Reproducible Spatiotemporal Urban Heat Island Modeling for Bengaluru Using Earth Observation and Machine Learning

Pranjal Prakash
School of Computer Science and Applications
REVA University
Bengaluru, India
24130500752@reva.edu.in

Vinag G.
School of Computer Science and Applications
REVA University
Bengaluru, India
vinay.g@reva.edu.in

***Abstract*—The present research proposes a workflow for Urban Heat Island analysis for Bengaluru, utilizing multiple years of Earth Observation and interpretable machine learning. The methodology ensures a reproducible workflow with spatial leakage control, multiple validation levels, and fixed artifacts. The approach uses stacked Sentinel-2 and Landsat imagery for 2020-2025, followed by physics-informed feature verification and XGBoost with spatial block cross-validation.. The data consists of 10,202,847 data points over 164 tiles. Held-out validation achieves an RMSE of 3.487 K and an R^2 of 0.527, while independent validation achieves an RMSE of 3.31 K and an R^2 of 0.554. Regional validation indicates mean urban and rural temperatures of 310.8 K and 308.8 K, respectively, with an Urban Heat Island Intensity of 1.97 K. Seasonal consistency is observed with a mean of 1.65 K and a range of 1.31-2.29 K. The novelty of this study is an interpretable geospatial machine learning approach with a focus on reproducibility and leakage control constraints.**

***Keywords*—Urban Heat Island, Land Surface Temperature, Earth Observation, XGBoost, Explainable AI, Reproducibility**

I. Introduction

Urban Heat Island (UHI) effects in rapidly growing urban areas play an important role in the context of thermal comfort, energy consumption, and environmental sustainability. Remote sensing techniques help to effectively monitor the surface temperatures in the region, hence facilitating the analysis of the Urban Heat Island effects in the region.

Bengaluru is known to experience considerable seasonality as well as heterogeneity in land cover due to vegetation as well as the growth of built-up areas. This calls for the necessity of spatially aware validation.

Studies on Urban Heat Island effects have contributed significantly towards the predictability of the phenomenon; however, little information is available on the spatial leakage controls, traceability of executions, and reproducibility of the experiments. The assessment of the strength of the reported results is difficult in the context of spatiotemporal geospatial regression modeling.

A. Problem Statement and Gap

However, most of the research articles on Earth Observing (EO) based urban heat island (UHI) analysis have shown predictive capabilities but lack information on strategies for controlling spatial leakage, traceability of execution, and reproducibility. This is causing uncertainty in geospatial regression analysis over multiple seasons due to the presence of masking effects, autocorrelation, and temporal drift.

B. Contribution and Scope

This work provides an interpretable geospatial machine learning framework in the context of a controlled evaluation scenario. It integrates spatial block validation, physics-informed feature analysis, and multi-level consistency checks into one single execution protocol. It is methodological in nature and not architectural, focusing on reliability, traceability, and leakage awareness in the context of large-scale geospatial regression. XGBoost is used as the non-linear baseline for tabular Earth Observation features, with the intention to expand to other models in the future.

II. Related Work

In the earlier literature on UHI using the remote sensing data, temperature data is combined with other data on vegetation and built-up areas to carry out contrast analyses between urban and rural areas [1]–[4]. Thermal remote sensing and the modeling of the relationship between LST and vegetation are well established in the field of urban studies [5]–[8]. Recent studies have also included machine learning approaches to the estimation and interpretation of LST data [11], [14]. Tree-based ensemble approaches, including Random Forest and gradient boosting, share the same theoretical basis for regression tasks [9], [10], while XGBoost is the most popular choice for tabular geospatial modeling [12]. SHAP is the most popular feature attribution tool for model interpretability [13].

The contribution of this work is not a new backbone architecture, but a reproducible and leakage-aware evaluation framework for EO-based UHI regression.

III. Methodology

A. Methodology Flowchart

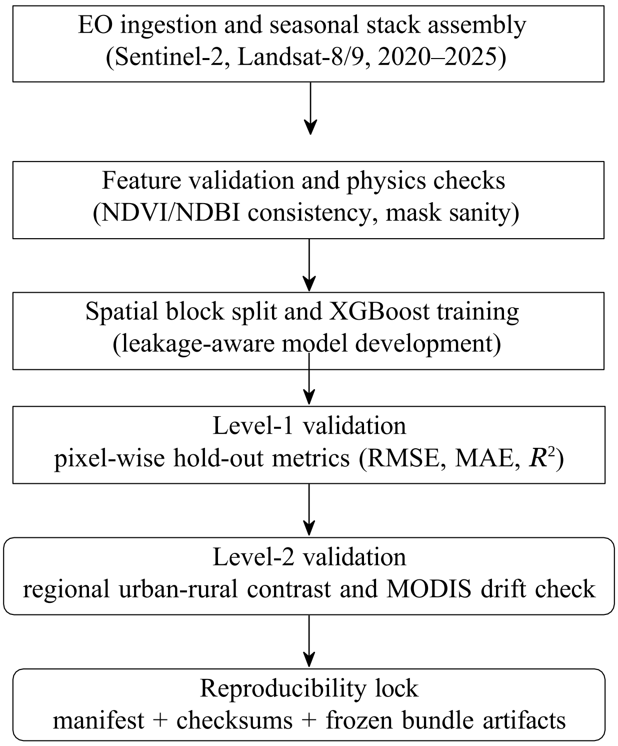

Fig. 1. Methodology flow used for final reporting. It links EO ingestion, seasonal alignment, model training, two-level validation, and checksum-based artifact locking.

B. Primary Pipeline

The primary path taken in the reported results is: Primary workflow: preprocess, align_seasonal, validate_features, model_plan, and validate_modis. Predicted land-surface temperature at year y, season s, and pixel p is modeled as:

Eq. (1): T_pred(y,s,p) = f(x(y,s,p))

In other words, the feature vector, i.e., x(y, s, p), represents the features corresponding to year y, season s, and pixel p. The feature vector is of d-dimensional nature, i.e., x(y, s, p) is a d-dimensional vector.

Seasonal alignment has been chosen as the main processing path, while the output from feature alignment has been treated as an auxiliary product.

C. Validation Design

Two-level validation was applied. Level-1 used pixel-wise hold-out evaluation with model predictions compared against held-out Landsat-derived targets. Level-2 evaluated regional urban-rural thermal contrast and seasonal consistency against MODIS seasonal means.

Eq. (2): Delta_T_UHI = T_urban_mean - T_rural_mean

IV. Experimental Setup

The training dataset consisted of 10,202,847 samples created from 164 aligned tiles. A spatial block hold-out method was used to prevent spatial leakage. The training set consisted of 6,795,225 samples, and the validation set consisted of 3,407,622 samples.

The XGBoost regressor was trained on the dataset by making use of NDVI, NDBI, latitude, longitude, seasonal variables, and WorldCover-derived class variables. The evaluation of the external dataset made use of RMSE, MAE, and R^2 metrics. The external validation consisted of a pixel-level hold-out evaluation of the dataset at level 1 and a level 2 regional evaluation compared to seasonal MODIS averages.

V. Experiments and Results

A. Dataset and Split Summary

This subsection reports the size, feature scope, and split structure of the final analysis table used for training and validation. It establishes the fixed data basis for all subsequent metrics and interpretability results.

TABLE I
DATASET AND SPLIT SUMMARY

Item	Value
Seasonal stacks	12 (2020–2025, summer/winter)
Aligned tiles for training table	164
Total training rows	10,202,847
Spatial CV train rows	6,795,225
Spatial CV validation rows	3,407,622
Features	NDVI, NDBI, lat/lon, land-cover flags, season flag

B. Core Performance Metrics

The metrics will be provided for both cross-validation and hold-out evaluation approaches.

TABLE II
MODEL AND VALIDATION METRICS

Evaluation Level	Metric	Value
Spatial CV (XGBoost)	RMSE	3.487 K
Spatial CV (XGBoost)	R^2	0.527
Level-1 hold-out	RMSE	3.31 K
Level-1 hold-out	MAE	2.34 K
Level-1 hold-out	R^2	0.554
Level-1 hold-out	Bias	+0.75 K
Level-1 hold-out	N	973,407 pixels
Level-2 regional	Urban mean LST	310.8 K
Level-2 regional	Rural mean LST	308.8 K
Level-2 regional	UHI intensity	1.97 K

C. Prediction and UHI Visualization

The following subsection discusses the patterns of spatial prediction behavior and the patterns of quantitative agreement for the interpretation of Urban Heat Islands (UHI). Figure 2 summarizes regional contrast behavior on evaluation data.

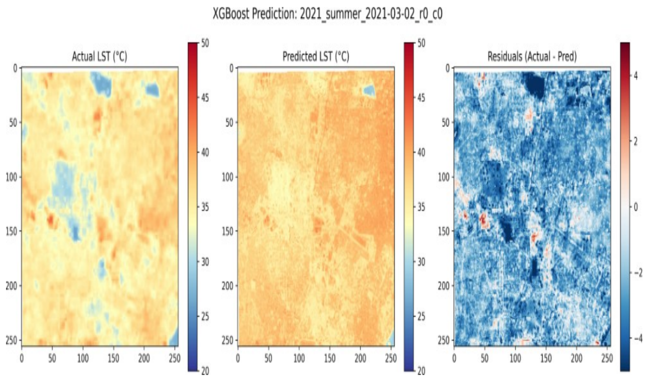

Fig. 2. Displays a prediction map for evaluating tiles based on spatial thermal variations. The warm and cool clusters help to validate the urban–rural urban heat island contrast.

Figure 3 indicates the relation between observed and predicted temperature data for evaluating datasets.

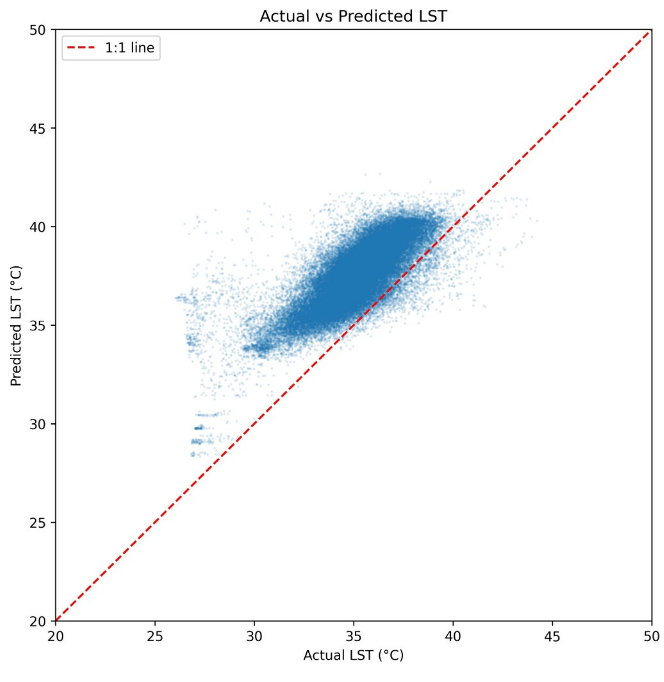

Fig. 3. Indicates a scatter plot for observed and predicted temperature data for evaluating datasets. The diagonal dots show consistency, while scattered dots show errors.

D. SHAP-Based Interpretability

The following subsections explore model behaviour based on global SHAP rankings and dependence plots. Fig. 4–8 show feature influence and direction for primary predictor variables.

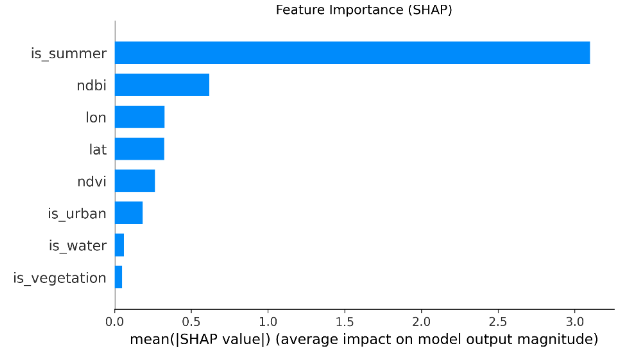

Fig. 4. Global SHAP bar plot ranking features by mean absolute contribution. Seasonality and spectral indices have high influence over all samples.

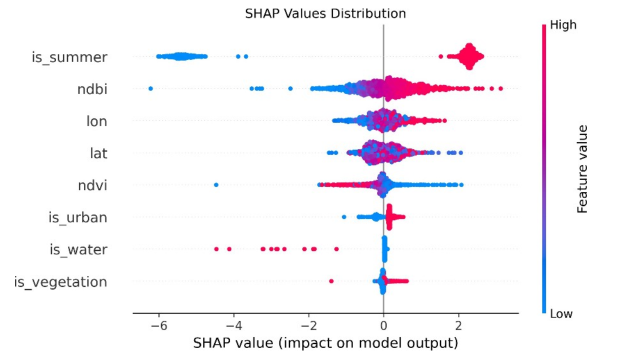

Fig. 5. SHAP beeswarm plot showing signed contribution distribution per feature. The spread shows the magnitude, and the colour shows the sign of the values, which are mixed locally.

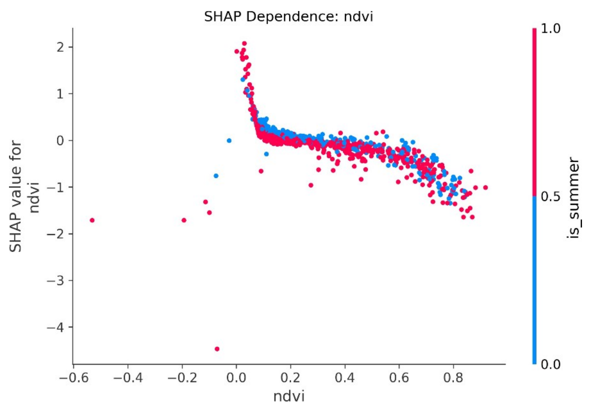

Fig. 6. SHAP dependence of NDVI on feature value. Increasing values of NDVI tend to decrease SHAP values, implying cooling effects associated with vegetation.

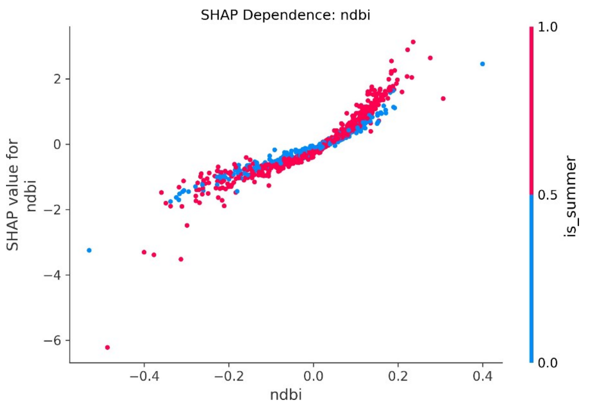

Fig. 7. SHAP dependence for NDBI depending on feature value. The higher the built-up intensity, the greater the positive SHAP contribution.

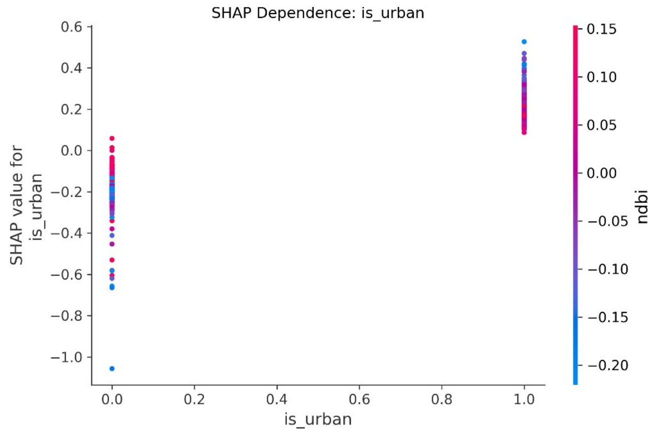

Fig. 8. SHAP dependence for the binary urban indicator. Urban-class samples have a positive SHAP shift compared to non-urban pixels, confirming the simulated thermal contrast.

E. Gate and Drift Consistency Results

This sub-section provides the pass/fail results for publication gates and drift envelope checks, as they relate to seasonal aggregates. The physical consistency and range stability are verified as they relate to the expected Urban Heat Island (UHI) behaviour. Table III provides the consolidated gate and drift results.

TABLE III
PUBLICATION GATES AND DRIFT CHECK

Check	Result	Status
Step-11 Objective Scorecard	5/5 (100%)	PASS (GREEN)
Physical consistency	corr(NDVI,LST) = -0.558	PASS
Physical consistency	corr(NDBI,LST) = 0.622	PASS
Seasonal UHI mean	1.65 K	PASS
Seasonal UHI range	1.31–2.29 K	PASS
Published envelope consistency	Within 0.6 to 2.6 K: 100%	PASS

F. Seasonal UHI Summary

The values provided are for seasonal urban and rural aggregates, used for the assessment of temporal consistency. The values provided can be used to interpret the UHI range and magnitude across different seasons. Table IV below shows the entire seasonal urban and rural summary

TABLE IV
SEASONAL URBAN-RURAL UHI FROM STEP-12 DRIFT RUN

Year-Season	Urban (K)	Rural (K)	UHI (K)
2020 Summer	312.8	310.9	1.95
2020 Winter	302.2	300.8	1.35
2021 Summer	310.9	309.6	1.37
2021 Winter	306.6	304.3	2.29
2022 Summer	310.4	308.8	1.54
2022 Winter	302.5	301.0	1.52
2023 Summer	312.2	310.3	1.85
2023 Winter	305.4	303.8	1.61
2024 Summer	312.3	311.0	1.31
2024 Winter	302.8	301.3	1.57
2025 Summer	311.9	310.0	1.87
2025 Winter	304.2	302.5	1.62

G. Interpretability

Expected physical behavior was observed in the SHAP analysis, with NDVI having a contribution to the cooling tendency and NDBI having a contribution to the warming tendency. The season had the greatest influence on the model inference, as expected due to the strong contrast between summer and winter.

For local interpretability, prediction for sample x is equal to the baseline prediction plus the sum of SHAP contributions from all input features.

Eq. (3): $T\_pred(x) = base\_value + shap\_1(x) + shap\_2(x) + \ldots + shap\_d(x)$

Here, base_value refers to a baseline prediction, and shap_j(x) refers to the contribution of feature j for a given sample x.

In global interpretability , the global importance of feature j is defined as the average absolute SHAP contribution of feature j over N evaluated samples.

Eq. (4): $I\_j = (abs(shap\_j(x\_1)) + abs(shap\_j(x\_2)) + \ldots + abs(shap\_j(x\_N))) / N$

VI. DISCUSSION

The results show a clear physical consistency in their trends, which are seasonal in nature and replicable in a controlled framework of evaluation. To be specific, the trends in sign and magnitude of NDVI and NDBI are in line with those expected in urban heat islands, where increased vegetation intensity is associated with reduced surface temperature, and increased built-up intensity contributes to increased surface temperature. Moreover, the UHI values over different seasons were found to fall within the ranges provided for Bengaluru in this entire exercise, which comprised 12 seasonal aggregates.

The held-out level-1 model's performance, with a coefficient of determination of 0.554 and a root mean square error of 3.31 K, compares favorably with, or is superior to, its in-training validation metrics of 0.527 and 3.487 K, which lends credence to the spatially aware split strategy employed in this exercise, although the error still remains higher than the 3 K goal.

The current work focuses on replicability and validation over architectural innovation, and hence, both RF and CNN-based models are identified as potential candidates for future evaluation using a fixed data split and replicability framework.

A. Threats to Validity

Three limitations are identified. First, there is a lower valid pixel count for winter season in some year/season combinations. Second, heuristic NDVI/NDBI visualization masks are used qualitatively rather than quantitatively. Finally, Level 2 MODIS validation is interpreted regionally rather than per-pixel supervisory ground truth for model learning.

B. Reproducibility Artifacts

This subsection discusses the artifacts used to ensure the traceability and reproducibility of the manuscripts. Each document within the bundle is designed to serve a certain purpose to ensure the traceability of the metrics used and the verification of the integrity during the experiments. Table V shows the contents of the reproducibility bundle.

TABLE V
REPRODUCIBILITY BUNDLE CONTENTS

Artifact	Purpose
`xgb_pilot_metrics.json`	Primary model metrics (RMSE, R^2).
`level1_pixel_metrics.json`	Independent Level-1 hold-out metrics output.
`level2_regional_metrics.json`	Regional UHI and MODIS consistency summary.
`final_metrics_table.csv`	Paper-ready compact metrics table.
`final_metrics_snapshot.json`	Locked metrics snapshot for manuscript traceability.
`manifest.json`	Run identity and packaged file listing.
`checksums.sha256`	Integrity verification for all bundled files.

C. Operational Reproduction Protocol

To minimize run-to-run variability, a fixed execution protocol was followed:

1) The primary run configuration was fixed, and the preprocessing was not repeated after the validation closure.
2) Seasonal alignment was performed for all combinations of the year and season, ensuring that the total number of seasonal stacks was 12.
3) Sanity and physical consistency checks are performed, such as value range checks, seam checks, and NDVI/NDBI correlation checks.
4) The training table is constructed, and tiles are stitched together while maintaining a fixed spatial hold-out split.
5) The XGBoost model has been trained, and the model artifacts along with the evaluation metrics are stored.
6) SHAP global importance and dependence plots are created using the final model.
7) Level-1 and Level-2 validation procedures were performed, and the results were stored in validation directories.
8) A reproducibility bundle has been completed, including a manifest and checksum records, which are the basis of the results reported here.

D. Error Sources and Mitigation

Table VI summarizes key risks and the mitigation strategy used in the final run.

TABLE VI
OBSERVED RISKS AND MITIGATION STRATEGY

Risk	Mitigation Used in Final Run
Seasonal stack drift	Step-12 envelope check against published Bengaluru UHI range
Spatial leakage in training	Spatial block hold-out by tile identifiers
Masking artifacts and NaN regions	Feature gates and documented warnings with downstream consistency checks
Overclaiming external validation	MODIS used only for regional consistency interpretation
Reproducibility loss across environments	Manifest plus SHA-256 checksum bundle for artifact integrity

E. Contextual Literature Comparison

This sub-section positions the given XGBoost performance within the context of some relevant EO-based urban thermal retrieval studies, without making any direct comparisons using different data sets and settings. Table VII shows the context.

TABLE VII
CONTEXTUAL LITERATURE COMPARISON

Study	Method	RMSE (K)	Notes
Chen et al., 2023	Machine-learning LST estimation review	3–5	Representative EO urban thermal retrieval range
Zhang et al., 2024	Landsat-8 random forest LST retrieval	3.2	Landsat-based random forest urban thermal retrieval
This Work	XGBoost with spatial block cross-validation	3.31	Reproducible multi-season baseline with fixed split and artifact traceability

Table VII is a contextual comparison, not a direct benchmark. Data sources, spatial scales, and validation protocols differ across studies.

VII. Controlled Benchmark Extension

This paper proposes a reproducible framework for integrating EO data and machine learning for the analysis of the Urban Heat Island effect in Bengaluru. The framework is validated for consistency. The major contribution of the paper is the interpretable geospatial machine learning framework that focuses on reproducibility and incorporates leakage control constraints. The major contribution of the paper is the interpretable geospatial machine learning framework that focuses on reproducibility.

Constraints related to the control of leakage are included in the framework. There are certain limitations in the paper. Seasonal data imbalance is noticeable in the data, especially during specific winter seasons. The data is visualized using heuristic thresholds. Comparisons are made with the MODIS data, and the data is consistent on the regional scale.

VIII. Conclusion

This paper outlines a framework for reproducible earth observation and machine learning. The focus is on the Bengaluru urban heat island. The findings show that the framework is consistent on a seasonal level and that there are physically significant relationships. The novelty of the paper is that it offers an interpretable geospatial machine learning framework. The framework is reproducible and avoids data leakage.

There are some limitations to the framework, according to the authors. These limitations are that the framework is seasonal, especially with regard to winter. There is also the use of heuristic thresholds to produce visualization masks. The use of MODIS is also shown to be an example of regional consistency rather than ground truth. Future work includes expanding the framework to include Random Forest and CNN-based benchmarks with spatial cross-validation. There is also the possibility of expanding the spatial and temporal extent to evaluate the framework.

References

[1] T. R. Oke, "The energetic basis of the urban heat island," *Quarterly Journal of the Royal Meteorological Society*, vol. 108, no. 455, pp. 1–24, 1982.

[2] T. R. Oke, "City size and the urban heat island," *Atmospheric Environment*, vol. 7, no. 8, pp. 769–779, 1973.

[3] A. Buyantuyev and J. Wu, "Urban heat islands and landscape heterogeneity: linking spatiotemporal variations in surface temperatures to land-cover and socioeconomic patterns," *Landscape Ecology*, vol. 25, pp. 17–33, 2010.

[4] A. J. Arnfield, "Two decades of urban climate research: A review of turbulence, exchanges of energy and water, and the urban heat island," *International Journal of Climatology*, vol. 23, no. 1, pp. 1–26, 2003.

[5] J. A. Voogt and T. R. Oke, "Thermal remote sensing of urban climates," *Remote Sensing of Environment*, vol. 86, no. 3, pp. 370–384, 2003, doi: 10.1016/S0034-4257(03)00079-8.

[6] Q. Weng, D. Lu, and J. Schubring, "Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies," *Remote Sensing of Environment*, vol. 89, no. 4, pp. 467–483, 2004, doi: 10.1016/j.rse.2003.11.005.

[7] M. L. Imhoff, P. Zhang, R. E. Wolfe, and L. Bounoua, "Remote sensing of the urban heat island effect across biomes in the continental USA," *Remote Sensing of Environment*, vol. 114, no. 3, pp. 504–513, 2010.

[8] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bulletin of the American Meteorological Society*, vol. 93, no. 12, pp. 1879–1900, 2012, doi: 10.1175/BAMS-D-11-00019.1.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[11] M. Belgiu and L. Draˇgu¸t, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785–794.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NIPS*, 2017, pp. 4765–4774.

[14] Y. Chen *et al.*, "Machine learning for urban heat island assessment from remote sensing data: A review," *Remote Sensing*, 2023.

[15] W. Zhang *et al.*, "Research on Landsat 8 land surface temperature retrieval and spatial-temporal migration capabilities based on random forest model," *Advances in Space Research*, vol. 74, no. 2, pp. 610–627, Jul. 2024, doi: 10.1016/j.asr.2024.04.007.