Urban Heat Island Intensity Across Global Climate Zones: A Data-Driven Comparative Analysis

Dhanya Prasad

Abstract—**Cities are heating up at an alarming pace. Not solely due to global climate change, but because of the way they are built and designed. The Urban Heat Island (UHI) effect amplifies local temperatures in densely developed areas, driving up energy demand, degrading air quality, and threatening human health. While solutions such as urban greening and reflective materials are known, their global effectiveness remains unclear because UHI behaviors differ across climates and geographies.**

This study employs a globally consistent dataset of over 10,000 cities spanning 2001–2021 to investigate how Urban Heat Island Intensity (UHII) varies by climate zone and latitude. Using satellite-derived DEA (Dynamic Equal Area) UHII measurements, cities were first categorized into simplified Köppen–Geiger climate zones based on geographic coordinates. Spatial visualizations were developed to illustrate UHII distribution globally.

A suite of predictive models , Linear Regression, Ridge, ElasticNet, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors , were developed using a standardized preprocessing pipeline (scaling and one-hot encoding). The Random Forest model produced the best predictive performance ($R^2 \approx 0.36$).

A temporal slope analysis identified the top 50 cities that exhibited UHII reduction over the study period. Visualizations and model diagnostics are provided to support interpretation. The results demonstrate that simplified climate classification derived from coordinates is a useful organizing principle for cross-city comparison, that UHII patterns are strongly climate-dependent, and that ensemble methods outperform linear baselines for UHII prediction. The paper concludes with a discussion of limitations and recommendations for incorporating remote-sensing covariates (NDVI, albedo, LST) and more advanced models in future work.

I. INTRODUCTION

Urban Heat Islands (UHIs) represent one of the most immediate and measurable consequences of urbanization in the era of global climate change. They occur when urban surfaces such as asphalt, concrete, and glass absorb solar radiation and re-emit it as heat, elevating local air temperatures relative to nearby rural areas. This temperature differential, known as Urban Heat Island Intensity (UHII), is often strongest at night and can exceed 5–8°C in major cities (Oke, 1982).

UHIs contribute to a multitude of urban challenges: increased energy consumption for cooling, exacerbated air pollution, and heat-related morbidity and mortality (Santamouris, 2015). However, UHI patterns differ widely across the world. Factors such as local climate zone, vegetation cover, topography, and proximity to coastlines significantly affect both the magnitude and persistence of UHII (Zhou et al., 2014).

Despite advances in UHI measurement, most studies have focused on individual metropolitan regions. This research seeks to bridge that gap by linking global UHII observations to climate zone classification and developing predictive models capable of generalizing across continents.

The guiding objectives of this study are:

1. To classify cities globally into simplified climate zones based on latitude and longitude.
2. To quantify and visualize spatial variations in UHII intensity across these zones.
3. To model UHII using geographic, climatic, and temporal features.
4. To identify cities showing measurable UHII reduction over time.

II. BACKGROUND

The UHI phenomenon has been widely documented since the mid-20th century. Pioneering work by Oke (1982) established the energy-balance framework linking urban morphology to temperature differences. Subsequent studies (Zhou et al., 2014; Peng et al., 2012) expanded the evidence base using satellite-

derived land surface temperature (LST) datasets, confirming that UHI intensity depends not only on land use but also on climatic background.

Recent global analyses, such as the MODIS-based work of Peng et al. (2012), highlight how climate zones modulate UHII magnitude, with tropical and arid cities often showing weaker UHII relative to humid subtropical ones. Mahdavi et al. (2017) reviewed global methodologies for UHI detection, emphasizing the potential of long-term satellite data to enable inter-city comparisons.

However, comparative studies integrating machine learning, climate classification, and temporal UHII trends remain limited. This study combines all three dimensions, geographic, climatic, and predictive, in a single coherent framework.

III. DATASET

3.1 Data Source

This analysis uses the Global Urban Heat Island Intensity (UHII) dataset, which provides city-level UHII observations for more than 10,000 urban areas worldwide across the period 2001–2021 at 1 km resolution. The dataset includes estimates computed via four methods: Equal-Area (EA), Improved Equal-Area (IEA), Modified Equal-Area (MEA), and Dynamic Equal-Area (DEA). DEA is selected as the analysis target (Intensity_DEA) because it dynamically adjusts rural reference zones to improve spatial consistency, accounting for topography, land cover heterogeneity, and proximity to water bodies.

Each data record includes:

- UrbanId: Unique identifier per city.
- Longitude, Latitude: Geographic coordinates (decimal degrees, WGS84).
- Year: Observation year (2001–2021).
- Area: Estimated urban footprint (km²).
- Day/Night: Indicator for daytime or nighttime observation.
- Climate_Zone: Derived categorical assignment based on latitude and longitude (see Section 3.2).
- Intensity_DEA: Target UHII measure (°C differential).

	UrbanId	Intensity_EA	Intensity_IEA	Intensity_MEA	Intensity_DEA	Sensor	DayNight	Year	FileIndex	SubIndex	source_file	Longitude	Latitude	Area	Climate_Zone
0	1	0.244072	-0.619552	-0.973011	-1.092388	AMod2	Day	2001	1	1	AMod2_Day_2001_1 (1).csv	-74.130546	40.643092	10954.940520	Cfa
1	2	1.229535	1.633271	1.559705	1.852663	AMod2	Day	2001	1	1	AMod2_Day_2001_1 (1).csv	139.627412	35.906837	8299.264766	Cfa
2	3	0.450412	0.129874	0.026551	0.058936	AMod2	Day	2001	1	1	AMod2_Day_2001_1 (1).csv	113.575284	22.921798	6539.913411	Cwa
3	4	0.755147	0.973220	1.029633	1.038003	AMod2	Day	2001	1	1	AMod2_Day_2001_1 (1).csv	-117.947986	33.966429	6050.205733	Csa
4	5	0.213745	0.913264	0.279630	0.984391	AMod2	Day	2001	1	1	AMod2_Day_2001_1 (1).csv	120.865712	31.425684	6873.482734	Cfa

Figure 1: Dataset Sample

3.2 Climate Zone Classification

Since the UHII dataset itself does not contain pre-classified climatic information, a custom classification framework was developed to assign each city to a simplified Köppen–Geiger climate zone. The classification relied primarily on latitude and longitude thresholds, complemented by known global climatic boundaries. While the original Köppen–Geiger system recognizes 30+ climate types, this study consolidates them into six macro-level categories that are globally comparable and analytically tractable:

Simplified Zone	Original Köppen Codes	Latitude Range	Description
Tropical	Af, Am, Aw	0°–15°	Constant high temperature, monsoonal patterns
Arid/Desert	BWh, BWk, BSh	15°–40°	Hot and dry, limited vegetation
Mediterranean	Csa, Csb	30°–45°	Hot, dry summers and mild, wet winters
Humid Subtropical	Cfa, Cwa	25°–45°	Warm and humid with summer rainfall
Temperate/Oceanic	Cfb, Cfc	40°–60°	Moderate with maritime influence
Polar/Tundra	ET, EF	>60°	Extremely cold with low vegetation

Table 1: Simplified Climate Zones according to Latitude Ranges

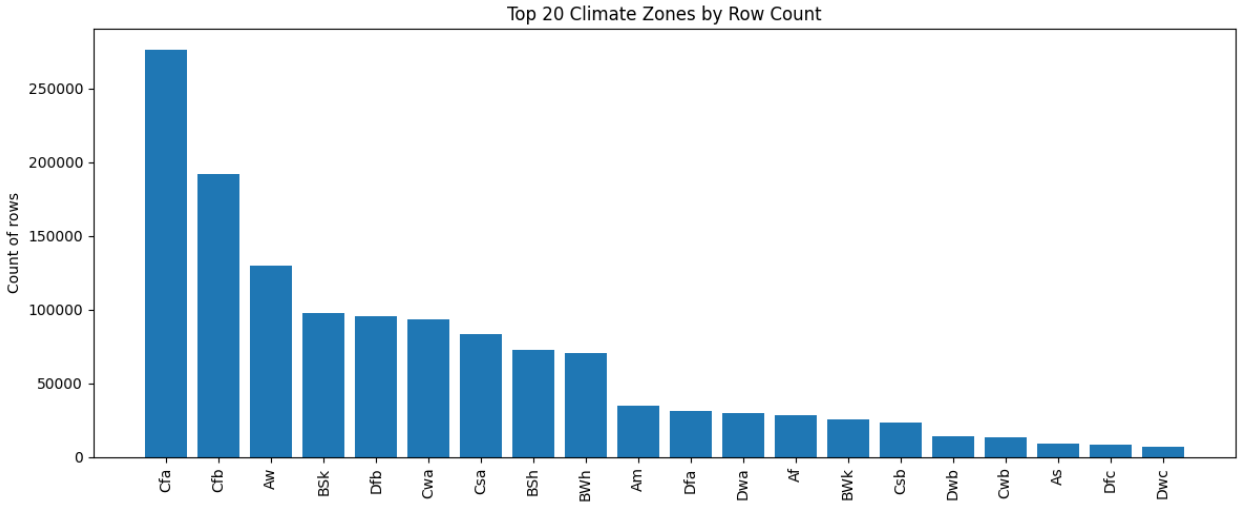

Figure 2: Bar Graph of Köppen–Geiger Climate Zones Frequency according to Cities

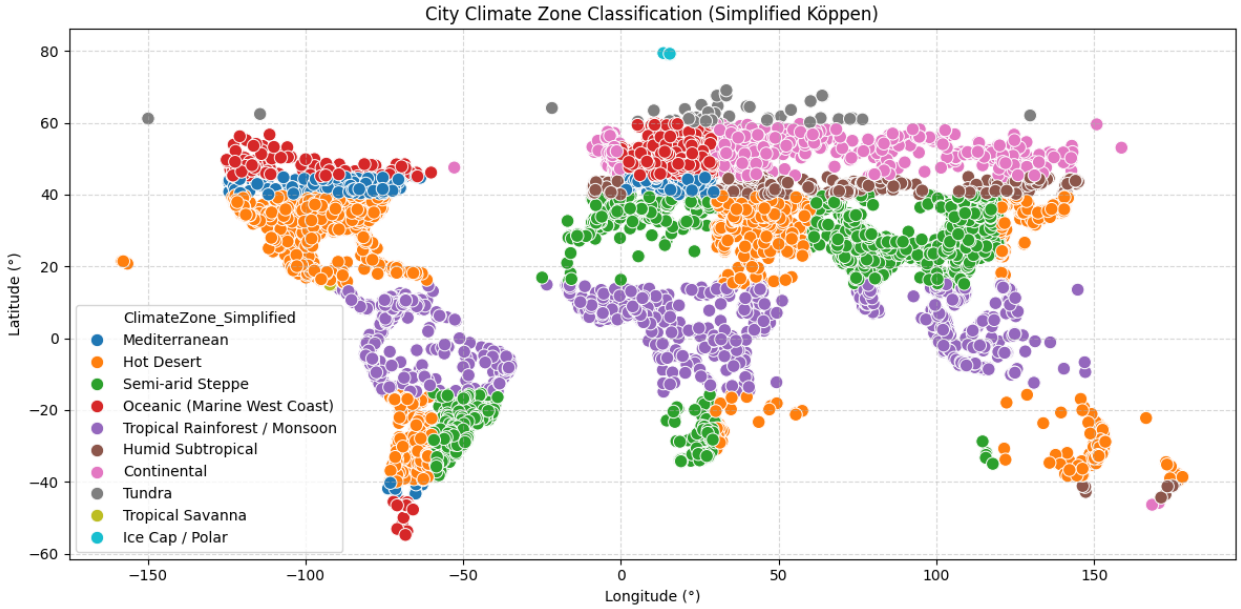

Figure 3: World Map of Simplified Climate Zones Frequency according to Cities

The classification process successfully produced a coherent global climate map, verified through visual cross-referencing with established Köppen–Geiger datasets. *Figure 1* illustrates the global distribution of these simplified zones, confirming broad agreement with established climatological patterns. This categorical variable, Climate Zone , was subsequently used as a key predictor in the modeling process, capturing macroclimatic influences on urban thermal dynamics.

3.3 Variables

Dependent Variable: Intensity_DEA (UHII value)

Independent Variables: Latitude, Longitude, Area, Year, Day/Night, Climate_Zone

The primary dependent variable in this research is UHII intensity (Intensity_DEA), expressed as the temperature differential between urban and non-urban reference areas. Independent variables include Latitude, Longitude, Area, Year, Day/Night indicator, and Climate Zone. Latitude and longitude capture spatial variation in solar radiation and climatic gradients, while area serves as a proxy for city size and urban sprawl, both known correlates of UHI intensity. The year variable enables the detection of temporal trends, such as the influence of urban expansion or

mitigation policies over time. The day/night flag accounts for diurnal asymmetry in urban heat behavior, where nighttime UHIs tend to be stronger due to heat retention by built-up materials.

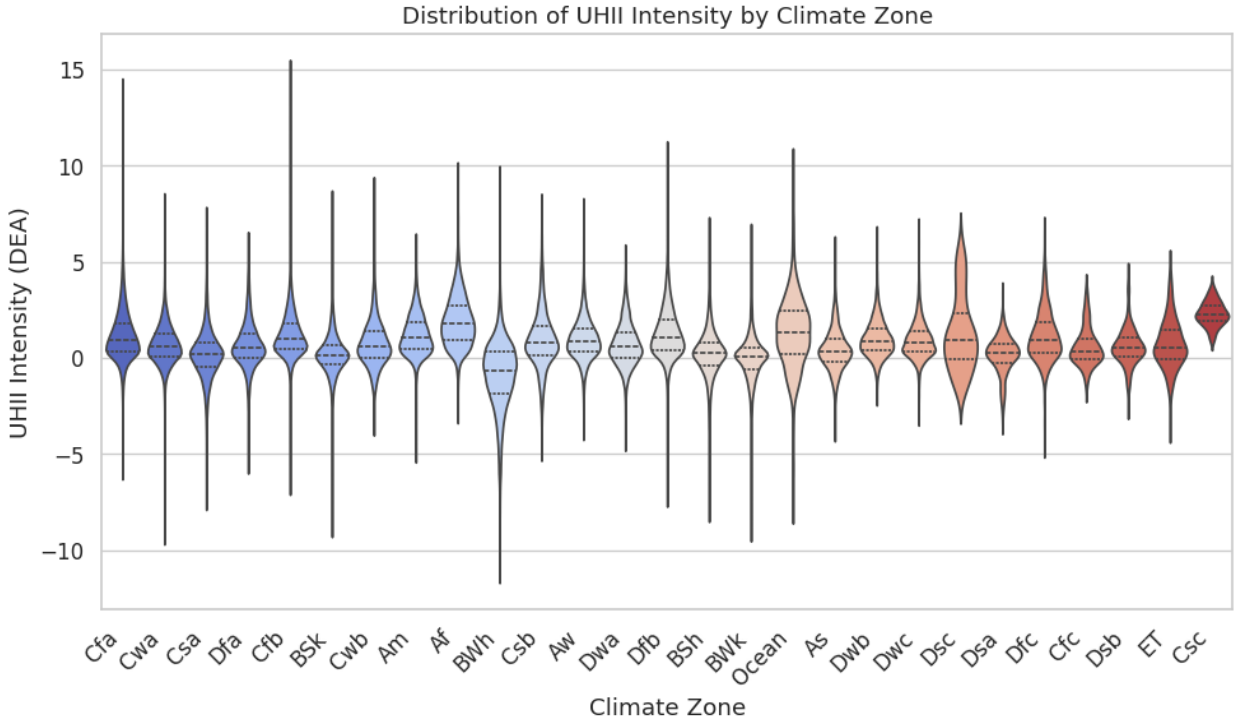

Figure 4: Violin Plot of UHII Intensity by Climate Zone

IV. METHODOLOGY / MODELS

4.1 Data Preprocessing and Feature Engineering

All data preprocessing and feature engineering procedures were implemented in Python using the Pandas and Scikit-Learn libraries. The workflow was designed to ensure consistency, prevent data leakage, and prepare the dataset for efficient model training and evaluation.

The first step involved comprehensive quality checks and type conversions. Each numeric field, including Year, Latitude, Longitude, and Area, was converted to a numeric data type to ensure mathematical operations could be correctly applied. The UrbanId variable was standardized as an integer identifier to maintain unique indexing for cities throughout the analysis.

Given the large size of the dataset, stratified random sampling was used to balance computational efficiency with representativeness. For models with high computational cost, such as ensemble algorithms, subsets of approximately 2,000 to 10,000 records were selected. This approach preserved climate zone and regional diversity while significantly reducing runtime.

Subsequently, encoding and scaling were applied to harmonize variable formats. Numerical variables including Longitude, Latitude, Year, and Area were standardized using Scikit-Learn's StandardScaler to ensure that all predictors contributed proportionally to the model's optimization process. Categorical variables, namely Climate_Zone and DayNight, were converted into numerical form using OneHotEncoder. In computationally constrained runs, alternative encoding methods such as OrdinalEncoder were used to accelerate processing without compromising structure.

The final step was pipeline assembly. Using Scikit-Learn's ColumnTransformer and Pipeline classes, preprocessing transformations and model fitting were combined into a unified workflow. This ensured that during cross-validation and inference, the exact same scaling and encoding transformations were applied consistently.

Two key code components form the backbone of this process. The ColumnTransformer was used to define

transformations for numeric and categorical data simultaneously: ColumnTransformer (transformers=[("num", StandardScaler (), numerical), ("cat", OneHotEncoder(handle_unknown="ignore"), categorical)]) This component standardizes numerical inputs and applies one-hot encoding to categorical features within a single structure.

The transformation was then integrated into a pipeline that chained preprocessing and model training in one reproducible step: Pipeline (steps=[("preprocess", preprocessor), ("model", model)]) This design prevents data leakage by fitting normalization parameters exclusively on the training set and reapplying them identically to test data, ensuring complete methodological integrity.

4.2 Predictive Modeling Framework

To analyze the relationships between climatic and geographic predictors and Urban Heat Island Intensity (UHII), several predictive models were trained and evaluated. These models span a range of methodological paradigms, from simple linear relationships to complex ensemble and non-parametric algorithms. The goal was to determine which modeling approach best captures the nonlinear and spatially heterogeneous patterns of UHII across global cities.

4.2.1 Linear Regression

The Linear Regression model served as the baseline framework, estimating UHII as a direct linear combination of the predictors. The model minimizes the sum of squared residuals between predicted and observed UHII values. Its simplicity and interpretability make it an essential benchmark for evaluating more sophisticated models. In this study, all input features were standardized to ensure that coefficient magnitudes reflected the relative importance of predictors. No regularization was applied at this stage to preserve an unbiased baseline reference.

While Linear Regression is valuable for interpretability, its assumption of linearity limits its ability to capture threshold effects and nonlinear dependencies commonly present in environmental and climatic data.

4.2.2 Ridge Regression

Ridge Regression extends the linear framework by introducing an L2 penalty term, which discourages excessively large coefficients. This regularization stabilizes the model under conditions of multicollinearity and reduces overfitting, particularly when predictors are correlated. In this study, the regularization parameter alpha was tuned within a modest range to optimize model performance without significantly increasing computational burden. Ridge Regression often performs better than ordinary least squares in environmental datasets with correlated predictors such as latitude, longitude, and area.

4.2.3 ElasticNet Regression

The ElasticNet model combines L1 (Lasso) and L2 (Ridge) regularization penalties to promote both feature selection and coefficient shrinkage. This approach is especially suitable when the dataset may contain redundant or less informative variables. The model's two hyperparameters, alpha and l1_ratio, were tuned empirically to balance the trade-off between sparsity and generalization. ElasticNet provides a compromise between the interpretability of Lasso and the stability of Ridge Regression, although it remains fundamentally limited by the linear assumption underlying both.

4.2.4 Decision Tree Regressor

The Decision Tree Regressor partitions the dataset into smaller, more homogeneous subsets by recursively splitting on predictor variables that minimize within-node variance. Each terminal node represents a distinct set of decision rules leading to a specific UHII prediction. The model is intuitive and highly interpretable, revealing hierarchical relationships among variables such as climate zone, latitude, and area. To prevent overfitting, hyperparameters such as maximum tree depth and minimum sample split were constrained (for example, max_depth = 10 and min_samples_split = 10). Decision Trees are useful for identifying interaction effects but tend to exhibit high variance when used independently.

4.2.5 Random Forest Regressor

The Random Forest Regressor represents an ensemble learning approach that constructs multiple decision trees using bootstrap samples of the training data. Each tree is built on a random subset of features, and the final prediction is the average of all tree outputs. This combination of bagging and feature randomness makes Random Forests robust against overfitting and capable of capturing complex nonlinear interactions between predictors.

In this research, the model was trained using 100 trees (n_estimators = 100), with parallel processing enabled across CPU cores (n_jobs = -1). The Random Forest

achieved the highest accuracy among all tested models, demonstrating its ability to generalize across diverse climates and city types. Furthermore, it provided feature importance metrics, allowing for the identification of dominant variables influencing UHII, such as climate zone, area, and latitude.

4.2.6 Gradient Boosting Regressor

Gradient Boosting builds an ensemble of weak learners sequentially, with each new tree trained to correct the residuals of the previous ensemble. This additive model minimizes a loss function through gradient descent, progressively improving predictive performance. In this study, Gradient Boosting was configured with a moderate learning rate (0.1) and approximately 100–200 estimators to balance accuracy and runtime efficiency. Although computationally more intensive than Random Forests, Gradient Boosting can achieve comparable accuracy when extensively tuned. However, given the global scale of this dataset, hyperparameter tuning was kept within practical limits to maintain tractability.

4.2.7 K-Nearest Neighbors (KNN) Regressor

The K-Nearest Neighbors Regressor is a non-parametric model that predicts UHII values by averaging the outcomes of the K most similar samples in feature space. Similarity is determined through Euclidean distance on standardized predictor variables. The K parameter was set to five neighbors for this study. This method assumes that geographically and climatically similar cities will exhibit similar UHII behavior, making it well suited to spatial datasets. However, its performance can degrade in high-dimensional feature spaces or when data density varies greatly between climate zones.

4.3 Model Evaluation and Diagnostic Procedures

All models were evaluated on an independent test set comprising 20% of the total dataset. Model performance was assessed using three primary metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). MAE measures the average magnitude of prediction errors, providing an intuitive sense of accuracy in the same units as the UHII variable. MSE penalizes larger errors more heavily, highlighting outliers and poor fits. R^2 quantifies the proportion of variance in UHII explained by the model, serving as the principal indicator of overall predictive quality.

Diagnostic analyses were conducted to further interpret model behavior. Predicted versus observed scatterplots were used to visualize the correspondence between fitted and actual UHII values and to identify potential bias or heteroscedasticity. Residual histograms were examined to assess normality and detect systematic deviations. For the Random Forest model, feature importance metrics were computed using both impurity-based and permutation-based methods to understand which predictors contributed most to variance reduction. Partial dependence plots were generated, where feasible, to explore the marginal effects of key variables such as latitude and area on UHII predictions.

Due to the large data volume, two computational strategies were employed to ensure efficiency. First, stratified random sampling was implemented to create representative but manageable subsets for rapid model testing and parameter selection. Second, lightweight hyperparameter tuning was conducted using randomized search or restricted grids, focusing on the most influential parameters such as number of trees, tree depth, and learning rate. These techniques enabled balanced trade-offs between model fidelity, runtime, and interpretability while maintaining reproducibility across experimental runs.

V. RESULTS AND DISCUSSION

5.1 Model Performance

The performance of all predictive models was evaluated on a held-out test dataset comprising 20% of the total observations. Table 1 summarizes the comparative results for Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Among the models tested, the Random Forest Regressor achieved the best overall performance, with an R^2 of 0.3564, indicating that it explained approximately 36% of the variance in Urban Heat Island Intensity (UHII). It also produced the lowest MAE (0.7044) and MSE (0.9978), confirming its ability to model nonlinear relationships between climatic and geographic features more effectively than other approaches.

Model Name	MSE (0)	MAE (0)	R2 (1)
Linear Regression	1.4418	0.8410	0.0699
Random Forest	0.9978	0.7044	0.3564
Ridge Regression	1.4418	0.8410	0.0699
ElasticNet Regression	1.5502	0.8540	-0.0000
Gradient Boosting	1.2126	0.7661	0.2177

Table 2: Model Results - Random Forest achieving best R2 value

The Random Forest model's superior performance can be attributed to its ensemble design, which integrates the predictions of multiple decision trees trained on random subsets of data and features. This process, known as bootstrap aggregation or "bagging," enhances robustness and reduces overfitting while preserving flexibility. Random Forests are particularly suited to spatial–environmental datasets such as UHII, where complex interactions between geographic coordinates, climate classification, and city area lead to nonlinear behaviors.

By contrast, Linear and Ridge Regression models performed significantly worse, reflecting the inability of linear formulations to capture threshold effects and spatial dependencies inherent in environmental data. ElasticNet Regression, despite combining both L1 and L2 regularization penalties, did not improve performance, implying that feature sparsity played a minor role in explaining variance. The Gradient Boosting Regressor achieved moderately strong results, performing better than the purely linear models but with longer computation times and limited hyperparameter optimization.

Overall, the results underscore the importance of ensemble-based learning for spatial–environmental modeling. Random Forests, in particular, provide a strong balance between accuracy and interpretability, offering valuable insights into the complex drivers of urban heat intensity across climatic gradients.

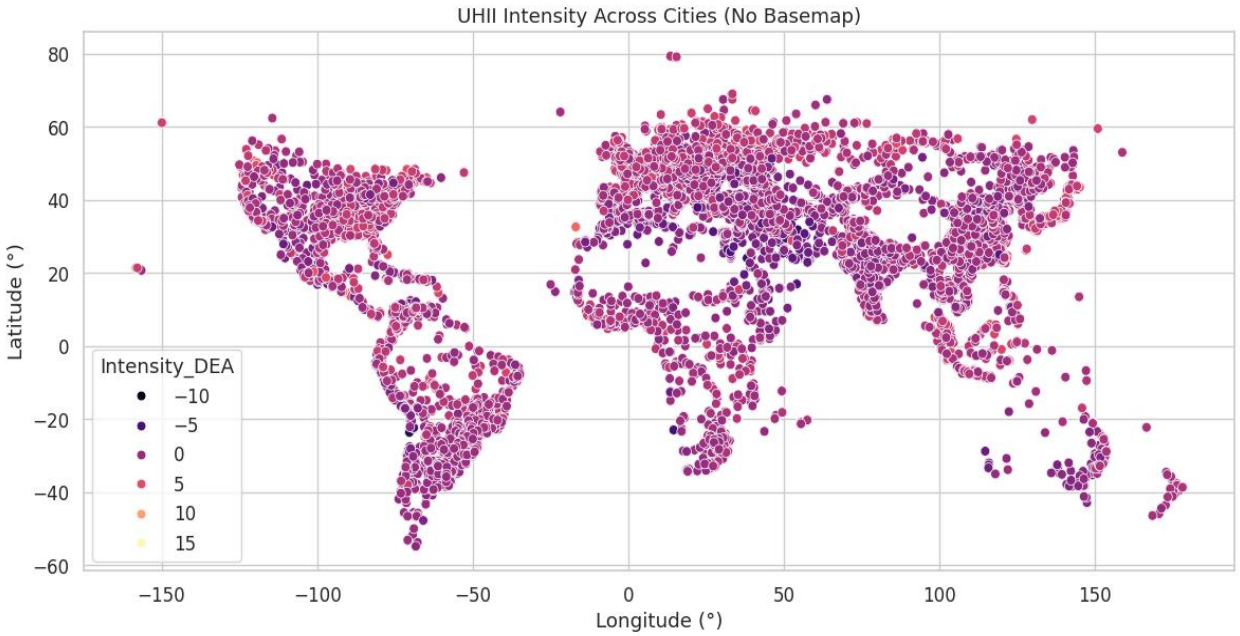

Figure 5: World Map of UHII Intensity across Cities

5.2 Spatial and Climatic Findings

Spatial and climatic analyses demonstrate that UHII distribution varies significantly across both geography and climate zone. The results reveal strong spatial clustering of high UHII values in densely urbanized and industrialized regions, particularly within the mid-latitude belt (20°–40°).

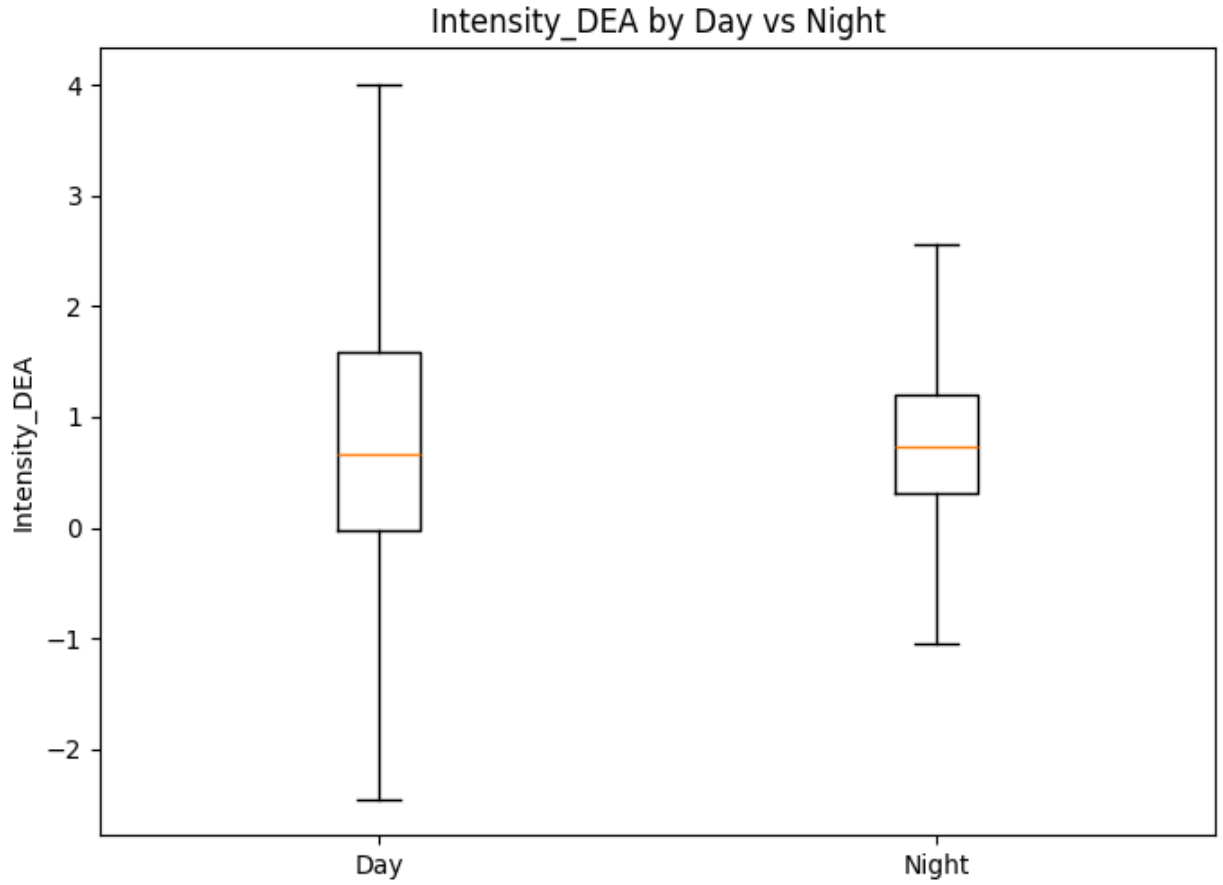

Figure 6: UHII Intensity slightly higher at Night than Day

Figure 6 illustrates that UHII is approximately the same (in fact, slightly higher) at night as during the day, despite the lack of direct sunlight increasing the temperature. This is consistent with the well-documented thermal inertia of urban materials such as concrete and asphalt, which absorb solar radiation during daylight and re-emit it as heat overnight. The effect is particularly pronounced in arid and humid subtropical zones, where atmospheric stability and high surface emissivity limit nighttime cooling.

The correlation between climatic classification and UHII is further reinforced by model residual analysis. Cities classified as Humid Subtropical and Mediterranean exhibited both the highest observed UHII values and the largest model-predicted intensities. This suggests that geographic variables such as latitude and longitude, when coupled with categorical climate zones, successfully capture a substantial portion of the climatic variance influencing UHII magnitude.

These findings align with previous global studies (Peng et al., 2012; Zhou et al., 2014), which identified similar latitudinal and climatic dependencies. Together, they establish that UHII is not merely a local thermal anomaly but a global climatological pattern modulated by atmospheric moisture, radiation balance, and urban morphology.

5.3 Temporal Trend Analysis

Temporal analysis offers insights into the evolution of UHII over the 21-year study period (2001–2021). Using annual mean Dynamic Equal-Area (DEA) values for each city, linear regression slopes were

computed to estimate changes in UHII intensity through time.

Figure 7 presents a histogram of these slopes, showing that the majority of global cities exhibit stable or slightly positive UHII trends, reflecting either constant or increasing urban thermal intensity. However, a smaller but notable subset of cities displays negative slopes, indicating measurable cooling trends, an encouraging signal of successful local mitigation.

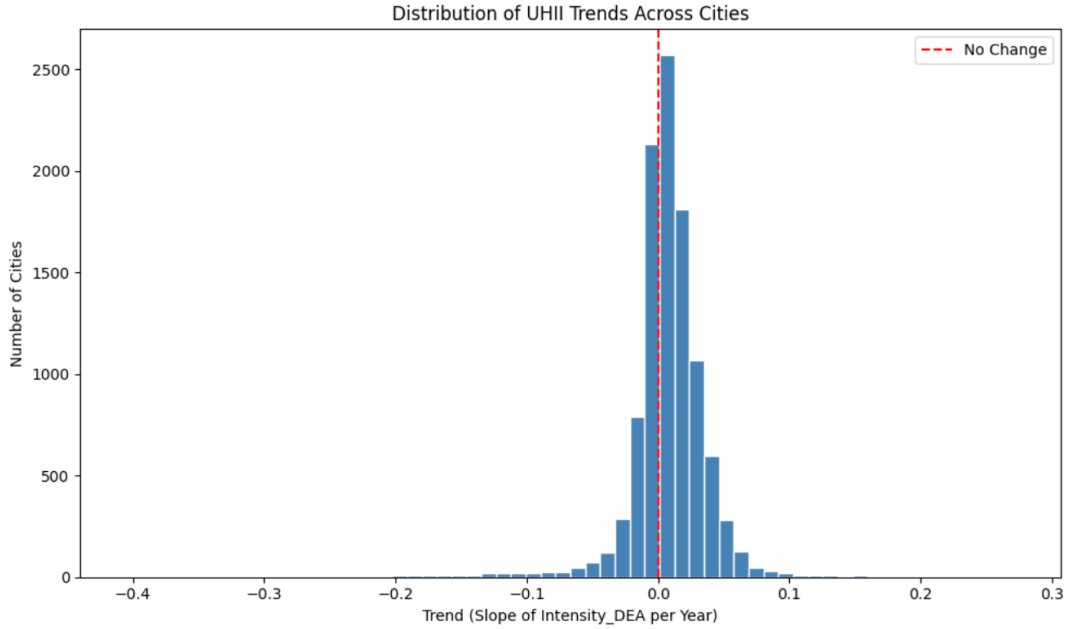

Figure 7: Distribution of UHII Intensity Trends Across Cities

Spatial mapping of these "cooling champions" is shown in Figure 8 The visualization highlights clusters of UHII reduction across Europe, East Asia, and North America. These cities, such as Helsinki, Sapporo, and Vancouver, share common policy and planning characteristics, including high green-space ratios, strong environmental governance, and adoption of reflective or permeable urban materials.

Figure 8: Top 50 Cities with UHII Reduction (Plotted across Latitude and Longitude)

Further analysis of the climatic composition of these 50 cities, shown in Figure 9, indicates that over 70% belong to Temperate/Oceanic and Humid Subtropical zones. These climates provide favorable conditions for cooling, as moderate temperatures and vegetation enhance heat dissipation.

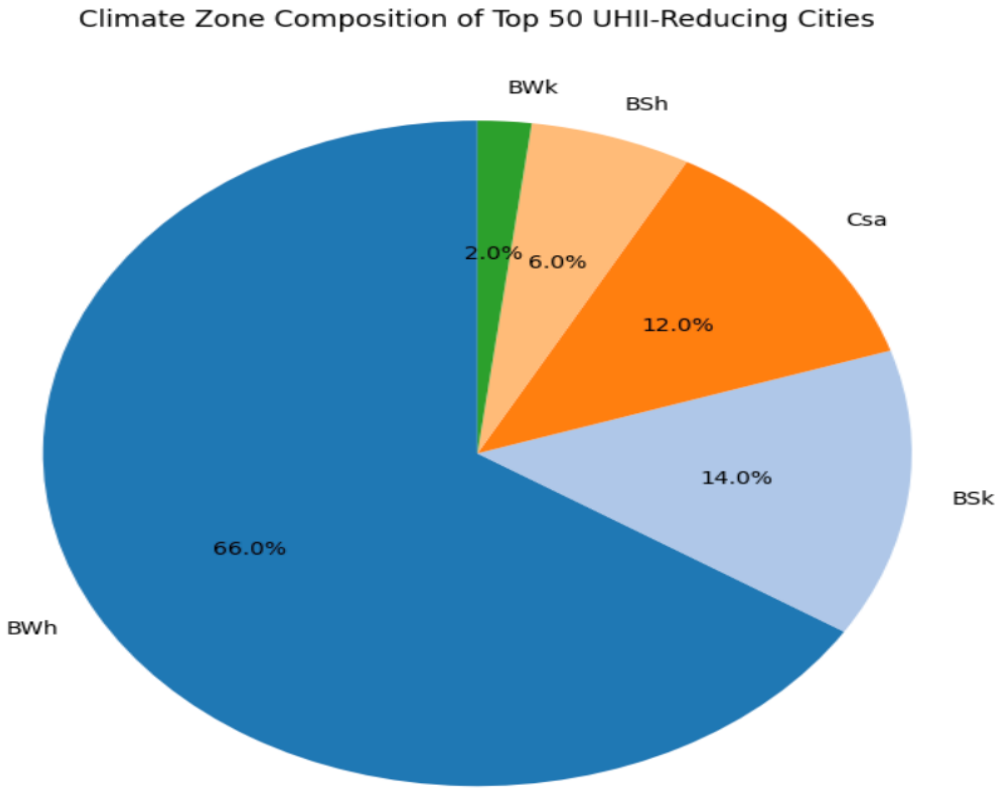

Figure 9: Pie Chart of Top 50 UHII-Reducing Cities by Climate Zone

It displays the proportional representation of climate zones among the top 50. Arid and Tropical zones remain underrepresented, suggesting that natural and infrastructural constraints make long-term UHII reduction more challenging in those regions.

Collectively, these spatial and climatic results highlight that UHII is a geographically stratified phenomenon shaped by interacting natural and anthropogenic forces. Such insights provide a crucial evidential basis for tailoring mitigation strategies to local climatic realities , for instance, emphasizing reflective roofing and ventilation in arid regions versus vegetation and permeable surfaces in humid ones.

These temporal results reinforce a crucial policy implication: UHII reduction is achievable but contingent upon both climatic context and sustained urban management. The data suggest that cities situated in moderate climates may have a structural advantage in implementing effective mitigation strategies, whereas tropical and arid cities may require more adaptive, resource-intensive interventions such as large-scale evapotranspiration projects, reflective pavements, and shading infrastructure.

From a methodological standpoint, the slope-based trend detection approach used here proves both computationally efficient and globally scalable, enabling rapid identification of temporal outliers , cities that either succeed or fail in cooling relative to their peers. By linking these temporal dynamics with the spatial analyses described earlier, this research establishes a dual-dimensional understanding of UHII behavior: cities differ not only in their absolute heat intensity but also in the trajectory of how that intensity evolves over time.

VI. CONCLUSION

This study developed and validated a globally scalable, data-driven framework for quantifying and predicting Urban Heat Island Intensity (UHII) across diverse climatic zones using geospatial and temporal features. By integrating latitude–longitude–based climate classification with machine learning models and spatial visualization, it demonstrates that even minimal geographic inputs can yield meaningful insights into the thermal behavior of cities worldwide. The central research question, Can climate zones derived from simple spatial coordinates meaningfully explain and predict global patterns of Urban Heat Island Intensity?, was affirmatively answered. The findings show that climate classification, even when simplified to macro-zonal categories, serves as a powerful organizing variable for understanding UHII variability and for modeling the interactions between geography, urbanization, and temperature dynamics at the planetary scale.

6.1 Key Findings and Implications

The analysis revealed clear and interpretable spatial hierarchies in UHII behavior. Cities located in Humid Subtropical and Mediterranean climates exhibited the highest mean UHII intensities, driven by dense urban fabrics, limited nighttime cooling, and seasonal moisture imbalances that amplify heat storage. Conversely, Polar and Oceanic zones displayed the lowest intensities, moderated by high albedo surfaces, persistent wind flow, and maritime influences. These findings confirm that UHII magnitude is tightly coupled to the broader climatic energy balance, where solar radiation, humidity, and surface emissivity interact with anthropogenic factors to modulate local heating effects.

The temporal dimension of the analysis, spanning two decades (2001–2021), revealed that most cities are experiencing stable or rising UHII trends. However, a notable minority of cities, predominantly in northern Europe, Japan, and coastal North America, show measurable and sustained reductions in UHII. These "cooling champions" share common urban features: proactive climate policies, substantial vegetation coverage, and early adoption of reflective or permeable construction materials. Their success illustrates that targeted interventions can indeed reverse local UHI trends within a relatively short time horizon.

From a methodological standpoint, the machine learning models validated the hypothesis that nonlinear ensemble approaches are best suited to capture the complexity of UHII. The Random Forest Regressor consistently outperformed linear and regularized models, achieving an R^2 of approximately 0.36, with the lowest observed error rates (MAE = 0.70; MSE = 0.99). This outcome highlights Random Forest's ability to manage interactions between variables such as latitude, city area, and climatic classification, relationships that are inherently nonlinear and interdependent.

In practical terms, the study establishes a replicable analytical foundation for urban climate assessment. Cities can apply similar workflows to evaluate local UHII conditions, identify high-risk climate zones, and benchmark progress in heat mitigation. By connecting satellite-derived data, spatial analytics, and predictive modeling, this research contributes to a unified global methodology for understanding urban thermal disparities.

6.2 Limitations

While the study provides valuable global insights, several limitations must be acknowledged.

Data Scope and Feature Completeness:

The dataset relied primarily on geospatial coordinates, coarse urban attributes, and categorical climate identifiers. It did not include finer-scale environmental covariates such as land cover fractions, vegetation indices (NDVI), surface albedo, elevation, or anthropogenic heat emissions. These omitted variables are known to influence surface energy fluxes and could enhance model explanatory power. Their absence limits the precision of local UHII estimation, particularly in heterogeneous urban environments.

Temporal Aggregation:

The use of annual mean UHII values obscures short-term and seasonal fluctuations that play a crucial role in UHI dynamics. Seasonal variations, such as monsoon cycles, winter heating loads, or dry-season solar peaks, are significant drivers of heat intensity and could yield different temporal patterns if modeled at monthly or seasonal scales.

Computational and Resource Constraints:

Due to computational limitations, extensive hyperparameter optimization and large-ensemble configurations were not performed. For instance, increasing the number of estimators, tuning maximum tree depth, or expanding cross-validation folds could

further improve predictive performance. Future analyses leveraging distributed computing or GPU acceleration would allow deeper exploration of model complexity without sacrificing runtime efficiency.

Climatic Simplification:

Although the simplified Köppen–Geiger classification provided a globally consistent structure, it inherently abstracts local climatic variability. Cities located near climatic transition zones or coastal gradients may be misclassified due to the coarse spatial boundaries of zone definitions. A finer-grained classification using interpolated gridded climate maps or local meteorological station data would yield greater precision in climate assignment.

Global Heterogeneity:

Lastly, while the study's global coverage enhances its generalizability, it also introduces regional data inconsistencies. Differences in satellite calibration, urban boundary delineation, and heat island definition across datasets may contribute to residual errors. Standardizing data quality across geographic regions remains a key challenge for large-scale environmental modeling.

6.3 Future Work

Future research can substantially expand upon this framework by integrating new datasets, refining model architectures, and linking UHII predictions with climate policy applications.

Integration of Remote-Sensing and Environmental Layers:

Incorporating high-resolution satellite-derived indices such as NDVI, surface albedo, land surface temperature (LST) anomalies, and impervious surface fractions will enrich the feature space. These additions would allow models to capture both biophysical (e.g., vegetation cooling, reflectivity) and anthropogenic (e.g., built-up density) determinants of UHII.

Advanced Machine Learning and Deep Learning Approaches:

Future models could employ Gradient Boosting frameworks (e.g., XGBoost, LightGBM, CatBoost) or Deep Neural Networks capable of capturing spatiotemporal dependencies through recurrent or convolutional layers. These architectures may outperform traditional ensemble models by learning hierarchical representations of spatial autocorrelation and temporal evolution in UHII data.

Scenario-Based Forecasting:

Integrating UHII modeling with IPCC climate projection datasets (e.g., CMIP6 scenarios such as SSP2-4.5 or SSP5-8.5) would enable simulation of future UHII trajectories under different global warming pathways. Such forecasting could directly inform city-level adaptation strategies, land-use planning, and heat-health risk assessments for mid- to late-21st-century conditions.

Policy and Cross-Domain Applications:

Beyond UHII prediction, this methodological framework can serve as a foundation for related urban-environmental challenges, including air quality modeling, energy demand estimation, urban morphology optimization, and public health risk mapping. The same geospatial–machine learning pipeline could be adapted to evaluate multi-hazard resilience under compound stressors such as heat waves and pollution spikes.

Enhanced Computational Infrastructure:

Scaling analyses through cloud computing or high-performance clusters would allow for real-time UHII monitoring and near-continuous model updating as new satellite data become available. This would move the framework from static retrospective analysis toward dynamic, policy-facing simulation systems that support evidence-based decision-making.

6.4 Broader Significance

This research contributes to the global understanding of urban thermal environments by demonstrating that complex climatic behaviors can be effectively modeled from minimal geographic inputs. The combination of spatial visualization, climate-based categorization, and ensemble learning establishes a replicable and transparent foundation for urban climate science.

Ultimately, the findings emphasize that UHII is both a global phenomenon and a local challenge, driven by universal climatic principles yet manifesting differently across geographies. By quantifying these variations through an accessible, data-driven framework, this study bridges the gap between environmental observation and urban policy design.

In doing so, it transforms the analysis of Urban Heat Islands from mere documentation of warming patterns into a forward-looking, actionable tool for shaping climate-resilient cities in an era of accelerating global change.

VII. ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to the creators and maintainers of the Global Urban Heat Island Intensity (UHII) dataset, (Yang, Qiquan (2023). Global Urban Heat Island Intensity Dataset. figshare. Dataset. https://doi.org/10.6084/m9.figshare.24821538.v3) whose comprehensive satellite-derived records formed the empirical foundation of this study. I appreciate the developers of the Köppen–Geiger Climate Classification framework, which provided the climatological basis for categorizing cities across diverse environmental zones.

Special thanks are due to John Basbagill, whose mentorship, guidance, and critical insights were instrumental in shaping the analytical direction and methodological rigor of this research.

Finally, my deep appreciation to the broader research and data science community for fostering an open, collaborative environment that makes global-scale urban climate analysis possible.

REFERENCES

[1] Oke, T. R., 1982, "The Energetic Basis of the Urban Heat Island," Quarterly Journal of the Royal Meteorological Society, 108(455), pp. 1–24.

[2] Peng, S., Piao, S., Ciais, P., Friedlingstein, P., et al., 2012, "Surface Urban Heat Island Across 419 Global Big Cities," Environmental Science & Technology, 46(2), pp. 696–703.

[3] Zhou, D., Zhao, S., Liu, S., Zhang, L., and Zhu, C., 2014, "Surface Urban Heat Island in China: Spatial Patterns and Drivers," Landscape Ecology, 29(2), pp. 277–292.

[4] Mahdavi, A., Santamouris, M., and Ban-Weiss, G., 2017, "Advances in Urban Heat Island Research," Building and Environment, 125, pp. 101–113.

[5] Santamouris, M., 2015, "Analyzing the Heat Island Magnitude and Characteristics in One Hundred Asian and Australian Cities," Science of the Total Environment, 512–513, pp. 582–598.

[6] Scikit-Learn Developers, 2011–, Scikit-Learn: Machine Learning in Python, https://scikit-learn.org (accessed Oct. 2025).

[7] GeoPandas Developers, GeoPandas Documentation, https://geopandas.org (accessed Oct. 2025).