buildings

MDPI

Article

LCZ-Stratified Estimation of Near-Surface Air Temperature from Land Surface Temperature: Evidence from Central Xi'an, China

Mingzhi Zhang [1], Dian Zhou [1], Keju Liu [1], Jie Si [2,3], Zhaolin Gu [1], Siyuan Miao [4], Xin Lin [5], Tianyu Yang [1] and Duo Xu [1,*]

[1] School of Human Settlement and Civil Engineering, Xi'an Jiaotong University, Xi'an 710049, China; zmz120859@stu.xjtu.edu.cn (M.Z.)
[2] School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710049, China
[3] Xi'an City Planning Design and Research Institute, Xi'an 710082, China
[4] XJTU-POLIMI Joint School, Xi'an Jiaotong University, Xi'an 710049, China
[5] State Grid Shanghai Electric Power Engineering Construction Consulting Company, Shanghai 200120, China
* Correspondence: duo_xu@xjtu.edu.cn

Abstract

Near-surface air temperature (Ta) is essential for urban heat assessment, but ground observations are spatially sparse. This study combined ECOSTRESS land surface temperature (LST), concurrent Ta and built-environment (BE) indicators in four local climate zone (LCZ) categories in central Xi'an. The analysis comprised 176 site-overpass records from 123 sites across six events. Under ten repetitions of site-grouped five-fold cross-validation, LST-only CatBoost yielded RMSEs of 2.69, 1.63, 4.41 and 2.22 °C in LCZ5-II, LCZ4, LCZ2 and LCZ6-II, respectively. OLS performed better in LCZ4 (1.41 °C). Adding 14 screened BE variables increased the equal-weight mean RMSE across LCZs from 2.74 to 3.29 °C; this disadvantage persisted across the tested model capacity settings in LCZ5-II, LCZ4 and LCZ6-II. In event-held-out tests, the LST-only RMSE rose to 5.84, 6.57 and 7.12 °C in those three LCZs. The dominant LCZ2 event lacked sufficient training support for evaluation. LCZ2's inverse prediction–observation association arose predominantly between event means, and all tested site-grouped CatBoost configurations retained negative predictive R^2 values. Within this small sample, additional BE inputs offered limited predictive value, and errors increased substantially when transferring to new events.

Keywords: near-surface air temperature; land surface temperature; ECOSTRESS; local climate zones; built environment; grouped cross-validation; SHAP

Check for updates

Academic Editor: Francesco Nocera

Received: 27 August 2026
Revised: 7 October 2026
Accepted: 16 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Near-surface air temperature (Ta) is important for urban heat exposure, public health and building energy use [1,2]. Urban form and surface cover alter radiation, heat storage and ventilation, creating thermal differences that sparse ground networks may not resolve. Satellite land surface temperature (LST) can complement these observations, but its usefulness for estimating Ta requires evaluation against concurrent ground measurements.

Thermal-infrared products such as Landsat, MODIS and ECOSTRESS differ in spatial resolution and observation timing [3,4]. ECOSTRESS offers an effective native resolution of approximately 70 m and variable overpass times, but clouds, quality requirements and the availability of the concurrent Ta limit usable observations. These constraints are central to evaluating local LST–Ta relationships and their predictive value.

Satellite thermal-infrared observations retrieve radiative surface temperature, while ground instruments measure air temperature at their installation height. Sensible heat exchange, turbulent mixing and shading connect the two quantities, but their relationship varies with surface materials, building configuration, observation time and weather [4–8]. Estimating Ta from LST therefore requires concurrent ground measurements and validation under the relevant urban conditions.

Urban morphology modifies solar radiation, heat storage, longwave heat loss and ventilation through building height and density, street-canyon structure and surface cover. Studies examining LST and outdoor Ta separately have found distinct responses to these factors [5,9]. A single citywide LST–Ta relationship may consequently obscure differences among built environments.

Local climate zones (LCZs) classify urban form using building height and density, surface cover, and spatial openness, providing a standardized framework for comparing thermal conditions among built-environment types [10,11]. Previous studies have used LCZs to map parcel-level urban morphology in Xi'an [12] and to compare the LST or Ta among LCZs [13,14]. A multi-city observational study in eastern China also summarized air temperature and humidity characteristics across LCZs [15]. These studies show that LCZs can organize thermal comparisons among urban-form types and support the identification of type-specific temperature patterns.

Continuous differences remain within an LCZ, including variations in building footprints and vertical profiles, impervious surface fraction, and blue–green cover. Vaidya et al. quantified how surrounding spatial heterogeneity affected the LST within the LCZ framework, while Bai et al. identified threshold and scale dependence in the responses of Ta and LST to surrounding landscapes during extreme heat [16,17]. Other studies have shown that the relationships between built-environment (BE) variables and urban thermal conditions vary with urban context and that combining planimetric and vertical morphology can help characterize parcel-scale temperature differences [18–22]. LCZs and continuous BE variables therefore represent complementary levels of spatial information: LCZs describe differences among urban-form types, whereas BE variables characterize local heterogeneity within a type.

Li et al. [23] combined MODIS LST with 29 ancillary factors at 292 stations across China, trained the models on 2017–2020 data and tested them on 2021 data. Their LightGBM and SHAP analysis identified contributions from meteorological, terrain, urban morphology and anthropogenic variables. Our study uses concurrent ECOSTRESS–Ta pairs within four LCZ categories to examine local prediction, spatial and temporal transfer, and the incremental value of BE inputs and predictor reduction.

Repeated observations from a monitoring site can produce optimistic validation if they enter both the training and test sets [24]. We grouped all records from each site in the same fold. Site-grouped metrics therefore evaluate withheld sites under the sampled event conditions; supplementary event-held-out tests examine transfer across the six overpasses.

We combined ECOSTRESS L2 LST&E, concurrent fixed-site Ta measurements and an existing parcel-level LCZ map in central Xi'an. We compared CatBoost with training-mean and OLS benchmarks, tested whether BE variables improved LST-only predictions, and examined predictor-count sensitivity and LCZ2 prediction failure. The analysis covers the available sites and six summer overpasses.

2. Materials and Methods

2.1. Research Framework and Study Area

2.1.1. Research Framework

The analytical framework combined ECOSTRESS LST, concurrent Ta, LCZ categories and spatial indicators (Figure 1). The nonlinear model and CatBoost input comparisons

used identical site-grouped partitions, with training-mean and OLS benchmarks. Supplementary analyses tested event transfer with and without test-site exclusion, varied model capacity and input definitions, and summarized site-level BE distributions. Event-level residuals and within-/between-event covariance were used to diagnose LCZ2 prediction failure. The SHAP-based predictor selection procedure is described in Section 2.4.4.

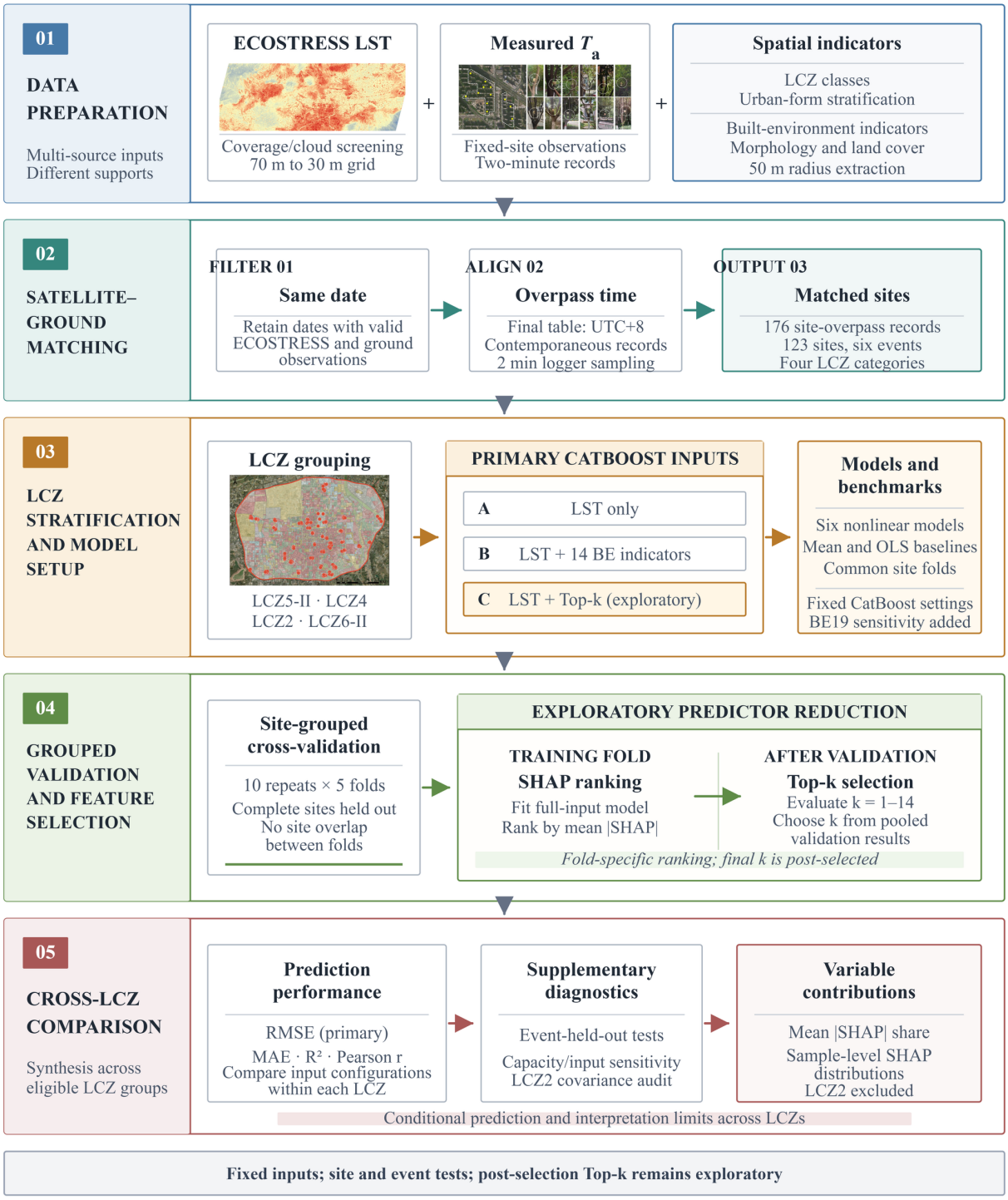


Figure 1. LCZ-stratified LST–Ta analysis framework. ECOSTRESS uses a 70 m source grid and a 30 m processing grid; spatial summaries use a 50 m radius. Site-grouped tests are supplemented by event-held-out and joint event/site exclusion tests, capacity/input sensitivity and LCZ2 diagnostics. SHAP ranking occurs within training folds; k is selected from aggregated validation results.

2.1.2. Study Area and LCZ Classification

Xi'an is located in the central Guanzhong Plain of Shaanxi Province in northwestern China, with the Wei River to the north and the Qinling Mountains to the south. The study area covers the central built-up area of Xi'an. It lies mainly on an open plain, where the urban land surface is shaped by rapid urbanization and the transition to mountainous terrain in the south. Xi'an has a warm-temperate, semi-humid continental monsoon climate, with hot, rainy summers and precipitation concentrated mainly in summer and autumn.

Spatial grouping used the parcel-level LCZ map of Xu et al. [12], derived from spatial indicators and supervised classification. The scheme adapts the Stewart–Oke framework [10] to Xi'an and distinguishes local density subtypes. Figure 2c locates the study area within China, Shaanxi Province and Xi'an; Figure 2d shows the LCZ parcels and monitoring sites.

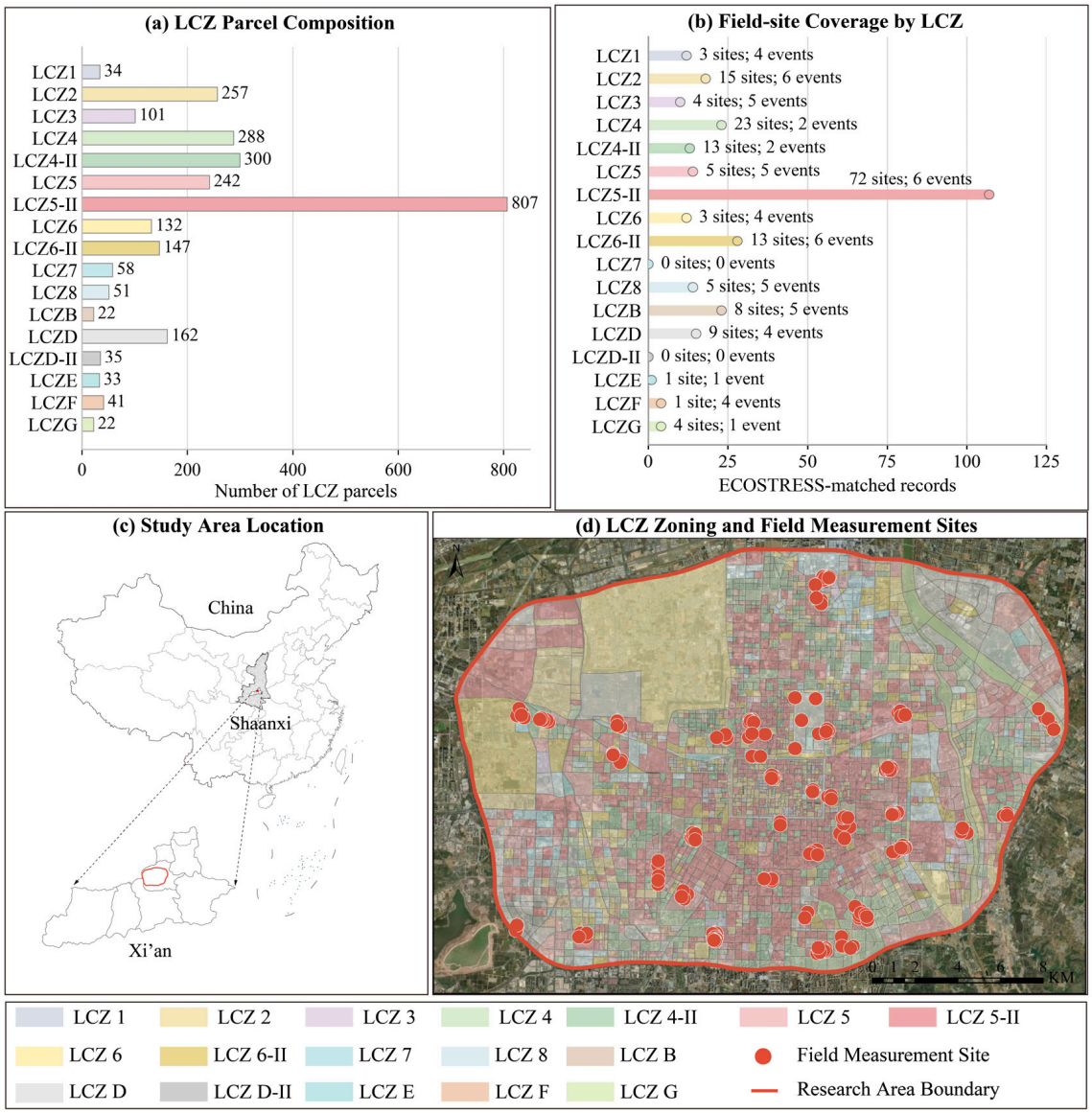


Figure 2. Study area, LCZ distribution, and field measurement coverage. (**a**) Number of parcels in each LCZ category; (**b**) number of ECOSTRESS-matched records in each LCZ category, with the numbers of monitoring sites and overpass events indicated; (**c**) location of the study area in Xi'an, Shaanxi Province, China; (**d**) spatial distribution of LCZ parcels and field measurement sites within the study area.

LCZ2 corresponds broadly to compact mid-rise development and LCZ4 to open high-rise development. LCZ5-II and LCZ6-II are local medium-density refinements of open mid-rise and open low-rise development. Table 6 in Xu et al. [12] defines both II subtypes by a parcel-level building footprint fraction of $0.20 \leq BD < 0.40$, with $10 \leq BH < 30$ m for LCZ5-II and BH < 10 m for LCZ6-II. BD is the building footprint area divided by the parcel area; BH is the building volume divided by the building footprint area. These parcel-classification thresholds differ from the 50 m site buffer attributes used as predictors.

LCZ5-II, LCZ4, LCZ2 and LCZ6-II contained 107, 23, 18 and 28 matched records from 72, 23, 15 and 13 independent sites, respectively. LCZ4 covered two overpasses, and the other categories covered six, with uneven record counts. Figure 2 shows the parcel composition and observational coverage.

2.2. Near-Surface Air Temperature Observations and ECOSTRESS Matching

2.2.1. Ta Observations and ECOSTRESS LST Processing

Near-surface Ta was measured using Biaozhi GM1366 temperature and humidity loggers in naturally ventilated louvered radiation shields. Loggers were generally attached to tree trunks, lamp posts or fences at approximately adult eye level, retrospectively estimated as 1.6 m; exact installation heights were not recorded. Locations avoided air-conditioning exhaust, and housings shielded the sensors from direct sunlight. Wall and road distances and compass orientation were not standardized. Temperature and relative humidity were recorded every 2 min, after inter-instrument recording consistency was assessed. The network and part of the observations originated from earlier work by our group.

Sites were selected in three deployment batches using a preliminary summer Landsat LST map, LCZ zoning and the prevalence of land-use types. The first batch emphasized residential compounds in different LCZs within the ring road. Locations were distributed across each compound to broaden the spatial coverage and limit redundant placement. Two subsequent batches added other land-use and LCZ types. This purposive design sampled contrasting urban environments.

LST was obtained from the ECOSTRESS Tiled Land Surface Temperature and Emissivity product (ECO_L2T_LSTE, Version 002; NASA LP DAAC) [25,26]. Source tiles use a 70 m grid in WGS 84/UTM zone 49N (EPSG:32649), with floating-point values, scale factor 1 and offset 0. Kelvin values were converted to degrees Celsius by subtracting 273.15. Tiles were mosaicked and bilinearly resampled to a 30 m grid for spatial extraction. Quality control and cloud screening were completed before matching. Scenes with extensive cloud cover, incomplete study-area coverage or missing LST were excluded. Six overpasses in July–August 2024 and 2025 were retained.

2.2.2. Temporal Matching and Sample Composition

Satellite overpass times and logger timestamps were expressed in Beijing time (UTC+8). For each site and retained overpass, Ta was extracted from the 2 min logger series at the observation time nearest to the satellite overpass. No temporal averaging or interpolation was applied. Matching offsets were summarized relative to the recorded overpass window: observations within the start–end interval had zero offset, and observations outside it were assigned their distance to the nearer boundary.

The matched dataset contained 298 site–overpass records from 179 monitoring sites across six events. Of these, 158 observations fell within the recorded overpass windows; all remaining window offsets were at most 40 s. Restricting the sample to the four LCZ categories yielded 176 modeling records from 123 sites: 107 in LCZ5-II, 23 in LCZ4, 18 in LCZ2 and 28 in LCZ6-II. The other 122 records were excluded from modeling.

Figure 3a lists the six overpass dates and the number of matched records for each event. Figure 3b shows event-specific means and standard deviations (SDs) of the LST and concurrent Ta for the four LCZs combined in the main analysis. Figure 3c shows the distribution of time offsets for all 298 records.

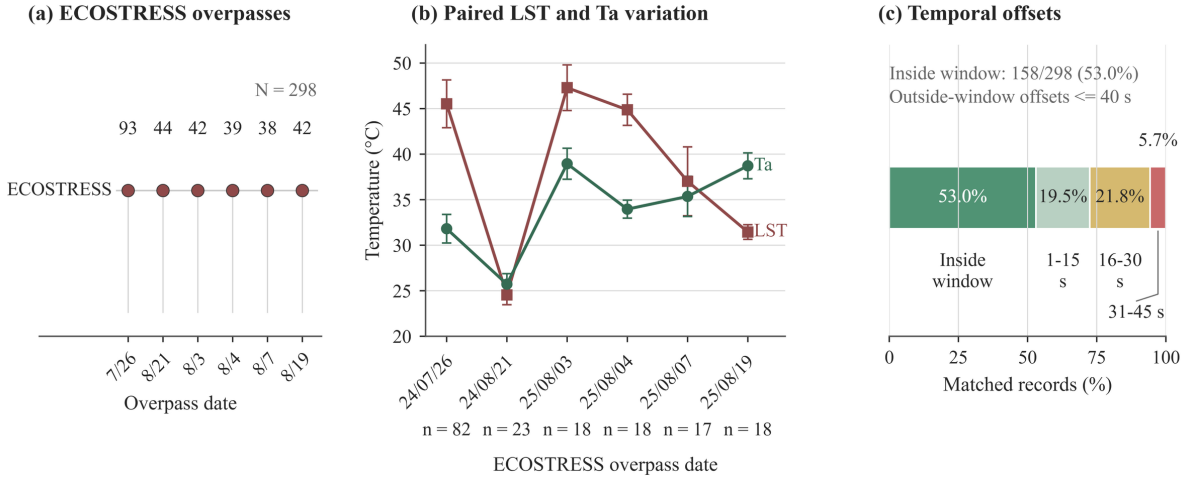


Figure 3. ECOSTRESS observations and temporal matching: (**a**) ECOSTRESS overpass dates and corresponding record counts; (**b**) variation in paired LST and Ta across overpass dates, with error bars indicating standard deviations and n indicating the number of paired records; (**c**) proportions of records within the matching window and in each outside-window temporal offset category. In panel (**a**), the red circles indicate ECOSTRESS overpasses. In panel (**b**), red squares and green circles represent LST and Ta, respectively. In panel (**c**), dark green indicates records within the matching window, while light green, yellow, and red indicate outside-window temporal offsets of 1–15 s, 16–30 s, and 31–45 s, respectively.

2.3. Built-Environment Indicators and Collinearity Assessment

2.3.1. Built-Environment Indicators and Extraction Extent

The BE indicators were classified into five groups according to their spatial meaning: surface cover, building footprint pattern, building vertical profile, blue–green cover, and urban activity intensity. The initial set contained 19 candidate variables. Their full names, abbreviations, calculations, physical interpretations, and consolidated data sources are provided in Table A5.

The LST and the BE indicators were summarized within a 50 m radius around each monitoring site. This neighborhood was motivated by previous studies [27,28] and local building dimensions: 86.5% of valid footprints were no longer than 50 m, and 87.3% were no wider than 20 m. The resulting 100 m diameter defines the extraction extent; the source-product and processing grids are 70 m and 30 m, respectively. Differences in physical support among the LST, BE and point-measured Ta are discussed in Section 4.4.

The 123 modeling site coordinates were compared with the original tiled product grids. All covering tiles yielded 270 site–tile associations, including 105 sites within overlapping tile extents. Site-to-containing-cell-center distances ranged from 2.37 to 46.88 m. Each 50 m radius buffer intersected 4–7 cells of the corresponding 70 m grid, and its containing cell covered 27.41–62.35% of the buffer area. These measures describe the delivered grid geometry. Tile overlap and mosaicking prevent assigning a unique source cell from the site coordinate alone.

2.3.2. Pearson Correlation and Variance Inflation Factor Assessment

Pearson correlation analysis and variance inflation factor (VIF) assessment were both based on 123 independent sites across the four LCZs. The BE attributes were constant across

repeated records of each site in the analysis table, so one spatial attribute record was retained for each site. Pearson correlations were calculated using pairwise complete observations.

VIF screening used a threshold of 5 [29] on the combined set of 123 independent sites and removed MBV, AWH, SDH, BD and Hmax. The retained variables were MBS, BVD, BSI, MH, SVF, NDVI, NHV, MNDWI, ISF, BVCV, NTL, MRD, RG and FD, with VIFs of 1.12–4.67. These 14 variables defined the fixed BE input for subsequent comparisons. Screening used all sites rather than training folds alone. Figure 4 reports the Pearson correlations and VIF values.



Figure 4. Pearson correlations and variance inflation factor (VIF) screening of built-environment indicators at the 50 m scale: (**a**) Pearson correlation matrix; (**b**) stepwise VIF screening, with green bars indicating retained variables and red hatched bars indicating removed variables. The dashed line indicates the VIF threshold of 5.

2.4. Machine-Learning Models and Evaluation

2.4.1. Candidate Algorithms and Selection Criterion

Algorithms were compared separately within each of the four LCZs. The prediction target was Ta matched to the ECOSTRESS overpass time, and the concurrent LST was the predictor. All six algorithms used the same inputs and cross-validation splits, so algorithm selection was not affected by differences in the BE variables.

The nonlinear comparison included SVR [30], random forest [31], ExtraTrees [32], XGBoost [33], LightGBM [34] and CatBoost [35], with settings in Table A2. Algorithms were compared by the equal-weight mean RMSE across LCZs, alongside MAE, predictive R^2 and prediction–observation Pearson r. CatBoost provided the common model for subsequent input comparisons, using 140 iterations, depth 3, learning rate 0.045, L2 leaf regularization 8 and RMSE loss. Training-mean Ta and OLS regression of Ta on LST with an intercept served as simple benchmarks. Both were fitted within each training fold and evaluated on the same held-out records as CatBoost.

2.4.2. Cross-Validation and Performance Metrics

The evaluation used ten repetitions of site-grouped five-fold cross-validation, keeping all records from each site together [24]. Within each repetition, predictions from the five test folds were pooled before calculating the metrics. The means and SDs were then calculated across repetitions. SDs quantify the sensitivity to site partitions. Sites from the same event, and nearby sites with shared conditions, could occur in both the training and test sets.

RMSE, MAE, predictive R^2 and Pearson r were calculated from pooled out-of-fold predictions within each repetition. Predictive R^2 uses the mean of the test observations as its reference; negative values indicate a squared error exceeding the observed total sum of squares. The training-mean benchmark instead uses the training-fold Ta mean to predict each held-out record. Prediction–observation r describes the predictions, while the raw LST–Ta correlation describes the paired measurements.

2.4.3. Three Ta Prediction Models

Within each LCZ, fixed CatBoost settings were used for LST alone, LST plus the 14 globally screened BE variables, and LST plus training-fold SHAP-ranked Top-k BE variables. All configurations used identical site-grouped partitions, alongside the training-mean and OLS benchmarks. Predictor ranking and k selection follow Section 2.4.4.

2.4.4. Model-Associated Contributions and Reduced Predictor Selection

Model-associated variable contributions were analyzed using SHapley Additive exPlanations (SHAP) [36]. During each cross-validation run, CatBoost was fitted to the training data using LST and all 14 BE variables. The BE variables were then ranked by mean absolute SHAP value. The top k variables were combined with LST for k = 1–14, and out-of-fold performance was calculated on the corresponding validation data.

Within each LCZ, the smallest k was selected for which R^2 decreased by at most 0.01 and RMSE increased by at most 0.10 °C relative to BE14. SHAP ranking used training folds, but k was chosen from the aggregated validation results. The resulting Top-k metrics are exploratory and may contain selection bias. LCZ2 SHAP rankings were excluded from environmental interpretation because its models failed validation. The CatBoost and SHAP analyses were conducted using Python 3.12.14, CatBoost 1.2.10, and SHAP 0.52.0.

2.4.5. Supplementary Validation and Descriptive Diagnostics

Each event was held out in turn within each LCZ. A second protocol also excluded all test-site identities from the remaining training events. Eligible folds required at least

five training records from three sites and two distinct LST values. This computational guard was applied to all models; ineligible folds were reported as unavailable. Training-mean, OLS-LST and CatBoost LST-only/BE14/BE19 models used identical tests. CatBoost retained its original settings with seed 42. Table A6 reports the fold support and site overlap; Table A7 reports the errors pooled over eligible records.

Capacity sensitivity reused the ten saved repetitions of five site-grouped folds. Three fixed settings were evaluated: depth 3/L2 8 (original), depth 1/L2 8 and depth 3/L2 20. All used 140 iterations, learning rate 0.045, RMSE loss and seed 42 + 101 × (repeat − 1) + fold. Inputs were LST-only, LST + BE14 and LST + all 19 candidate BE variables. Predictors were complete in the modeling sample. The training and test RMSEs pooled the corresponding five folds within each repetition (Table A8). A separate 200-fit comparison added 12 building footprint, height and volume indicators to the LST under the original settings (Morph12; Table A9). The RMSE gain was calculated as the mean LST-only RMSE minus the augmented-model mean RMSE; percentage gain used the LST-only mean as the denominator. Positive gain indicates lower error. All input comparisons used identical held-out records.

BE distributions were summarized using one record per site (123 sites); repeated records had identical BE values in the finalized table. For each indicator, the mean, sample SD and range were calculated by LCZ. Descriptive eta-squared was the between-LCZ sum of squares divided by the total sum of squares across sites; no significance or causal interpretation was assigned (Appendix G). For LCZ2, prediction–observation covariance was decomposed separately within each repetition into record-weighted within-event covariance and covariance between event means, both using denominator N = 18. A singleton event contributes zero within-event covariance but can contribute to the between-event term. Appendix H reports the event means and error shares.

3. Results

3.1. LST-Ta Differences and Ta Prediction Across LCZs

3.1.1. Direct Differences Between LST and Ta

Figure 5 summarizes the 176 paired observations. The mean LST exceeded the mean Ta in all four LCZ groups. The mean ± record-level SD of LST − Ta was 7.79 ± 8.02, 8.65 ± 4.88, 11.98 ± 7.08 and 3.33 ± 7.28 °C for LCZ5-II, LCZ4, LCZ2 and LCZ6-II, respectively. These distributions include repeated sites across events, and group differences may partly reflect event composition.

Most observations in Figure 5a fall below the 1:1 line, indicating a systematic difference between LST and Ta and the need for empirical conversion. Changes across overpass dates were also asynchronous. As shown in Figure 3b, from 7 to 19 August 2025, the mean Ta across the four LCZs increased from 35.36 to 38.72 °C, whereas the mean LST decreased from 37.02 to 31.45 °C.

3.1.2. Comparison of Candidate Machine-Learning Algorithms

Figure 6 compares the equal-weight mean RMSE of six algorithms across the four LCZs. Points and error bars show the mean and SD from 10 repetitions of site-grouped five-fold cross-validation. CatBoost and ExtraTrees yielded RMSEs of 2.74 ± 0.04 and 2.78 ± 0.16 °C, respectively, a difference of 0.04 °C. This difference was smaller than the variation across repeated splits and negligible for practical temperature applications. ExtraTrees nevertheless produced a higher R^2 than CatBoost (0.331 vs. 0.236), so the two metrics ranked these algorithms differently. The RMSEs for the other four algorithms ranged from 2.86 to 3.39 °C. The complete results are provided in Table A1.

CatBoost was retained for consistent input comparisons. Its predictions were also evaluated against training-mean and OLS benchmarks.

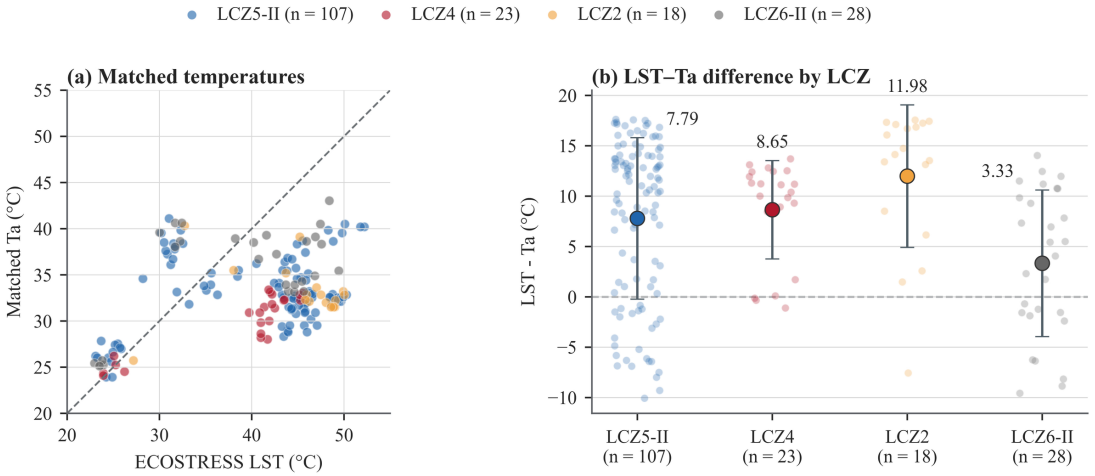


Figure 5. Matched LST–Ta patterns across LCZs: (**a**) paired ECOSTRESS LST and measured Ta, with the dashed line indicating the 1:1 relationship; (**b**) distributions of LST minus Ta by LCZ. Small points represent individual observations, large circles indicate group means, and error bars indicate ±1 standard deviation. Numerical annotations show the group means, and the dashed horizontal line indicates zero difference.

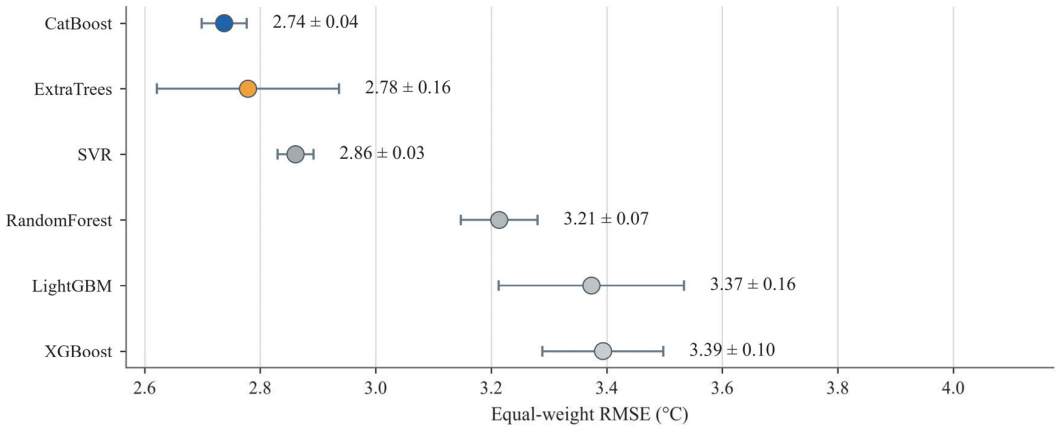


Figure 6. Cross-validated comparison of candidate algorithms. Blue, orange, and gray markers represent CatBoost, ExtraTrees, and the remaining candidate algorithms, respectively.

3.1.3. Ta Prediction Performance Using LST Alone

Table 1 reports the LST-only CatBoost results and observational support. The mean RMSE was 1.63 °C in LCZ4, 2.22 °C in LCZ6-II and 2.69 °C in LCZ5-II, with predictive R^2 values of 0.714, 0.821 and 0.503, respectively.

LCZ2 had an RMSE of 4.41 °C, predictive R^2 of −1.092 and prediction–observation r of −0.818. The raw LST–Ta r was −0.027 across all 18 records and −0.190 within the 13-record event on 26 July 2024. The dominant event contributed 2.98% of the squared prediction error across ten repetitions, while the single records on 21 August 2024 and 19 August 2025 together contributed 76.73%. Covariance decomposition located the inverse prediction association between events: the mean total covariance was −3.79 °C^2, comprising −3.82 °C^2 between events and 0.03 °C^2 within events. The between-event component was negative in every repetition (Appendix H).

OLS had a lower RMSE than LST-only CatBoost in LCZ4 (1.41 vs. 1.63 °C), while CatBoost had lower RMSEs in LCZ5-II (2.69 vs. 3.75 °C) and LCZ6-II (2.22 vs. 4.60 °C). In LCZ2, the training-mean benchmark yielded 3.62 °C, below all three original CatBoost configurations (Table 2). The Morph12 and regularization comparisons are reported separately in Appendix F.

Table 1. Performance of LCZ-specific Ta~LST models based on 10 repeated site-grouped five-fold cross-validation runs. All metrics are means across repetitions; the reported ± value is the between-repetition SD of RMSE.

LCZ	Records	Sites	Events	RMSE, Mean ± SD (°C)	MAE (°C)	R^2	Pearson r
LCZ5-II	107	72	6	2.69 ± 0.02	2.08	0.503	0.712
LCZ4	23	23	2	1.63 ± 0.10	1.33	0.714	0.851
LCZ2	18	15	6	4.41 ± 0.03	2.61	−1.092	−0.818
LCZ6-II	28	13	6	2.22 ± 0.14	1.93	0.821	0.922

Table 2. Simple benchmarks and LST-only CatBoost under identical site-grouped partitions. RMSE is reported as mean ± split SD (°C).

LCZ	Training Mean RMSE (°C)	OLS (LST) RMSE (°C)	CatBoost (LST) RMSE (°C)	Sites/Events
LCZ5-II	3.86 ± 0.03	3.75 ± 0.04	2.69 ± 0.02	72/6
LCZ4	3.23 ± 0.09	1.41 ± 0.04	1.63 ± 0.10	23/2
LCZ2	3.62 ± 0.07	5.63 ± 0.07	4.41 ± 0.03	15/6
LCZ6-II	5.61 ± 0.19	4.60 ± 0.19	2.22 ± 0.14	13/6

Benchmarks used training-fold data only. Metrics pool each repetition's five validation folds; SDs summarize variation across the ten partitions.

3.2. Performance of the Three Ta Models Across LCZs

Figure 7 and Table A4 compare the original CatBoost input configurations. The equal-weight mean RMSE across LCZs was 2.74 ± 0.04 °C for LST-only, 3.29 ± 0.10 °C for BE14 and 3.19 ± 0.19 °C for the retrospectively selected Top-k models. BE14 increased the mean MAE from 1.99 to 2.53 °C and reduced the mean LCZ-specific R^2 and r. These are averages across LCZs; Top-k selection follows Section 2.4.4.

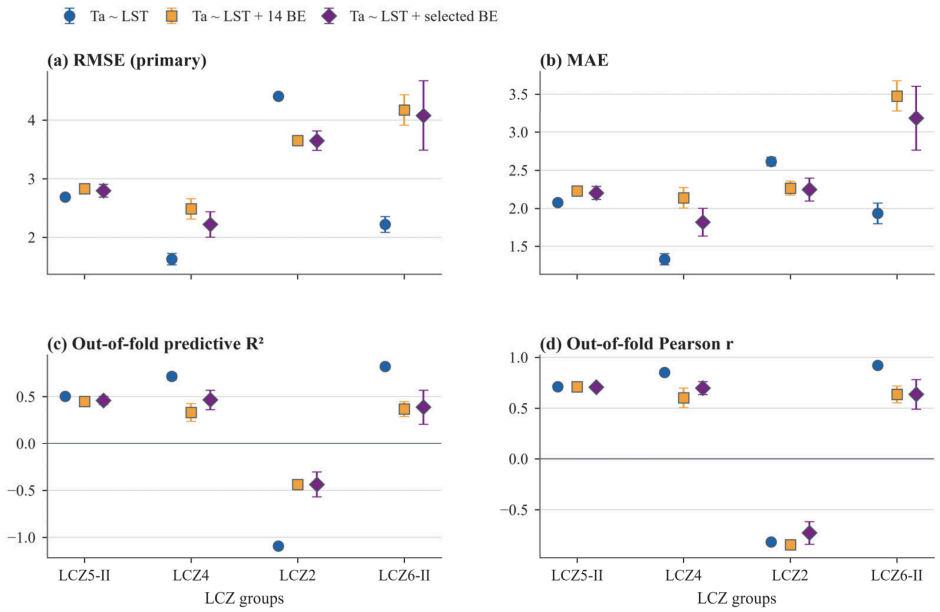


Figure 7. CatBoost input comparisons under repeated site-grouped validation: (**a**) RMSE; (**b**) MAE; (**c**) out-of-fold predictive R^2; (**d**) Pearson correlation coefficient between out-of-fold predictions and observations. Blue circles, orange squares, and purple diamonds represent Ta ~ LST, Ta ~ LST + 14 BE, and Ta ~ LST + selected BE, respectively. Selected-BE configurations use k chosen from the same cross-validation results. Error bars show between-partition SD.

Relative to LST-only CatBoost, selected-k configurations increased the RMSE by 0.11, 0.59 and 1.86 °C in LCZ5-II, LCZ4 and LCZ6-II. In LCZ2, the RMSE decreased from 4.41 to 3.65 °C but remained above the training-mean benchmark of 3.62 °C, with negative R^2 and prediction–observation r.

Figure 7 reports the RMSE, MAE, predictive R^2 and prediction–observation r as means across ten repetitions, with error bars indicating split SD.

3.3. SHAP Patterns Across LCZs

Figure 8 and Table A3 summarize the full-data SHAP rankings for LCZ5-II, LCZ4 and LCZ6-II. LCZ2 is excluded because its models failed predictive validation. These rankings describe variable contributions to the fitted models.

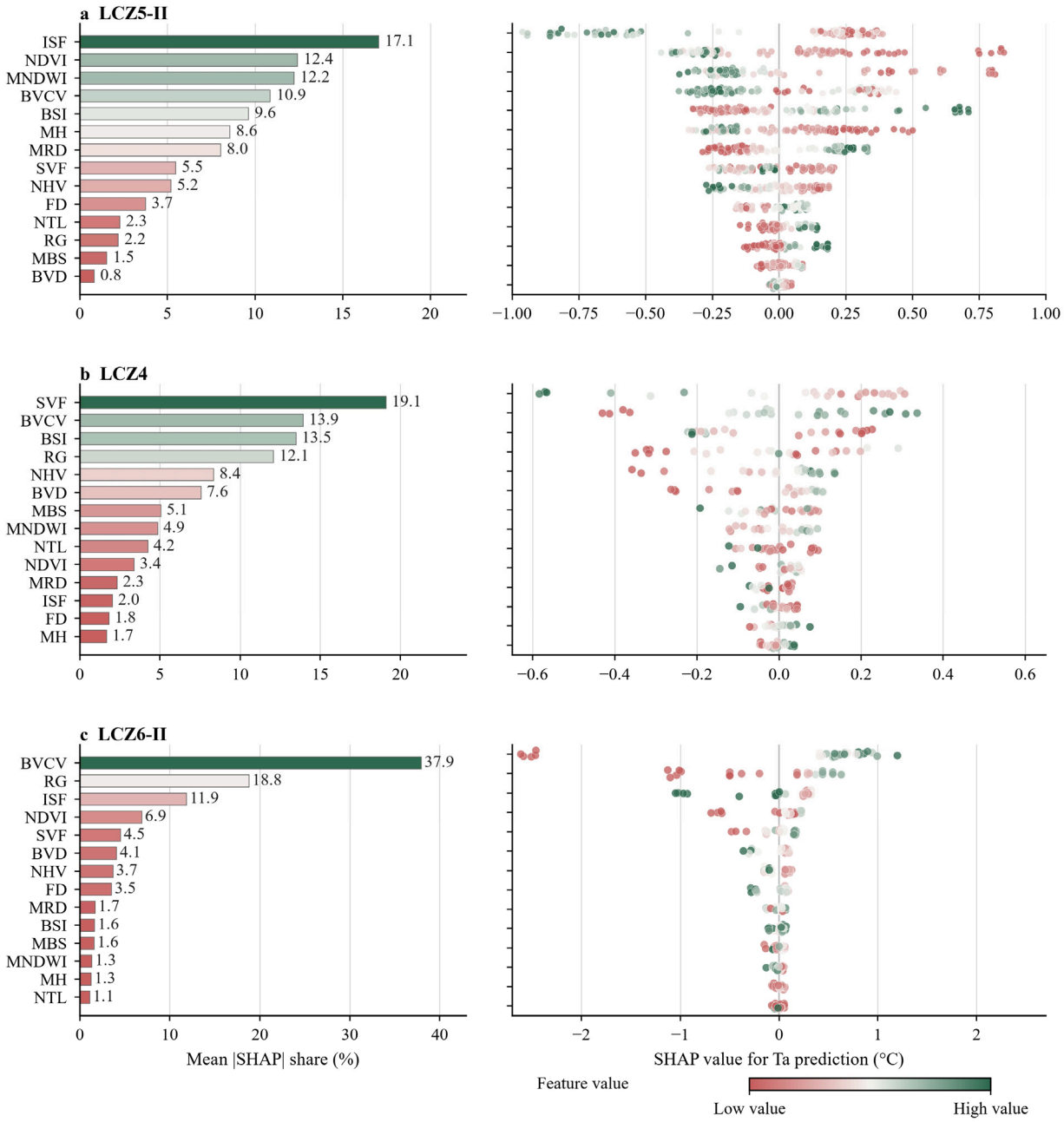


Figure 8. Full-data BE SHAP patterns: (**a**) LCZ5-II; (**b**) LCZ4; (**c**) LCZ6-II. For each LCZ, the left panel shows the percentage shares of mean absolute SHAP values, and the right panel shows the distributions of individual SHAP values. LCZ2 is excluded because predictive validation failed. BE percentage shares exclude LST.

Bars report each variable's share of the mean absolute SHAP values summed over BE14, excluding LST. Points show the record-level SHAP values in temperature units, colored by the normalized predictor value. Correlated inputs and limited site/event coverage can affect the stability of these attributions.

3.4. Exploratory Predictor-Count Sensitivity

Figure A1 shows the Top-k curves with shared RMSE axes and LST-only/BE14 references. Applying the tolerances in Section 2.4.4 selected k = 5, 1, 2 and 1 for LCZ5-II, LCZ4, LCZ2 and LCZ6-II, respectively. These retrospective counts summarize the sensitivity within the observed sample.

Reducing the BE input count maintained or modestly improved the error relative to BE14 but did not outperform LST-only CatBoost in LCZ5-II, LCZ4 or LCZ6-II. LCZ2 retained a negative predictive R^2 and an RMSE above the training-mean benchmark (Table A4).

3.5. Supplementary Robustness and BE Distributions

Holding events out increased the LST-only CatBoost RMSE to 5.84 °C in LCZ5-II, 6.57 °C in LCZ4 and 7.12 °C in LCZ6-II, compared with 2.69, 1.63 and 2.22 °C under site-grouped validation. Joint event/site exclusion increased the first and third values further to 6.53 and 9.25 °C; LCZ4 had no repeated sites across its two events, so its two protocols were identical. OLS retained a lower event-held-out RMSE in LCZ4 (2.30 °C), although each fit was supported by only one training event. All CatBoost event-held-out R^2 values were negative in these three LCZs. The dominant LCZ2 event was unavailable, so its aggregate event-test result covered only five singleton records; four of these involved the same site. BE14/BE19 had a positive R^2 on that restricted event-only subset but a negative R^2 after test-site exclusion. Appendix E reports the eligible tests.

BE14 and BE19 increased the test RMSE in LCZ5-II, LCZ4 and LCZ6-II under all three capacity settings (Table A8). In the original setting, BE14 reduced the training RMSE from 2.33 to 1.75 °C in LCZ5-II and from 1.03 to 0.81 °C in LCZ4, while the test RMSE increased from 2.69 to 2.83 °C and from 1.63 to 2.49 °C. Stronger L2 regularization reduced the LCZ2 BE14/BE19 RMSE to about 3.55 °C, below the training-mean benchmark of 3.62 °C, but the R^2 remained about −0.356. Morph12 increased the RMSE by 0.14, 0.82 and 1.48 °C in LCZ5-II, LCZ4 and LCZ6-II (5.36%, 50.13% and 66.38%). In LCZ2, it reduced the RMSE by 0.82 °C (18.56%) to 3.59 °C, with $R^2 = -0.388$. The equal-weight LCZ mean rose from 2.74 to 3.14 °C for Morph12, compared with 3.29 °C for BE14 and 3.34 °C for BE19 (Table A9).

Across 123 unique sites, descriptive eta-squared ranged from 0.007 (RG) to 0.233 (BSI), with values of 0.176 for ISF, 0.162 for NDVI and 0.155 for SVF (Table A10). Thus, between-LCZ mean differences accounted for at most 23.3% of any indicator's total sum of squares; substantial variation remained within categories.

4. Discussion

4.1. Conditional LST–Ta Prediction and LCZ Differences

The LST and Ta means changed in different directions between some overpasses, consistent with their distinct spatial and temporal behavior [4–8]. Event-specific site composition also differed, preventing attribution of these contrasts to weather or urban form alone. Close temporal matching leaves differences in spatial support and unmeasured meteorological conditions as potential sources of prediction error.

LCZ stratification separates urban-form categories for local diagnosis [12–16]. OLS outperformed LST-only CatBoost in LCZ4, whereas CatBoost had lower RMSEs in LCZ5-II and LCZ6-II. These differences indicate that model complexity has different values

across the sampled groups, although uneven site and event coverage prevents isolating an LCZ effect.

Uneven event support helps explain the LCZ2 prediction failure. Thirteen of its eighteen records came from one overpass, while four of the other five records came from the same site. Covariance decomposition showed that the inverse prediction association arose mainly between event means. The mean LST/Ta was 48.04/32.29 °C in the dominant event, compared with 32.70/40.30 °C in the single record on 19 August 2025. Radiation, shading, heat storage and ventilation can alter these relationships [10,37], but site-level wind and cloud observations were not collected, preventing attribution to a specific process.

All three original LCZ2 CatBoost configurations exceeded the training-mean RMSE. Morph12 and stronger regularization reduced the error slightly below that benchmark, but the predictive R^2 remained negative. Thus, the RMSE improvement offered by additional inputs was insufficient to produce useful LCZ2 predictions.

Li et al. [23] reported test RMSEs of 3.03 °C by day and 2.64 °C by night using MODIS LST and 29 ancillary factors at 292 Chinese stations, with training in 2017–2020 and testing in 2021. Their SHAP results highlighted meteorological background, terrain, urban morphology and anthropogenic factors. Colaninno and Morello [7] obtained RMSEs of 1.45–1.77 °C using MODIS, station observations, elevation and geographically weighted regression over two consecutive heatwave days. These studies cover different spatial supports and validation targets from our LCZ-specific ECOSTRESS tests. ECOSTRESS research in New York City by Hu and Uejio [38] instead estimated ground radiant temperature and its morphological controls; its target differs from near-surface Ta. Across these studies, morphology can describe thermal variation without guaranteeing lower held-out Ta error after LST is included. Here, the direct input comparisons and the rise in error under event-held-out testing show that the value of additional predictors depends on both observational support and the prediction target.

4.2. Incremental BE Information and Model Complexity

The Morph12 comparison isolates the predictive contribution of building geometry from vegetation, water, impervious cover and activity indicators. Adding those 12 variables increased the error in LCZ5-II, LCZ4 and LCZ6-II, as did the broader BE14 and BE19 sets. LCZ2 showed an error reduction but remained poorly predicted. Physical associations documented in earlier studies [9,18–23] therefore do not translate automatically into useful additional information for these small LCZ-specific models.

More predictors offer more candidate tree splits and can increase sample-specific fitting. In LCZ5-II and LCZ4, BE14 reduced the training error while increasing the test error. Shallower trees and stronger regularization did not reverse the BE disadvantage in LCZ5-II, LCZ4 or LCZ6-II, suggesting limited incremental signal relative to model flexibility. Pearson correlation and VIF characterize linear redundancy; additional nonlinear dependencies may affect tree-based selection and attribution.

Between-LCZ differences explained at most 23.3% of the site-level variance in any BE indicator (Appendix G). LCZ categories and continuous BE variables thus contain overlapping information while retaining substantial within-category variation. The input comparisons show that this residual variation did not consistently improve Ta prediction beyond LST.

The SHAP rankings in LCZ5-II, LCZ4 and LCZ6-II suggest hypotheses about cover and neighborhood-scale effects [17,27,28]. Testing these relationships on additional sites and events would help distinguish stable patterns from sample-specific model attributions [36].

4.3. Predictor Reduction and Application Boundaries

Top-k reduction retained or modestly improved BE14 performance. However, choosing k from the reported validation results introduces selection bias. Independent evaluation requires an inner selection loop and untouched outer test folds; the present curves describe predictor-count sensitivity (Figure A1).

Studies of green-space configuration and blue–green cooling [27,39] provide hypotheses for extending the predictor comparisons. For citywide heat-exposure mapping and planning, these comparisons should be evaluated under the spatial and temporal conditions of the intended application.

4.4. Limitations and Future Work

Site grouping prevented repeated records from crossing training and test folds [24], but nearby sites could share weather and source pixels. The event-held-out errors were substantially higher, and excluding test-site identities further increased the LST-only error in LCZ5-II and LCZ6-II. The dominant LCZ2 event could not be evaluated under the training-support rule, and LCZ4 provided only two event tests. More seasons, overpasses and spatially separated sites are needed to evaluate transferability.

Spatial support also limits interpretation. The 50 m buffers intersected multiple 70 m source cells, and many sites lay within overlapping tile extents. Bilinear resampling to 30 m cannot resolve subpixel thermal composition. Mixed surfaces, geolocation uncertainty and shared source information can therefore affect LST–Ta pairing. The circular extraction buffer also differs from the atmospheric source area of point-measured Ta; alternative neighborhood scales require evaluation [17].

Purposive site selection and nonstandardized exposure conditions limit citywide representativeness. Concurrent wind and cloud measurements around the monitoring sites were unavailable, so ventilation and radiative mechanisms remain unresolved. BE14 also used full-sample screening, and Top-k selection reused validation outputs. Future input selection should occur entirely within training folds and should be assessed on independent sites and events.

5. Conclusions

This exploratory study evaluated ECOSTRESS-based Ta prediction using 176 records from 123 sites and six overpasses in four LCZ categories of central Xi'an. Under site-grouped validation, LST-only CatBoost outperformed OLS in LCZ5-II and LCZ6-II, while OLS performed better in LCZ4. The event-held-out errors were substantially larger, limiting transfer across observation conditions.

Adding BE14, BE19 or the 12 building morphology indicators increased the original-setting RMSE in LCZ5-II, LCZ4 and LCZ6-II. The BE14/BE19 disadvantage persisted under all three capacity settings. LCZ2 benefited from additional inputs in RMSE, but all tested site-grouped CatBoost configurations retained negative predictive R^2 values. Its inverse prediction association was concentrated between event means, and its SHAP rankings were excluded from environmental interpretation. The Top-k results remain exploratory because k was chosen from the reported validation outputs.

These results support comparing complex models with simple benchmarks and testing the incremental value of local predictors under the intended spatial and temporal validation target. Broader site and event coverage is needed for citywide mapping and heat-exposure assessment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings16214262/s1, Supplementary File S1: Public supple-

mentary analysis workbook containing pooled event-validation metrics, capacity/input sensitivity summaries, paired RMSE gain statistics, and aggregated built-environment summaries.

Author Contributions: Conceptualization, D.Z., K.L. and D.X.; Methodology, M.Z. and Z.G.; Software, M.Z.; Validation, M.Z.; Formal Analysis, M.Z.; Investigation, M.Z. and T.Y.; Resources, J.S.; Data Curation, J.S.; Writing—Original Draft Preparation, M.Z.; Writing—Review and Editing, D.Z., K.L., D.X., X.L. and S.M.; Visualization, M.Z., X.L. and S.M.; Supervision, D.Z., K.L., Z.G. and D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [grant number 52300236], the Xi'an Municipal Social Science Planning Fund Project [grant number Y26FZ27], and the Fundamental Research Funds for the Central Universities of Xi'an Jiaotong University [grant number xzy012026039].

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable privacy and site-access restrictions. The data are not publicly available because they contain site-specific information from residential communities and privately owned industrial premises where field measurements were conducted with permission, but public disclosure of the associated site information is restricted.

Conflicts of Interest: Author Xin Lin was employed by the company State Grid Shanghai Electric Power Engineering Construction Consulting Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Nonlinear Model Comparison Metrics

Table A1. Equal-weight performance of six nonlinear algorithms fitted separately within each LCZ using LST.

Model	RMSE, Mean ± SD (°C)	MAE (°C)	R^2	Pearson r
CatBoost	2.74 ± 0.04	1.99	0.236	0.417
ExtraTrees	2.78 ± 0.16	2.17	0.331	0.417
SVR	2.86 ± 0.03	2.00	0.197	0.422
Random forest	3.21 ± 0.07	2.43	0.183	0.348
LightGBM	3.37 ± 0.16	2.54	0.116	0.320
XGBoost	3.39 ± 0.10	2.30	−0.087	0.333

Table A2. Hyperparameters used to compare the six nonlinear algorithms.

Model	Fixed Settings Used in Model Screening
SVR	RBF kernel; C = 10; epsilon = 0.1; gamma = scale; standardized LST.
Random forest	120 trees; maximum depth = 5; minimum samples per leaf = 4.
ExtraTrees	120 trees; maximum depth = 5; minimum samples per leaf = 3.
XGBoost	120 trees; maximum depth = 3; learning rate = 0.045; subsample = 0.85; column subsample = 0.85; L2 regularization = 6.
LightGBM	120 trees; maximum depth = 3; 15 leaves; learning rate = 0.045; minimum child samples = 6; L2 regularization = 6.
CatBoost	140 iterations; depth = 3; learning rate = 0.045; L2 leaf regularization = 8; RMSE loss.

Appendix B. Descriptive SHAP Rankings and Exploratory Input Comparisons

Table A3 summarizes the full-data BE rankings for the three LCZs with positive predictive R^2. Table A4 reports the original input comparisons; selected-k rows are exploratory because k was chosen after validation.

Table A3. Descriptive top-five BE rankings from full-data LST + 14 BE fits. Shares are normalized over BE variables only and exclude LST. LCZ2 is excluded from ranking interpretation because validation failed.

LCZ	Top Five Indicators (Mean |SHAP| Share)	Ranking Scope
LCZ5-II	ISF (17.1%); NDVI (12.4%); MNDWI (12.2%); BVCV (10.9%); BSI (9.6%)	Full-data LST + 14-variable model
LCZ4	SVF (19.1%); BVCV (13.9%); BSI (13.5%); RG (12.1%); NHV (8.4%)	Full-data LST + 14-variable model
LCZ6-II	BVCV (37.9%); RG (18.8%); ISF (11.9%); NDVI (6.9%); SVF (4.5%)	Full-data LST + 14-variable model

Table A4. Consolidated CatBoost input configurations and repeated site-grouped validation metrics. Top-k inputs were re-ranked within training folds, but k was chosen from the aggregated results; these rows are exploratory.

LCZ	Scheme	Inputs	RMSE ± SD (°C)	MAE (°C)	R^2	r
LCZ5-II	LST only	LST	2.69 ± 0.02	2.08	0.503	0.712
LCZ5-II	BE14	LST + 14 globally screened BE variables	2.83 ± 0.08	2.23	0.449	0.712
LCZ5-II	Top-5 (exploratory)	LST + training-fold Top-5; k selected after CV	2.80 ± 0.11	2.20	0.461	0.707
LCZ4	LST only	LST	1.63 ± 0.10	1.33	0.714	0.851
LCZ4	BE14	LST + 14 globally screened BE variables	2.49 ± 0.17	2.14	0.333	0.603
LCZ4	Top-1 (exploratory)	LST + training-fold Top-1; k selected after CV	2.22 ± 0.22	1.82	0.466	0.699
LCZ2	LST only	LST	4.41 ± 0.03	2.61	−1.092	−0.818
LCZ2	BE14	LST + 14 globally screened BE variables	3.65 ± 0.06	2.27	−0.436	−0.845
LCZ2	Top-2 (exploratory)	LST + training-fold Top-2; k selected after CV	3.65 ± 0.17	2.25	−0.436	−0.728
LCZ6-II	LST only	LST	2.22 ± 0.14	1.93	0.821	0.922
LCZ6-II	BE14	LST + 14 globally screened BE variables	4.17 ± 0.26	3.48	0.368	0.637
LCZ6-II	Top-1 (exploratory)	LST + training-fold Top-1; k selected after CV	4.08 ± 0.59	3.18	0.388	0.637

All configurations used 140 iterations, depth 3, learning rate 0.045, L2 leaf regularization 8 and RMSE loss, with ten repeated site-grouped five-fold partitions. The 14 globally screened BE variables were MBS, BVD, BSI, MH, SVF, NDVI, NHV, MNDWI, ISF, BVCV, NTL, MRD, RG and FD. Reported RMSE ± SD summarizes the mean and between-partition SD. Top-k variable identities vary across training folds; sensitivity curves appear in Figure A1.

Appendix C. Built-Environment Indicators and Data Sources

Table A5. Built-environment indicators, definitions, and consolidated data sources.

Category	Indicator	Formula	Definition	Data Source
Surface cover	Impervious surface fraction (ISF)	$ISF = \frac{A_{impervious}}{A}$	Fraction of impervious surface within the buffer.	Sentinel-2 Level-2A (L2A) surface reflectance via Google Earth Engine
Footprint pattern	Building density (BD)	$BD = \frac{\sum_{i=1}^{n} A_i}{A}$	Plan-area fraction covered by building footprints.	Building footprint and height attributes from Baidu Maps and OpenStreetMap (OSM) 3D building data
	Fractal dimension (FD)	$FD = \frac{1}{n}\sum_{i=1}^{n} \frac{2\ln(\frac{P_i}{4})}{\ln(A_i)}$	Plan-form complexity of building footprints; larger values indicate more complex outlines.	
Vertical profile	Mean building height (MH)	$MH = \frac{\sum_{i=1}^{n} H_i}{n}$	Mean vertical height of buildings within the buffer.	
	Maximum building height (Hmax)	$H_{\max} = \max_{1 \le i \le n} H_i$	Maximum building height within the buffer.	
	Standard deviation of height (SDH)	$SDH = \sqrt{\frac{\sum_{i=1}^{n}(H_i - \overline{H})^2}{n}}$	Dispersion of building heights.	
	Normalized height variation (NHV)	$NHV = \frac{SDH}{\overline{H}}$	Relative variation in building height normalized by mean height.	
	Area-weighted height (AWH)	$AWH = \frac{\sum_{i=1}^{n} A_i \cdot H_i}{\sum_{i=1}^{n} A_i}$	Mean building height weighted by footprint area.	
	Building volume coefficient of variation (BVCV)	$BVCV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \overline{v})^2}}{\overline{v}}$	Relative dispersion of individual building volumes.	
	Mean building surface area (MBS)	$MBS = \frac{1}{n}\sum_{i=1}^{n}(H_i \cdot P_i + A_i)$	Mean exterior surface area of individual buildings.	
	Mean building volume (MBV)	$MBV = \frac{1}{n}\sum_{i=1}^{n}(A_i \cdot H_i)$	Mean volume of individual buildings.	
	Building structure index (BSI)	$BSI = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{H_i}$	Relationship between building footprint area and height.	
	Building volume density (BVD)	$BVD = \frac{\sum_{i=1}^{n} V_i}{A}$	Total building volume per unit buffer area.	
	Sky view factor (SVF)	$SVF = \frac{1}{n}\sum_{i=1}^{n} cos^2\beta_i$	Fraction of the sky hemisphere visible from the observation surface.	TerraSAR-X add-on for Digital Elevation Measurement (TanDEM-X) digital elevation model (DEM), canopy-height data, and building height data
Blue–green cover	Green-space ratio (RG)	$RG = \frac{\sum_{i=1}^{n} G_i}{A}$	Fraction of green-space area within the buffer.	1 m urban green-space/land-cover data
	Normalized difference vegetation index (NDVI)	$NDVI = \frac{NIR - R}{NIR + R}$	Vegetation cover and growth condition.	Sentinel-2 Level-2A (L2A) surface reflectance
	Modified normalized difference water index (MNDWI)	$\mathrm{MNDWI} = \frac{\mathrm{Green} - \mathrm{SWIR}}{\mathrm{Green} + \mathrm{SWIR}}$	Enhances open-water information while suppressing built-up and bare-surface backgrounds.	

Table A5. *Cont.*

Category	Indicator	Formula	Definition	Data Source
Activity intensity	Nighttime light intensity (NTL)	$NTL = \frac{1}{n}\sum_{i=1}^{n} DN_i$	Nighttime light intensity as a proxy for human activity.	Sustainable Development Science Satellite-1 (SDGSAT-1) Glimmer Imager (GLI) nighttime-light data
	Main-road density (MRD)	$MRD = \frac{\sum L_i}{A}$	Density of main roads and a proxy for traffic-related anthropogenic heat.	OpenStreetMap (OSM) road-network data

Note: A is the buffer area; Ai, Pi, Hi, and Vi are the footprint area, perimeter, height, and volume of building i; Gi and Li denote the green-space area and road length; n is the number of buildings, pixels, or angular samples used by the corresponding indicator.

Appendix D. Exploratory Predictor-Count Sensitivity

SHAP rankings were recalculated within training folds, while k was selected from the aggregated validation results. All LCZ panels use a common RMSE scale.

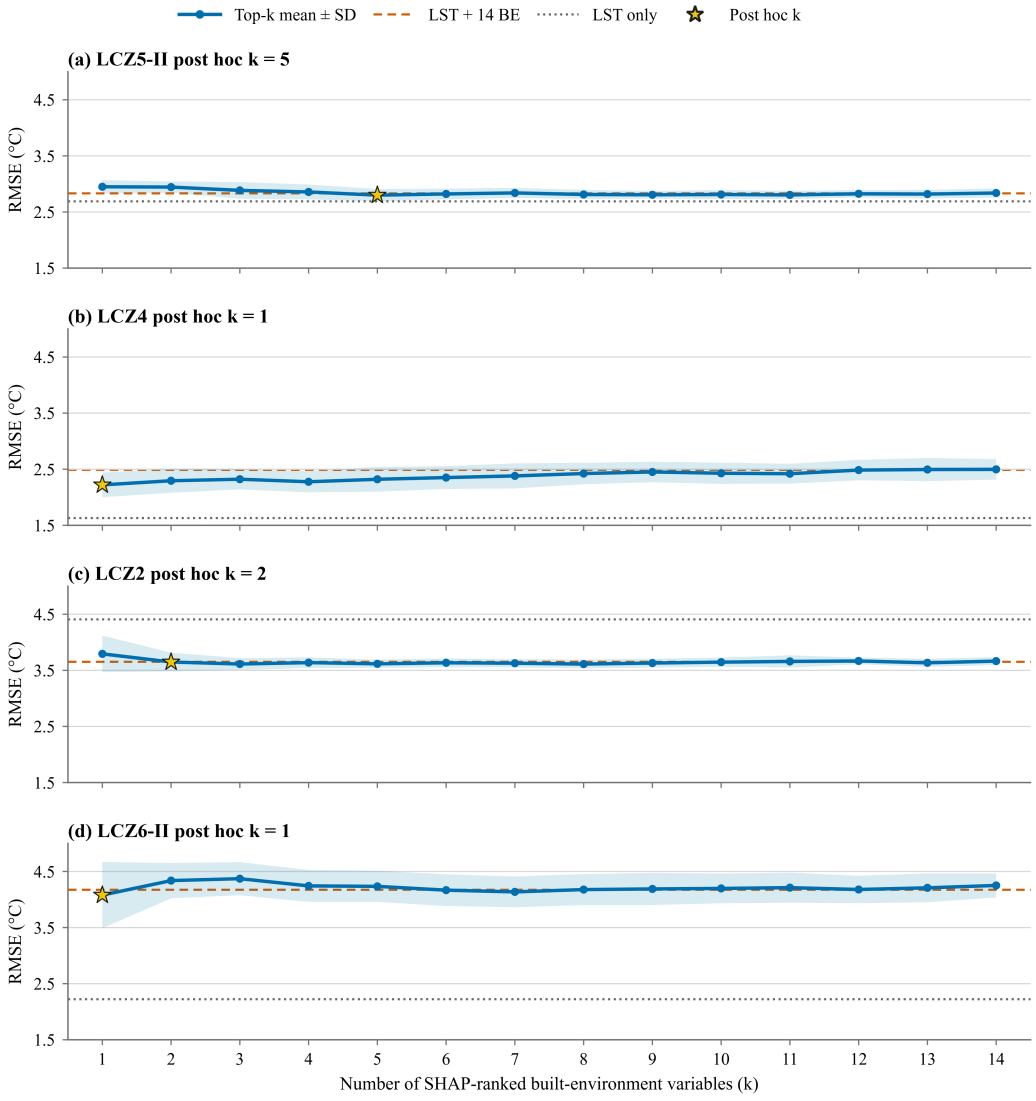


Figure A1. Exploratory Top-k sensitivity under the original site-grouped partitions: (**a**) LCZ5-II; (**b**) LCZ4; (**c**) LCZ2; (**d**) LCZ6-II. All panels use the same RMSE limits and ticks. Shading shows ±1 SD across ten repeated partitions; horizontal lines identify the LST-only and BE14 references. Stars indicate k selected retrospectively from these results.

Appendix E. Event-Held-Out and Joint Event/Site Tests

Table A6. Support for event tests. Train/test tuples give records/unique sites/events for training and records/sites for testing. E holds out one event; ES also excludes test-site identities. Overlap counts test sites present in E training. Eligibility required at least five training records from three sites and two distinct LST values.

LCZ	Held-Out Event	Test n/Sites	E Train n/Sites/Events	ES Train n/Sites/Events	E Overlap	Scored
LCZ5-II	2024-07-26	48/48	59/24/5	59/24/5	0	Yes
LCZ5-II	2024-08-21	12/12	95/60/5	95/60/5	0	Yes
LCZ5-II	2025-08-03	12/12	95/72/5	60/60/2	12	Yes
LCZ5-II	2025-08-04	12/12	95/72/5	60/60/2	12	Yes
LCZ5-II	2025-08-07	11/11	96/72/5	63/61/5	11	Yes
LCZ5-II	2025-08-19	12/12	95/72/5	60/60/2	12	Yes
LCZ4	2024-07-26	18/18	5/5/1	5/5/1	0	Yes
LCZ4	2024-08-21	5/5	18/18/1	18/18/1	0	Yes
LCZ2	2024-07-26	13/13	5/2/5	5/2/5	0	No
LCZ2	2024-08-21	1/1	17/14/5	17/14/5	0	Yes
LCZ2	2025-08-03	1/1	17/15/5	14/14/2	1	Yes
LCZ2	2025-08-04	1/1	17/15/5	14/14/2	1	Yes
LCZ2	2025-08-07	1/1	17/15/5	14/14/2	1	Yes
LCZ2	2025-08-19	1/1	17/15/5	14/14/2	1	Yes
LCZ6-II	2024-07-26	3/3	25/10/5	25/10/5	0	Yes
LCZ6-II	2024-08-21	5/5	23/8/5	23/8/5	0	Yes
LCZ6-II	2025-08-03	5/5	23/13/5	8/8/2	5	Yes
LCZ6-II	2025-08-04	5/5	23/13/5	8/8/2	5	Yes
LCZ6-II	2025-08-07	5/5	23/13/5	8/8/2	5	Yes
LCZ6-II	2025-08-19	5/5	23/13/5	8/8/2	5	Yes

Table A7. RMSE (°C) pooled over eligible event-held-out records. E and ES denote the protocols in Table A6; n/events gives the evaluated coverage. LCZ2 excludes the dominant 13-record event in both protocols, so its rows concern only five singleton records, with four from one site. Values are single pooled estimates, without split SD or an event-sampling confidence interval. All five models in each row use identical records.

Test	LCZ	n/Events	Mean	OLS	LST-Only	BE14	BE19
E	LCZ5-II	107/6	4.41	6.43	5.84	5.15	4.94
E	LCZ4	23/2	6.60	2.30	6.57	6.49	6.40
E	LCZ2	5/5	5.86	9.20	6.40	4.82	4.71
E	LCZ6-II	28/6	6.36	8.13	7.12	6.51	6.45
ES	LCZ5-II	107/6	5.15	6.90	6.53	6.22	6.11
ES	LCZ4	23/2	6.60	2.30	6.57	6.49	6.40
ES	LCZ2	5/5	6.50	10.43	8.23	6.69	6.56
ES	LCZ6-II	28/6	9.92	8.92	9.25	9.42	9.58

Pooled event-held-out validation metrics and evaluated coverage are provided in Supplementary File S1. R^2 and r are undefined within singleton test events; pooled values can be dominated by between-event contrasts. Positive LCZ2 BE14/BE19 R^2 under E (0.115/0.157) became negative under ES (−0.702/−0.635), indicating sensitivity to the limited repeated-site support.

Appendix F. Fixed Capacity and Input Sensitivity

Table A8. Training and test RMSE (°C) under the saved site-grouped partitions. Cells give training RMSE; test RMSE ± split SD. Original: depth 3/L2 8; shallow: depth 1/L2 8; stronger L2: depth 3/L2 20. Within a repetition, each record appears in four training folds and one test fold. Metrics pool the respective residuals before summarizing ten repetitions.

LCZ	Setting	LST-Only: Train; Test	BE14: Train; Test	BE19: Train; Test
LCZ5-II	Original	2.33; 2.69 ± 0.02	1.75; 2.83 ± 0.08	1.75; 2.86 ± 0.06
LCZ5-II	Shallow	2.74; 2.93 ± 0.02	2.35; 3.05 ± 0.07	2.34; 3.06 ± 0.08
LCZ5-II	Stronger L2	2.45; 2.71 ± 0.02	2.10; 2.88 ± 0.06	2.12; 2.94 ± 0.07
LCZ4	Original	1.03; 1.63 ± 0.10	0.81; 2.49 ± 0.17	0.81; 2.57 ± 0.17
LCZ4	Shallow	1.20; 1.61 ± 0.12	1.12; 2.36 ± 0.22	1.11; 2.41 ± 0.19
LCZ4	Stronger L2	1.42; 1.90 ± 0.15	1.47; 2.61 ± 0.15	1.47; 2.65 ± 0.15
LCZ2	Original	1.58; 4.41 ± 0.03	1.38; 3.65 ± 0.06	1.38; 3.63 ± 0.05
LCZ2	Shallow	2.13; 4.50 ± 0.03	1.58; 3.92 ± 0.09	1.58; 3.91 ± 0.07
LCZ2	Stronger L2	2.10; 4.05 ± 0.03	1.86; 3.55 ± 0.05	1.86; 3.55 ± 0.05
LCZ6-II	Original	1.68; 2.22 ± 0.14	1.82; 4.17 ± 0.26	1.83; 4.31 ± 0.30
LCZ6-II	Shallow	2.64; 2.89 ± 0.15	2.30; 4.55 ± 0.27	2.28; 4.53 ± 0.24
LCZ6-II	Stronger L2	2.50; 3.00 ± 0.19	2.66; 4.64 ± 0.27	2.70; 4.77 ± 0.28

Capacity/input sensitivity comprised 1800 fits. Original-setting LST-only and BE14 predictions reproduced archived values to numerical precision (maximum absolute difference < 10^{-12} °C). The separate Morph12 comparison comprised 200 fits on the same partitions. Summary capacity/input sensitivity metrics and paired gain statistics are included in Supplementary File S1.

Table A9. Incremental RMSE from building morphology and BE inputs under the original settings. Gain = RMSE(LST-only) − RMSE(augmented); positive values indicate lower error. Percentages use mean LST-only RMSE as the denominator.

LCZ	LST-Only RMSE (°C)	Morph12 RMSE (°C)	Morph12 Gain (°C; %)	BE14 Gain (°C; %)	BE19 Gain (°C; %)	Morph12 R^2
LCZ5-II	2.69	2.83	−0.14; −5.36%	−0.14; −5.29%	−0.17; −6.38%	0.447
LCZ4	1.63	2.45	−0.82; −50.13%	−0.86; −52.77%	−0.94; −57.61%	0.356
LCZ2	4.41	3.59	+0.82; +18.56%	+0.76; +17.15%	+0.78; +17.71%	−0.388
LCZ6-II	2.22	3.70	−1.48; −66.38%	−1.95; −87.79%	−2.09; −93.86%	0.501

Morph12 comprises BD, MH, Hmax, SDH, NHV, AWH, MBV, MBS, FD, BSI, BVD and BVCV. SVF is excluded because its source includes canopy as well as building geometry. Vegetation, water, impervious cover and activity indicators are also excluded. LCZ2 Morph12 R^2 remains negative despite a positive RMSE gain. Paired split SDs of RMSE gains are reported in Supplementary File S1.

Appendix G. Independent-Site BE Distributions

Table A10. BE summaries at 123 unique sites: LCZ5-II n = 72, LCZ4 n = 23, LCZ2 n = 15 and LCZ6-II n = 13. Each cell gives mean ± sample SD followed by [minimum, maximum], in the units and definitions of Table A5. Values use three significant figures; eta^2 uses three decimal places. Eta^2 is the between-LCZ sum of squares divided by the total site-level sum of squares for that indicator. Summaries use the fixed attributes assigned to each site's 50 m buffer in the analysis table.

Indicator	LCZ5-II	LCZ4	LCZ2	LCZ6-II	eta^2
MRD	7.02 ± 8.87 [0, 36.3]	5.83 ± 7.82 [0, 25.4]	6.05 ± 7.68 [0, 22.6]	12.7 ± 6.7 [0, 24.6]	0.052
NDVI	0.289 ± 0.138 [0.0929, 0.711]	0.356 ± 0.137 [0.127, 0.7]	0.219 ± 0.0872 [0.0504, 0.332]	0.152 ± 0.139 [−0.0104, 0.44]	0.162
ISF	0.274 ± 0.171 [0.00731, 0.815]	0.2 ± 0.133 [0, 0.53]	0.263 ± 0.113 [0.145, 0.526]	0.499 ± 0.292 [0.132, 0.977]	0.176
SVF	0.748 ± 0.076 [0.615, 0.985]	0.806 ± 0.0848 [0.702, 0.994]	0.704 ± 0.0536 [0.642, 0.831]	0.801 ± 0.084 [0.671, 0.919]	0.155
BD	0.262 ± 0.125 [0, 0.52]	0.173 ± 0.109 [0, 0.391]	0.339 ± 0.114 [0.121, 0.57]	0.338 ± 0.164 [0.0921, 0.697]	0.156
MH	15.9 ± 5.01 [0, 30.1]	17.9 ± 8.25 [0, 44.7]	16.4 ± 3.51 [12.8, 28.4]	12 ± 2.22 [10, 16.8]	0.078
Hmax	24.7 ± 15 [0, 76.2]	26.3 ± 16.3 [0, 83.6]	21.7 ± 6.79 [16.9, 43.4]	17.4 ± 5.86 [11.8, 30.6]	0.034
SDH	4.51 ± 4.33 [0, 19.9]	6.06 ± 6.52 [0, 32]	3.22 ± 1.86 [1.56, 9.33]	2.73 ± 2.3 [0.756, 8.03]	0.048
NHV	0.252 ± 0.184 [0, 0.935]	0.282 ± 0.163 [0, 0.716]	0.191 ± 0.0704 [0.0998, 0.334]	0.209 ± 0.127 [0.071, 0.494]	0.029
AWH	16.7 ± 5.93 [0, 40.6]	18.1 ± 7.4 [0, 36.8]	17.1 ± 4.43 [13.9, 32.5]	12.7 ± 2.97 [10.2, 19.8]	0.059
MBV	$1.5 \times 10^3 \pm 1.33 \times 10^3$ [0, 9.72×10^3]	$1.57 \times 10^3 \pm 1.12 \times 10^3$ [0, 3.86×10^3]	$1.65 \times 10^3 \pm 1.11 \times 10^3$ [807, 5.58×10^3]	$2.51 \times 10^3 \pm 2.47 \times 10^3$ [570, 9.24×10^3]	0.045
MBS	673 ± 387 [0, 2.6×10^3]	721 ± 352 [0, 1.53×10^3]	741 ± 340 [482, 1.94×10^3]	848 ± 589 [337, 2.45×10^3]	0.019
FD	1.18 ± 0.474 [0, 3.7]	1.09 ± 0.295 [0, 1.76]	1.22 ± 0.467 [0.937, 2.89]	2.37 ± 3.33 [0.955, 12.6]	0.100
BSI	5.78 ± 3.63 [0, 14.9]	5.87 ± 5.09 [0, 21.8]	5.57 ± 0.972 [3.56, 7.49]	14.8 ± 11.5 [5.48, 41.2]	0.233
BVD	4.36 ± 2.24 [0, 10.8]	3.2 ± 1.91 [0, 5.89]	5.75 ± 2.02 [1.69, 8.52]	4.36 ± 2.63 [0.943, 10.6]	0.094
BVCV	1.19 ± 0.522 [0, 1.99]	1.16 ± 0.6 [0, 2.06]	1.21 ± 0.374 [0, 1.7]	0.835 ± 0.742 [0, 1.95]	0.039
NTL	64.2 ± 183 [1.83, 1.39×10^3]	42.2 ± 81.7 [1.84, 389]	33.7 ± 23.5 [6.95, 74.3]	32.4 ± 47.3 [1, 146]	0.009
RG	0.187 ± 0.229 [0, 1]	0.223 ± 0.196 [0, 0.684]	0.219 ± 0.26 [0, 0.997]	0.162 ± 0.238 [0, 0.604]	0.007
MNDWI	−0.269 ± 0.0793 [−0.506, −0.0961]	−0.321 ± 0.0547 [−0.429, −0.21]	−0.268 ± 0.0941 [−0.51, −0.153]	−0.274 ± 0.0883 [−0.436, −0.164]	0.065

Appendix H. LCZ2 Event-Level Prediction Diagnostics

Prediction–observation covariance was calculated within each of the ten original repetitions and decomposed as C_total = C_within + C_between. C_within sums event-centered cross-products divided by 18; C_between sums n_event × (mean observed Ta_event −

overall mean observed Ta) × (mean predicted Ta_event − overall mean predicted Ta), also divided by 18. Singleton events contribute zero within-event covariance. Across repetitions, the mean C_total, C_within and C_between were −3.79, 0.03 and −3.82 °C^2, respectively. C_between was negative in all ten repetitions; within-event covariance ranged from −0.02 to 0.06 °C^2.

Table A11. LCZ2 event means and residual contributions for the original LST-only model. Predicted means and covariance contributions are averaged over ten repetitions; SSE shares pool squared errors over those repetitions. Covariance contributions use the record weights defined above and sum to C_between across events. Dates follow the finalized matching table.

Event	n	Observed Ta (°C)	LST (°C)	Predicted Ta (°C)	SSE Share (%)	Between Covariance (°C^2)
26 July 2024	13	32.29	48.04	32.66	2.98	−0.03
21 August 2024	1	25.70	27.14	37.29	38.43	−1.93
3 August 2025	1	39.10	45.21	32.07	14.13	−0.18
4 August 2025	1	35.20	43.67	32.07	2.80	−0.06
7 August 2025	1	35.50	38.04	32.07	3.37	−0.07
19 August 2025	1	40.30	32.70	28.73	38.30	−1.55

Raw LST–Ta r was −0.027 across all 18 records and −0.190 within the dominant event. These weak observed associations contrast with the strong inverse prediction association between events. Additional event coverage and concurrent local meteorological observations are needed to investigate the underlying processes.

References

1. Tuholske, C.; Caylor, K.; Funk, C.; Verdin, A.; Sweeney, S.; Grace, K.; Peterson, P.; Evans, T. Global urban population exposure to extreme heat. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2024792118. [CrossRef]
2. Singh, M.; Sharston, R. Quantifying the dualistic nature of urban heat Island effect (UHI) on building energy consumption. *Energy Build.* **2022**, *255*, 111649. [CrossRef]
3. Weng, Q. Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 335–344. [CrossRef]
4. Li, Z.; Wu, H.; Duan, S.; Zhao, W.; Ren, H.; Liu, X.; Leng, P.; Tang, R.; Ye, X.; Zhu, J.; et al. Satellite Remote Sensing of Global Land Surface Temperature: Definition, Methods, Products, and Applications. *Rev. Geophys.* **2023**, *61*, e2022RG000777. [CrossRef]
5. Cao, J.; Zhou, W.; Zheng, Z.; Ren, T.; Wang, W. Within-city spatial and temporal heterogeneity of air temperature and its relationship with land surface temperature. *Landsc. Urban Plan.* **2021**, *206*, 103979. [CrossRef]
6. Xiang, Y.; Zheng, B.; Bedra, K.B.; Ouyang, Q.; Liu, J.; Zheng, J. Spatial and seasonal differences between near surface air temperature and land surface temperature for Urban Heat Island effect assessment. *Urban Clim.* **2023**, *52*, 101745. [CrossRef]
7. Colaninno, N.; Morello, E. Towards an operational model for estimating day and night instantaneous near-surface air temperature for urban heat island studies: Outline and assessment. *Urban Clim.* **2022**, *46*, 101320. [CrossRef]
8. Voogt, J.; Oke, T. Thermal remote sensing of urban climates. *Remote Sens. Environ.* **2003**, *86*, 370–384. [CrossRef]
9. Hong, T.; Heo, Y. Exploring the impact of urban factors on land surface temperature and outdoor air temperature: A case study in Seoul, Korea. *Build. Environ.* **2023**, *243*, 110645. [CrossRef]
10. Stewart, I.D.; Oke, T.R. Local Climate Zones for Urban Temperature Studies. *Bull. Am. Meteorol. Soc.* **2012**, *93*, 1879–1900. [CrossRef]
11. Ching, J.; Mills, G.; Bechtel, B.; See, L.; Feddema, J.; Wang, X.; Ren, C.; Brousse, O.; Martilli, A.; Neophytou, M.; et al. WUDAPT: An Urban Weather, Climate, and Environmental Modeling Infrastructure for the Anthropocene. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 1907–1924. [CrossRef]
12. Xu, D.; Zhang, Q.; Zhou, D.; Yang, Y.; Wang, Y.; Rogora, A. Local climate zone in Xi'an city: A novel classification approach employing spatial indicators and supervised classification. *Buildings* **2023**, *13*, 2806. [CrossRef]
13. Eldesoky, A.H.; Gil, J.; Pont, M.B. The suitability of the urban local climate zone classification scheme for surface temperature studies in distinct macroclimate regions. *Urban Clim.* **2021**, *37*, 100823. [CrossRef]

14. Emery, J.; Pohl, B.; Crétat, J.; Richard, Y.; Pergaud, J.; Rega, M.; Zito, S.; Dudek, J.; Vairet, T.; Joly, D.; et al. How local climate zones influence urban air temperature: Measurements by bicycle in Dijon, France. *Urban Clim.* **2021**, *40*, 101017. [CrossRef]
15. Yang, X.; Xu, S.; Peng, L.L.; Chen, Y.; Yao, L. General air temperature and humidity features of local climate zones: A multi-city observational study in eastern China. *Urban Clim.* **2023**, *51*, 101652. [CrossRef]
16. Vaidya, M.; Keskar, R.; Kotharkar, R. Quantifying the effect of surrounding spatial heterogeneity on land surface temperature based on local climate zones using mutual information. *Sustain. Cities Soc.* **2024**, *107*, 105455. [CrossRef]
17. Bai, X.; Yu, Z.; Wang, B.; Zhang, Y.; Zhou, S.; Sha, X.; Li, S.; Yao, X.; Geng, X. Quantifying threshold and scale response of urban air and surface temperature to surrounding landscapes under extreme heat. *Build. Environ.* **2024**, *247*, 111029. [CrossRef]
18. Guo, A.; Yue, W.; Yang, J.; He, T.; Zhang, M.; Li, M. Divergent impact of urban 2D/3D morphology on thermal environment along urban gradients. *Urban Clim.* **2022**, *45*, 101278. [CrossRef]
19. Zhang, L.; Nikolopoulou, M.; Guo, S.; Song, D. Impact of LCZs spatial pattern on urban heat island: A case study in Wuhan, China. *Build. Environ.* **2022**, *226*, 109785. [CrossRef]
20. Zhang, Z.; Luan, W.; Yang, J.; Guo, A.; Su, M.; Tian, C. The influences of 2D/3D urban morphology on land surface temperature at the block scale in Chinese megacities. *Urban Clim.* **2023**, *49*, 101553. [CrossRef]
21. Wei, X.; Guan, F.; Zhang, X.; Van de Weghe, N.; Huang, H. Integrating planar and vertical environmental features for modelling land surface temperature based on street view images and land cover data. *Build. Environ.* **2023**, *235*, 110231. [CrossRef]
22. Lin, X.; Lin, X.; Yan, C. Inter- and intra-LCZ thermal heterogeneity: The dominant role of external environments in shaping local land surface temperature. *Sustain. Cities Soc.* **2025**, *121*, 106188. [CrossRef]
23. Li, S.; Wong, M.S.; Zhu, R.; Shi, G.; Yang, J. Impacts of land surface temperature and ambient factors on near-surface air temperature estimation: A multisource evaluation using SHAP analysis. *Sustain. Cities Soc.* **2025**, *122*, 106257. [CrossRef]
24. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
25. Hulley, G.C.; Gottsche, F.M.; Rivera, G.; Hook, S.J.; Freepartner, R.J.; Martin, M.A.; Cawse-Nicholson, K.; Johnson, W.R. Validation and Quality Assessment of the ECOSTRESS Level-2 Land Surface Temperature and Emissivity Product. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5000523. [CrossRef]
26. Hook, S.; Hulley, G. *ECOSTRESS Tiled Land Surface Temperature and Emissivity Instantaneous L2 Global 70 m V002 [Data Set]*; NASA Land Processes Distributed Active Archive Center: Sioux Falls, SD, USA, 2022. [CrossRef]
27. Beele, E.; Aerts, R.; Reyniers, M.; Somers, B. Spatial configuration of green space matters: Associations between urban land cover and air temperature. *Landsc. Urban Plan.* **2024**, *249*, 105121. [CrossRef]
28. Ziter, C.D.; Pedersen, E.J.; Kucharik, C.J.; Turner, M.G. Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 7575–7580. [CrossRef]
29. O'Brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]
30. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
33. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
34. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed on 26 August 2026).
35. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6638–6648. Available online: https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html (accessed on 26 August 2026).
36. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774.
37. Oke, T.R. The energetic basis of the urban heat island. *Q. J. R. Meteorol. Soc.* **1982**, *108*, 1–24. [CrossRef]
38. Hu, L.; Uejio, C. Ground Urban Heat Island: Strengthening the Connection Between Spaceborne Thermal Observations and Urban Heat Risk Management. *GeoHealth* **2024**, *8*, e2024GH001114. [CrossRef]
39. Zhou, W.; Cao, W.; Wu, T.; Zhang, T. The win-win interaction between integrated blue and green space on urban cooling. *Sci. Total Environ.* **2023**, *863*, 160712. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.