividing the X and Y coordinates of each city into a regular grid of 196 bins (14×14), based on 30-m Landsat pixel coordinates. This block size was chosen to separate training and testing sets and reduce data leakage across block boundaries. This ensured that training and testing sets were spatially independent, minimizing data leakage across block boundaries. The minimum sample threshold of 9000 points per city was maintained to ensure sufficient representation of diverse urban land surface conditions across heterogeneous urban landscapes.

2.3. Methods

An overview of the methodology adopted in this study is as follows. An ANN model was created to learn each city's LST pattern using satellite data and assess cross-city transferability. Performance metrics such as coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and concordance correlation coefficient (CCC) were examined, and the models were further verified by 10-fold spatial cross-validation. Geostatistical validations such as Moran's I and Getis-Ord G_i^* were

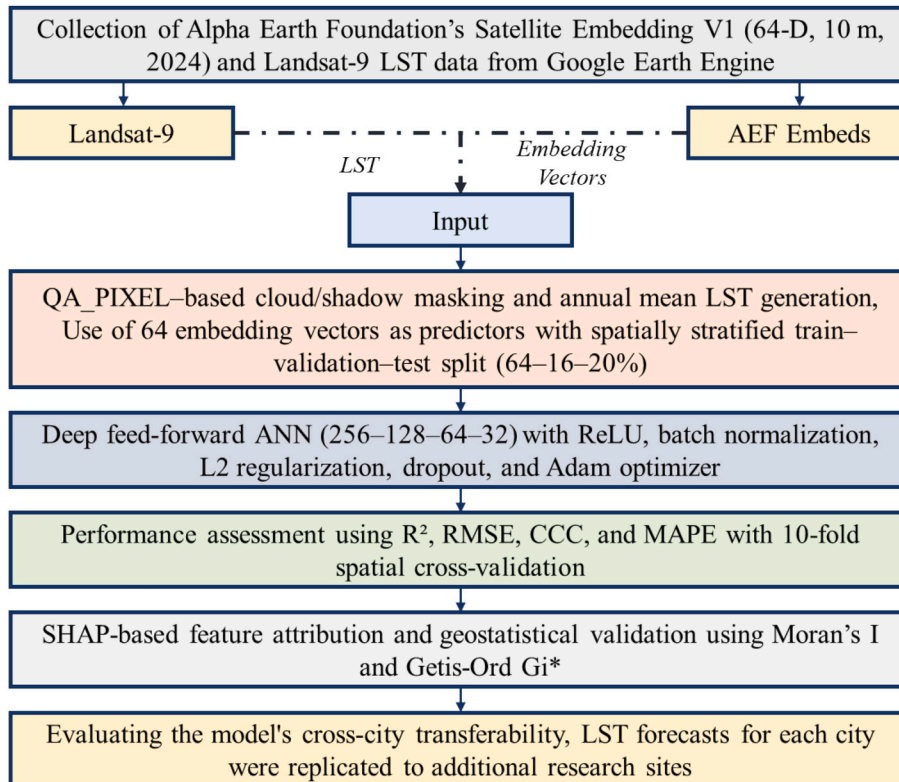


Fig. 3. Methodological Framework of the experimental GeoAI based LST prediction.

used to identify any possible geographic clusters of prediction errors. SHAP analysis was used to find the exact embedding layers required for LST prediction. Fig. 3 presents the study's methodological framework.

2.3.1. ANN model architecture and development

ANNs are hierarchical networks of many parallel, linked adaptive processing units that interact with the environment, similar to biological neural networks (Kohonen, 1988; Wu and Feng, 2017). ANN development results from simplified, brain-physiology-based mathematical models, and nodes process data simply. In particular, processing units are artificial neurons, the computational equivalent of real cells (Kohonen, 1988). We employed an artificial neural network for LST modeling and prediction. The input layer of the network comprises 64 nodes, and the output layer has only one node corresponding to the predicted LST. The network, featuring a regularized deep feed-forward architecture, was trained using backpropagation on spatially stratified samples.

Three-layer backpropagation neural networks are the most common connection pattern (Haykin, 2009; Wu and Feng, 2017). However, our experimental dataset comprises all 64 bands from the AlphaEarth embeddings and at least 9000 data points per city, which required an additional layer (Table 1). Consequently, we used four dense hidden layers (256–128–64–32 units) connected to a single output layer. The number of hidden-layer nodes was determined by trial and error. Moreover, for each hidden layer, ReLU activations, batch normalization, L2 regularization, and dropout were added to reduce overfitting. L2 regularization penalizes large weights and improves generalization. In each validation phase, backpropagation computes the mean squared error (MSE) loss gradients for all network weights using reverse automatic differentiation and updates them using the Adam optimizer. Early stopping was triggered if the MSE loss on the validation set did not improve for 30 consecutive epochs, although the overall training epoch count was set to 400. The entire ANN pipeline was developed using the Scikit-learn and Keras libraries in Python 3.9.

The AEF embeddings are constructed using representations obtained from satellite images that are not associated with specific physical factors. This reduces the model's physical interpretability, which is recognized as a natural trade-off of embedding-based techniques (Brown et al., 2025; M. Rahman, 2026). However, the fundamental goal of this work is to provide accurate high-resolution LST predictions rather than to explain processes. Spatially stratified sampling, dropout regularization, L2 weight penalization, batch normalization, and early stopping were used to mitigate the risk of spurious correlations, resulting in a robust generalization pipeline across a wide range of land surfaces.

2.3.2. Model performance evaluation and spatial cross validation

To evaluate model performance and to identify any potential overfitting or underfitting, each of the sample sets were utilized to compute model performance metrics e.g., coefficient of determination (R^2), root mean squared error (RMSE), concordance correlation coefficient (CCC), and mean absolute percentage error (MAPE). R^2 can be computed following Eq. (1).

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

R^2 measures how well the model predictions explain the variance in the observed values, where $\hat{y}_i$ and y_i are the predicted and observed values at location i , respectively, and $\bar{y}$ is the mean of all observed values.

RMSE is a statistical measure that quantifies the average difference between predicted and observed values by taking the square root of the average of the squared differences (Eq. (2)).

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

CCC assesses the agreement between the predicted and observed values, which may be used to assess consistency or interdependence. It can be computed following Eq. (3).

Table 1
GeoAI based ANN model architecture summary.

Layer	Output Shape	Number of Parameters
dense (<i>Dense</i>)	None, 256	16,640
batch_normalization (<i>BatchNormalization</i>)	None, 256	1024
dropout (<i>Dropout</i>)	None, 256	0
dense_1 (<i>Dense</i>)	None, 128	32,896
batch_normalization_1 (<i>BatchNormalization</i>)	None, 128	512
dropout_1 (<i>Dropout</i>)	None, 128	0
dense_2 (<i>Dense</i>)	None, 64	8256
batch_normalization_2 (<i>BatchNormalization</i>)	None, 64	256
dropout_2 (<i>Dropout</i>)	None, 64	0
dense_3 (<i>Dense</i>)	None, 32	2080
batch_normalization_3 (<i>BatchNormalization</i>)	None, 32	128
dropout_3 (<i>Dropout</i>)	None, 32	0
dense_4 (<i>Dense</i>)	None, 1	33

$$CCC = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\bar{\hat{y}} - \bar{y})^2} \quad (3)$$

Where, $\bar{\hat{y}}$ and $\bar{y}$ are the means for the predicted and observed values and $\sigma_{\hat{y}}^2$ and σ_y^2 are the corresponding variances, ρ is the Pearson's correlation coefficient between the two variables. Finally, MAPE measures mean prediction accuracy as a ratio specified by the following Eq. (4).

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

In addition to evaluating model performance on the training, validation, and testing datasets, a 10-fold spatial cross-validation was performed to identify model consistency and to prevent over-fitting by testing on geographically distinct areas. The ANN model was trained on data points in any nine spatial block groups and tested on data points in the one held-out block group for LST prediction. All performance metrics (R^2 , RMSE, CCC, and MAPE) were computed on each of the 10 held-out folds to evaluate prediction accuracy and consistency. A general spatial cross validation can be assessed by Eq. (5).

$$\hat{E}_{spatial} = \frac{1}{K} \sum_{k=1}^K \mathcal{L}(yG_k, \hat{f}(xG_k; D - G_k)) \quad (5)$$

Where G_k denotes the k -th spatial group, and $\hat{f}$ is trained on data excluding G_k . The equation averages the loss $\mathcal{L}$ across K spatial blocks, where each block yG_k is a geographic subset held out for testing while the model $\hat{f}$ is trained on the remaining data $D - G_k$, using the true labels yG_k and input features xG_k from that block.

All cross-validation metrics were computed on the log-transformed LST scale used for model training. For each city, model metrics (e.g., R^2 , RMSE, MAPE, and CCC) were calculated across the 10 folds, and the standard deviation of these fold-level model metric values was used to assess spatial consistency.

2.3.3. SHAP analysis

Explainable Artificial Intelligence (XAI) provides insights into black-box models, where concerns, e.g., transparency and confidence in the models, are prevalent. In recent years, technologies have been widely used to explain how black-box models operate. Furthermore, research has demonstrated that these methods can be used as feature selection tools, improving model accuracy while reducing computational costs during training (Lundberg and Lee, 2017).

The Shapley values, which represent the optimal solution in game theory, are the basis for SHAP (SHapley Additive exPlanations) analysis, a method for elucidating the importance of individual predictor variables (Lundberg and Lee, 2017). SHAP is a post hoc approach, meaning it is used after a model has completed its training phase. The SHAP explainers are interfaces that enable efficient computation of SHAP values and can be used with the trained model's output vector. With these SHAP values, we may understand the data using various graphs.

SHAP technique provides a straightforward and obvious way to see how various characteristics affect the model's predictions. SHAP leverages cooperative game theory to analyze model predictions (Lundberg and Lee, 2017; Shapley, 1953). In this technique, Shapley values are generated, in which the independent variables are understood as players collaborating to obtain the payoff, which, in this instance, is the model's specific forecast minus the average of all predictions (Shapley, 1953). SHAP values may help in explaining why an ANN model makes a particular prediction or which features influence the choice. SHAP values measure the relevance of each feature, allowing identification of the characteristics that drive the model's predictions, thereby aiding both model comprehension and debugging.

2.3.4. Spatial pattern in prediction errors

After the modeling, we further evaluated whether the sample prediction errors (measured by absolute prediction error) convey any spatial autocorrelation or localized clusters. Therefore, Global Moran's I and Getis-Ord G_i^* were evaluated for the absolute error scores for each set of samples (the training, validation, and testing sets). Spatial autocorrelation assesses the correlation between a variable and its geographic location, reflecting the relationship between geographic proximity and attribute similarity (Cox et al., 1984; Moran, 1950). Moran's I is the predominant test statistic for assessing spatial autocorrelation across an entire area. The value of Moran's I typically ranges from 1 to -1 . Positive autocorrelation in the data results in positive values of Moran's I , whereas negative autocorrelation yields negative values. A lack of autocorrelation yields a value around 0 (Getis, 2009; Griffith et al., 2003).

The geographical cluster of LST prediction errors is defined by hotspot and cold-spot analysis utilizing Getis-Ord G_i^* statistic. The Z -score indicated whether cities had high or low cluster spatial values (Getis and Ord, 1992; Ord and Getis, 1995). Statistically significant large positive Z scores showed intensive clustering of high values (hot spot), whereas large negative Z scores indicated intense clustering of low values (cold spot). No geographic grouping is indicated by a Z score around zero. The hot spot analysis tool calculates the Getis-Ord G_i^* statistics for each dataset (Getis and Ord, 1992; Ord and Getis, 1995).

2.3.5. Model transferability evaluation

For this study, we developed separate ANN models for each of the 26 cities and performed a cross-city transferability evaluation to assess how efficiently these models perform in diverse urban settings. We employed the trained model from each city to predict the LST

for all the other cities. For example, the model developed for Denver was used to predict LST values in all the other 25 cities. The coefficient of determination and MAPE between the Landsat-derived LST and the model-predicted LST were then calculated for each city. The resulting coefficient of determination and MAPE matrices demonstrate how accurately a model can predict LST in other cities. Here, each row shows the transferability of a 'source' city model to all other locations, while each column shows how well a 'target' city was predicted by each model in the study. The diagonal components show how each city predicts its own LST. With this approach, we were able to systematically evaluate model transferability and determine which cities' models generalized better. The entire North America is divided into 15 major homogeneous ecoregions based on topography, land cover, and climatic characteristics (Bailey, 1976; Omernik and Griffith, 2014). After the inter-city transferability assessment, the models were further evaluated by ecoregion. 5 ecoregions were identified, e.g., Eastern Temperate Forests (ETF), Great Plains (GP), North American Deserts (NAD), Northwestern Forested Mountains (NFM), and Tropical Wet Forests (TWF).

A transfer efficiency matrix was calculated by dividing the average MAPE when other cities' models predict it by the model's own MAPE. This indicates how well the model performs across different cities. The efficiency can be achieved as follows in Eq. (6):

$$TE_{city} = \frac{\sum_{j \neq i} MAPE_{j \rightarrow city}}{(n-1)MAPE_{self \rightarrow city}} \quad (6)$$

Here, $\sum_{j \neq i} MAPE_{j \rightarrow city}$ is sum of MAPEs from all other cities models predicting target city, n is the number of cities (25), and $MAPE_{self \rightarrow city}$ presents when the city uses its own model. A transfer efficiency scores close to 1 presents that the target city model has the highest capability for generalization for other cities.

3. Results

3.1. Model development and evaluations

3.1.1. Model performance

The inter-city model performance for predicting high-resolution LST is presented in Table 2. The highest R^2 on the testing set was obtained for New Orleans ($R^2 = 0.89$), followed by Minneapolis ($R^2 = 0.85$) and Tampa ($R^2 = 0.83$). Similar high CCC scores were also observed for New Orleans (0.94) and Minneapolis (0.92). On the other hand, the lowest model performance was observed for Charlotte, with an R^2 of 0.52, an RMSE of 0.0326, and a MAPE of 1.79%. Although New York had a comparatively high R^2 (0.78), it also had the highest RMSE (0.228). The models trained for all these cities showed minimal differences in their metrics across the training, validation, and testing sets.

An interesting observation was noted for Phoenix. Although its R^2 was among the lowest (0.63, 0.67, and 0.58 on the training,

Table 2

Model performance metrics across training, validation, and testing datasets for all 26 study cities. All reported metrics (R^2 , RMSE, CCC, and MAPE) are computed on log-transformed LST values, consistent with the log-scaled target variable used during ANN model training.

	Training				Validation				Testing			
	R^2	RMSE	CCC	MAPE (%)	R^2	RMSE	CCC	MAPE (%)	R^2	RMSE	CCC	MAPE (%)
Albuquerque	0.8265	0.0132	0.9088	0.67	0.8258	0.013	0.9043	0.68	0.7632	0.0153	0.8627	0.79
Baltimore	0.8335	0.0274	0.9074	1.48	0.7585	0.0313	0.8625	1.7	0.8319	0.0351	0.9085	1.86
Boston	0.7765	0.0279	0.8671	1.53	0.6415	0.0379	0.7942	2.06	0.6971	0.0296	0.8138	1.57
Charlotte	0.578	0.0303	0.7253	1.71	0.5946	0.0291	0.7443	1.64	0.5278	0.0326	0.6977	1.79
Chicago	0.7444	0.0272	0.8443	1.5	0.6367	0.0313	0.7772	1.69	0.5618	0.0268	0.7242	1.49
Cincinnati	0.6515	0.039	0.7839	2.22	0.66	0.0368	0.7953	2.12	0.5678	0.0435	0.7323	2.57
Cleveland	0.8695	0.0243	0.9284	1.32	0.5209	0.0312	0.6761	1.46	0.7192	0.0263	0.8235	1.41
Denver	0.8073	0.02	0.8897	1.1	0.6159	0.0229	0.763	1.24	0.7627	0.0224	0.8677	1.25
Des Moines	0.8243	0.019	0.905	1.09	0.7811	0.0221	0.8781	1.3	0.6991	0.0212	0.8387	1.25
Detroit	0.6504	0.0327	0.8232	1.7	0.8232	0.0327	0.9098	1.8	0.7778	0.0337	0.8974	1.87
Jackson	0.868	0.0149	0.9269	0.78	0.8626	0.0176	0.9281	0.93	0.79	0.0239	0.8821	1.16
Little Rock	0.8174	0.0213	0.8989	1.15	0.7705	0.0254	0.8717	1.39	0.8014	0.0255	0.8873	1.37
Louisville	0.8083	0.0262	0.8928	1.45	0.799	0.0263	0.8869	1.44	0.8018	0.0278	0.886	1.51
Memphis	0.854	0.0223	0.9181	1.19	0.8296	0.025	0.9058	1.35	0.8124	0.0246	0.8958	1.3
Miami	0.8101	0.0158	0.8946	0.8	0.7237	0.0185	0.8437	0.92	0.7608	0.0187	0.8704	0.95
Minneapolis	0.8137	0.0226	0.8946	1.24	0.645	0.0256	0.7852	1.42	0.8592	0.0228	0.9204	1.23
Nashville	0.7826	0.0312	0.8732	1.77	0.6777	0.0326	0.8133	1.81	0.6607	0.0338	0.8006	1.9
New Orleans	0.8944	0.02	0.9457	1.06	0.8532	0.0217	0.9247	1.16	0.889	0.0217	0.9431	1.17
New York	0.8943	0.0231	0.9413	1.23	0.8655	0.0237	0.928	1.25	0.787	0.0228	0.8861	1.18
Philadelphia	0.8513	0.0207	0.9201	1.09	0.8383	0.0215	0.9098	1.11	0.8659	0.0257	0.9277	1.36
Phoenix	0.6386	0.0145	0.7819	0.67	0.6723	0.0141	0.8046	0.65	0.5815	0.0141	0.7386	0.67
Pittsburgh	0.7355	0.0307	0.8479	1.65	0.6769	0.0334	0.8038	1.86	0.6969	0.0316	0.8256	1.8
Raleigh	0.8372	0.0209	0.9086	1.19	0.8151	0.0224	0.8977	1.28	0.7773	0.0242	0.874	1.39
San Antonio	0.768	0.0213	0.8679	1.09	0.5685	0.0238	0.7369	1.19	0.5316	0.0255	0.7002	1.28
Tampa	0.8456	0.0173	0.9161	0.88	0.7767	0.0216	0.8731	1.12	0.8389	0.0206	0.9115	1.05
Washington	0.8512	0.0244	0.9169	1.24	0.8325	0.0209	0.9068	1.16	0.8117	0.0211	0.8956	1.15

validation, and testing sets, respectively), Phoenix also exhibited the lowest RMSE (0.014, 0.014, and 0.014) and MAPE (0.67%, 0.65%, and 0.67%). Prominent model overfitting was evident in San Antonio and Charlotte, where model performance dropped significantly on the validation (R^2 of 0.56 and 0.59, respectively) and testing (R^2 of 0.53 and 0.52, respectively) datasets compared to the training dataset (R^2 of 0.76 and 0.74, respectively). For Cleveland and Denver, both R^2 and CCC dropped significantly on the validation datasets, while the differences between the training and testing datasets for these metrics remained comparatively stable.

The ANN model was benchmarked against Random Forest (RF) and XGBoost (XGB) using data with similar distributions. Here, ANN outperformed in most cities, achieving higher R^2 and CCC values and lower RMSE and MAPE. In contrast, RF and XGB showed significant overfitting, as shown by their outstanding training performance but much worse validation and testing outcomes (Supp Table 5).

Fig. 4 shows the scatter density plot of the observed versus predicted LST values for the testing set. Here, most models predicted LST values in good agreement with observed LST values, with a few under- or overpredicting. New Orleans, Philadelphia, Minneapolis, Tampa, etc., have high R^2 (>0.83), CCC (>0.9), low RMSE (<0.21 °C), and MAPE ($<1.05\%$). In San Antonio, the majority of the LST values were over-predicted (for Landsat LST values between 30 °C and 32 °C), while some were under-predicted (for Landsat LST values over 40 °C). In Phoenix and Chicago, many of the LST values in the 37°–40 °C and 20°–25 °C ranges, respectively, were also over-predicted. Study areas including Des Moines, Little Rock, Denver, and Washington, despite achieving comparatively good model performance, also exhibited underprediction when LST values exceeded 38 °C.

The distribution of MAPE was further analyzed across the training, validation, and testing datasets for each of the 26 models (e.g., Fig. 2). All MAPE distribution graphs across the three sets were right-skewed and showed no substantial discrepancy between the mean and the median. However, localized errors were observed in some cities. Across the training, validation and testing sets, the few statistically significant Moran's I values are very low (e.g., <0.2), and comparatively high Moran's I values between 0.4 and 0.6 are not statistically significant (Supp Table. 1). The hotspot analysis using Getis-Ord G_i^* further confirms that no hot or cold prediction error clusters are present in the training, validation and testing sets. The prediction errors across all study cities are generally random and exhibit no significant spatial autocorrelation.

3.1.2. SHAP analysis

The SHAP analysis for each of the inter-city models reveal each AEF embedding band's relative importance and direction (positive vs. negative) of contribution in predicting LST (Fig. 5). From a general overview (Supp Fig. 5.1–5.26), Bands 31, 23, 38, 1, 46, 27, 44, and 11 were among the top 3 features with the highest relative importance across different models. Band 31 was the top feature for 4 cities, followed by Band 23 for 3 cities and Band 38 for 2 cities.

The top 4 models' (ordered by testing R^2) SHAP plots, each including the top 20 most important features, are shown in Fig. 5. Bands 23, 28, and 30 were the top 3 features for Minneapolis, Bands 52, 36, and 51 for Tampa, Bands 31, 23, and 44 for Philadelphia, and Bands 61, 56, and 35 for New Orleans. The most common features, e.g., Bands 23 and 31, had high feature values concentrated in negative SHAPs, indicating that these bands negatively influence LST prediction, whereas the opposite was observed for Bands 28, 36, 44, and 61. Other top features, e.g., Bands 38, 11, and 46, had a positive influence on LST predictions for cities such as Albuquerque, Chicago, and Nashville. In contrast, Bands 1 and 27 had a negative impact in cities such as Jackson, New York, and Washington. While individual AEF embedding bands lack direct physical labels, recent interpretability analyses demonstrate that embedding dimensions collectively map onto specific land surface properties, including temperature, vegetation, hydrology, and terrain (Rahman, 2026). The positive influence of Bands 38, 11, and 46 in arid and dense urban areas likely reflects latent representations of impervious surface density, albedo, and sparse vegetation, which are known LST drivers in these environments.

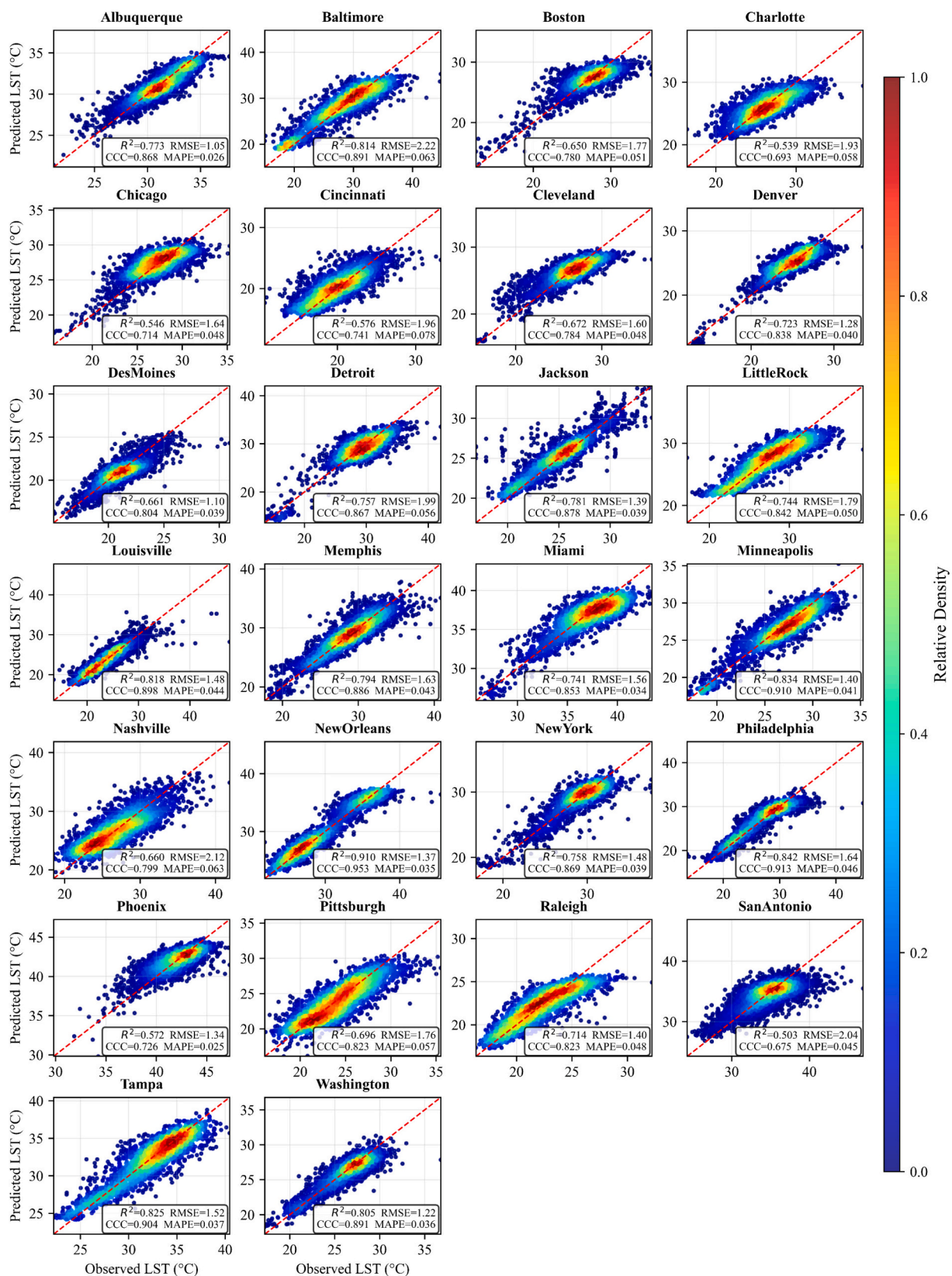
A co-occurrence network (Fig. 6) was constructed from the top 20 important features across all 26 models. This network demonstrates which individual features and feature pairs were most frequent across all models. Band 18 was one of the noteworthy features, appearing among the top 20 important features for 18 models, followed by Bands 31 and 1, which appeared for 17 models.

Furthermore, Bands 12 and 31, and Bands 18 and 31, were the most co-occurring band pairs across 12 models, followed by Bands 1 and 18 and Bands 31 and 44, which appeared in 11 models. Further, Bands 31 and 23 appeared among the top 5 features for 9 models, followed by Band 1 for 8 models and Band 46 for 6 models. Band 31 was also one of the most essential features for the four models. Band 00 and Band 58, however, were not among the top 20 features in any of the 26 models.

3.1.3. Spatial cross-validation

From the intra-city spatial cross-validation (Fig. 7), it is evident that LST prediction for New Orleans was the most consistent, with high accuracy across the spatial folds. For New Orleans, the mean R^2 and CCC were both above 0.9, with standard deviations (SDs) of 0.008 and 0.004, respectively. Other high-accuracy models, e.g., Detroit and Philadelphia, although they had comparatively high accuracy ($R^2 > 0.8$), had higher SDs (0.03 and 0.018, respectively) across folds. The most inconsistent model was observed in Cincinnati, where the SD scores for all the performance metrics were the highest among the 26 models. RMSE standard deviations across the 26 city models ranged from ~ 0.0002 (e.g., Jackson, Miami) to ~ 0.0069 (Cincinnati), with a mean of ~ 0.0012 on the log-transformed scale of spatial cross-validation.

A high average coefficient of variation is evident in Pittsburgh and Phoenix, at 0.046 and 0.044, respectively. Interestingly, Phoenix, although it had low R^2 (mean = ~ 0.6 ; SD = 0.042) and CCC ($\sim 0.65 \pm 0.027$), also had low RMSE ($\sim 0.015 \pm 0.0007$) and MAPE ($\sim 0.006 \pm 0.0002\%$), which were comparatively consistent across the folds.



(caption on next page)

Fig. 4. Scatter-density plots between observed and predicted LST on the testing dataset for all 26 cities. RMSE values are reported in $^{\circ}\text{C}$ scale following back-transformation from log-scaled model outputs, enabling direct physical interpretation of prediction errors.

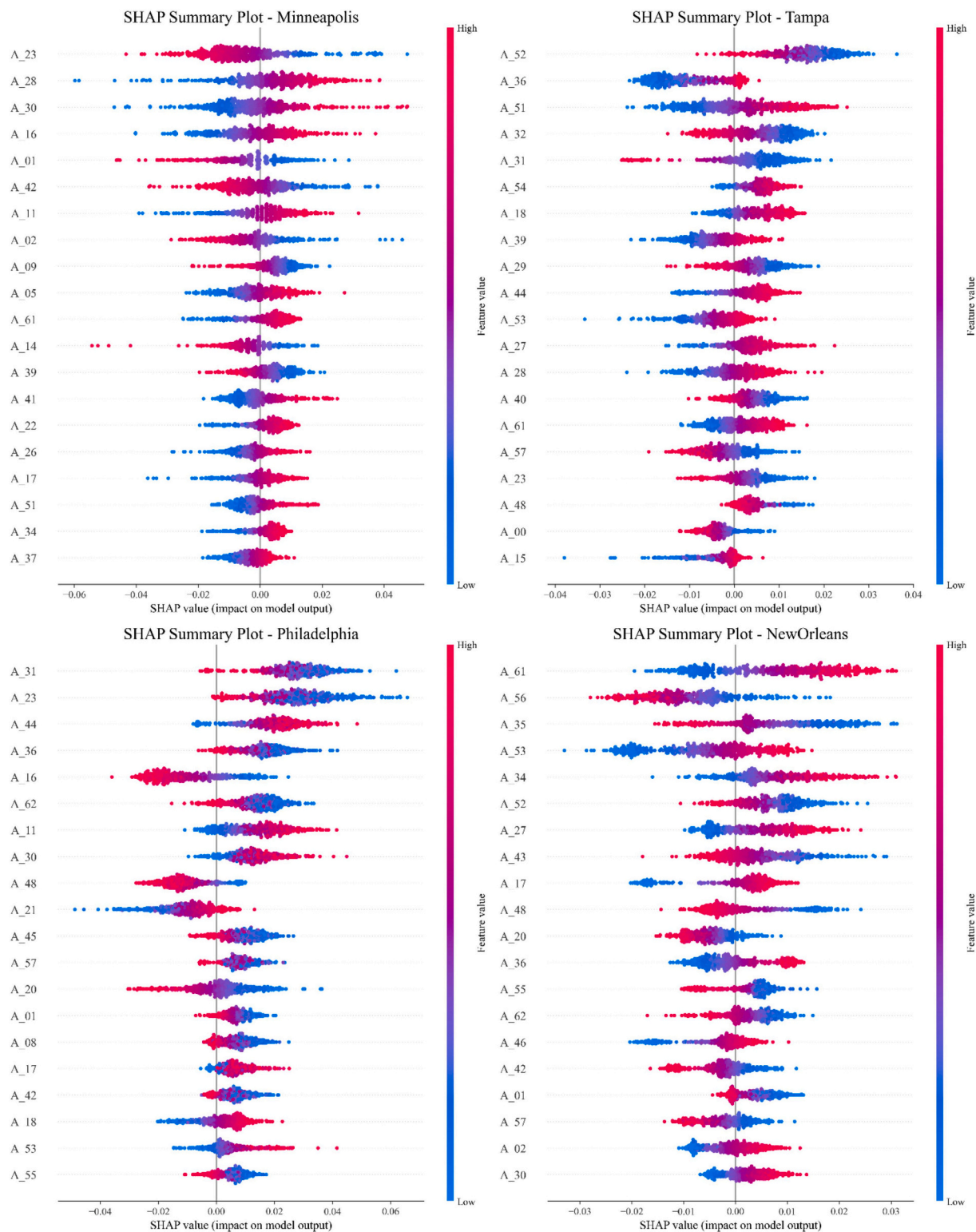


Fig. 5. Representative SHAP summary plots for the four best-performing models, showing the top 20 most important bands to predict LST.

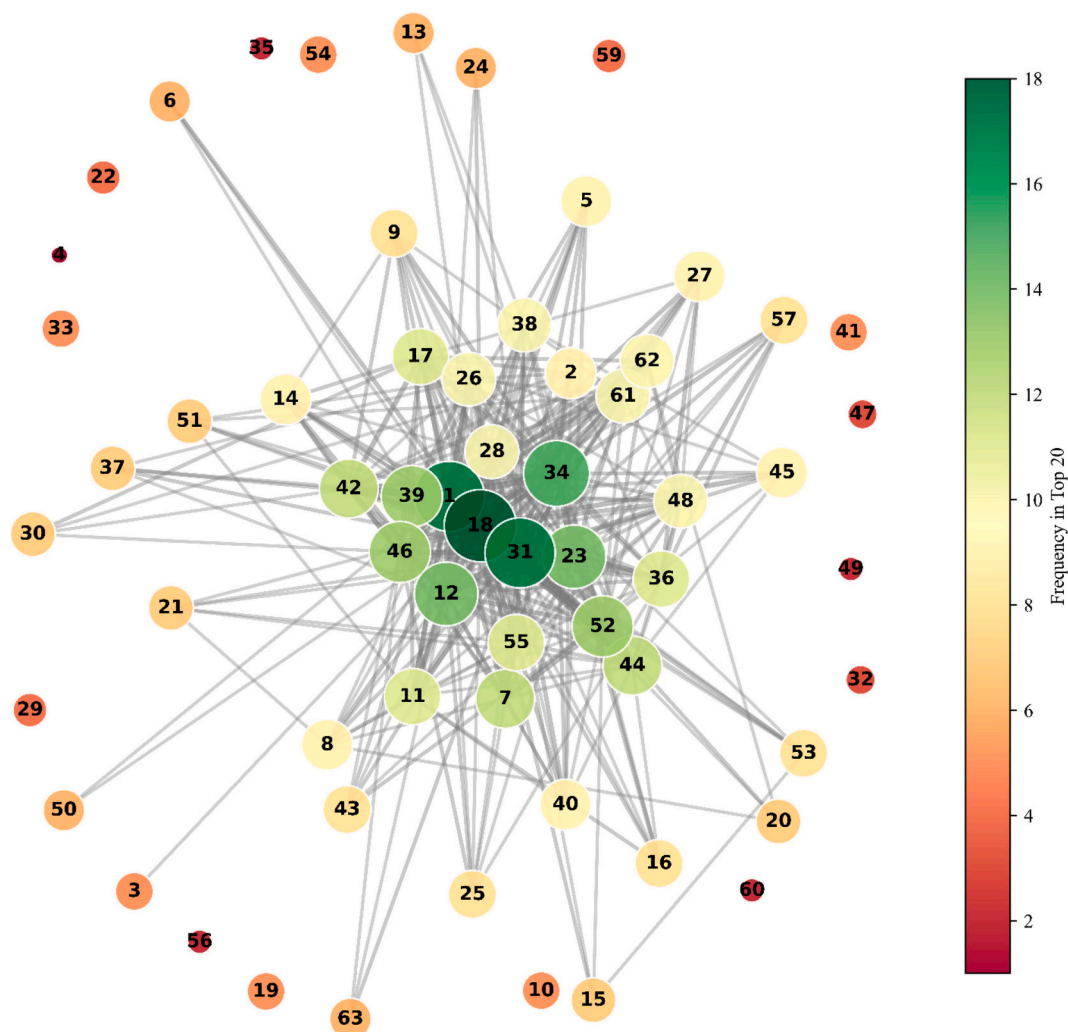


Fig. 6. A band co-occurrence network visualizing the top SHAP features from all 26 models. Node size and darkness represent a feature's frequency across cities, while edge thickness indicates co-occurrence strength; band numbers are provided as labels.

3.2. Model transferability evaluation

The inter-city model transferability test reveals that the model trained in Philadelphia had the lowest MAPE across the prediction target cities (mean = 10.47%; SD = 5.98%), while the model trained in Louisville was most consistently predicted LST across the target cities ($12.78 \pm 6.30\%$).

Although none of the models had a transfer MAPE lower than that achieved in its own city, the best transfer was observed for Cleveland to Jackson with an MAPE of 4.35%, followed by Tampa to New Orleans with an MAPE of 4.68% (Fig. 8). On the contrary, the worst transfer was recorded for Baltimore to Denver, where the MAPE is 113.75%. LST in New York was most accurately predicted by models trained in other cities, with an average MAPE of 13.76% (SD = 8.37%). LST predictions in Baltimore by models trained in other cities, however, were more consistent ($14.61 \pm 7.53\%$) than in New York. The model for Tampa had unusually low validation-set performance compared with high test-set metrics (Table 2). After transferability assessment, it is evident that it had an MAPE of 19.65% as a target city but an MAPE of 29.74% as a source city (Fig. 8). New York, on the other hand, had a comparatively stable bidirectional error ratio where the mean MAPE scores for other models predicting LST of NY and NY's model predicting LST for other cities are 13.91% and 13.88%, respectively.

To assess the transferability based on geography, we have further grouped the cities based on their North American ecoregions, e.g., ETF, GP, NAD, NFM, and TWF (Fig. 9). Here, NFM comprised the models with the lowest mean transfer MAPE of $15.80 \pm 8.724\%$, while ETF comprised the most consistent model transfer MAPE of $19.08 \pm 3.891\%$. The best cross-region transfer was NFM to GP, with a mean MAPE of 11.1%, followed by TWF to NAD, with a MAPE of 14%. Predicting LST for target cities in tropical wet forests seems least challenging, as the source models achieved a low mean transfer MAPE ($20.58 \pm 9.892\%$). However, predicting LST for target cities in GP was the most difficult, with the worst transfer MAPE from source models trained on cities in NAD (45.9) and TWF (46.5).

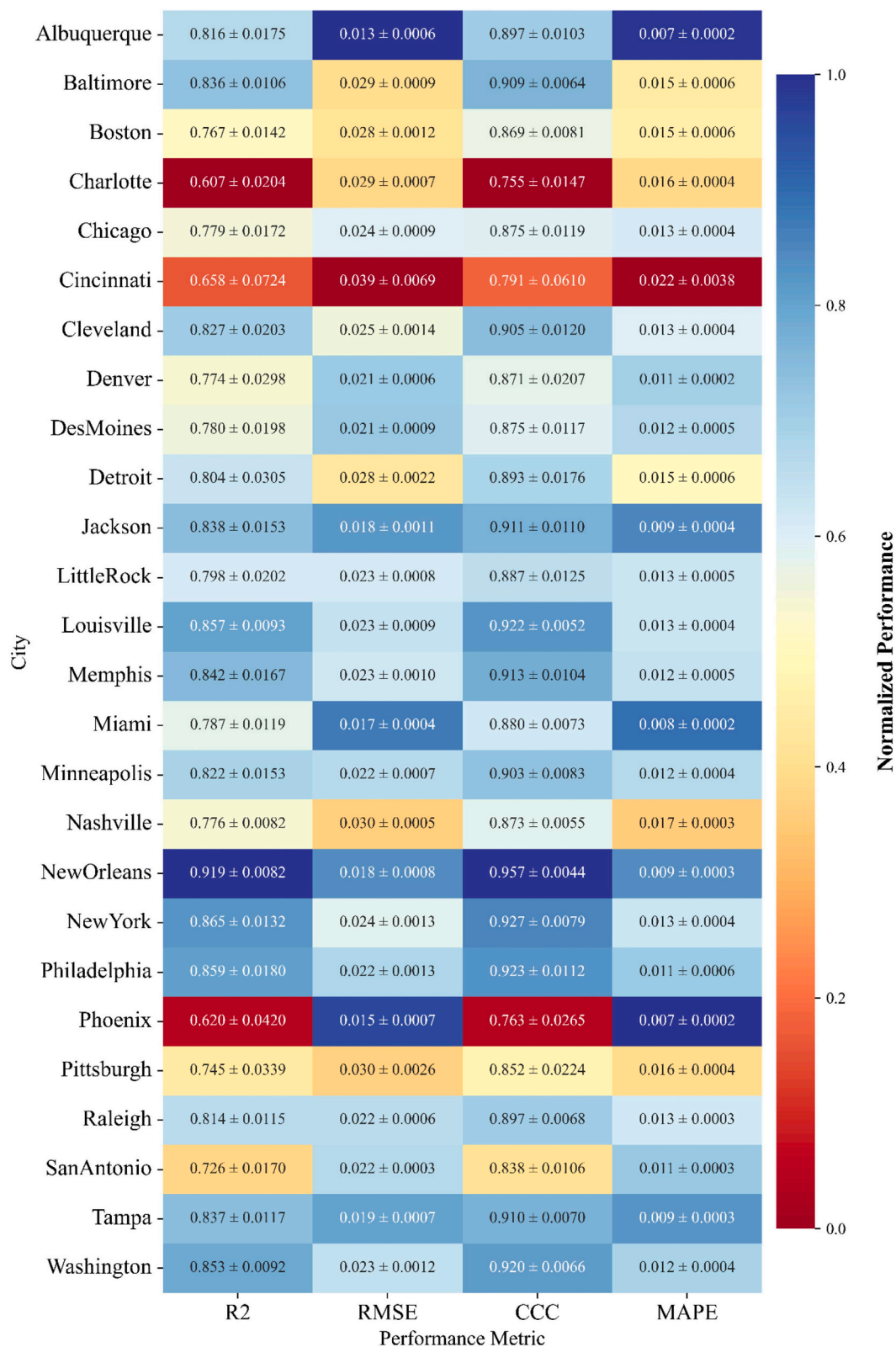


Fig. 7. Distribution of model performance metrics (R², RMSE, MAPE, CCC) computed on 10-fold spatial cross-validation for each of the 26 cities.

From Fig. 10, it is evident that Little Rock comprises the highest mean coefficient of determination of 0.569, followed by Raleigh ($R^2 = 0.559$) and Washington ($R^2 = 0.558$). On the contrary, San Antonio ($r = 0.28$) and Denver ($R^2 = 0.352$) had the lowest R^2 values when predicting other cities' LST. Washington, New York, New Orleans, Philadelphia, and Raleigh generalize models very efficiently, as they have comparatively higher prediction-accuracy coefficient and are also accessible for transfer.

Denver, while having comparatively low accuracy in predicting LST, was also among the least transferable, with a mean R^2 of 0.164 when predicting LST for Denver. However, it is a very interesting finding for Phoenix: when the model trained in Phoenix was used to predict LST for other cities, it had a very high mean MAPE, but the mean R^2 was 0.538. And even when using the other models to predict LST in Phoenix, the mean R^2 is 0.057 (e.g., Phoenix to predict Washington: $R^2 = 0.708$, MAPE = 51.4%). This indicates that the models trained in Phoenix and Denver are highly location-specific, and other models are not suitable to predict LST in these climates.

The transfer efficiency matrix (Fig. 11) shows that cities with Humid Subtropical climates (e.g., Charlotte, Louisville, Memphis, Nashville) transfer most effectively among themselves. In contrast, cities in the Mid-Atlantic region (Baltimore, Philadelphia, Washington) and in the South region (Jackson, Little Rock, New Orleans) show consistent performance when transferring models. Baltimore

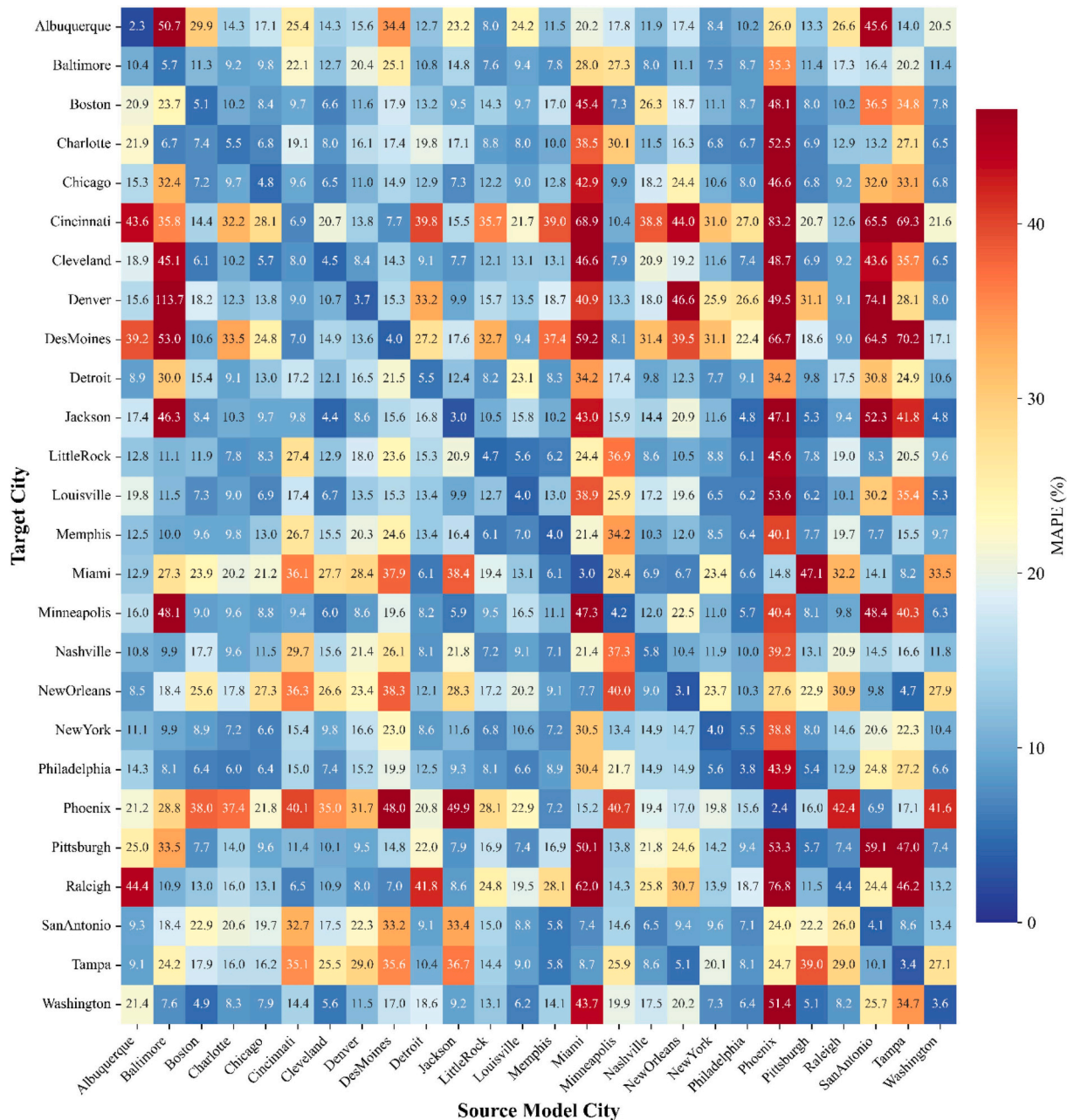


Fig. 8. MAPE evaluation matrix based on selected source (model) cities and target cities.

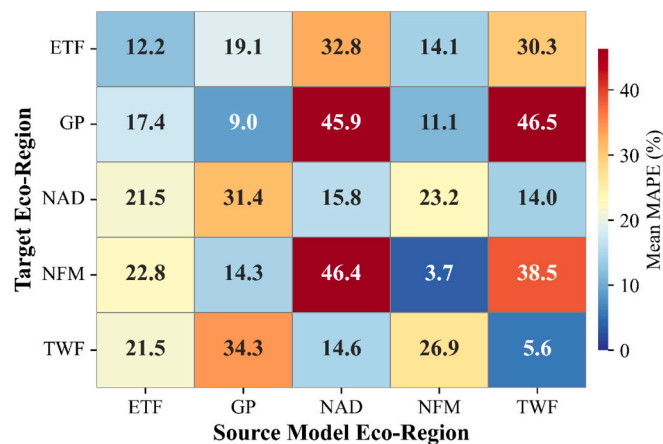


Fig. 9. Mean transfer performance (MAPE %) between five North American ecoregions, with diagonal elements representing within-region predictions and off-diagonal elements representing cross-region model transfers. Here, ETF: Eastern Temperate Forests, GP: Great Plains, NAD: North American Deserts, NFM: Northwestern Forested Mountains, TWF: Tropical Wet Forests.

had the highest transfer efficiency with a score of 2.63, followed by Nashville (2.85) and Charlotte (2.9). Furthermore, Baltimore's transfer learning had a standard deviation of 7.5. On the other hand, Raleigh (efficiency = 5.39, SD = 17.6), Jackson (efficiency = 6.16, SD = 14.6), and Denver (efficiency = 7.20, SD = 23.5) had among the poorest transferability, with low consistency. Based on Phoenix's and Denver's transfer efficiency, unique geographic features, such as desert basins or high elevations, reduce transferability to other cities.

4. Discussion

4.1. City-level model development assessment

We trained, validated, and assessed the models to identify the best-performing model, assess how well the models are transferable, and determine in which aspect the models increase prediction error. Firstly, spatial stratification was applied to split each city's sample data into 3 sets (training, validation, and testing). This robust spatial splitting helps avoid potential data leakage and reduce the risk of model overfitting. Excellent model performance on the training dataset was achieved in Minneapolis, New Orleans, and Philadelphia, with high R^2 ($R^2 > 0.85$) and minimal MAPE (MAPE $< 0.8\%$). On the testing dataset, the mean RMSE across all cities was $< 1.35^\circ\text{C}$ (ranging from 1.028 to 2.214°C), indicating higher model prediction accuracy than existing studies. Kartal and Sekertekin (2022) and Shwetha and Kumar (2016) presented spatial interpolation of the MODIS LST dataset using ANN, achieving mean RMSE values of 3.5°C and 3.03 K , respectively. Several decision tree-based ensemble regression models were also employed for LST modeling, with the authors reporting high R^2 ($R^2 > 0.85$) but lower accuracy (RMSE $> 3^\circ\text{C}$) (He et al., 2025; Tanoori et al., 2024).

The predicted LST from AEF closely resembles the spatial thermal pattern of Landsat LST in New Orleans (Fig. 12). High-temperature zones over dense urban and industrial sectors are well represented, whereas cooler vegetated and water-adjacent areas are reliably replicated. The Landsat-derived LST does not capture fine-scale thermal anomalies between built-up areas and surrounding land uses, whereas the AEF-predicted 10-m LST resolves these intra-urban thermal contrasts at the neighborhood scale. The residual map reveals mostly random, low-magnitude errors throughout the city, with no systematic geographical bias, supporting the spatially balanced model performance.

Examining the spatial pattern of LST prediction errors is crucial as it indicates the presence of random noise or failure to capture the spatial process (Ploton et al., 2020). While MAPE were assessed by many LST prediction studies, very few evaluated whether the errors were randomly distributed. This study applied global Moran's I and Getis Ord G_i^* to assess the spatial pattern of LST prediction errors and found that prediction error has no significant spatial autocorrelation or clustering, indicating that the errors are not biased toward certain areas. The predicted Moran's I value indicates considerable positive spatial autocorrelation ($p < 0.001$), consistent with the observed LST clustering across all cities (Supp Table 6). Moreover, in spatial cross-validation, all models achieved similar performance across folds, with minimal deviation. This further confirms our model robustness, which may be attributed to the spatial stratification strategy used to split data for model training, validation, and testing. The low fold-to-fold RMSE variability (SD ~ 0.0012) suggests that each city model learned a spatially consistent LST-embedding relationship, rather than fitting to block-specific noise. The log-transformed scale on which these metrics were computed further compresses the numerical range of RMSE, reinforcing the appearance of low variability alongside this genuine spatial consistency.

The worst model performance was observed in Charlotte, Cincinnati, and San Antonio. They exhibited low model performance (i.e., comparatively high RMSE and MAPE). While our ANN architecture was efficient at capturing complex urban structures, these cities exhibit unequal thermal distributions within their structures (Gambino et al., 2021; Hashemi and Adib, 2024; Zhang et al., 2022b) potentially leading to noisy learned embedding outputs. The model for Cincinnati had the highest and most variable MAPE scores

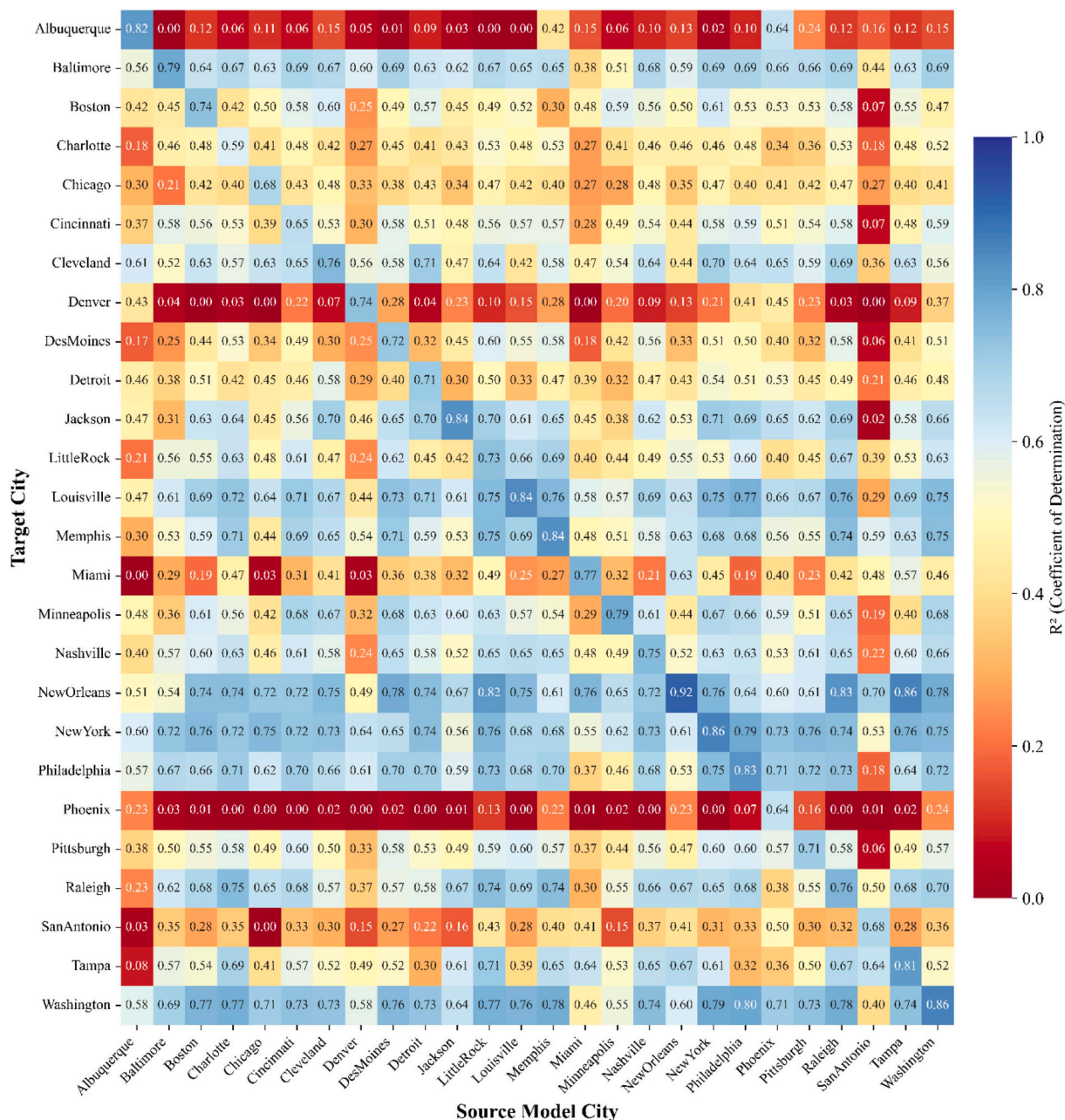


Fig. 10. coefficient of determination (R²) of the predictions based on selected source (model) cities and target cities.

across the training, validation, and testing datasets, as well as in the spatial cross-validation. This aligns with Gambino et al. (2021)'s finding that the LST in Cincinnati was non-stationary across the region, potentially owing to unique land-cover features.

The model for Phoenix achieved relatively low R² (<0.63) but relatively high CCC (>0.73). It also had the lowest MAPE (<0.68%) and RMSE (<0.015). This instance can be attributed to underprediction due to flattening values distant from the central tendency. While severe temperatures (>38–40 °C) are infrequent, models tend to underpredict at extremes and overpredict in the moderately high range (30–32 °C). While RMSE and MAPE are usually considered more reliable metrics for evaluating spatial predictions (Hodson, 2022), a high CCC at this location confirms the model's ability to capture the spatial trend. ANN regression models are known to underperform at the tails of the training distribution due to sparse sample density at extreme values (Khosravi et al., 2011). This produces systematic mean reversion: predictions are pulled toward central tendency, underestimating very high LST and overestimating moderately high LST values (Schober et al., 2018). Cities like Phoenix, San Antonio, and Des Moines demonstrate this pattern clearly, where observed LST exceeding 38 °C is consistently underpredicted. Embedding-based inputs may further amplify this effect, as extreme thermal signatures are underrepresented in the learned feature space (Goodfellow et al., 2016). Urban heat

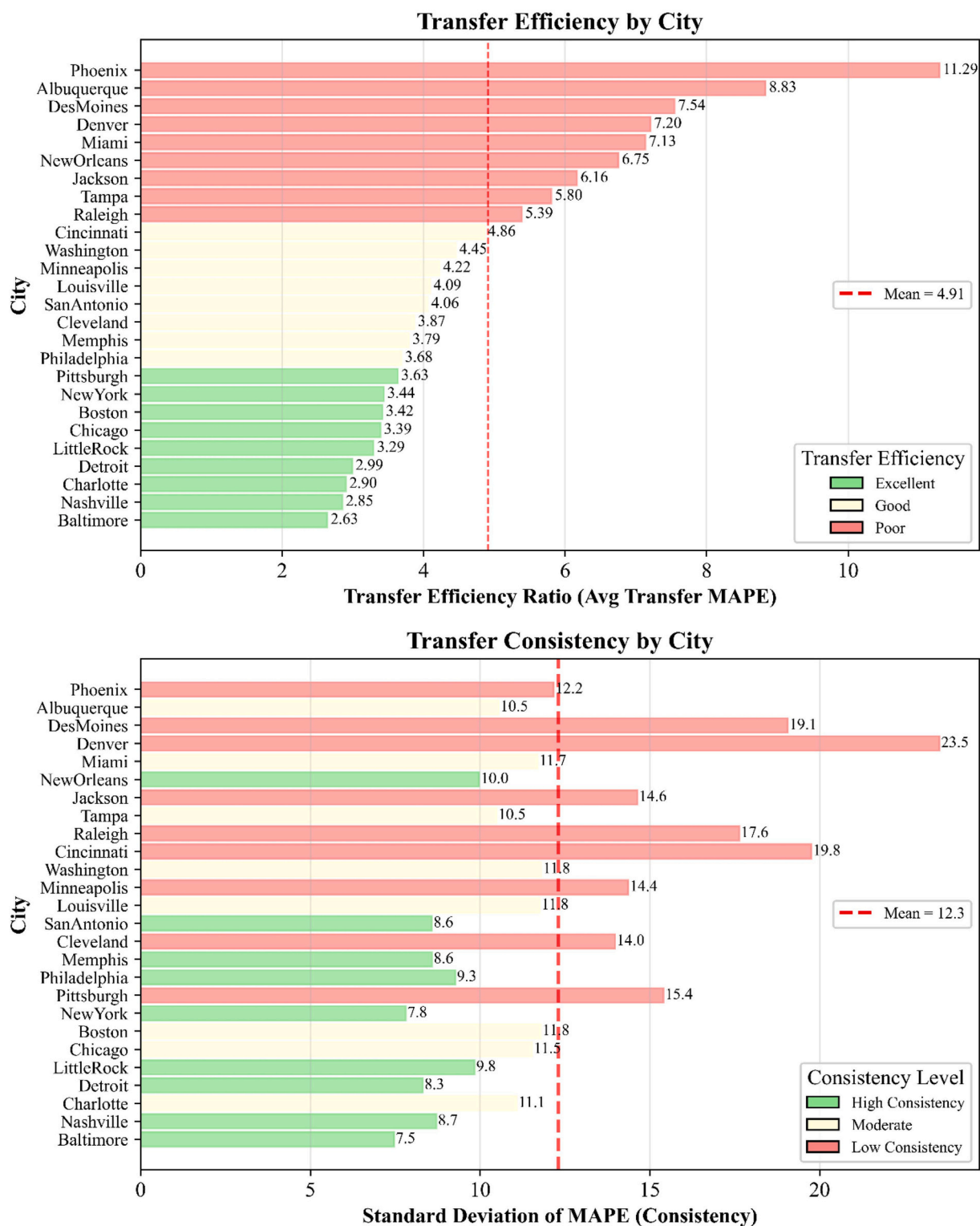


Fig. 11. Transfer efficiency ratios for LST prediction models, color coding represents percentile-based performance categories: excellent (green, top 33%), good (yellow, middle 33%), and poor (red, bottom 33%) transferability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

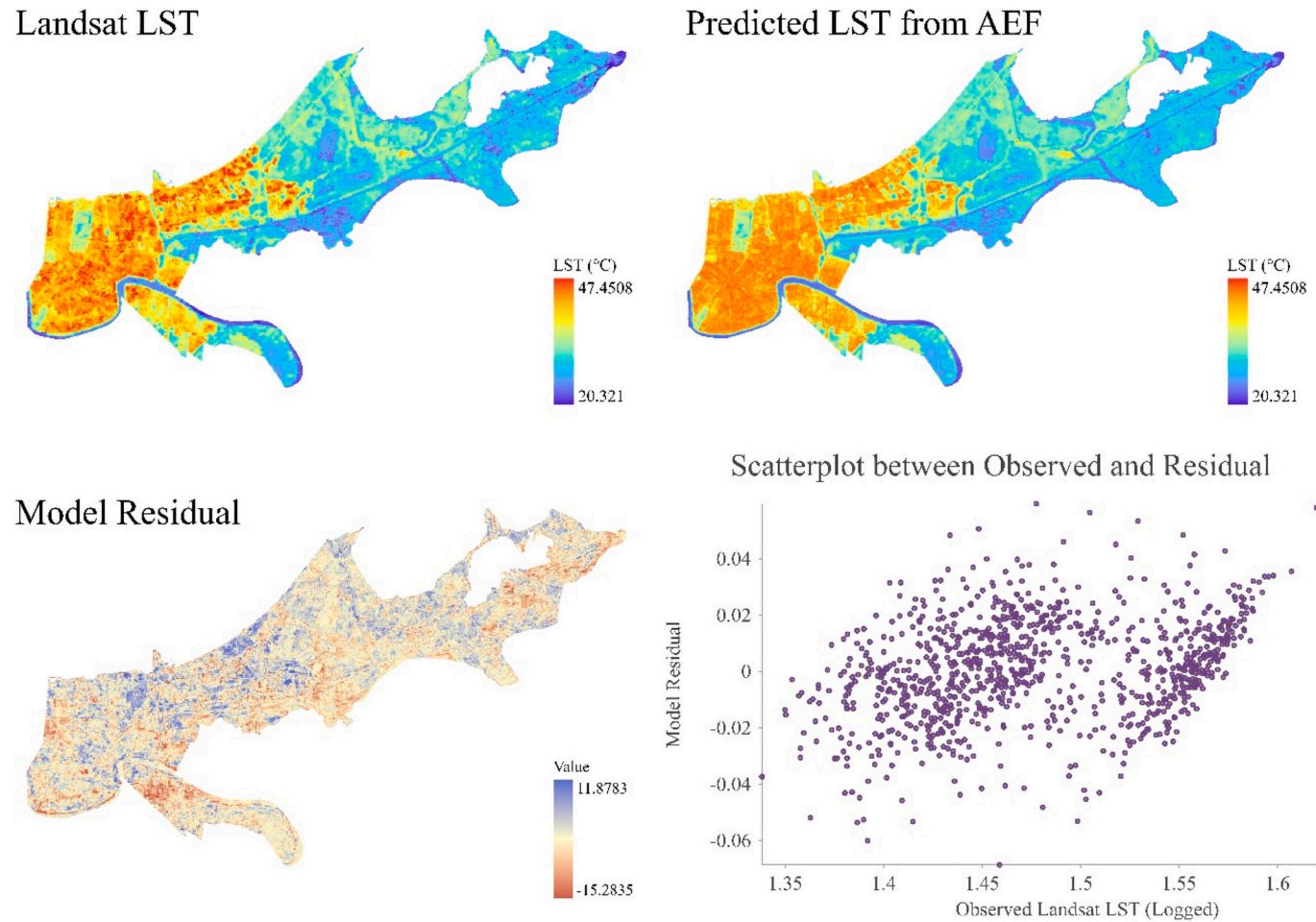


Fig. 12. Spatial comparison of observed Landsat LST and AEF-predicted LST for New Orleans, the best-performing city model. Top panels show the observed and predicted LST maps. Bottom left shows the model residual map, and bottom right shows the scatterplot between observed log-scaled LST and model residuals.

intervention studies should therefore treat model outputs at LST extremes as conservative estimates requiring ground-truth verification.

4.2. Cross-city model transferability assessment

We assessed how accurately a model trained in a city could predict LST in other cities (i.e., cross-city model transferability). Models trained in cities such as Philadelphia, Memphis, and Louisville exhibited the lowest mean transfer MAPE (<4%). LST in New York, Philadelphia, and Washington was most accurately predicted by models trained in other cities (MAPE <4.5%), which might be due to the fact that these cities had comparatively homogeneous LST distributions across the geography (Frey, 2017; McPhearson et al., 2014). In addition, models trained in Baltimore, Nashville, and New York exhibited one of the highest levels of transferability with minimal deviations, indicating consistent transfer capacity.

Phoenix and Cincinnati had the poorest model transferability, but this may be attributed to distinct causes. Cincinnati's model had the highest RMSE (0.043) and MAPE (2.57%) across the trained models (Table 2). It also had the highest MAPE (9.2%) among the target cities for LST prediction during inter-city transferability. This indicates that Cincinnati is a critical zone with high error in predicting LST from AEF's embedded dataset due to spatial heterogeneity in LST distribution (Gambino et al., 2021). Cincinnati's fragmented urban-rural interface, mixed industrial land cover, and irregular topography likely amplify LST heterogeneity beyond what AEF embeddings can adequately represent. On the other hand, the model trained in Phoenix suffers from over-regularization on such a complex data structure, excessively constraining its parameters and preventing it from learning meaningful patterns from the data. Phoenix, as both a target and a source city, has mean MAPEs of 8.9% and 11.15%, respectively. Its mean transfer MAPE was the highest among all source cities. This implies that the city has a unique geographic configuration, making it difficult for the model trained there to transfer to other geographies. Phoenix's extremely arid urban landscape, dominated by sparse vegetation, highly reflective surfaces, and intense heat island signatures, creates a distinct thermal environment that is poorly represented in the broader training distribution.

Overall, model transferability is highly asymmetric and dependent on climatic homogeneity. (Chouaib et al., 2018) pointed out the significance of geographical homogeneity to analyze hydroclimatic characteristics in eastern US (Chouaib et al., 2018). We found that model transferability is better across regions with similar climates (e.g., NFM to GP, MAPE = 11.1%) than across climatically distinct regions (e.g., TWF to GP, MAPE = 46.5%). Moreover, prediction R^2 from these cities, Cleveland to Jackson ($R^2 = 0.702$, MAPE = 4.4%), Tampa to New Orleans ($R^2 = 0.856$, MAPE = 4.7%), and Boston to Washington ($R^2 = 0.765$, MAPE = 4.9%), support the hypothesis of generalizing a specific model for cross-city models.

4.3. Limitations

The key limitation of this study was the model's learned statistical relationships between 10-m AEF embeddings and 30-m Landsat LST targets. The resulting 10-m output represents statistically inferred sub-pixel thermal detail rather than independently verified fine-scale thermal structure. Furthermore, it cannot assess seasonal contrasts in LST across the region. Currently, Google AEF is only providing annual embedding vectors. Consequently, we used the annual mean Landsat LST as the base dataset. This reduced the specific temporal or seasonal patterns of LST at a location to a single annual mean LST value. Episodic heat events, monsoon cooling, and seasonal land-cover changes are therefore not captured by the model. AEF embedding layers do not interpret physical or environmental characteristics. Therefore, the LST model may capture correlations without understanding the underlying causal mechanisms, such as land cover, albedo, or surface moisture, that influence LST. This is an inherent trade-off of embedding-based downscaling approaches.

Causal drivers of LST cannot be explicitly identified from embedding dimensions alone. The RMSE and MAPE were consistently low across all cities, but a few models had a poor R^2 value, suggesting that the forecasts were underperforming. While no significant spatial clusters were identified, spatial cross-validation revealed considerable ambiguity in certain folds. This shows a very complicated LST distribution in locations that AEF's embeddings cannot represent. No independent ground-based LST validation was conducted. The training target relies entirely on Landsat-derived LST, which carries its own retrieval uncertainties. Future studies should incorporate physically interpretable feature sets and ground-based thermal validation to address these constraints. Studies should further explore alternative temporally resolved datasets alongside AEF embeddings to capture seasonal and diurnal LST variability beyond what annual composites can represent. Moreover, individual embedding bands showed no consistent correlation with biophysical indices (Supp Table 5), leaving embedding-landcover relationships an open research gap.

5. Conclusion

This study integrates Google AEF's geospatially embedded dataset to predict urban LST at 10 m resolution using an ANN model across prominent US cities. The trained models showed promising results, achieving high R^2 scores for cities such as New Orleans (0.89) and Philadelphia (0.86). The mean error ratio (i.e., MAPE) ranged between 0.67 and 2.57%, while the RMSE ranged between 1.02 and 2.2 °C. The models were validated using a 10-fold spatial cross-validation approach, yielding very stable model performance across all spatial folds (except Cincinnati). Moreover, prediction errors appear to be randomly distributed, without spatial clustering. Google AEF's embedding dataset comprises 64 bands, and we found several concurring bands, e.g., Bands 31, 15, and 1, which remained among the most important features for LST prediction across cities.

We also assessed these models' cross-city geographic transferability. The model trained in Philadelphia had the lowest mean

transfer MAPE when predicting LST in other cities. On the other hand, Baltimore had the best transferability, with the lowest standard deviation in transfer MAPE values. However, cross-city prediction accuracy began to degrade as geographical distance increased. Results indicate that models transfer well between cities in similar climates (e.g., between ETF and NFM), while severe climatic mismatches (between GP and NAD/TWF) degrade model transferability (with MAPEs surpassing 30–45%). Asymmetric patterns suggest that target climate similarity is more important than training climate variety, because models perform best when trained and tested in climatically similar regions (low diagonal MAPE values). On the other hand, training in extreme locations such as NAD's deserts fails to provide better generalization to dissimilar climates.

LST predictions have been widely studied to assess future climate scenarios and fill missing values. Satellite instruments, e.g., Landsat and Sentinel, can provide LST with 30 m resolution, while instruments like MODIS can provide daily LST insights. In contrast, the AEF dataset is the only openly available dataset with a 10 m resolution. Therefore, while not ideal for temporal resolution, as the dataset cannot support a rapid thermal stress study, it can provide indicative long-term thermal characteristics at a fine spatial scale. However, input data limitations and model interpretability constraints should be considered before operational application. Future studies should incorporate additional GeoAI models to benchmark the best-performing models for predicting LST from AEF embedding layers. The use of inter-city transferability, coupled with physiographic traits, should be investigated further to determine the likely reasons for the underperforming cities. Further, we encourage researchers to evaluate the model's ability to model LST from rural or developing areas to better understand its generalizability. However, urban planning applications should treat model outputs as exploratory rather than definitive, given the acknowledged limitations in input data validation and embedding interpretability.

CRedit authorship contribution statement

Bijoy Mitra: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Guiming Zhang:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Funding

The authors did not receive any financial assistance to conduct this research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support received from the Department of Geography & the Environment, University of Denver, US to conduct this research. The authors gratefully acknowledge the University of Denver and Elsevier for covering the open access publication fee under their institutional agreement.

Appendix A. Supplementary data

The spatial autocorrelation analysis for each set, feature importance percentages, and spatial cross validation metrics for all the cities are presented in *SuppDataAnalysis*. Error metrics, learning curves, and SHAP analysis for all the models were presented in *SuppFigs*. Supplementary data to this article can be found online at <https://doi.org/10.1016/j.uclim.2026.103058>.

Data availability

The satellite data from Landsat, Sentinel5-p, VIIRS, and NASADEM are openly accessible from Google Earth Engine. The code for this study is available from Bijoy Mitra upon reasonable request.

References

- Artmann, M., 2014. Assessment of soil sealing management responses, strategies, and targets toward ecologically sustainable urban land use management. *AMBIO* 43 (4), 530–541. <https://doi.org/10.1007/S13280-014-0511-1>.
- Bailey, R.G., 1976. Ecoregions of the United States (Map). US Department of Agriculture, US Forest Service, Intermountain Region, Ogden, UT. <https://stormwater.pca.state.mn.us/sites/stormwater/files/P-gen3-12a.pdf>.
- Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Lydia Zhang, L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., DeepMind, G., 2025. AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data. <https://arxiv.org/pdf/2507.22291>.
- Chouaib, W., Alila, Y., Caldwell, P.V., 2018. Parameter transferability within homogeneous regions and comparisons with predictions from a priori parameters in the eastern United States. *J. Hydrol.* 560, 24–38. <https://doi.org/10.1016/J.JHYDROL.2018.03.018>.
- Cox, T.F., Cliff, A.D., Ord, J.K., 1984. Spatial processes: models and applications. *J. R. Stat. Soc. Ser. A (Gen.)* 147 (3), 515. <https://doi.org/10.2307/2981590>.

- Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Joseph Hughes, M., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* 194, 379–390. <https://doi.org/10.1016/j.rse.2017.03.026>.
- Frey, N., 2017. Equity in the distribution of urban environmental amenities: the case of Washington, D.C. *Urban Geogr.* 38 (10), 1534–1549. <https://doi.org/10.1080/02723638.2016.1238686>.
- Fu, H., Shao, Z., Fu, P., Huang, X., Cheng, T., Fan, Y., 2022. Combining ATC and 3D-CNN for reconstructing spatially and temporally continuous land surface temperature. *Int. J. Appl. Earth Obs. Geoinf.* 108, 102733. <https://doi.org/10.1016/J.JAG.2022.102733>.
- Gambino, C., LaJoie, P., Cronin-Golomb, O., Feibel, S., Rokosz, K., 2021. Assessing urban heat in the Cincinnati and Covington area to inform local decision-makers and environmental justice initiatives. In: AGU Fall Meeting Abstracts, 2021. A42C-07.
- Getis, A., 2009. Spatial autocorrelation. In: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Springer, pp. 255–278.
- Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* 24 (3), 189–206. <https://doi.org/10.1111/J.1538-4632.1992.TB00261.X>.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- Griffith, D.A., Wong, D.W.S., Whitfield, T., 2003. Exploring relationships between the global and regional measures of spatial autocorrelation. *J. Reg. Sci.* 43 (4), 683–710. <https://doi.org/10.1111/J.0022-4146.2003.00316.X>.
- Han, J., Fang, S., Mi, Q., Wang, X., Yu, Y., Zhuo, W., Peng, X., 2024. A time-continuous land surface temperature (LST) data fusion approach based on deep learning with microwave remote sensing and high-density ground truth observations. *Sci. Total Environ.* 914, 169992. <https://doi.org/10.1016/J.SCIOTENV.2024.169992>.
- Hashemi, F., Adib, M., 2024. Examining thermal inequities: land surface temperature, social vulnerability, and historical redlining in San Antonio, TX. *Urban Clim.* 55, 101960. <https://doi.org/10.1016/J.UCLIM.2024.101960>.
- Haykin, S., 2009. *Neural Networks and Learning Machines*, 3/E. Pearson Education India.
- He, P., Chen, Z., Zhang, L., Ma, C., Luo, C., 2025. Machine learning prediction of future land surface temperature from SAR optical fusion under urban expansion in Changsha, China. *Sci. Rep.* <https://doi.org/10.1038/s41598-025-30976-5>.
- Heinl, M., Hammerle, A., Tappeiner, U., Leitinger, G., 2015. Determinants of urban–rural land surface temperature differences – a landscape scale perspective. *Landsc. Urban Plan.* 134, 33–42. <https://doi.org/10.1016/J.LANDURBPLAN.2014.10.003>.
- Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci. Model Dev.* 15 (14), 5481–5487. <https://doi.org/10.5194/GMD-15-5481-2022>.
- Hollmann, R., Merchant, C.J., Saunders, R., Downy, C., Buchwitz, M., Cazenave, A., Chuvieco, E., Defourny, P., De Leeuw, G., Forsberg, R., Holzer-Popp, T., Paul, F., Sandven, S., Sathyendranath, S., Van Roozendael, M., Wagner, W., 2013. The ESA climate change initiative: satellite data records for essential climate variables. *Bull. Am. Meteorol. Soc.* 94 (10), 1541–1552. <https://doi.org/10.1175/BAMS-D-11-00254.1>.
- IPCC, 2022. Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-Industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. Cambridge University Press. <https://doi.org/10.1017/9781009157940>.
- IPCC, 2023. Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. <https://doi.org/10.1017/9781009325844>.
- Jia, H., Yang, D., Deng, W., Wei, Q., Jiang, W., 2021. Predicting land surface temperature with geographically weighed regression and deep learning. *Wiley Interdisc. Rev.* 11 (1), e1396. <https://doi.org/10.1002/WIDM.1396>.
- Kafy, A., Dey, N.N., Al Rakib, A., Rahaman, Z.A., Nasher, N.M.R., Bhatt, A., 2021. Modeling the relationship between land use/land cover and land surface temperature in Dhaka, Bangladesh using CA-ANN algorithm. *Environ. Challenges* 4, 100190. <https://doi.org/10.1016/J.ENVC.2021.100190>.
- Kartal, S., Sekertekin, A., 2022. Prediction of MODIS land surface temperature using new hybrid models based on spatial interpolation techniques and deep learning models. *Environ. Sci. Pollut. Res.* 29 (44), 67115–67134. <https://doi.org/10.1007/S11356-022-20572-9>.
- Karunakaran, K., Sudalaimuthu, K., 2025. GWR-ANN modelling for spatiotemporal prediction of land surface temperature in Thanjavur delta: evaluating environmental impacts on climate action and SDGs. *Environ. Dev. Sustain.* 27 (8), 20181–20228. <https://doi.org/10.1007/S10668-025-06157-9>.
- Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.F., 2011. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Netw.* 22 (9), 1341–1356. <https://doi.org/10.1109/TNN.2011.2162110>.
- Kohonen, T., 1988. An introduction to neural computing. *Neural Netw.* 1 (1), 3–16. [https://doi.org/10.1016/0893-6080\(88\)90020-2](https://doi.org/10.1016/0893-6080(88)90020-2).
- Laforceza, R., Carrus, G., Sanesi, G., Davies, C., 2009. Benefits and well-being perceived by people visiting green spaces in periods of heat stress. *Urban For. Urban Green.* 8 (2), 97–108. <https://doi.org/10.1016/J.UFUG.2009.02.003>.
- Li, Z.L., Wu, H., Duan, S.B., Zhao, W., Ren, H., Liu, X., Leng, P., Tang, R., Ye, X., Zhu, J., Sun, Y., Si, M., Liu, M., Li, J., Zhang, X., Shang, G., Tang, B.H., Yan, G., Zhou, C., 2023. Satellite remote sensing of global land surface temperature: definition, methods, products, and applications. *Rev. Geophys.* 61 (1), e2022RG000777. <https://doi.org/10.1029/2022RG000777>.
- Lu, D., Weng, Q., 2009. Extraction of urban impervious surfaces from an IKONOS image. *Int. J. Remote Sens.* 30 (5), 1297–1311. <https://doi.org/10.1080/01431160802508985>.
- Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pp. 4768–4777.
- Lute, A.C., Luce, C.H., 2017. Are model transferability and complexity antithetical? Insights from validation of a variable-complexity empirical snow model in space and time. *Water Resour. Res.* 53 (11), 8825–8850. <https://doi.org/10.1002/2017WR020752>.
- Marshall, J.D., 2007. Urban land area and population growth: a new scaling relationship for metropolitan expansion. *Urban Stud.* 44 (10), 1889–1904. <https://doi.org/10.1080/00420980701471943>.
- Masek, J.G., Vermote, E.F., Saleous, N.E., Wolfe, R., Hall, F.G., Huemmrich, K.F., Gao, F., Kutler, J., Lim, T.-K., 2006. A Landsat surface reflectance dataset for North America, 1990–2000. *IEEE Geosci. Remote Sens. Lett.* 3 (1), 68–72. <https://doi.org/10.1109/LGRS.2005.857030>.
- McPhearson, T., Hamstead, Z.A., Kremer, P., 2014. Urban ecosystem services for resilience planning and management in New York City. *AMBIO* 43 (4), 502–515. <https://doi.org/10.1007/S13280-014-0509-8>.
- Mentaschi, L., Duveiller, G., Zulian, G., Corbane, C., Pesaresi, M., Maes, J., Stocchino, A., Feyen, L., 2022. Global long-term mapping of surface temperature shows intensified intra-city urban heat island extremes. *Glob. Environ. Chang.* 72, 102441. <https://doi.org/10.1016/J.GLOENVCHA.2021.102441>.
- Morabito, M., Crisci, A., Messeri, A., Orlandini, S., Raschi, A., Maracchi, G., Munafo, M., 2016. The impact of built-up surfaces on land surface temperatures in Italian urban areas. *Sci. Total Environ.* 551–552, 317–326. <https://doi.org/10.1016/J.SCIOTENV.2016.02.029>.
- Moran, P.A., 1950. A test for the serial independence of residuals. *Biometrika* 37 (1–2), 178–181. <https://doi.org/10.1093/BIOMET/37.1-2.178>.
- Munafo, M., Assennato, F., Congedo, L., Luti, T., Marinosci, I., Monti, G., Riitano, N., Sallustio, L., Strollo, A., Tombolini, I., 2015. *Il consumo di suolo in Italia-Edizione 2015*. Ispra.
- Nguyen, L.H., Nghiem, S.V., Henebry, G.M., 2018. Expansion of major urban areas in the US Great Plains from 2000 to 2009 using satellite scatterometer data. *Remote Sens. Environ.* 204, 524–533. <https://doi.org/10.1016/J.RSE.2017.10.004>.
- Olafsson, H., Roustia, I., Dalvi, M., Krzyszcak, J., 2025. Spatiotemporal dynamics and ANN-based projection of daytime and nighttime land surface temperature in Iceland (2001–2035). *Model. Earth Syst. Environ.* 11 (5), 353. <https://doi.org/10.1007/S40808-025-02558-X>.
- Omerini, J.M., Griffith, G.E., 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. *Environ. Manag.* 54 (6), 1249–1266. <https://doi.org/10.1007/s00267-014-0364-1>.
- Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geogr. Anal.* 27 (4), 286–306. <https://doi.org/10.1111/J.1538-4632.1995.TB00912.X>.

- Pande, C.B., Egbueri, J.C., Costache, R., Sidek, L.M., Wang, Q., Alshehri, F., Din, N.M., Gautam, V.K., Chandra Pal, S., 2024. Predictive modeling of land surface temperature (LST) based on Landsat-8 satellite data and machine learning models for sustainable development. *J. Clean. Prod.* 444, 141035. <https://doi.org/10.1016/J.JCLEPRO.2024.141035>.
- Peng, S.S., Piao, S., Zeng, Z., Ciais, P., Zhou, L., Li, L.Z.X., Myneni, R.B., Yin, Y., Zeng, H., 2014. Afforestation in China cools local land surface temperature. *Proc. Natl. Acad. Sci.* 111 (8), 2915–2919. <https://doi.org/10.1073/PNAS.1315126111>.
- Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Péliissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* 11 (1), 4540-. <https://doi.org/10.1038/S41467-020-18321-Y>.
- Priyatikanto, R., Lu, Y., Dash, J., Sheffield, J., 2023. Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation. *Agric. For. Meteorol.* 341 (1), 109652. <https://doi.org/10.1016/j.agrformet.2023.109652>.
- Rahman, M., 2026. Physically interpretable AlphaEarth foundation model embeddings enable LLM-based land surface intelligence. *Remote Sens. Appl.: Soc. Environ.* 42, 102045. <https://doi.org/10.1016/j.rsase.2026.102045>.
- Rahman, M.T.U., Esha, E.J., 2022. Prediction of land cover change based on CA-ANN model to assess its local impacts on Bagerhat, southwestern coastal Bangladesh. *Geocarto Int.* 37 (9), 2604–2626. <https://doi.org/10.1080/10106049.2020.1831621>.
- Rahman, M.T., Aldosary, A.S., Mortoja, M.G., 2017. Modeling future land cover changes and their effects on the land surface temperatures in the Saudi Arabian eastern coastal city of Dammam. *Land* 6 (2), 36. <https://doi.org/10.3390/LAND6020036>.
- Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* 126 (5), 1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>.
- Seto, K.C., Fragkias, M., Gopal, B., Reilly, M.K., 2011. A meta-analysis of global urban land expansion. *PLoS One* 6 (8), e23777. <https://doi.org/10.1371/JOURNAL.PONE.0023777>.
- Shapley, L.S., 1953. Stochastic games. *Proc. Natl. Acad. Sci.* 39 (10), 1095–1100.
- Shwetha, H.R., Kumar, D.N., 2016. Prediction of high spatio-temporal resolution land surface temperature under cloudy conditions using microwave vegetation index and ANN. *ISPRS J. Photogramm. Remote Sens.* 117, 40–55. <https://doi.org/10.1016/J.ISPRSJPRS.2016.03.011>.
- Solomon, S., 2007. IPCC (2007): climate change the physical science basis. In: *Agu Fall Meeting Abstracts*, 2007. U43D-01.
- Tanoori, G., Soltani, A., Modiri, A., 2024. Machine learning for urban Heat Island (UHI) analysis: predicting land surface temperature (LST) in urban environments. *Urban Clim.* 55, 101962. <https://doi.org/10.1016/J.UCLIM.2024.101962>.
- Tollefson, J., 2025. Google AI model mines trillions of images to create maps of Earth ‘at any place and time.’. *Nature* 644 (8076), 313–314. <https://doi.org/10.1038/D41586-025-02412-1>.
- Tran, D.X., Pla, F., Latorre-Carmona, P., Myint, S.W., Caetano, M., Kieu, H.V., 2017. Characterizing the relationship between land use land cover change and land surface temperature. *ISPRS J. Photogramm. Remote Sens.* 124, 119–132. <https://doi.org/10.1016/J.ISPRSJPRS.2017.01.001>.
- van Dijk, M., Morley, T., Rau, M.L., Saghai, Y., 2021. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* 2 (7), 494–501. <https://doi.org/10.1038/S43016-021-00322-9>.
- Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* 185, 46–56. <https://doi.org/10.1016/j.rse.2016.04.008>.
- Weng, Q., Fu, P., Gao, F., 2014. Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. *Remote Sens. Environ.* 145, 55–67. <https://doi.org/10.1016/J.RSE.2014.02.003>.
- Wu, Y. chen, Feng, J. wen, 2017. Development and application of artificial neural network. *Wirel. Pers. Commun.* 102 (2), 1645–1656. <https://doi.org/10.1007/S11277-017-5224-X>.
- Xian, G., 2023. Characterizing urban heat islands across 50 major cities in the United States. In: *Fact Sheet*. <https://doi.org/10.3133/fs20233048>.
- Yan, Y., Mao, K., Shi, J., Piao, S., Shen, X., Dozier, J., Liu, Y., Ren, H. li, Bao, Q., 2020. Driving forces of land surface temperature anomalous changes in North America in 2002–2018. *Sci. Rep.* 10 (1), 6931-. <https://doi.org/10.1038/S41598-020-63701-5>.
- Zhang, Y., Guindon, B., 2012. Multispectral analysis for manmade surface extraction from RapidEye and SPOT5. *Can. J. Remote. Sens.* 38 (2), 180–196. <https://doi.org/10.5589/m12-032>.
- Zhang, T., Zhou, Y., Zhu, Z., Li, X., Asrar, G.R., 2022a. A global seamless 1km resolution daily land surface temperature dataset (2003–2020). *Earth Syst. Sci. Data* 14 (2), 651–664. <https://doi.org/10.5194/ESSD-14-651-2022>.
- Zhang, Y., Chen, G., Myint, S.W., Zhou, Y., Hay, G.J., Vukomanovic, J., Meentemeyer, R.K., 2022b. UrbanWatch: a 1-meter resolution land cover and land use database for 22 major cities in the United States. *Remote Sens. Environ.* 278, 113106. <https://doi.org/10.1016/J.RSE.2022.113106>.