International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 1309-1317
https://doi.org/10.70917/ijcisim-2026-6087
Article

AI-Based Prediction of Urban Air Temperature in Baghdad Using Long-Term Climate Reanalysis and Urbanization Proxy Features (1940–2025)

Assis. Prof. Dr. Taif Sami Hasan

Cybersecurity Engineering Department, Al-Ma'moon University, Baghdad, Iraq
Email: taif.s.hasan@almamonuc.edu.iq

Abstract: This study develops a machine-learning framework for daily urban air-temperature prediction in Baghdad by combining long-term meteorological reanalysis with transparent, engineered proxy variables representing progressive urban transformation. Daily weather data for 1940–2024 were obtained for central Baghdad from the Open-Meteo Historical Weather API, while 2025 was reserved as a temporal holdout period. The predictor set includes mean, maximum, and minimum air temperature, precipitation, wind speed, cyclic seasonal encodings, and heuristic indices for cement expansion, vegetation decline, asphalt growth, anthropogenic heat, and drought. Missing values are handled by median imputation and invalid observations are removed before model fitting. A Random Forest regressor with 400 trees and a maximum depth of 18 is used because of its ability to model nonlinear interactions among heterogeneous predictors. The source implementation reports approximate 2025 performance ranges of 1.5–2.5 °C for MAE, 2–3 °C for RMSE, and 0.85–0.95 for R^2. Feature-importance results indicate that seasonal terms dominate predictive skill, whereas urbanization proxies provide comparatively smaller marginal contributions. The framework is therefore best interpreted as a climate-aware predictive prototype rather than a causal urbanization model. The study contributes a reproducible structure for long-horizon urban temperature modelling in data-scarce hot-arid cities and identifies clear directions for improvement using satellite-derived land-cover, humidity, radiation, and independently measured urban indicators.
Keywords: Baghdad; urban heat island; machine learning; Random Forest; temperature prediction; ERA5; urbanization proxies; climate reanalysis

1. Introduction

Urban warming has become a major environmental and planning concern because dense built surfaces, reduced vegetation, anthropogenic heat release, and altered surface-energy exchange can raise temperatures within cities relative to their surroundings. The physical basis of the urban heat island (UHI) has long been associated with differences in radiative storage, surface moisture, aerodynamic roughness, and heat release between urban and non-urban surfaces [1]. Continued urban expansion also changes land cover at large scales, frequently replacing vegetated or permeable surfaces with impervious materials that store sensible heat [2]. These processes are particularly important in hot-arid cities, where summer solar loads are intense and water availability for evaporative cooling is limited.

Baghdad is a useful case study because its urban fabric has changed substantially during recent decades while the city remains exposed to a severe hot-dry climate. Previous local investigations have shown that urban form, building geometry, vegetation, pavement characteristics, and land-use transformation can affect the city's thermal environment. Numerical work on Baghdad's urban form found meaningful differences among compact, attached, and detached urban configurations [3], while later studies connected microclimatic comfort to urban geometry, shading, and vegetation [4], [6]. Remote-sensing studies likewise reported expansion of built-up land and loss of green areas associated with increasing surface and air temperatures [5], [7], [18]. A recent spatiotemporal analysis covering 2003–

Copyright: © 2026 by the authors

2023 further documented a strong relationship between built-up expansion, vegetation change, and land-surface temperature [8].

At the same time, machine learning (ML) is increasingly used to predict urban air temperature because it can capture nonlinear relationships without solving the full atmospheric equations. A recent review of urban-canopy temperature prediction found that tree-based models are widely used and can perform well when predictors and validation procedures are reported transparently [11]. Random Forest (RF), in particular, is attractive because it is robust to nonlinearities, interactions, and mixed predictor scales [10]. Studies of UHI determinants in other cities also demonstrate the utility of RF for ranking environmental drivers such as vegetation, urban morphology, and albedo [12].

The present work extends the original system concept by organizing it as a reproducible research workflow. Its main objective is to evaluate whether long-term meteorological data, combined with simple urbanization proxy features, can provide useful daily temperature predictions for Baghdad in a temporally held-out year. The contribution is methodological rather than causal: the engineered urbanization variables are deterministic proxies and are not direct measurements of land cover, construction volume, traffic, or anthropogenic heat. Accordingly, their role is assessed as supplemental predictive information, and the discussion avoids interpreting RF importance as evidence of causal urban effects.

The study addresses three questions: (i) can a Random Forest trained on 1940–2024 data reproduce the seasonal and short-term structure of 2025 daily temperature; (ii) which predictor groups contribute most to the fitted model; and (iii) what limitations must be resolved before the framework can be used for urban-climate attribution or future scenario analysis?

2. Related Work

2.1 Urban heat and urbanization in Baghdad

Baghdad-specific research consistently indicates that urban morphology and land-cover change are relevant to thermal conditions. Alobaydi et al. [3] used ENVI-met simulations to compare urban configurations and showed that street geometry influences air, surface, and radiant temperatures. Abaas [4] similarly found that compact urban geometry, shade, vegetation, and high-albedo materials can improve thermal comfort. Using remote sensing, Hassoon and Ali [5] linked urban expansion between 2000 and 2015 to higher surface and air temperatures, while Abdulateef and Al-Alwan [18] identified loss of vegetation and water, low-albedo materials, modern grid forms, and anthropogenic heat as contributors to Baghdad's surface UHI. Salman and Saleem [6] evaluated mitigation scenarios involving vegetation, cool materials, and urban geometry, and Najah et al. [7] quantified the thermal consequences of transforming green spaces into built land. Most recently, Abood et al. [8] reported continued urban expansion between 2003 and 2023 and strong associations between built-up indices, vegetation indices, and land-surface temperature.

These studies provide strong motivation for including urban environmental information in predictive models. However, most of the evidence is derived from satellite land-surface temperature, numerical microclimate simulation, or selected field periods rather than a continuous daily air-temperature series spanning several decades. The present framework therefore explores a complementary long-term prediction perspective.

2.2 Machine learning for urban temperature prediction

ML-based urban temperature research has expanded rapidly. Wang et al. [11] reviewed 45 studies published during 2011–2022 and emphasized the importance of transparent metadata, quality control, diverse predictor sources, and rigorous evaluation. Hou et al. [12] used RF to prioritize environmental determinants of UHI across major Chinese cities, demonstrating the value of tree-based models for nonlinear attribution, while also highlighting the importance of directly observed land-surface characteristics such as NDVI and albedo. Broader reviews of UHI and heat-wave research have likewise shown increasing reliance on remote sensing and multi-source data [13].

This literature also suggests an important methodological caution: predictor importance is sensitive to correlated variables, data resolution, temporal autocorrelation, and the definition of the target. In the present dataset, several urbanization features are linear functions of a common time-progress variable. They are therefore highly collinear and should not be interpreted as independent physical drivers. This limitation is explicitly considered in the interpretation of the results.

3. Materials and Methods

3.1 Study area and meteorological data

The study area is represented by a central Baghdad coordinate at approximately 33.3152° N and 44.3661° E. Daily historical weather data were obtained through the Open-Meteo Historical Weather API [14]. Open-Meteo provides long-term reanalysis-based historical weather records, including ERA5-derived products that extend back to 1940. ERA5 is a global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts and is widely used for climate and environmental research [9].

The model-development period covers 1 January 1940 through 31 December 2024. The calendar year 2025 is excluded from training and used as a temporal holdout set. This split is preferable to random row-wise partitioning because it preserves chronological order and provides a more realistic evaluation of performance on unseen future dates. The source implementation retrieves data year by year to reduce API request size and includes retry logic for failed requests, timeouts, and rate-limit responses.

Data component	Specification
Location	Baghdad, Iraq (33.3152° N, 44.3661° E)
Training period	1940–2024
Temporal holdout	2025
Primary data source	Open-Meteo Historical Weather API / reanalysis-based data
Meteorological variables	Mean, maximum, and minimum temperature; precipitation/rain; wind speed
Temporal variables	Year, month, day-of-year, sine and cosine seasonal encodings
Engineered environmental variables	Cement, green-area, asphalt, human-heat, and drought indices

Table 1. Data scope and predictor groups used in the study.

3.2 Data preprocessing

Data preprocessing follows four steps. First, the API response is converted into a daily tabular structure and dates are parsed into calendar attributes. Second, numeric missing values are replaced using median imputation. Third, infinite values and structurally invalid rows are removed. Fourth, the 2025 observations are separated before model fitting so that they do not influence training. The implementation should retain the same transformation logic for both training and holdout data to prevent data leakage.

Because the source data are reanalysis-derived rather than a dense network of independent weather stations, the 2025 evaluation should be interpreted as temporal generalization against the same data family, not as independent station validation. A future version should compare model outputs with observations from the Iraqi Meteorological Organization or another quality-controlled in-situ source.

3.3 Feature engineering

Three predictor families are used: meteorological variables, seasonal variables, and urban/environmental proxies. The seasonal cycle is represented using sine and cosine transformations of day-of-year d, which avoids the artificial discontinuity between 31 December and 1 January:

$$\sin_d = \sin(2\pi d / 365.25), \quad \cos_d = \cos(2\pi d / 365.25).$$

Urban transformation is represented by a normalized progress variable p spanning the training interval. To ensure that the final training year maps to 1, p is defined as:

$$p = (\text{year} - 1940) / (2024 - 1940) = (\text{year} - 1940) / 84, \quad 0 \le p \le 1.$$

The four urbanization features are heuristic indices rather than observed physical quantities. They encode monotonic expansion or decline over time and are defined as follows:

$$\text{Cement} = 10 + 60p; \quad \text{Green} = 70 - 60p; \quad \text{Asphalt} = 8 + 55p; \quad \text{HumanHeat} = 5 + 50p.$$

A simple drought proxy is derived from daily rainfall R (mm) as:

$$\text{Drought} = \text{clip}(100 - 2R, 0, 100).$$

Feature	Definition	Interpretation / limitation
Cement index	$10 + 60p$	Proxy for increasing built infrastructure; not measured cement area.
Green index	$70 - 60p$	Proxy for vegetation decline; not satellite NDVI or field vegetation cover.
Asphalt index	$8 + 55p$	Proxy for road/impervious expansion; not measured paved area.
Human-heat index	$5 + 50p$	Proxy for growth in vehicles, generators, and urban metabolism.
Drought index	$\text{clip}(100 - 2R, 0, 100)$	Rainfall-deficit proxy; does not represent soil-moisture drought.

Table 2. Engineered environmental proxy features. These variables are heuristic and should not be interpreted as measured urban indicators.

A critical consequence of this construction is that Cement, Green, Asphalt, and HumanHeat are all deterministic transformations of the same time-progress variable. Their individual RF importance values are therefore unstable under multicollinearity and do not identify independent causal effects. The model can use them as trend proxies, but causal urban-climate interpretation requires observed time-varying inputs such as Landsat/Sentinel NDVI, NDBI, impervious-surface fraction, population density, traffic activity, and anthropogenic heat inventories.

3.4 Random Forest model

Random Forest regression combines predictions from an ensemble of decision trees trained on bootstrap-resampled data and randomized feature subsets [10]. For an ensemble of N trees, the regression prediction can be written as:

$$\hat{y}(x) = [\text{Tree}_1(x) + \text{Tree}_2(x) + \cdots + \text{Tree_N}(x)] / N.$$

The implementation uses 400 trees and a maximum tree depth of 18, with parallel processing enabled. These settings are retained from the source system. RF is suitable for the present problem because it can approximate nonlinear seasonal relationships and interactions without requiring explicit functional forms. The model can be implemented using the RandomForestRegressor estimator in scikit-learn [15].

Hyperparameter	Value
Estimator	Random Forest Regressor
Number of trees	400

Maximum depth	18
Parallel jobs	All available cores
Training window	1940–2024
Evaluation window	2025

Table 3. Random Forest configuration used in the source implementation.

3.5 Evaluation metrics

Performance is evaluated using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). These are standard regression measures and are more interpretable than a custom percentage "accuracy" for a continuous temperature target. The metrics are defined as:

$$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{[(1/n)\ \Sigma\ (y_i - \hat{y}_i)^2]}$$

$$R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2].$$

The original manuscript reports approximate metric ranges rather than exact scalar values. For scientific reproducibility, the final submission should export exact MAE, RMSE, and R^2 values directly from the evaluation script and, ideally, include confidence intervals or bootstrap uncertainty estimates.

4. Results

4.1 Temporal prediction performance

The 2025 holdout curve shows that the RF model reproduces the dominant annual temperature cycle and tracks many shorter-duration changes. The predicted series follows the winter-to-summer warming phase, the summer maximum, and the autumn cooling phase. The largest visible discrepancies occur around abrupt short-term fluctuations and some extreme warm or cool events, which is consistent with the smoothing behavior expected from an ensemble model trained on long-term patterns.

Metric	Reported source range	Interpretation
MAE	≈ 1.5–2.5 °C	Average absolute daily prediction error.
RMSE	≈ 2–3 °C	Places greater penalty on larger deviations.
R^2	≈ 0.85–0.95	High proportion of variance explained in the reported implementation.

Table 4. Approximate evaluation ranges reported in the source implementation. Exact values should replace these ranges before journal submission.

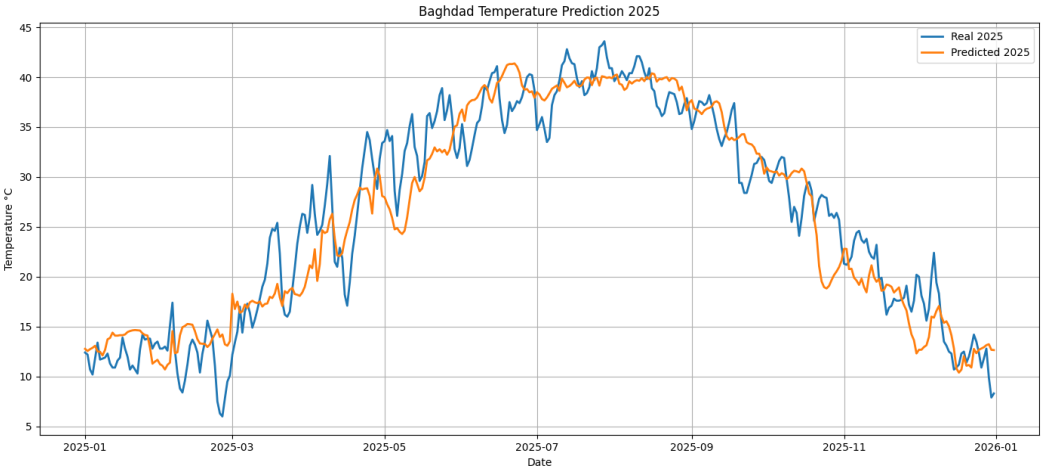


Figure 1. Daily predicted and reference temperatures for the 2025 temporal holdout period, as reported in the source system.

4.2 Feature importance

The feature-importance chart is dominated by the cyclic seasonal variables, especially cos_day and sin_day. Wind contributes a much smaller share, while the engineered urbanization proxies and drought index have low marginal importance in the displayed model. This pattern is scientifically plausible for daily temperature because seasonality controls a large fraction of annual variance. It also means that the current results do not support the claim that urbanization is the dominant predictive factor.

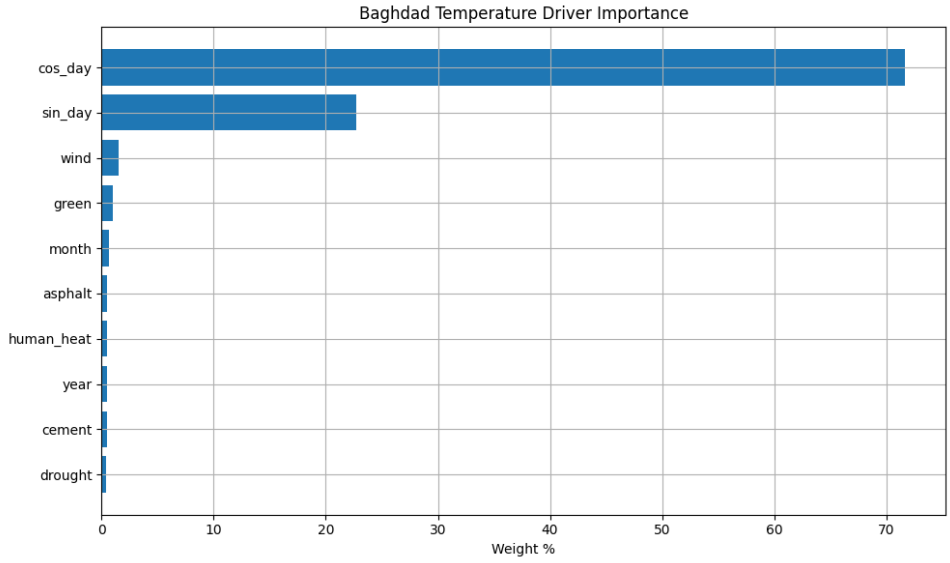


Figure 2. Random Forest feature-importance ranking from the source implementation. Seasonal encodings dominate the model.

Because multiple urbanization proxies are derived from the same progress variable, their individual importance scores should not be compared as if they represented distinct independent mechanisms. Furthermore, conventional impurity-based RF importance can be biased by correlated variables. Permutation importance or SHAP analysis, applied after replacing heuristic proxies with measured urban indicators, would provide a stronger interpretation in future work.

5. Discussion

5.1 Main interpretation

The principal result is that a Random Forest trained on a long historical series can reproduce Baghdad's strong seasonal temperature structure with relatively small reported daily errors. This finding aligns with the broader literature showing that tree-based ML methods can perform effectively for urban temperature prediction when the target is strongly controlled by recurring seasonal and meteorological patterns [11]. The holdout design also improves the credibility of the evaluation relative to random train-test splitting because the model is tested on an unseen future calendar year.

However, the model should not be described as demonstrating that urbanization is the dominant cause of Baghdad's temperature increase. The displayed importance ranking shows the opposite from a predictive standpoint: cyclic seasonal encodings carry most of the model importance. In addition, the four long-term urbanization indices are constructed from the same linear time variable, which prevents clean attribution among cement expansion, vegetation decline, asphalt growth, and human heat. The urban-climate literature for Baghdad provides independent evidence that these mechanisms matter [3]–[8], [18], but the present RF experiment does not isolate their causal effects.

5.2 Relation to Baghdad urban-climate evidence

The value of the urban proxy variables is therefore conceptual: they encode the expectation that Baghdad's urban environment has changed substantially over the 1940–2024 period. This expectation is supported by local remote-sensing and simulation research. Hassoon and Ali [5] reported measurable urban expansion and concurrent thermal change, Najah et al. [7] demonstrated the thermal implications of green-space transformation, and Abood et al. [8] found strong relationships between built-up and vegetation indices and land-surface temperature. These studies support the hypothesis that physically measured urban variables could improve long-term air-temperature prediction beyond meteorological seasonality alone.

The next methodological step is thus not to increase the complexity of the current synthetic indices, but to replace them with observable time-varying covariates. Candidate inputs include annual or seasonal NDVI, NDBI, land-cover fractions, impervious-surface percentage, surface albedo, urban morphology, population density, night-time lights, traffic proxies, and anthropogenic heat estimates. These data can be combined with humidity, pressure, solar radiation, soil moisture, and wind direction available from reanalysis or other data sources. Such a design would allow the model to test whether measured urban change improves out-of-time prediction and would make feature-attribution results physically interpretable.

5.3 Practical implications

Even in its current prototype form, the framework is useful for smart-city and environmental analytics because it demonstrates a complete pipeline from automated data acquisition to preprocessing, feature engineering, model fitting, evaluation, and visualization. A refined version could support heat-risk planning, infrastructure operations, energy-demand anticipation, and evaluation of greening scenarios. For policy applications, however, predictive accuracy must be paired with uncertainty quantification and physically grounded explanatory variables.

6. Limitations

The study has several limitations that should be stated explicitly. First, the urbanization features are synthetic proxies rather than observations. Second, Cement, Green, Asphalt, and HumanHeat are deterministic functions of the same time-progress variable, creating strong multicollinearity and limiting interpretability. Third, the study does not use satellite-derived NDVI, NDBI, albedo, impervious-surface fraction, or land-cover data. Fourth, important meteorological predictors such as relative humidity, surface pressure, solar radiation, cloud cover, and soil moisture are omitted. Fifth, the 2025 holdout comes from the same reanalysis/API family as the training data and therefore is

not independent station validation. Sixth, the reported performance values are approximate ranges rather than exact exported metrics. Seventh, no formal comparison with baseline methods such as climatology, linear regression, gradient boosting, XGBoost, LSTM, or Transformer models is reported. Finally, the present framework is predictive and should not be used for causal claims about urbanization without additional experimental design and observed urban covariates.

7. Conclusions

This study reorganizes and strengthens an AI-based framework for daily urban air-temperature prediction in Baghdad using long-term reanalysis data and engineered urbanization proxies. The Random Forest model, trained on 1940–2024 and evaluated on a 2025 temporal holdout, captures the dominant annual temperature cycle and yields strong reported predictive performance. The feature-importance analysis shows that seasonal encodings dominate model behavior, while the current urbanization proxies have relatively small marginal contributions.

The main scientific conclusion is therefore not that urbanization dominates the model, but that long-term climate prediction can be enhanced by a structured ML pipeline and that physically measured urban variables should be incorporated in the next version. Replacing heuristic indices with satellite and socioeconomic observations, adding humidity and radiation variables, reporting exact metrics with uncertainty, and benchmarking multiple algorithms would substantially increase the rigor and publication value of the framework. With these extensions, the system could become a useful component of climate-aware urban planning and heat-risk management for Baghdad and other hot-arid cities.

8. Future Work

Replace synthetic urbanization proxies with annual or seasonal satellite-derived NDVI, NDBI, impervious-surface, albedo, and land-use variables.

Add humidity, atmospheric pressure, shortwave radiation, cloud cover, soil moisture, and wind-direction predictors.

Benchmark Random Forest against climatology, linear models, gradient boosting/XGBoost, LSTM, and Transformer architectures using identical temporal splits.

Use blocked or rolling-origin cross-validation to quantify year-to-year generalization and report uncertainty intervals.

Evaluate permutation importance and SHAP values rather than relying only on impurity-based feature importance.

Validate predictions against independent station observations and, where available, high-resolution urban sensor data.

Extend the framework to scenario-based projections only after introducing physically grounded urban and climate predictors.

9. Data Source and Reproducibility Statement

The meteorological data source described in this manuscript is publicly accessible through the Open-Meteo Historical Weather API [14], which provides reanalysis-based historical weather data. The engineered proxy variables are fully defined in Section 3.3. For final journal submission, the authors should archive the exact data-query parameters, preprocessing script, model configuration, random seed, and exported evaluation metrics to enable independent reproduction of the reported results.

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, doi: 10.1002/qj.49710845502.
2. K. C. Seto, B. Güneralp, and L. R. Hutyra, “Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools,” Proceedings of the National Academy of Sciences, vol. 109, no. 40, pp. 16083–16088, 2012, doi: 10.1073/pnas.1211658109.
3. D. Alobaydi, M. A. Bakarman, and B. Obeidat, “The Impact of Urban Form Configuration on the Urban Heat Island: The Case Study of Baghdad, Iraq,” Procedia Engineering, vol. 145, pp. 820–827, 2016, doi: 10.1016/j.proeng.2016.04.107.
4. Z. R. Abaas, “Impact of development on Baghdad’s urban microclimate and human thermal comfort,” Alexandria Engineering Journal, vol. 59, no. 1, pp. 275–290, 2020, doi: 10.1016/j.aej.2019.12.040.
5. A. F. Hassoon and A. B. Ali, “Irregular urban Expansion and Its Effects on Air Temperature over Baghdad City using Remote Sensing Technique,” Iraqi Journal of Science, vol. 62, no. 6, pp. 2110–2121, 2021, doi: 10.24996/ijs.2021.62.6.36.
6. A. M. Salman and Y. M. Saleem, “The effect of Urban Heat Island mitigation strategies on outdoor human thermal comfort in the city of Baghdad,” Frontiers of Architectural Research, vol. 10, no. 4, pp. 838–856, 2021, doi: 10.1016/j.foar.2021.07.002.
7. F. T. Najah, S. F. K. Abdullah, and T. A. Abdulkareem, “Urban Land Use Changes: Effect of Green Urban Spaces Transformation on Urban Heat Islands in Baghdad,” Alexandria Engineering Journal, vol. 66, pp. 555–571, 2023, doi: 10.1016/j.aej.2022.11.005.
8. K. I. Abood, A. B. Ali, and S. M. Ahmed, “Urban Heat Island Dynamics and Urban Expansion in Baghdad: A Spatiotemporal Remote Sensing Analysis (2003–2023),” Iraqi Geological Journal, vol. 59, no. 2A, 2026, doi: 10.46717/igj.2026.59.2A.2.
9. H. Hersbach et al., “The ERA5 global reanalysis,” Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.
10. L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
11. H. Wang, J. Yang, G. Chen, C. Ren, and J. Zhang, “Machine learning applications on air temperature prediction in the urban canopy layer: A critical review of 2011–2022,” Urban Climate, vol. 49, Art. no. 101499, 2023, doi: 10.1016/j.uclim.2023.101499.
12. H. Hou, Q. Longyang, H. Su, R. Zeng, T. Xu, and Z.-H. Wang, “Prioritizing environmental determinants of urban heat islands: A machine learning study for major cities in China,” International Journal of Applied Earth Observation and Geoinformation, vol. 122, Art. no. 103411, 2023, doi: 10.1016/j.jag.2023.103411.
13. S. Cheval et al., “A systematic review of urban heat island and heat waves research (1991–2022),” Climate Risk Management, vol. 44, Art. no. 100603, 2024, doi: 10.1016/j.crm.2024.100603.
14. Open-Meteo, “Historical Weather API,” historical weather documentation, accessed Aug. 27, 2026.
15. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
16. C. D. Ziter, E. J. Pedersen, C. J. Kucharik, and M. G. Turner, “Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer,” Proceedings of the National Academy of Sciences, vol. 116, no. 15, pp. 7575–7580, 2019, doi: 10.1073/pnas.1817561116.
17. B. Zhou, D. Rybski, and J. P. Kropp, “The role of city size and urban form in the surface urban heat island,” Scientific Reports, vol. 7, Art. no. 4791, 2017, doi: 10.1038/s41598-017-04242-2.
18. M. F. Abdulateef and H. A. S. Al-Alwan, “Assessment of surface urban heat island intensity and its causes in the city of Baghdad,” IOP Conference Series: Materials Science and Engineering, vol. 745, Art. no. 012162, 2020, doi: 10.1088/1757-899X/745/1/012162.