Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hot School Grounds, Mismatched Evidence: Facility-Scale Urban Heat in Barcelona and Madrid

Jiarui Shen

JiaruiShen1976@163.com

University of Chicago

Research Article

Keywords: school grounds, land surface temperature, Local Climate Zones, Landsat, spatial cross-validation, urban heat

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10736277/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hot School Grounds, Mismatched Evidence: Facility-Scale Urban Heat in Barcelona and Madrid

Jiarui Shen
PhD Candidate
Department of Sociology, Division of the Social Sciences, University of Chicago
1126 East 59th Street, Chicago, IL 60637, United States
Email: JiaruiShen1976@163.com
ORCID: https://orcid.org/0009-0007-6166-2963

Abstract

Urban heat is often mapped at pixel or neighborhood scale, although adaptation decisions are implemented through bounded public assets. We test a facility-aligned screening framework using Landsat 8 Collection 2 Level-2 imagery, Local Climate Zone context, and five surface indicators for 158 public primary schools in Barcelona and 276 in Madrid during July-August 2014-2024 (21,438 complete pixel-year observations). Median summer school-ground LST was 31.8 °C in Barcelona and 37.2 °C in Madrid; median Theil-Sen slopes were 0.09 and 0.18 °C/year, while standard Mann-Kendall tests with within-city FDR control classified positive monotonic trends in 23.4% and 40.6% of schools, respectively. Under five-fold school-grouped cross-validation, Extra Trees performed best for held-out schools (R^2=0.78±0.05; RMSE=2.41±0.18 °C), compared with R^2=0.89 under a random pixel split. NDMI and NDVI ranked first and second in pooled predictive importance. Aligning measurement, validation, and interpretation with the campus changes the evidential basis for screening. The results support field-audit prioritization, not causal retrofit effects or children's thermal comfort.

Keywords: *school grounds; land surface temperature; Local Climate Zones; Landsat; spatial cross-validation; urban heat*

1. Introduction

Extreme summer heat is now a routine design and management problem for cities. Remote sensing has made intra-urban thermal contrasts visible at scale, while the Local Climate Zone (LCZ) framework provides a standardized description of urban form and land cover (Stewart & Oke, 2012; Ching et al., 2018; Demuzere et al., 2022). Machine-learning models further capture nonlinear relationships between land surface temperature (LST) and environmental predictors. Tahiri et al. (2026), for example, combined Landsat 8, LCZs, five spectral/environmental indicators and seven regressors to compare Casablanca and Barcelona over 2014–2024. That architecture is useful because it links morphology, long-term thermal evidence and model comparison. Yet technical resolution alone does not guarantee decision relevance: a detailed raster can still be poorly aligned with the asset that a municipality must inspect, fund and redesign.

A more consequential gap therefore lies between diagnostic scale and implementation scale. City-wide rasters and neighborhood averages can identify broad thermal contrasts, but municipal budgets, maintenance responsibilities and redesign projects are attached to managed places such as school grounds, parks, streets and plazas. Recent implementation-oriented urban-heat research similarly identifies the transition from scientific diagnosis to

actionable planning instruments as a persistent bottleneck (Lau et al., 2025). We describe this as a decision-unit mismatch: a model may be spatially detailed while remaining poorly aligned with the facility where action is actually taken. Public primary-school grounds provide a stringent test because they are bounded, numerous, routinely managed and sufficiently large for repeated Landsat observation. School-climate-shelter and naturalization studies also show that adaptation is implemented through site-specific design, maintenance and governance rather than through an abstract city-average pixel (López Plazas et al., 2023; Sekulova & Ruiz-Mallén, 2024).

School grounds are an increasingly important urban-climate setting. Greening studies show that school environments can contribute to urban resilience and access to nature (Flax et al., 2020; Baró et al., 2021), while thermal-environment research identifies shade, vegetation and surface configuration as central design variables (Antoniadis et al., 2020). Cross-city evidence documents unequal access to school-related green and blue infrastructure (Gallez et al., 2024). Barcelona has become a prominent implementation setting: climate-shelter programs combine green, blue and grey interventions, with published work covering design and monitoring, mixed-method evaluation, children's well-being and community use (López Plazas et al., 2023; Sanz-Mas et al., 2024; Sanz-Mas, Continente, Brugueras, et al., 2025; Sanz-Mas, Continente, Marí-Dell'Olmo, & López, 2025). Governance research shows that naturalization depends on institutional conditions as well as physical design (Sekulova & Ruiz-Mallén, 2024), and recent Tokyo evidence highlights the operational pressure created by extreme schoolyard heat (Terada et al., 2026). LCZ-based work around elementary schools further shows that surrounding urban structure is associated with LST variation (He et al., 2024). What remains less developed is a decade-long facility-scale design that combines contextual morphology, repeated satellite observations, school-level trend evidence and leakage-aware prediction.

The study addresses three linked gaps. First, school-heat studies are often local snapshots, intervention evaluations or exposure assessments; fewer combine a decade of repeated satellite observations with formal school-level trend estimation. Second, LCZ is commonly treated as the principal spatial unit or neighborhood context even though the school ground is the managed asset. Treating LCZ as context and the campus as the decision unit separates surrounding morphology from facility conditions. Third, pixel-level machine-learning studies can report optimistic performance when observations from the same site appear in both training and testing. The present design therefore uses school identity as the primary holdout unit and retains a random pixel split only as a supplementary interpolation benchmark.

Barcelona and Madrid provide an analytically useful comparison within one national setting while differing in urban form and summer climate. Barcelona is a dense coastal Mediterranean city with a visible school-climate-shelter program; Madrid is an inland city with stronger summer heating and a large public-school inventory. The comparison is not intended to rank which city is "better adapted." Instead, it asks whether a common facility-scale workflow yields comparable or context-dependent thermal signals under contrasting urban climates, and whether pooled surface relationships remain informative without being treated as universal design rules.

The outcome is strictly school-ground LST. Satellite LST is not equivalent to near-surface air temperature, mean radiant temperature, Universal Thermal Climate Index or children's physiological heat strain. A lower satellite LST therefore cannot, by itself, establish lower human heat exposure. Throughout the paper, "school-ground surface thermal

environment" denotes satellite-observed surface conditions, and current thermal level is kept distinct from longitudinal change.

Three research questions structure the analysis. RQ1: How do summer LST levels and 2014–2024 longitudinal trends vary across public primary-school grounds in Barcelona and Madrid, with LCZ retained as urban-form context? RQ2: Which vegetation, moisture, built-surface, bare-surface and shortwave-reflectance proxy indicators show the strongest bivariate associations and predictive importance? RQ3: Which of seven machine-learning regressors performs best for pixel-year observations from schools not represented in training, and what bounded site-audit priorities follow from the selected model?

The central contribution is decision-unit alignment between urban-heat evidence and the public facility where adaptation decisions are made. LCZ supplies contextual morphology, the campus polygon defines the managed unit for repeated LST measurement, annual summaries define longitudinal evidence, and school_id defines the prediction holdout. The same facility then anchors site-audit translation. Methodologically, the design tests how performance changes when validation matches the management unit; empirically, it builds a decade-long two-city facility panel; practically, it translates predictive attribution into bounded audit domains without treating association as retrofit effect. Figure 1 summarizes this alignment framework, while Figure 2 situates Barcelona and Madrid within the comparative design.

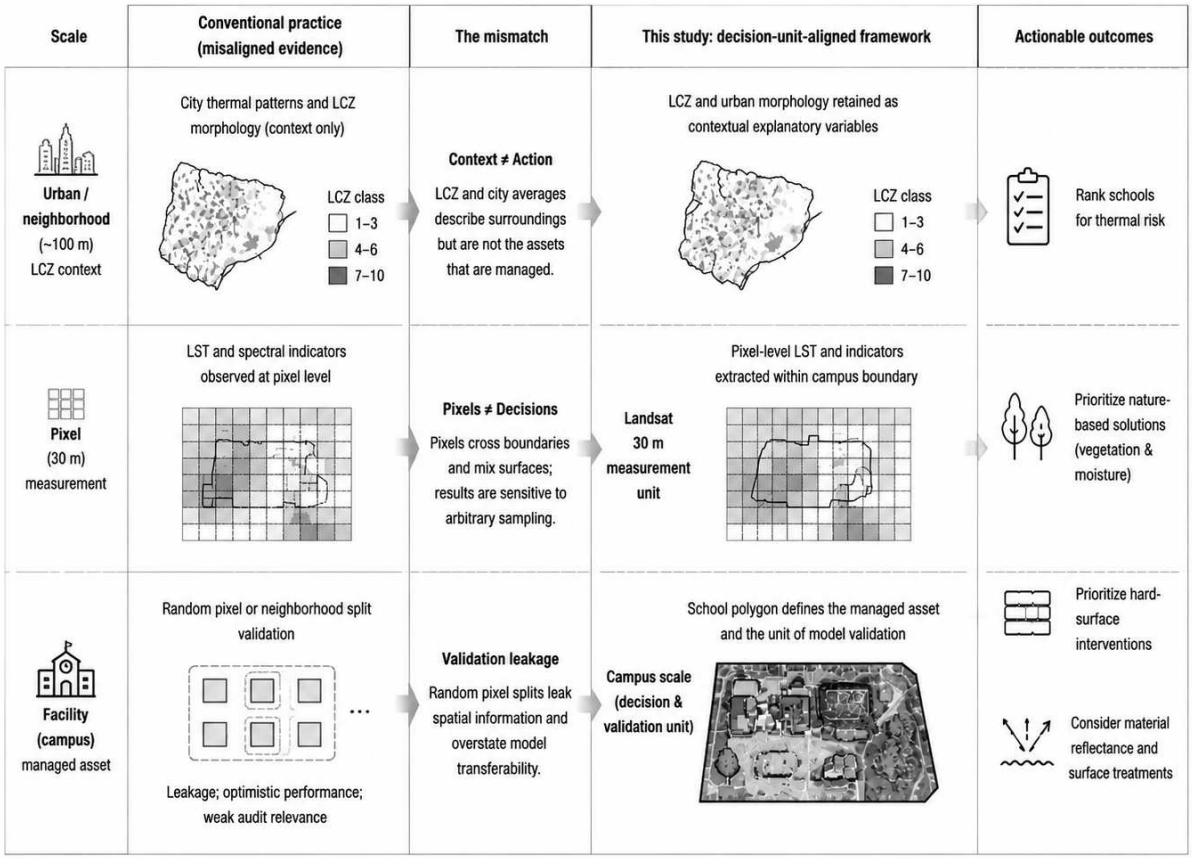


Figure 1. Decision-unit mismatch and facility-scale alignment framework for school-ground heat evidence

The framework illustrates the transition from conventional urban-scale or pixel-based thermal evidence to a decision-unit-aligned approach. LCZ describes surrounding urban morphology as contextual information, Landsat pixels provide thermal measurements, and the school campus polygon defines the managed decision and validation unit that connects evidence to site-audit and adaptation prioritization. The framework does not imply causal effects of specific interventions but identifies bounded domains for follow-up assessment.

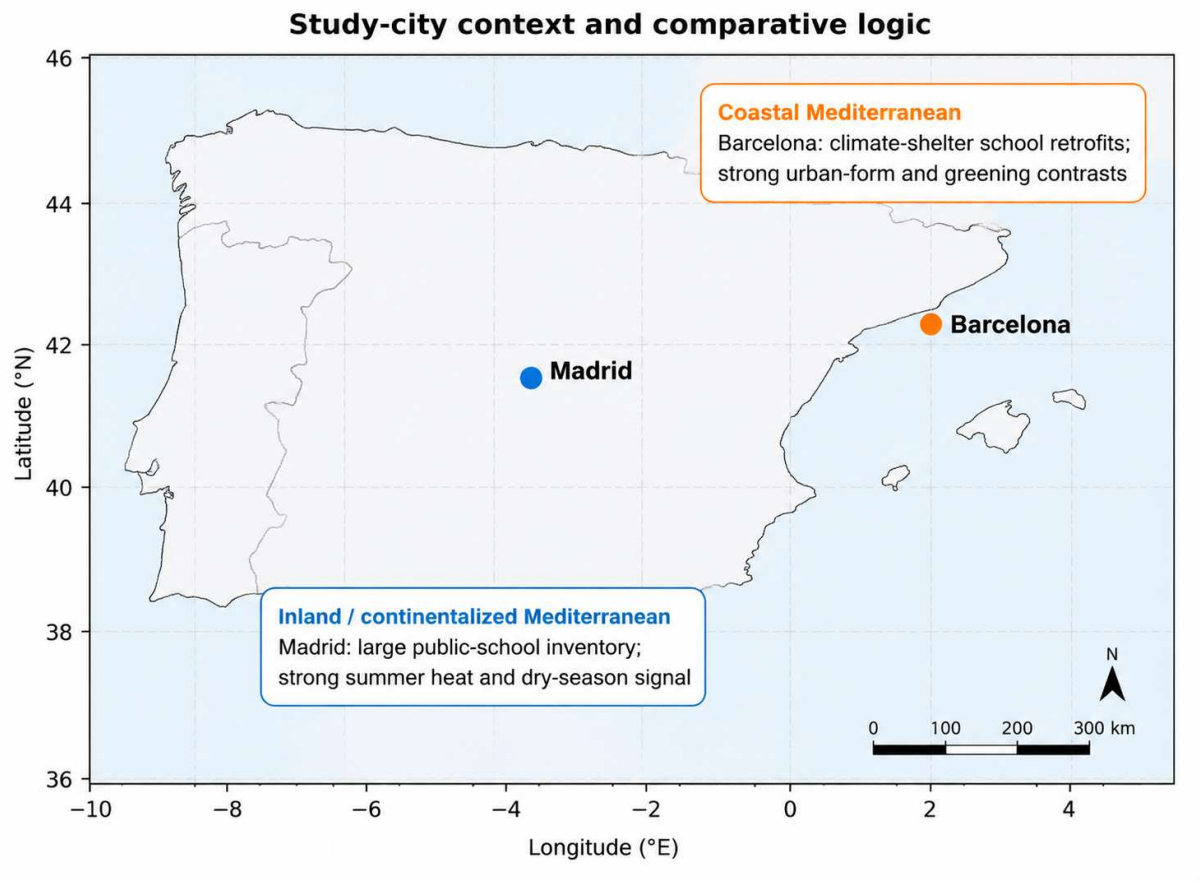


Figure 2. Study-city context and comparative logic. City points provide geographic orientation; the public primary-school campus polygons used for Landsat extraction are the analytical units but are not plotted in this locator.

Note. The figure provides geographic orientation only; all analytical inference uses the campus, Landsat and LCZ data defined in Section 2.

Three nested spatial roles are therefore kept distinct. The Landsat pixel is the measurement unit, the campus is the decision and validation unit, and LCZ is the contextual morphology unit. This separation matters because repeated pixel rows do not themselves define independent facilities, while campus boundaries determine where school-level summaries, management responsibility and prioritization claims attach. LCZ adds information about surrounding urban form without being treated as a school-level intervention target. The hierarchy also clarifies the validation logic: high pixel-level accuracy is not sufficient for facility screening if observations from the same school appear in both training and testing, whereas school-grouped validation still does not establish geographic external validity because different schools may remain spatially proximate. The design

therefore narrows the claim while aligning it more closely with the municipal object of action.

2. Materials and methods

2.1. Study design and analytical unit

The design is a longitudinal, comparative, remote-sensing study in which the public primary-school ground is the functional analytical unit and the Landsat pixel is the measurement unit. The observation window is fixed to 1 July–31 August for each year from 2014 through 2024. Restricting the analysis to the same seasonal window reduces confounding by large seasonal changes in solar geometry, vegetation phenology and moisture conditions and concentrates the analysis on the period when heat-mitigation design is most consequential. Annual summer composites are used rather than a single scene so that results are not driven by one acquisition date.

A “school ground” is operationalized as the quality-controlled campus polygon associated with a public primary school. The polygon may contain buildings, paved play space, vegetated space and other surfaces that are functionally part of the school site. The study therefore estimates a campus-scale surface thermal environment rather than attempting to isolate only the playground. This distinction is necessary because the 30-m gridded Level-2 surface-temperature product can integrate multiple materials near campus boundaries and is derived from coarser native TIRS thermal measurements. The approach trades fine microclimatic detail for consistent cross-city, cross-year observation.

Schools enter the analytical cohort when: (1) the official registry identifies a public primary school within the municipal boundary; (2) a campus polygon can be matched unambiguously; (3) the polygon contains at least three valid Landsat pixel centroids; and (4) at least 7 of 11 annual summer composites meet the completeness rule. The same thresholds apply in both cities. Verifiable relocations, irreconcilable boundaries, or inseparable mixed land uses are excluded where records permit. Because the archive does not support complete reason-coded reconstruction for every screened record, we report fixed eligibility rules, screened frames, final cohorts, and retention rates rather than reconstructing missing intermediate categories. Longitudinal extraction uses the retained quality-controlled footprint; undocumented historical boundary changes are not inferred.

City-specific official open-data inventories define the sampling frames. Madrid sources identify ownership, location, educational stage, and public schoolyard context; Barcelona's municipal inventory identifies education facilities and primary-school records. These records establish school identity and location, not outcomes. Campus geometries are matched to open education polygons and manually quality-checked against current orthophotography or satellite basemaps, producing the derived campus-pixel panel used in this study.

2.2. Data sources

2.2.1. Landsat 8 Collection 2 Level-2 imagery

Landsat 8 Operational Land Imager/Thermal Infrared Sensor Collection 2, Tier 1, Level-2 scenes are accessed through Google Earth Engine (LANDSAT/LC08/C02/T1_L2). The product provides surface reflectance and surface-temperature bands, allowing one spectral-index/LST workflow across the observation period. Earth Engine is used for filtering, masking, compositing, and export (Gorelick et al., 2017); persistent asset identifiers and the scripted transformations provide data provenance.

For each city-year, images are restricted to 1 July-31 August and the municipal study boundary. QA_PIXEL removes clouds, shadows, and other contaminated observations. Collection 2 scale/offset factors are applied to surface reflectance and ST_B10, with temperature converted from Kelvin to degrees Celsius. LST and all five predictors are calculated from the same masked collection, followed by a per-pixel summer median composite. This reduces dependence on any single acquisition while preserving a common seasonal window.

The Level-2 surface-temperature product is distributed on a 30-m grid, although native TIRS measurements are coarser. Narrow vegetation strips, small courtyards, and shade structures can therefore be unresolved or mixed with adjacent surfaces. We consequently interpret values at campus scale, apply minimum valid-area rules, and do not treat individual cells as exact playground microclimates.

2.2.2. Local Climate Zones

LCZ context is assigned from the global 100-m map in Earth Engine (RUB/RUBCLIM/LCZ/global_lcz_map/latest), using the recommended LCZ_Filter band. The product implements the 17-class Stewart-Oke typology and provides a consistent cross-city description of urban form (Stewart & Oke, 2012; Demuzere et al., 2022).

LCZ is used as contextual urban-form metadata, not as a causal predictor. Each school receives the modal intersecting LCZ, with class proportions retained where campuses cross boundaries. Subclass summaries are descriptive or exploratory when school counts permit; the primary predictive models exclude LCZ from the five-feature input set.

2.2.3. School inventories and campus geometry

School identity is based on authoritative open records: the Comunidad de Madrid education-centre inventory, Madrid's public-school schoolyard dataset as a consistency check, and Barcelona's municipal education-equipment inventory. Download dates and file hashes are retained in the data-lineage log.

Official point or address records are matched to campus polygons hierarchically: normalized name/address match, nearest candidate within 100 m when unique, then manual adjudication. Ambiguous, missing, or multi-institution polygons enter boundary review and are included only when a traceable boundary can be established from open, date-stamped geospatial imagery. Each accepted polygon receives a stable school_id across yearly extracts.

The data lineage is therefore explicit: official records define which schools exist; campus polygons define where pixels are sampled; Landsat provides the thermal and spectral measurements; the global LCZ product provides surrounding urban-form context. Published articles are used only to motivate the study and justify methods. No numerical result, temperature estimate or feature-importance value is copied from prior publications into the empirical dataset.

The source-to-variable mapping and provenance requirements are summarized in Table 1.

Table 1. Data sources and analytical role.

Source	Product / record	Spatial / temporal scale	Role	Documentation key
USGS / Google Earth Engine	Landsat 8 C2 T1 L2	30 m; summers 2014–2024	LST and five surface predictors	LANDSAT/LC08/C02/T1_L2; QA_PIXEL mask; scripted scale factors
Ruhr University / Earth Engine	Global LCZ map, LCZ_Filter	100 m; nominal 2018	Urban-form context	RUB/RUBCLIM/LCZ/global_lcz_map/latest
Comunidad de Madrid	Centros educativos de la comunidad de Madrid	Official school inventory; annual update	Madrid school identity, ownership, stage	Downloaded resource + hash + extraction date
Ayuntamiento de Madrid	Public-school schoolyard study	Public infant/primary schools	Geometry/QC context and schoolyard attributes	Municipal open-data resource
Open Data Barcelona	Education equipment inventory	City education facilities	Barcelona school identity and location	Dataset resource + hash + extraction date
Open education polygons / manual QC	Campus boundaries	School-site polygon	Pixel sampling frame	Name/address matching log; polygon version; manual-review flag

Note. Source products define measurements and context separately; published articles motivate methods but do not supply empirical outcomes.

Together, these sources provide a traceable source-to-variable record without implying public release or exact reconstruction of the derived linkage and analytical files; detailed provenance is reported in Supplementary Table S1.

2.3. Environmental indicators and LST

The predictor set is deliberately parsimonious and physically interpretable. LST is the dependent variable. NDVI represents photosynthetically active green cover; NDMI describes vegetation and surface-moisture conditions; NDBI captures the spectral signal of built-up and impervious surfaces; NDBSI captures exposed bare or dry soil-like surfaces; and a shortwave-reflectance transformation is retained as an albedo proxy for relative material reflectance. The NDVI formulation follows Tucker (1979); the near-infrared–shortwave-infrared moisture formulation follows Gao (1996) and is reported here as NDMI; NDBI follows Zha et al. (2003); NDBSI follows Liu et al. (2022). The fifth predictor is intentionally

interpreted more narrowly than an absolute surface-albedo product because the objective is within-study predictive comparison rather than standalone albedo retrieval.

LST is calculated from ST_B10 using the Collection 2 scale and offset and converted to degrees Celsius. The spectral indices use scaled surface reflectance, where BLUE, RED, NIR, SWIR1, and SWIR2 denote the relevant Landsat 8 bands. Equation (6) preserves the archived mapped-band shortwave-reflectance transformation. Its conceptual origin is Liang's (2001) narrowband-to-broadband approach, but those coefficients were developed for TM/ETM+, and validated OLI albedo retrieval can require sensor-specific spectral-response, atmospheric, and BRDF treatment (He et al., 2018). We therefore label Eq. (6) an albedo proxy for relative material-reflectance screening rather than an absolute Landsat-8 OLI broadband-albedo product; residual physically implausible values are removed under the documented quality rule.

$$LST = (ST_{B10} \times 0.00341802 + 149.0) - 273.15 \quad (1)$$

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \quad (2)$$

$$NDMI = \frac{\rho_{NIR} - \rho_{SWIR1}}{\rho_{NIR} + \rho_{SWIR1}} \quad (3)$$

$$NDBI = \frac{\rho_{SWIR1} - \rho_{NIR}}{\rho_{SWIR1} + \rho_{NIR}} \quad (4)$$

$$NDBSI = \frac{(\rho_{RED} + \rho_{SWIR1}) - (\rho_{NIR} + \rho_{BLUE})}{(\rho_{RED} + \rho_{SWIR1}) + (\rho_{NIR} + \rho_{BLUE})} \quad (5)$$

$$\alpha_{proxy} = 0.365\rho_{BLUE} + 0.130\rho_{RED} + 0.373\rho_{NIR} + 0.085\rho_{SWIR1} + 0.072\rho_{SWIR2} - 0.0018 \quad (6)$$

Table 2 links the six analytical variables to their physical interpretation and to the bounded design domain used later in the paper.

Table 2. Variables, physical interpretation and site-audit translation.

Variable	Analytical role	Physical interpretation	Site-audit interpretation (not causal claim)
LST	Dependent variable	Surface radiometric temperature	Outcome used to rank school-ground thermal conditions
NDVI	Predictor	Green vegetation signal	Tree/vegetation cover and planted surfaces
NDMI	Predictor	Moisture in vegetation/surface system	Moisture-management conditions, moisture retention, rain gardens, planted soils
NDBI	Predictor	Built/impervious spectral signal	Hardscape and built intensity; potential de-paving audit
NDBSI	Predictor	Bare/dry soil-like surface signal	Exposed dry ground and low-cover areas
Albedo proxy	Predictor	Relative shortwave-reflectance signal	Material-reflectance audit; evaluate alongside shade and glare

Note. Design translations are screening domains only; they are not intervention coefficients or estimates of temperature reduction.

2.4. Spatio-temporal trend analysis

For each school and year, the principal descriptive outcome is the median summer LST across all valid pixels whose centroids fall inside the campus polygon. Median rather than mean aggregation reduces sensitivity to edge pixels and extreme retrievals. Current thermal level and longitudinal change are analyzed separately: a school may have recurrently high summer LST without a steep trend, whereas another may begin near the city distribution but show a positive monotonic trend. The first dimension informs current screening priority; the second identifies statistically supported change.

Long-term monotonic change is estimated with the Theil-Sen slope and standard Mann-Kendall test (Mann, 1945; Sen, 1968). Each school contributes 7-11 annual summer composites. Theil-Sen β is the median pairwise yearly slope; positive values indicate increasing observed LST. Benjamini-Hochberg adjustment is applied to school-specific Mann-Kendall p values within each city, and reported q values therefore refer to standard-MK inference with within-city multiplicity control. No autocorrelation-robust Mann-Kendall sensitivity is claimed. The >=7-of-11 completeness rule is fixed across cities, and trends use observed annual school-ground LST without city-year background normalization; Sen slopes therefore describe total observed campus trajectories rather than campus-specific physical change isolated from interannual climate variability.

$$\beta = \text{median}_{i<j}\left(\frac{x_j - x_i}{t_j - t_i}\right) \quad (7)$$

$$S = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{sign}\left(x_j - x_i\right) \quad (8)$$

2.5. Correlation analysis

Pearson correlations between LST and the five predictors provide a descriptive bridge to nonlinear modeling. Primary correlations are city-level summaries; LCZ- or year-specific matrices remain exploratory when complete, adequately populated strata are unavailable. All five predictors are retained regardless of bivariate significance, so correlations are used to describe direction, magnitude, and inter-predictor dependence rather than for variable selection.

2.6. Machine-learning model development

Seven supervised regression algorithms are compared under an identical five-predictor input set: Decision Tree (DT), Random Forest (RF), Extra Trees Regressor (ET), Extreme Gradient Boosting (XGBoost), CatBoost, LightGBM and Gradient Boosting Regressor (GBR). LST is the dependent variable and NDVI, NDMI, NDBI, NDBSI and the albedo proxy are the only core predictors. The five-variable design keeps the analysis parsimonious, enables year-by-year attribution, and avoids a common problem in urban machine-learning studies in which dozens of correlated variables produce high predictive scores but weak design interpretation.

The seven algorithms span a simple decision-tree baseline, bagged/randomized tree ensembles and three modern boosting implementations, providing a controlled comparison across commonly used nonlinear model families (Breiman, 2001; Friedman, 2001; Geurts et al., 2006; Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018). The objective is not algorithmic novelty but a like-for-like test under an identical predictor set and a validation rule aligned with the school unit.

Model settings are selected only from each outer training partition; held-out schools are excluded from setting selection. Supplementary Table S2 reports the compact candidate grids. The retained archive verifies this training-only boundary but does not preserve the exact inner resampling rule or selected setting for every outer fold, so we do not claim a fully nested-CV procedure or reconstruct unverified details.

Model selection prioritizes the highest mean school-grouped R^2, with RMSE, MAE, and absolute MBE used to assess near-ties and error behavior. After model-family selection, feature importance is summarized across grouped folds. For Extra Trees, values are normalized impurity-based feature_importances_ (Geurts et al., 2006; Pedregosa et al., 2011), interpreted conditionally on the five-predictor specification rather than as separable physical effects.

The complete analytical sequence is summarized in Figure 3 to make the separation between data assembly, prediction and design interpretation explicit.

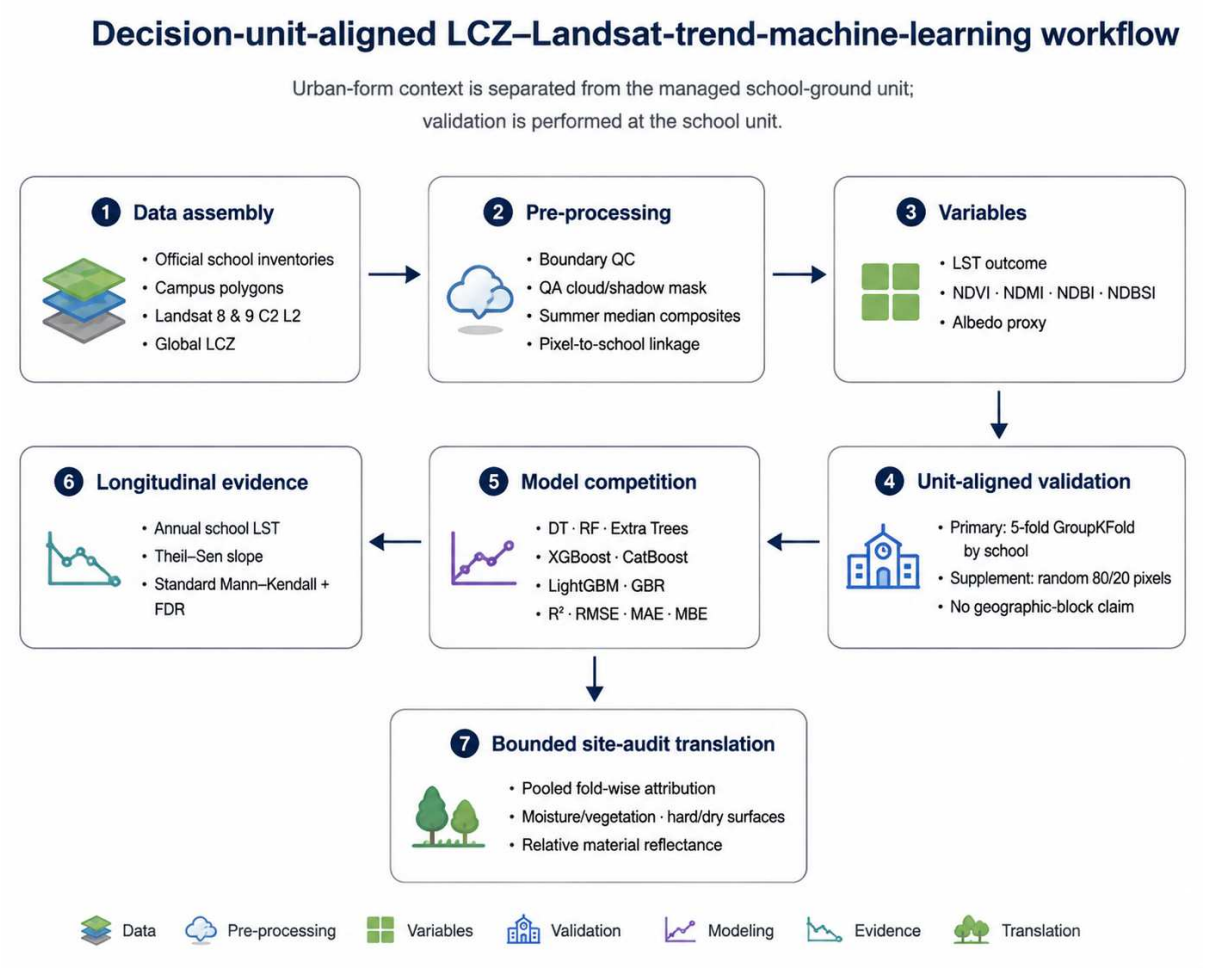


Figure 3. Decision-unit-aligned analytical workflow. LCZ describes surrounding urban-form context, the campus is the managed analytical unit, and primary predictive validation is performed on held-out schools rather than randomly mixed pixels.

Note. Solid workflow steps represent the main analysis; school-grouped validation is primary, the random 80/20 pixel split is supplementary, and no geographic-block or unseen-year transfer claim is made.

The workflow keeps predictive evaluation upstream of design translation so that applied recommendations cannot feed back into model selection.

2.7. School-grouped validation and model evaluation

Validation is organized by school rather than pixel because neighboring rows from one campus share vegetation, materials, morphology, and acquisition conditions. Following geospatial validation guidance, assessment should reflect both the prediction domain and data dependence (Roberts et al., 2017; Koldasbayeva et al., 2024; Stock, 2025). School grouping removes within-school train-test overlap but does not enforce geographic separation between schools; the claim is therefore held-out-school prediction, not geographic transfer.

Primary evaluation uses five-fold GroupKFold with school_id as the grouping variable. All rows from a school remain in one fold; city and LCZ are metadata rather than

stratification variables. Metrics are calculated over held-out pixel-year rows, so larger campuses or school-years with more valid cells contribute more observations. The estimand is therefore prediction for rows from unseen schools, not equal-school-weighted accuracy or geographic-block performance.

A random 80/20 pixel split is retained only as a supplementary interpolation benchmark. It is executed with a fixed seed to quantify how apparent performance changes when within-school dependence is not controlled. No separate geographic-block or blocked-year validation is used for inference; accordingly, the study does not claim transfer to geographically separated school clusters or to unseen years. Longitudinal inference is confined to the school-level Theil–Sen/Mann–Kendall trend analysis.

Four complementary metrics quantify predictive performance. R^2 measures explained variation; RMSE gives greater weight to large errors; MAE reports average absolute error in degrees Celsius; and MBE indicates systematic over- or underprediction. MBE is defined as predicted minus observed, so positive values indicate average overprediction. The primary performance claim uses the school-grouped folds and reports both means and fold dispersion.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (11)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (12)$$

2.8. Feature importance and site-audit translation

After model selection, feature importance is computed within each cross-validation fold and summarized across folds. The primary reporting object is the pooled fold distribution and resulting rank. City- and LCZ-stratified summaries are considered exploratory unless the relevant source table contains enough distinct schools to support reliable estimation. Because the five predictors are moderately intercorrelated (maximum observed $|r|=0.68$), individual impurity-based importance values may share attribution across correlated signals. This hierarchy prevents sparse subgroup patterns or small rank differences from being presented as universal design rules.

Interpretation is explicitly predictive. A high NDVI importance does not prove that planting a given number of trees will cause a specified temperature reduction, because vegetation is spatially co-determined with school age, neighborhood morphology, irrigation and other unobserved factors. Instead, it identifies vegetation signal as a strong predictor of

observed LST variation conditional on the five-variable model. The site-audit translation therefore uses 'priority for evaluation' language. Higher pooled NDVI importance identifies tree canopy and vegetated surface as a priority for follow-up audit; higher pooled NDMI importance points toward planted-soil moisture, water-sensitive landscaping and moisture-management conditions for follow-up inspection; NDBI/NDBSI importance points toward hard or dry exposed surfaces; the albedo-proxy importance indicates that relative material reflectance warrants attention together with shade, glare and radiant-exposure considerations. These interpretations concern relative predictive attribution within the fitted model, not uniquely separable physical effects.

Site-audit translation is based on the pooled feature ranking and, where fully documented, on descriptive contextual patterns. The confirmatory analysis uses the selected model's fold-wise impurity-based importance only; permutation importance and SHAP are not reported as independent attribution estimators. Consequently, the paper does not claim uniquely separable feature effects. The applied contribution is the stable hierarchy reported by the implemented model and its bounded translation into follow-up audit domains, with correlated-predictor attribution treated as a limitation rather than converted into a causal interpretation. Figure 4 provides a field-ground illustration of the heterogeneous surface conditions that motivate this facility-scale interpretation, showing why campus-level assessment is needed when translating remotely sensed indicators into practical audit domains. Supplementary Figure S1 provides documentary context from public-school environments in both study cities.



Figure 4. Field-ground example illustrating heterogeneous school-ground surfaces relevant to facility-scale heat assessment

Note. The photograph shows a publicly accessible view of a typical school campus environment in Barcelona and is provided for contextual illustration rather than empirical measurement. The image highlights heterogeneous surface conditions within a school boundary, including shaded vegetation, paved areas and built structures, which may not be fully resolved by 30 m Landsat pixels. The purpose of the figure is to illustrate why facility-scale assessment and site-audit interpretation are required when translating remotely sensed heat evidence into management-relevant priorities.

2.9. Validity, dependence and inference boundaries

The analysis is predictive rather than causal. The estimand is the association between five remotely sensed surface indicators and school-ground LST under the stated sampling and validation structure. Common environmental drivers—including drought, irrigation, renovation history and surrounding morphology—can influence both predictors and LST, so feature importance is not interpreted as an intervention effect.

Validity safeguards align the unit of analysis with the unit of claim: the same five predictors are used across models; school_id defines validation groups; annual campus summaries define trend series; and LCZ remains contextual metadata. These choices reduce avoidable train–test overlap and scope expansion but do not remove residual cross-school spatial dependence or unmeasured common causes.

Dependence is handled at the school level for descriptive uncertainty and temporal analysis. School-level contrasts use school-cluster bootstrap intervals, and longitudinal trends use annual campus summaries rather than treating all pixels as independent time-series observations.

Supplementary Table S3 consolidates the principal validity threats, controls and unresolved limits; Supplementary Table S5 separately maps each major claim to the evidence actually implemented.

2.10. Transparency and software environment

Transparency is provided through explicit source identifiers and workflow documentation rather than a public replication package. The Supplement records data lineage, QA and scaling rules, formulas, software versions, tuning grids, validation boundaries, and the claim-scope audit. Derived linkage files, analytical panels, and internal working artifacts are not deposited; Supplementary Table S4 lists the software environment.

The underlying Landsat, LCZ and official school-registry products can be obtained from the public source identifiers cited in the manuscript and Supplement. No previously published temperature estimate, model output or feature-importance value is reused as an empirical outcome, and no individual-level, student-level or confidential personal data are included.

2.11. Uncertainty and sensitivity reporting

Uncertainty is represented with school-cluster bootstrap 95% confidence intervals (2,000 resamples), fold dispersion for predictive metrics, and within-city Benjamini-Hochberg control for standard Mann-Kendall tests. Unretained analyses - including geographic blocking, city-specific predictive validation, equal-school weighting, alternative trend specifications, a second attribution estimator, and albedo-coefficient sensitivity - are not reconstructed post hoc; Supplementary Tables S3 and S5 define the resulting claim boundaries.

3. Results

Results report the completed analysis outputs and are interpreted within the observational, satellite-based design described above. Stratum-level estimates that could not be reconstructed from the retained source tables were not reported.

3.1. School cohort and LCZ context

The official screening frame contained 180 Barcelona records and 380 Madrid records meeting the public-primary-school definition. Application of the eligibility, boundary-quality, minimum-pixel and longitudinal-completeness rules produced a final cohort of 158 Barcelona schools and 276 Madrid schools (434 schools in total), corresponding to retention of 87.8% and 72.6% of the screened frames. The retained campuses covered 3.34 km² and contributed 21,438 complete pixel-year rows over 2014–2024. LCZ 2 (compact mid-rise) was the modal contextual class in Barcelona and LCZ 6 (open low-rise) in Madrid. LCZ is used here to describe surrounding urban form; no confirmatory class-specific effect is inferred from the modal-class summaries. Because a complete reason-coded intermediate disposition table was not available for all screened records, the analysis reports the fixed eligibility rules, screened frames, final cohorts and retention rates without reconstructing incomplete exclusion subcategories. The ≥7-of-11 annual-composite rule defines the longitudinal cohort; alternative completeness thresholds are not used to generate additional confirmatory trend claims.

Table 3 establishes the analytical denominators used for all subsequent school-level and pixel-year summaries.

Table 3. Analytical cohort, retention, campus coverage, and valid remote-sensing observations. B = Barcelona; M = Madrid.

City	Official records screened	Final schools	Retention (%)	Campus area (km²)	Valid pixel-year rows
B	180	158	87.8	0.96	8,532
M	380	276	72.6	2.38	12,906
Total	560	434	77.5	3.34	21,438

Note. Retention is calculated from the screened official frames. Intermediate match-stage counts are omitted because a complete reason-coded disposition table was not available for all screened records.

3.2. Annual summer LST and school-level trends

Across the 2014–2024 observation window, the median of school-level annual summer LST values was 31.8 °C (IQR 29.7–33.5) in Barcelona and 37.2 °C (IQR 34.7–39.1) in Madrid. The between-city median difference was 5.4 °C (school-cluster bootstrap 95% CI 4.9–5.9). Median Theil–Sen slopes were 0.09 °C/year in Barcelona and 0.18 °C/year in Madrid, with school-level slopes ranging from -0.08 to 0.45 °C/year and -0.05 to 0.56 °C/year, respectively. Under the implemented standard Mann–Kendall test with within-city Benjamini–Hochberg correction, 37 of 158 Barcelona schools (23.4%) and 112 of 276 Madrid schools (40.6%) had FDR-positive monotonic trends ($q<0.05$); 3 (1.9%) and 5 (1.8%)

schools, respectively, had FDR-negative trends. These classifications describe the implemented standard-MK evidence and are not presented as autocorrelation-robust estimates of campus-specific physical change. Current thermal level and longitudinal trend are therefore retained as separate evidence dimensions, with no additional rapid-warming threshold imposed.

The level and trend components of school-ground LST are reported together in Table 4 but interpreted as distinct evidence dimensions.

Table 4. School-ground summer LST level and 2014–2024 standard-MK trend summary. B = Barcelona; M = Madrid.

City	Median LST (°C)	IQR (°C)	Median Sen slope (°C/yr)	Slope range	Standard-MK FDR-positive	Standard-MK FDR-negative	Positive-trend share (%)
B	31.8	29.7–33.5	0.09	-0.08 to 0.45	37	3	23.4
M	37.2	34.7–39.1	0.18	-0.05 to 0.56	112	5	40.6

Note. Positive/negative trend classification uses the standard Mann–Kendall test with Benjamini–Hochberg $q<0.05$ within city. Positive-trend share equals FDR-positive schools divided by the final city cohort; no autocorrelation-robust trend significance is claimed.

3.3. Environmental indicators and bivariate relationships

The five surface indicators showed clear city-level differences. In Barcelona, the median NDVI, NDMI, NDBI, NDBSI and albedo-proxy values were 0.21, 0.04, -0.13, -0.11 and 0.18, respectively; the corresponding Madrid medians were 0.17, -0.02, -0.09, -0.09 and 0.21. The strongest reported bivariate association with LST in Barcelona was NDVI (r=-0.58), whereas in Madrid it was NDMI (r=-0.62). City-level descriptive results showed negative associations for NDVI and NDMI and generally positive associations for NDBI and NDBSI; the albedo proxy showed a mixed pattern in exploratory stratification. Because inter-predictor correlations reached |r|=0.68, these coefficients are treated as descriptive and are not substituted for multivariable model importance.

3.4. Model performance under school-grouped validation

Under school-grouped cross-validation, Extra Trees achieved the highest mean R^2 for pixel-year observations from held-out schools (0.78±0.05), with RMSE 2.41±0.18 °C, MAE 1.83±0.15 °C and MBE -0.05±0.21 °C. Fold-level R^2 ranged from 0.72 to 0.83, and Random Forest ranked second (R^2=0.75±0.05). In the supplementary random 80/20 pixel split, Extra Trees achieved R^2=0.89 and RMSE=1.72 °C, a 0.11 increase in R^2 relative to the school-grouped estimate. The change in apparent performance is larger than the 0.03 gap between Extra Trees and Random Forest under grouped validation, indicating that the evaluation structure matters more than the modest difference between the two leading algorithms. Because metrics are computed over held-out pixel-year rows, they should not be read as equal-weighted school-level accuracy estimates.

Model comparison under the primary school-grouped validation scheme is summarized in Table 5.

Table 5. School-grouped cross-validation performance for pixel-year observations from held-out schools.

Model	R^2 mean±SD	RMSE (°C)	MAE (°C)	MBE (°C)	Rank
DT	0.51±0.09	3.45±0.25	2.68±0.21	0.02±0.31	7
RF	0.75±0.05	2.52±0.17	1.91±0.14	-0.02±0.19	2
ET	0.78±0.05	2.41±0.18	1.83±0.15	-0.05±0.21	1
XGB	0.71±0.06	2.68±0.20	2.05±0.18	0.03±0.22	4
CAT	0.73±0.06	2.60±0.19	1.97±0.16	0.00±0.20	3
LGB	0.70±0.07	2.72±0.21	2.08±0.19	0.01±0.23	5
GBR	0.65±0.07	2.95±0.23	2.27±0.20	-0.01±0.25	6

Note. Values are mean±SD across grouped folds and are calculated over pixel-year observations belonging to held-out schools; R^2 is the primary ranking criterion, with RMSE, MAE and absolute MBE used to assess error magnitude and bias. These are not equal-weighted school-level accuracy estimates.

Extra Trees ranks first, although its margin over Random Forest is modest, so the technical contribution rests on validation discipline rather than algorithm novelty.

3.5. Feature importance and pooled ranking

For the selected Extra Trees model, the normalized impurity-based importance summary ranked NDMI highest (median 0.29; IQR 0.24–0.35), followed by NDVI (0.27; IQR 0.22–0.32), NDBI (0.19; IQR 0.15–0.23), the albedo proxy (0.13; IQR 0.10–0.17) and NDBSI (0.12; IQR 0.09–0.16). The close first- and second-ranked values support a combined moisture–vegetation screening emphasis rather than a single deterministic driver. Because inter-predictor correlations reached $|r|=0.68$ and no independent attribution estimator is reported, individual importance values are interpreted as relative predictive attribution within the five-predictor Extra Trees model. Primary inference is based on the pooled fold distribution; city×LCZ×year-specific patterns are not used for confirmatory interpretation because complete and sufficiently populated strata were not consistently available. Table 6 summarizes the pooled ranking and bounded audit interpretation.

Table 6. Pooled feature importance and bounded site-audit interpretation.

Var.	Median importance	IQR	Pooled rank	Site-audit domain	Interpretation boundary
NDVI	0.27	0.22–0.32	2	Vegetation/canopy audit	Predictive priority; not causal
NDMI	0.29	0.24–0.35	1	Moisture/planted-soil audit	Predictive priority; not causal
NDBI	0.19	0.15–0.23	3	Hardscape/built-surface audit	Predictive priority; not causal
NDBSI	0.12	0.09–0.16	5	Dry/bare-surface audit	Predictive priority; not causal
Albedo proxy	0.13	0.10–0.17	4	Material-reflectance audit	Predictive priority; not causal

Note. Importance values are normalized impurity-based predictive summaries from the Extra Trees model. Because predictors are correlated, pooled ranks are conditional on the five-variable specification and do not estimate intervention effects or universal LCZ-specific priorities.

3.6. Design-oriented site-audit priorities

The evidence-to-audit translation remains conditional. The pooled prominence of NDMI and NDVI makes planted-soil moisture, moisture-management conditions, tree canopy and vegetated-surface coverage the first domains for follow-up site audit. NDBI and NDBSI provide secondary screening signals for hardscape and exposed dry-surface review; the albedo proxy identifies relative material reflectance as a context-dependent audit domain that should be assessed together with shade and radiant exposure. These are site-audit priorities only. They indicate where field measurement, detailed microclimate simulation or intervention evaluation should be concentrated and do not estimate the causal temperature effect of any retrofit.

Figure 5 consolidates the four quantitative result layers used in the subsequent discussion.

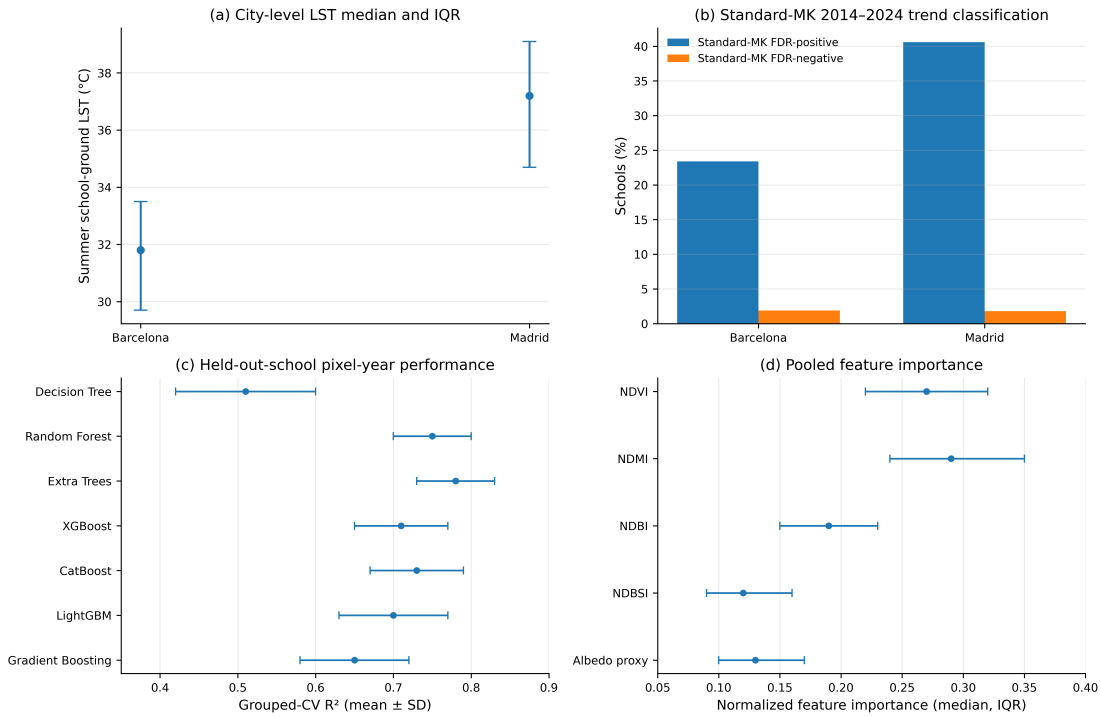


Figure 5. Core empirical results: (a) city median school-ground summer LST (IQR); (b) standard-MK FDR-positive/negative trend shares; (c) grouped-CV R^2 (mean±SD) for held-out-school pixel-year observations; and (d) Extra Trees feature importance (median, IQR).

Note. Panels correspond to Tables 4–6. Panel (b) reports standard-MK classifications under within-city FDR control, not autocorrelation-robust trend detections; panel (c) reports pixel-year performance within held-out schools, not equal-school-weighted accuracy; no panel estimates a causal intervention effect.

4. Discussion

4.1. Main empirical findings

Three results anchor the discussion. Madrid school grounds were hotter and showed steeper observed summer trajectories than Barcelona; moisture and vegetation formed the leading pooled attribution group; and Extra Trees ranked first under school-grouped validation while random pixel splitting produced materially higher apparent R^2. The most consequential finding is therefore not the winning algorithm. It is that the substantive evidence, the validation design and the applied interpretation change when the school is treated consistently as the managed decision unit.

4.2. RQ1: School-ground thermal levels and longitudinal trends

RQ1 asked how summer LST levels and longitudinal trends varied across public primary-school grounds in the two cities. Madrid schools had a higher median summer LST (37.2 vs 31.8 °C; difference 5.4 °C, 95% CI 4.9–5.9) and a higher share of standard-MK

FDR-positive trends (40.6% vs 23.4%). Median Theil–Sen slopes were 0.18 °C/year in Madrid and 0.09 °C/year in Barcelona. These results justify treating current thermal level and longitudinal change as separate screening dimensions: a school can be currently hot without a statistically supported monotonic increase, and vice versa. LCZ 6 was modal in Madrid and LCZ 2 in Barcelona, but LCZ is used only as surrounding urban-form context. Because annual LST was not normalized against city-year meteorological or background-temperature anomalies, the Sen slopes represent total observed campus trajectories rather than campus-specific physical change. These q-based classifications refer to the implemented standard Mann–Kendall test; no autocorrelation-robust trend significance is inferred.

4.3. RQ2: Surface indicators and predictive attribution

RQ2 concerned the direction of surface relationships and the relative predictive contribution of the five measured indicators. Barcelona's strongest reported city-level bivariate association with LST was NDVI (r=-0.58), whereas Madrid's was NDMI (r=-0.62). In pooled Extra Trees attribution, NDMI (0.29) and NDVI (0.27) ranked first and second, followed by NDBI (0.19), the albedo proxy (0.13) and NDBSI (0.12). Because inter-predictor correlations reached $|r|=0.68$ and the reported importance is impurity-based, this ordering is model-conditional; it does not identify uniquely separable physical mechanisms or causal effects. The appropriate applied interpretation is therefore a combined moisture–vegetation site-audit emphasis, with hard/dry surfaces and relative material reflectance as secondary domains for follow-up inspection.

4.4. RQ3: Held-out-school predictive performance under grouped validation

RQ3 asked which regressor most reliably predicted pixel-year observations from schools not represented in a training fold. Extra Trees achieved $R^2=0.78\pm0.05$ under school-grouped validation, compared with $R^2=0.89$ under random pixel splitting. The 0.11 difference is larger than the 0.03 gap between Extra Trees and Random Forest under the grouped design, making validation structure more consequential than the small algorithm ranking difference. This interpretation is consistent with current geospatial-ML guidance that evaluation should match the intended prediction domain and dependence structure (Roberts et al., 2017; Koldasbayeva et al., 2024; Stock, 2025). Here the separation unit is school identity, not geographic blocks; different training and test schools can remain spatially proximate. Metrics also remain row-weighted by valid pixel-year observations and pooled across the two cities. The technical contribution is therefore prevention of within-school train–test overlap, not proof of geographic, city-specific, external or school-equalized transfer.

This validation refinement extends the city-scale workflow of Tahiri et al. (2026), who reported strong tree-ensemble performance while noting a training–testing gap and the value of spatio-temporal validation. The present study keeps the same broad model family but changes the primary validation unit to the school. Extra Trees is selected because it performed best under that rule, not because the analysis was designed to reproduce a predetermined winning algorithm.

4.5. Facility-scale interpretation and site-audit relevance

The facility-scale contribution is best understood as decision-unit alignment. Urban heat mapping typically produces evidence for pixels, neighborhoods or morphological zones, whereas public adaptation is implemented through assets with identifiable boundaries, managers and budgets. Recent implementation-oriented urban-heat scholarship argues that the persistent bottleneck is often the translation of scientific diagnosis into planning instruments and operational action (Lau et al., 2025). A school-ground screening layer can therefore function as a practical interface between remote-sensing evidence and municipal asset management: it does not prescribe an intervention, but identifies where finer measurement or design review should occur. In this study, LCZ describes surrounding morphology, the campus polygon defines the managed facility, annual summaries define longitudinal evidence, school_id defines the prediction holdout, and the same unit anchors site-audit recommendations. This alignment avoids a scale jump in which city-level or randomly validated pixel evidence is translated directly into facility-level action. The pooled prominence of NDMI and NDVI prioritizes planted-soil moisture and vegetation/canopy for follow-up assessment, while NDBI/NDBSI and the albedo proxy identify hard/dry surfaces and relative material reflectance as secondary domains. The logic is compatible with school climate-shelter, naturalization and community-use programs, where design, maintenance and public use are coupled at the site level (López Plazas et al., 2023; Sekulova & Ruiz-Mallén, 2024; Sanz-Mas, Continente, Marí-Dell'Olmo, & López, 2025).

This alignment offers three practical advantages. First, it makes the evidence chain traceable from public school identity to campus geometry and then to repeated satellite observations, helping distinguish uncertainty in site definition from uncertainty in thermal measurement. Second, it ties model evaluation to the same unit used for operational prioritization, reducing the risk that apparent skill is driven by repeated observations from a site that also appears in training. Third, it establishes a defensible stopping rule for interpretation: remote sensing can identify facilities and audit domains, but higher-resolution field measurements, microclimate simulation or intervention studies are required before specifying design coefficients or human-exposure effects. Facility-scale screening is therefore neither a substitute for microclimate modeling nor a claim that Landsat resolves pedestrian heat exposure. It is an intermediate evidence layer that can reduce the search space for more expensive follow-up assessment. For municipalities managing large school portfolios, this distinction is operationally important: the immediate question is often not whether vegetation or moisture matters in general, but which facilities merit the next round of measurement, maintenance review or design attention.

4.6. Cross-city scope and terminology

Barcelona and Madrid differ in observed LST level, trend prevalence and leading bivariate correlation, so a single surface signature should not be assumed to transfer unchanged between cities. Moisture and vegetation nevertheless remain prominent in the pooled attribution, supporting a common first-stage audit emphasis while leaving city- or LCZ-specific design rules to separately validated analyses. Accordingly, Table 6 is used to define bounded audit domains rather than universal intervention prescriptions.

The contrast between cities also suggests a two-stage use of the workflow. A common first stage can apply identical data definitions, seasonal windows and validation rules to preserve comparability. A second stage can then introduce city-specific thresholds, background-climate normalization, morphology-sensitive covariates and local implementation criteria. This separation is preferable to forcing one pooled design rule across

climates or, at the other extreme, treating each city as methodologically unique. The present pooled model is therefore best understood as a common screening layer, with local calibration and external validation reserved for subsequent implementation studies.

Terminology remains deliberately narrow: 'school-ground surface thermal environment' refers to satellite-observed LST, while current thermal level is distinguished from standard-MK FDR-positive or FDR-negative monotonic trends. Neither term implies student heat stress, school resilience or retrofit effectiveness.

4.7. Limitations and validation needs

Several limitations bound interpretation. Landsat surface temperature is distributed on a 30-m grid but derives from coarser native TIRS measurements, so fine shade structures, narrow planting strips and human-scale thermal exposure cannot be resolved; the ~100-m LCZ product is likewise contextual rather than a record of every local morphological change. The five-predictor specification omits variables such as building height, sky-view factor, wind, anthropogenic heat, soil composition and direct irrigation. Eq. (6) is an archived shortwave-reflectance albedo proxy rather than a validated OLI broadband-albedo product, and no retained 0.356-versus-0.365 sensitivity run is available. Longitudinal inference uses unnormalized annual LST and standard rather than autocorrelation-robust Mann–Kendall testing. School-grouped validation removes within-school overlap but not inter-school spatial dependence and does not provide city-specific, equal-school-weighted or geographic-block performance; the exact inner tuning resampling rule is also not recoverable from the retained archive. Finally, impurity-based importance with correlated predictors remains model-conditional. Supplementary Tables S3 and S5 provide the complete claim-to-risk audit.

These limitations define validation needs rather than reasons to expand the claims. Higher-resolution field measurements would be required to evaluate air temperature, humidity, radiant exposure and human-scale thermal conditions, while documented green/blue/grey retrofits would require a separate quasi-experimental design to estimate intervention effects. Both questions remain outside the inferential scope of the present observational analysis.

5. Conclusions

Using 434 public primary-school grounds and 21,438 complete pixel-year observations, this study reorganizes an LCZ–Landsat–trend–machine-learning workflow around the public facility as the decision unit. Madrid had higher median school-ground summer LST than Barcelona (37.2 vs 31.8 °C; difference 5.4 °C, 95% CI 4.9–5.9), steeper median Theil–Sen slopes (0.18 vs 0.09 °C/year), and a larger share of standard-MK FDR-positive monotonic trends (40.6% vs 23.4%). Extra Trees provided the strongest held-out-school prediction under school-grouped validation (R^2=0.78±0.05; RMSE=2.41±0.18 °C), while random pixel splitting yielded R^2=0.89. NDMI and NDVI formed the leading pooled attribution group. Together, these results show that facility-scale screening depends not only on what is modeled but also on whether evidence and validation are aligned with the managed asset.

The practical implication is bounded to prioritization. Moisture-management conditions and vegetation/canopy are the first domains for follow-up audit, with hard/dry surfaces and relative material reflectance as secondary domains. These priorities indicate where higher-

resolution field verification or design assessment should begin; they do not quantify temperature reductions from planting, irrigation, de-paving or material changes, and they do not infer children's thermal comfort from satellite LST.

The conceptual contribution is decision-unit alignment: surrounding urban morphology, facility-level surface conditions, prediction validation and applied screening are organized around distinct but explicitly connected spatial units. The resulting evidence chain links city-scale thermal diagnosis to a managed public asset without treating contextual morphology, predictive attribution and retrofit effects as interchangeable. The present validation design prevents within-school train–test overlap but does not establish geographic-block or external transfer; broader design guidance requires independent field, spatial-validation or intervention evidence.

Declarations

Funding: No funding was received.

Competing interests: The author declares no competing interests.

Ethics approval: Ethics approval was not required because the study uses remote-sensing imagery and publicly available non-personal geospatial records only; no human participants or identifiable personal data were involved.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The underlying Landsat, LCZ, and official school-registry source products are publicly available through the identifiers cited in the manuscript and Supplementary Appendix. Derived campus-linkage files, analytical panels, fold-assignment files, executable scripts, and other internal working artifacts are not deposited with this submission.

Author contributions: J.S. is the sole author and is responsible for the conception, analysis, drafting, and revision of the manuscript.

Acknowledgements: None.

References

Antoniadis, D., Katsoulas, N., & Papanastasiou, D. K. (2020). Thermal environment of urban schoolyards: Current and future design with respect to children's thermal comfort. *Atmosphere*, *11*(11), 1144. https://doi.org/10.3390/atmos11111144

Baró, F., Camacho, D. A., Pérez del Pulgar, C., Triguero-Mas, M., & Anguelovski, I. (2021). School greening: Right or privilege? Examining urban nature within and around primary schools through an equity lens. *Landscape and Urban Planning*, *208*, 104019. https://doi.org/10.1016/j.landurbplan.2020.104019

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Ching, J., Mills, G., Bechtel, B., See, L., Feddema, J., Wang, X., Ren, C., Brousse, O., Martilli, A., Neophytou, M., Mouzourides, P., Stewart, I., Hanna, A., Ng, E., Foley, M., Alexander, P., Aliaga, D., Niyogi, D., Shreevastava, A., ... Theeuwes, N. (2018). WUDAPT: An urban weather, climate, and environmental modeling infrastructure for the Anthropocene. *Bulletin of the American Meteorological Society*, *99*(9), 1907–1924. https://doi.org/10.1175/BAMS-D-16-0236.1

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I. D., van Vliet, J., & Bechtel, B. (2022). A global map of local climate zones to support earth system modelling and urban-scale environmental science. *Earth System Science Data*, *14*, 3835–3873. https://doi.org/10.5194/essd-14-3835-2022

Flax, L., Korthals Altes, R., Kupers, R., & Mons, B. (2020). Greening schoolyards—An urban resilience perspective. *Cities*, *106*, 102890. https://doi.org/10.1016/j.cities.2020.102890

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gallez, E., Fraile Mujica, C. P., Gadeyne, S., Canters, F., & Baró, F. (2024). Nature-based school environments for all children? Comparing exposure to school-related green and blue infrastructure in four European cities. *Ecological Indicators*, *166*, 112374. https://doi.org/10.1016/j.ecolind.2024.112374

Gao, B.-C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

He, T., Liang, S., Wang, D., Cao, Y., Gao, F., Yu, Y., & Feng, M. (2018). Evaluating land surface albedo estimation from Landsat MSS, TM, ETM+, and OLI data based on the unified direct estimation approach. Remote Sensing of Environment, 204, 181–196. https://doi.org/10.1016/j.rse.2017.10.031

He, X., Gao, W., & Wang, R. (2024). Effect of urban structure on land surface temperature around elementary schools in Hangzhou based on local climate zones. *Sustainable Cities and Society*, *114*, 105724. https://doi.org/10.1016/j.scs.2024.105724

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and

practice. *Nature Communications*, *15*, 10700. https://doi.org/10.1038/s41467-024-55240-8

Lau, K. K. L., Yuan, C., & Ng, E. (2025). Urban heat island adaptation and mitigation in practice: Lessons from policy implementation in five cities. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *383*(2308), 20240581. https://doi.org/10.1098/rsta.2024.0581

Liang, S. (2001). Narrowband to broadband conversions of land surface albedo I: Algorithms. *Remote Sensing of Environment*, *76*(2), 213–238. https://doi.org/10.1016/S0034-4257(00)00205-4

Liu, Y., Meng, Q., Zhang, L., & Wu, C. (2022). NDBSI: A normalized difference bare soil index for remote sensing to improve bare soil mapping accuracy in urban and rural areas. *Catena*, *214*, 106265. https://doi.org/10.1016/j.catena.2022.106265

López Plazas, F., Crespo Sánchez, E., Llorca Pérez, R., & Santacana Albanilla, E. (2023). Schools as climate shelters: Design, implementation and monitoring methodology based on the Barcelona experience. *Journal of Cleaner Production*, *432*, 139588. https://doi.org/10.1016/j.jclepro.2023.139588

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, *13*(3), 245–259. https://doi.org/10.2307/1907187

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 6638–6648.

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., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Sanz-Mas, M., Continente, X., Brugueras, S., Marí-Dell'Olmo, M., Oliveras, L., & López, M. J. (2025). Evaluating the effect of green, blue, and gray measures for climate change adaptation on children's well-being in schoolyards in Barcelona. *Landscape and Urban Planning*, *253*, 105206. https://doi.org/10.1016/j.landurbplan.2024.105206

Sanz-Mas, M., Continente, X., Marí-Dell'Olmo, M., & López, M. J. (2025). Community use and perceptions of climate shelters in schoolyards in Barcelona. *International Journal of Public Health*, *70*, 1608083. https://doi.org/10.3389/ijph.2025.1608083

Sanz-Mas, M., Ubalde-López, M., Borràs, S., Brugueras, S., Continente, X., Daher, C., Marí-Dell'Olmo, M., & López, M. J. (2024). Adapting schools to climate change with green, blue, and grey measures in Barcelona: Study protocol of a mixed-method evaluation. *Journal of Urban Health*, *101*, 141–154. https://doi.org/10.1007/s11524-023-00814-y

Sekulova, F., & Ruiz-Mallén, I. (2024). The governance configurations of green schoolyards. *Environmental Science & Policy*, *156*, 103752. https://doi.org/10.1016/j.envsci.2024.103752

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12), 1879–1900. https://doi.org/10.1175/BAMS-D-11-00019.1

Stock, A. (2025). Choosing blocks for spatial cross-validation: Lessons from a marine remote sensing case study. *Frontiers in Remote Sensing*, *6*, 1531097. https://doi.org/10.3389/frsen.2025.1531097

Tahiri, M., Hidalgo-García, D., Rezapouraghdam, H., Mohsine, I., Bahi, H., Bouramtane, T., Karmouda, N., & Kacimi, I. (2026). What heats cosmopolitan cities in summer? A remote sensing and AI study of land surface temperature drivers: Casablanca vs Barcelona. *Sustainable Cities and Society*, *147*, 107545. https://doi.org/10.1016/j.scs.2026.107545

Terada, M., Ermilova, M., Noi, S., & Shikano, A. (2026). Adapting Tokyo's schoolyards to extreme heat and enrolment pressure: Perspectives from the frontline. *Urban Forestry & Urban Greening*, *120*, 129435. https://doi.org/10.1016/j.ufug.2026.129435

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, *24*(3), 583–594. https://doi.org/10.1080/01431160304987

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HotSchoolGroundsSupplementaryMaterial.docx