

Supplementary Information

Supplementary Note 1. Full robustness of the structure

Supplementary Table S1. Full robustness suite. The complete diagnostic suite behind the Robustness paragraph of the main text; each row is one check. Size-law slopes in °C per tenfold population, within-city and density terms in °C per log unit; 95% confidence intervals in parentheses, number of cities in italics. Sweep rows (start year, annulus, screen radius, lapse rate, income form) report the range of point estimates across settings. Intervals for the start-year (other than 1975 and 2000), country-clustered, climate-zone, exclusion, development-phase and cross-instrument rows of part (b) are Wald intervals from the reported test statistics; the central terms are the nocturnal size law (+0.66) and the within-city trend (+0.28 on 2000–2020, +0.22 on 1975–2020).

(a) Population-size law

check	result
Baseline (country-clustered standard errors, 26 clusters)	night +0.66 (+0.32 to +0.99) <i>506</i> ; daily mean +0.23 (+0.08 to +0.39) <i>685</i>
Extreme-bounds search (384 specifications)	robust in 100% of the 256 specifications in which city size appears
Mixed-effects model with a random country intercept	+0.80 (+0.56 to +1.04) <i>506</i>
Excluding the United States	+0.62 (+0.24 to +1.00) <i>302</i>
Excluding Europe / excluding Asia	+0.92 (+0.54 to +1.30) <i>325</i> / +0.67 (+0.31 to +1.04) <i>489</i>
City size sampled across the full range	6–14% of cities in every size class; not size-restricted
Incremental-covariate ladder (nine steps), size alone → full set	night +0.65 (+0.30 to +0.99) → +0.89 (+0.64 to +1.13) <i>502</i> ; daily mean +0.23 (+0.08 to +0.39) → +0.38 (+0.24 to +0.51) <i>679</i>

Urban-station distance bins and size × distance interaction	Supplementary Table S3
Rural-reference count ≥ 1 / ≥ 3 / ≥ 5 stations	+0.65 (+0.30 to +0.99) 502 / +0.65 (+0.30 to +0.99) 502 / +0.74 (+0.39 to +1.09) 453
Element-specific slope, day / daily mean / night	−0.16 (−0.34 to +0.01) 506 / +0.23 (+0.08 to +0.39) 685 / +0.66 (+0.32 to +0.99) 506
Income-control form (linear, income-group effects, quadratic)	night slope from +0.62 to +0.66; daily mean from +0.21 to +0.24
By climate zone (night) and size × zone joint test	Supplementary Table S3; joint test $p = 0.50$
RESET functional-form test	$F = 0.17$, $p = 0.68$; no misspecification
Breusch–Pagan test; HC3-robust slope	$p = 0.19$; +0.66 °C per tenfold population (+0.43 to +0.88 under HC3), unchanged
Cook’s distance, top 1% of cities dropped	+0.66 (+0.32 to +0.99) → +0.69 (+0.38 to +1.00) 501
Permutation of population (500 draws)	null 0.00 ± 0.13 against +0.66 observed

(b) Within-city trend

check	result
Start-year sweep, 1975 / 1980 / 1985 / 1990 / 1995 / 2000 / 2005	+0.22 (+0.14 to +0.29) / +0.22 (+0.15 to +0.30) / +0.22 (+0.13 to +0.32) / +0.22 (+0.12 to +0.32) / +0.21 (+0.10 to +0.32) / +0.28 (+0.15 to +0.42) / +0.27 (+0.12 to +0.43)

Rural annulus 25–75 ... 100–300 km (point estimates across annuli)	2000–2020 from +0.38 to +0.21; full record from +0.21 to +0.25
Rural-contamination screen radius 0–30 km (station-record rebuild, fitted reference)	2000–2020 from +0.25 (no screen) to +0.29, interval excluding zero at every radius; full record from +0.19 to +0.22
Lapse rate $\Gamma = 0 \dots 9.8 \text{ }^{\circ}\text{C km}^{-1}$	2000–2020 from +0.24 to +0.30; full record from +0.20 to +0.23
$\pm$ built-up fraction / $\pm$ GDP spline	full record from +0.22 to +0.24 with built-up fraction; 2000–2020 +0.23 with built-up fraction added (two-way fixed effects), +0.37 to +0.40 in the joint model with the income spline
+ built-up share and vegetation (NDVI) on the fixed footprint	Supplementary Table S6
+ building volume (GHS-BUILT-V) on the fixed footprint, 2000–2020	population +0.26 (+0.10 to +0.41), volume +0.02 (–0.02 to +0.06) 884; volume null at night, by day, in every region and climate zone (Supplementary Table S6)
Dropping any single country other than the United States, full record	from +0.20 to +0.24, every interval excluding zero
Excluding the United States, 2000–2020 / full record	+0.11 (–0.02 to +0.25) 595 / +0.09 (+0.01 to +0.18) 618
Excluding North America, 2000–2020 / full record	+0.05 (–0.07 to +0.18) 554 / +0.05 (–0.03 to +0.14) 579; collapses on both windows

Country-clustered instead of city-clustered standard errors, 2000–2020 / full record	+0.28 (+0.03 to +0.53) / +0.22 (+0.07 to +0.37)
City-cluster bootstrap, 2000–2020 / full record	+0.28 (+0.15 to +0.42) / +0.22 (+0.14 to +0.29)
Country-cluster bootstrap, 2000–2020 / full record	+0.28 (−0.01 to +0.48) / +0.22 (−0.04 to +0.30)
Single nearest station versus the mean of all stations within 25 km, full record	+0.22 versus +0.07; averaging re-introduces composition drift
Development phase, full record: not high-income / high-income	+0.90 (+0.44 to +1.35) 53 / +0.17 (+0.09 to +0.24) 750
By climate zone, full record: arid / temperate / continental	+0.23 (+0.07 to +0.39) 110 / +0.19 (+0.09 to +0.28) 357 / +0.15 (+0.00 to +0.30) 314
Driscoll–Kraay spatial-HAC standard errors, full record (Driscoll and Kraay 1998)	+0.19 (+0.15 to +0.22) 910
Placebo: density shuffled within city (300 draws)	2000–2020 0.00 ± 0.06 against +0.28 observed; full record 0.00 ± 0.02 against +0.22
Pesaran cross-sectional dependence, full record (Pesaran 2021)	mean pairwise residual $\rho \approx 0.00$
Independent archive: GHCN-M v4 QCU, the quality-controlled	+0.18 (−0.01 to +0.37) on GHCN-M against +0.19 (+0.01 to +0.37) on the matched GHCN-Daily series

unadjusted series (Menne et al. 2018)	681; this panel predates the duplicate-station rule, and with the cities that rule dropped removed it reads +0.21 (−0.02 to +0.44) against +0.17 (−0.05 to +0.39) 539
Cross-instrument: Yale Center for Earth Observation (YCEO) surface UHI v4, 2003–2018 (Chakraborty and Lee 2019, 2023)	within +0.13 (+0.06 to +0.20) night, −0.12 (−0.27 to +0.04) day; between +0.29 (+0.21 to +0.37) night, +0.18 (+0.03 to +0.33) day 1,049
Cross-instrument: Yang et al. (2024) canopy-layer urban heat island intensity (UHII) dataset	+0.07 (+0.03 to +0.11) daily mean, +0.10 (+0.06 to +0.14) night; night with density only +0.06 (+0.02 to +0.10) 4,904

(c) Between-city density (level), the mirror of the size law

check	result
Density alone / with size held, night	−0.35 (−0.83 to +0.12) / −0.63 (−1.08 to −0.18) 506
Nine-step covariate ladder	negative or null throughout, from −0.36 to −0.57; negative with interval excluding zero once size is held
By climate zone with size held, night: tropical / arid / temperate / continental	+0.06 (−1.43 to +1.54) 38 / −0.80 (−1.13 to −0.48) 95 / −0.52 (−1.11 to +0.06) 214 / −0.72 (−1.17 to −0.27) 159
By element with size held: daily mean / day	−0.29 (−0.62 to +0.05) 685 / −0.14 (−0.39 to +0.12) 506; the between-city term of the day/night joint model is positive by day, +0.85 (+0.49 to +1.21)

Level-versus-trend identification	2000–2020 +0.28 (+0.15 to +0.42) / –0.12; full record
(joint model), within / between	+0.22 (+0.14 to +0.29) / –0.01; equality rejected at $p = 1.4 \times 10^{-3}$ and 4.1×10^{-3}
Development phase	negative among mature cities; positive, co-scaling with size, only while urbanizing

S1.1 Regional dependence: the United States and North America

How much does the sample’s North American weight matter? For the size law, little; for the within-city trend, more than a country jackknife alone would suggest. North America holds 36% of the panel’s cities (31% United States). The size law survives excluding the United States and holds within each continent separately, and the within-city trend survives excluding the United States alone and any single other country. Dropping the whole of North America is a different matter: the trend collapses on both windows, not from lost precision but from an estimated relationship that genuinely differs by region. The size law’s regularities are properties of the urban sample broadly; the within-city trend’s are not.

Growth pace. One obvious alternative reading of the regional gap is that North American cities simply grew faster inside their existing footprints over the period, so a comparison restricted to comparably fast-growing cities would erase it. It does not: restricting to only those cities in each region whose within-city term grew at or above the North American median rate, North America still returns +0.14 °C per log unit (95% CI –0.01 to +0.28) while the rest of the world returns –0.02 °C (–0.16 to +0.11). On our panel the gap could not be attributed to pace.

What the regional gap could not be attributed to. Nine further explanations were tested on the primary window, each by re-estimating the within-city term for North America and for Europe under the design that would remove the gap if the explanation held. None accounts

for the gap on our panel. Supplementary Table S2 gives every estimate; the text follows the argument. The first explanation is a regional trend in the rural background, such as the slow mid-century warming of the rural United States, which would make North American heat islands appear to grow for reasons unrelated to the cities themselves. Country-by-year and climate-zone-by-year effects absorb any such trend, and the North American term survives them, strengthening rather than weakening when each climate zone carries its own trend; Europe remains null under the same designs. The second is station siting: if the term appeared only where urban stations sit close to the center, it might reflect siting rather than growth. Its point estimate is positive in North America whether the urban station sits within 5 km of the centroid or farther out, with intervals that widen in the halved samples, and it is absent in Europe in both bins. The third is the frozen footprint. Counting population on the moving GHS-WUP-MTUC boundary, so that residents a city acquires by extending its boundary are included, leaves Europe null and North America positive. The fourth is building mass, which S1.6 finds no support for in either region, as far as building volume measures it. The fifth is land cover: the built-up share of the footprint and the vegetation of the core, added to the panel, take almost none of the North American term (S1.6).

Two further explanations concern the settled landscape around a city rather than the city itself. Europe and populous Asia are, relative to North America, continuously settled, so a city there sits inside one large settled region rather than as an island in a sparse landscape, and an added resident might then change the urban–rural contrast less. The premise holds: on the full GHS-WUP-MTUC list of centers, the landscape within 500 km of a European city holds about three times the urban-center population per square kilometer of a North American one. The consequence the explanation predicts does not follow. Within North America the term is no smaller in the settled half of landscapes than in the sparse half, and European cities are null in both the sparse and the settled halves of the pooled distribution,

including those whose surroundings are no more settled than the settled North American cities. An interaction of the term with the log of landscape settlement is indistinguishable from zero in both regions. The narrower version of the same idea concerns the reference annulus itself: European countryside is more settled, and the 40–150 km annulus of a European city does hold more other urban centers, about twice as many with twice the population. If a reference warmed by neighboring cities were netting out the urban increment, the term should appear in European cities whose annulus is quiet and weaken in North American cities whose annulus is crowded. Neither pattern appears in the data. Europe is null in every tercile of annulus population, North America is positive in the middle tercile and not in the quiet one, and the interaction with annulus population is zero in both regions. The test counts urban centers, not diffuse settlement, so it does not exclude a slow, dispersed development of the European countryside below the urban-center threshold; the 12 km contamination screen (Methods) removes stations next to any center but not that.

Two more concern climate and activity. Within North America the term is largest in the temperate, hot-summer zone, with the continental and arid zones smaller and poorly identified, while Europe's temperate cities show none, so the same climate class gives different results on the two continents, and climate alone could not account for the gap. A night-time-lights test, intended to proxy activity per added resident, is inconclusive rather than negative. With variation restricted to within one sensor, night-time lights carry no coefficient anywhere and do not move the population term. But their within-city variation over the cores of these cities is very small, consistent with the known saturation of both sensors over bright developed-world cores. What the exclusions leave is a term that follows people rather than buildings, appears at night, and is concentrated where growth was fast. The Discussion of the main text sets out three hypotheses consistent with that description, anthropogenic heat per added resident, vegetation loss inside the footprint, and the materials

of the building stock, and S1.7 splits the term by season. A year-round night-time term speaks against cooling waste heat as the sole channel and against vegetation loss as the main one.

City-level energy, cooling and building-stock series are the data a follow-up needs to separate the two that remain.

Supplementary Table S2. The regional gap under each alternative design. Within-city coefficient on population added on the fixed footprint, °C per log unit, city-clustered standard errors, 2000–2020 unless stated. Rows are the design that would close the gap if the named explanation held; as pass/fail checks they retain p-values (Methods, Statistics).

design	North America	Europe
city and year effects (primary window)	+0.32 (p = 0.07), 331 cities	–0.03 (p = 0.84), 342
cities growing at or above the North American median only	+0.14 (p = 0.06), 166	rest of world –0.02 (p = 0.75), 88
city and country×year effects	+0.29 (p = 0.10)	+0.11 (p = 0.61)
city and climate-zone×year effects	+0.34 (p = 0.05)	–0.02 (p = 0.90)
the same, 1975–2020	+0.14 (p = 0.03)	–0.03 (p = 0.72)
United States only, climate-zone×year	+0.41 (p = 0.030), 279	
urban station ≤5 km from centroid	+0.37 (p = 0.19), 136	–0.11 (p = 0.65), 163
urban station 5–15 km	+0.16 (p = 0.50), 171	+0.10 (p = 0.53), 139
population on the moving MTUC boundary	+0.17 (p = 0.06), 232	–0.07 (p = 0.56), 321
building volume added (S1.6)	+0.27 (p = 0.12); volume +0.02 (p = 0.35)	–0.06 (p = 0.68); volume –0.05 (p = 0.36)
temperate zone only	+0.41 (p = 0.08), 113	–0.18 (p = 0.46), 162

continental zone only	+0.19 (p = 0.58), 137	+0.19 (p = 0.20), 156
arid zone only	+0.34 (p = 0.42), 69	
night-time lights added, within-sensor variation	+0.35 (p = 0.25); lights +0.02 (p = 0.85)	−0.22 (p = 0.22); lights −0.07 (p = 0.24)
reference annulus quiet / middle / crowded (terciles of other centers' population)	+0.02 / +0.66 / +0.12 (p = 0.96 / 0.04 / 0.69), 100 each	−0.03 / −0.01 / +0.14 (p = 0.86 / 0.98 / 0.71), 114 each
interaction with ln annulus population	−0.023 (p = 0.27)	+0.011 (p = 0.69)
surrounding 40–500 km landscape	+0.32 (p = 0.13), 224 /	+0.04 (p = 0.81), 110 /
sparse / settled (pooled median, 23 people per km ²)	+0.51 (p = 0.13), 75	−0.06 (p = 0.77), 232
interaction with ln landscape settlement	+0.019 (p = 0.83)	−0.089 (p = 0.20)
night term, summer / winter, 2000–2020	+0.42 (p = 0.13) / +0.17 (p = 0.58), 330	+0.27 (p = 0.30), 301 / −0.37 (p = 0.12), 238
night term, summer / winter, 1975–2020 (S1.7, Table 2 of the main text)	+0.44 (p < 0.001) / +0.33 (p = 0.004), 330	+0.10 (p = 0.57), 321 / −0.09 (p = 0.64), 251

S1.2 The nine-step incremental-covariate ladder

Is size a proxy for something the model leaves out? To ask, we added controls to the cross-city size relationship one at a time on a constant sample, in the order climate zone, absolute latitude, income group, GDP per capita, population density, built-up fraction, urban-station elevation, coastal distance and station siting. A size effect standing in for any of them would then be seen to decay as that control entered. None of the nine controls removes it. The slope strengthens, if anything, as correlated covariates are partialled out. The night slope rises from

+0.65 °C per tenfold population with size alone (95% CI +0.30 to +0.99) to +0.89 °C with all nine controls (+0.64 to +1.13), and the daily-mean slope from +0.23 °C (+0.08 to +0.39) to +0.38 °C (+0.24 to +0.51), with the interval excluding zero at every step. Size is therefore an independent driver rather than a proxy for climate, wealth, density or geography, though it explains a modest share of between-city variance, about 5% alone and 23% with the full set at night. Station siting behaves as physics predicts: the night-time level declines with the urban station's distance from the city center, from about +0.7 °C within 5 km to about +0.2 °C beyond 15 km, so distant stations *under*-read the island, while the size-law slope is indistinguishable across those distance bins. The slope is also positive, with an interval excluding zero, within every climate zone observed, and a joint test of the size-by-zone interactions does not reject a common slope, so the scaling form is stable across the sampled climates. Supplementary Table S3 gives the estimates.

Supplementary Table S3. The night-time size law and level by climate zone and by urban-station distance. Size-law slope in °C per tenfold population and mean night-time UHI in °C on the 506-city level sample, 95% confidence intervals in parentheses from country-clustered standard errors for slopes and from city standard errors for means, number of cities in italics.

group	size-law slope	mean night-time UHI
tropical (Köppen A)	+1.14 (+0.27 to +2.00)	−0.08 (−0.69 to +0.53)
	<i>38</i>	
arid (B)	+0.97 (+0.15 to +1.79)	+0.76 (+0.39 to +1.13)
	<i>95</i>	
temperate (C)	+0.51 (+0.15 to +0.88)	+0.67 (+0.50 to +0.84)
	<i>214</i>	
continental (D)	+0.62 (+0.26 to +0.99)	+0.52 (+0.35 to +0.69)
	<i>159</i>	

all non-tropical zones	+0.64 (+0.51 to +0.76)
	468
size × zone interactions, joint test	p = 0.50
urban station ≤5 km from the centroid	+0.72 (+0.56 to +0.89)
	227
5–15 km	+0.54 (+0.36 to +0.73)
	210
>15 km	+0.24 (−0.21 to +0.69)
	69
size × distance interaction, per km	+0.03 (−0.05 to +0.12)

S1.3 Income as a non-linear driver on top of size

Does income act on the heat island, and in what shape? As a straight line it does so only weakly: a linear term in GDP per capita survives in 44% of the extreme-bounds specifications that include it, short of the robustness bar. Once the relationship is allowed to bend it does. A four-knot restricted cubic spline in log GDP per capita is jointly significant for the daily mean and for the night and improves on a straight line by every measure, and its fitted shape is flat across most of the income range with an upturn at the top. Relative to a city at mean income, the richest 1% of cities carry +0.57 °C of additional daily-mean UHI (95% CI +0.31 to +0.87) and +0.47 °C at night (+0.09 to +0.88), and the top decile carries about +0.2 °C by day and +0.3 °C at night (Supplementary Table S4). The effect sits on top of the size effect without displacing it and, like size, is confined to the between-city level.

Did the gradient change over time? Across the five epochs of the primary window the cross-city income gradient roughly doubled between 2000 and 2015 and eased somewhat by 2020 (Supplementary Table S4). We do not read this as income causing UHI directly, because the

within-city GDP effect is null: a city growing richer did not itself acquire a larger heat island. The association is better read as growth-mediated, with income marking *which* cities added population inside their existing footprints over this period. The historical pattern runs opposite to projections that assume today's lower-income cities will intensify most, though the two differ in period and in which cities were growing, so this is a contrast rather than a contradiction.

S1.4 Development-phase dependence of the size-law slope

Does the size law's slope depend on where a city sits in its urbanization trajectory? Two checks point the same way, and neither settles it (Supplementary Table S4). Across cities, splitting the daily-mean cross-section by World Bank income group gives a slope in upper-middle-income cities, which are still urbanizing, more than three times that in high-income cities. Over time, on a fixed panel of 364 cities observed at every epoch from 1975 to 2020, so that the entry and exit of cities cannot confound it, the cross-city daily-mean slope falls by about half as those same cities mature. A size-by-time interaction puts the change at $-0.15\text{ }^{\circ}\text{C}$ per tenfold population over the 45 years (95% CI -0.26 to -0.04), so the decline is distinguishable from a constant slope. We nonetheless report the phase dependence as suggestive rather than established, for two reasons. The fixed panel consists almost entirely of mature, high-income cities, so what it shows is the maturation of one cohort rather than a contrast between phases. And the cross-sectional contrast is confounded with region: the upper-middle-income cohort is disproportionately North American and climatically atypical, so income proxies region and climate as much as it does phase.

What mature means, and what the proxy cannot separate. Development phase is proxied throughout by World Bank income group: high-income cities are taken as mature, the lower groups as still urbanizing. In this sample the proxy tracks growth, since the median high-income city added about a quarter of a log unit of population to its footprint over 1975–2020

against far more in the lower groups, but it cannot distinguish a city that has reached the top of its growth path from one that has stopped growing for other reasons while room remains. As a first look, we split the 748 high-income cities by the mean built-up share of their footprint, as a measure of room to grow, and by their own population change on the footprint. The within-city term is largest in the third with the most room, +0.24 °C per log unit (95% CI +0.15 to +0.33), about a tenth of a degree with intervals spanning zero in the two more built-up thirds, and positive among cities that both had room and grew (Supplementary Table S4). The direction agrees with the phase reading. The split is not stable, however: the cell of built-out cities that barely grew returns a large but poorly identified coefficient, because their population changed too little to identify one, so we report the exercise as a caveat rather than a result and leave it to follow-up work with a direct measure of remaining capacity.

Supplementary Table S4. Income and development-phase checks. Coefficients in °C per tenfold population (size law) or °C per log unit (within-city term), 95% confidence intervals in parentheses; country-clustered standard errors for cross-city slopes, city-clustered for panel terms, city bootstrap (1,000 draws) for the spline effects; number of cities in italics.

check	result
linear GDP per capita in the extreme- bounds search	significant in 44% of specifications
restricted cubic spline in log GDP per capita, joint test, daily mean / night	$p < 0.001$ / $p = 0.03$
income partial effect, top decile, daily mean / night	+0.19 (+0.04 to +0.33) 782 / +0.28 (+0.05 to +0.51) 554
income partial effect, richest 1%, daily mean / night	+0.57 (+0.31 to +0.87) / +0.47 (+0.09 to +0.88)

cross-city income gradient, °C per log	+0.18 (+0.08 to +0.28) / +0.28 (+0.02 to +0.53) /
unit of GDP per capita, 2000 / 2005 /	+0.38 (+0.11 to +0.66) / +0.39 (+0.17 to +0.61) /
2010 / 2015 / 2020	+0.33 (+0.06 to +0.61)
daily-mean size-law slope, upper-	+0.62 (+0.34 to +0.91) 114 / +0.17 (+0.08 to
middle-income / high-income cities	+0.25) 545
fixed panel of 364 cities, daily-mean	+0.27 (+0.04 to +0.50) / +0.12 (−0.01 to +0.25)
slope in 1975 / 2020	
size × time interaction, change over	−0.15 (−0.26 to −0.04)
1975–2020	
within-city term among 748 high-	+0.24 (+0.15 to +0.33) / about +0.1, intervals
income cities: most room / two built-up	span zero / +0.18 (+0.08 to +0.28)
thirds / room and grew	

S1.5 The within-city term on a delineation that moves

What does the within-city term mean when the boundary is allowed to move? The main text reports it as population added on a fixed footprint. GHS-UCDB holds each city’s delineated area constant across epochs, so within a city the density column is population divided by a constant and the two series are identical to six decimals. The consequence is that a within-city density effect cannot be identified from that record at all, an identity that follows from the delineation convention rather than from any one dataset.

GHS-WUP-MTUC R2025A re-delineates urban centers at every epoch and links them across epochs, which lifts the restriction. Matching the primary-window cities to it by centroid within 10 km, one city per lineage, joins 733 of the 885 cities. On that joined sample we hold the station UHI fixed as the dependent variable and change only how the covariate is built (Supplementary Table S5). The frozen-footprint term reproduces on the joined sample, and

population counted on the moving boundary carries a term of the same sign, $+0.19^{\circ}\text{C}$ per log unit (95% CI $+0.08$ to $+0.30$). Density on the moving boundary carries a term on its own whose interval spans zero, and once moving-boundary population is held it carries nothing, so the term follows population and not density; within cities the frozen column tracks moving-boundary population far more closely than it tracks moving-boundary density.

Which population count is the better specification? Decomposing the moving-boundary population into density and area gives two coefficients whose intervals overlap, $+0.16^{\circ}\text{C}$ per log unit (95% CI -0.12 to $+0.43$) for density and $+0.20^{\circ}\text{C}$ ($+0.07$ to $+0.33$) for area, which is the restriction under which the driver is population itself. Both series explain 1–2% of within-city variation.

Meaning: the frozen count is population measured with omission, since it misses the people added by territorial expansion, and it returns a proportionally larger coefficient on a proportionally smaller within-city standard deviation. The two agree on the sign, on significance and on the effect per standard deviation, and neither leaves room for a density term. The main text therefore keeps the fixed footprint, on which the full 1975–2020 record and the joint test are built, and reports the moving-boundary result here as the check that the choice does not carry the finding.

Supplementary Table S5. The within-city term on the frozen and on the moving delineation, 2000–2020. Two-way fixed-effects coefficients in $^{\circ}\text{C}$ per log unit with city-clustered 95% confidence intervals, on the 733 cities (2,844 city-epochs) matched to GHS-WUP-MTUC; the dependent variable is the same station UHI in every row.

construction of the covariate	coefficient
population on the frozen GHS-UCDB footprint (the density column)	$+0.30$ ($+0.08$ to $+0.51$)
population on the moving boundary	$+0.19$ ($+0.08$ to $+0.30$)

density on the moving boundary, alone	+0.20 (−0.09 to +0.49)
density on the moving boundary with moving	−0.04 (−0.36 to +0.28) / +0.20
population held: density / population	(+0.07 to +0.33)
moving population decomposed into density / area	+0.16 (−0.12 to +0.43) / +0.20
	(+0.07 to +0.33)

S1.6 Building volume and land cover do not carry the within-city term on our panel

Can building mass or land cover explain the within-city term? The Discussion lists three changes that could turn population added on a fixed footprint into a larger heat island: vegetation loss, building mass and anthropogenic heat. The first two can be tested with series on the same fixed footprint, and Supplementary Table S6 holds every estimate.

Building mass is proxied by building volume. GHS-BUILT-V (R2023A) reports the volume of buildings on the same fixed GHS-UCDB polygon at each epoch from 2000 to 2020, and we added its log, per unit area, to the primary panel; 884 of the 885 cities join. Volume moves more than population within a city over the window and is almost uncorrelated with it, so the test is limited neither by variation nor by collinearity. The population term does not move. It is +0.26 °C per log unit (95% CI +0.10 to +0.41) with volume in the model and the same without it, and barely lower with built-up fraction added as well. Volume itself carries no coefficient in the two-way fixed-effects model, nor within or between cities in the joint model, and the equality test for population is unchanged. Volume is also null on every split that gave it a chance to act, at night and by day, in North America, Europe and East Asia, and within each of the arid, temperate and continental Köppen groups, and dropping the cities whose polygon extraction may be truncated changes nothing (Table 1 of the main text; Supplementary Table S2).

Land cover is tested with two series: the built-up share of the polygon from GHS-UCDB, and the vegetation of the urban core and of its surrounding rings from MODIS NDVI, both at the five epochs. Their within-city variation is small, since a footprint that is already largely built changes its share slowly, but it is enough to ask whether either absorbs the population term. Because the within-city term is a night-time term, the night channel is the relevant test. On the 574 cities that carry all series, the built-up share takes about a tenth off the night-time population coefficient and the vegetation contrast between core and surroundings changes it not at all, and neither series carries a night-time coefficient of its own. In North America, where the term lives, the two series together move the night-time coefficient by a few hundredths of a degree, and in Europe it stays null. On the full 1975–2020 night record the built-up share takes about a fifth off the coefficient and its own interval spans zero. The daily mean gives the same picture. The built-up share takes about a sixth off the population coefficient without carrying a coefficient of its own, and core vegetation carries nothing. A widening vegetation deficit relative to the surroundings carries a negative coefficient whose interval spans zero, while leaving the population term unchanged, so vegetation loss is at most a small channel of its own rather than the one through which added population acts. Land cover therefore carries at most a fifth of the within-city term in the pooled panel and almost none of it where the term is largest.

Meaning: on the air-temperature record the within-city term does not behave as a proxy for building mass, and only a small part of it runs through land cover. The result is specific to the quantity measured. A station near the urban centroid reads the canopy-layer air, and the population it responds to may act through anthropogenic heat and activity rather than through the stored heat that building volume indexes; a polygon-wide volume mean may also be a poor summary of what changed around the one station that reports. The test does not say building mass is irrelevant to urban heat. It says that, on this record and for this quantity,

population added on a fixed footprint carries the within-city term and building volume does not.

Supplementary Table S6. Building volume and land cover on the fixed footprint. Two-way fixed-effects coefficients with city-clustered 95% confidence intervals; the population term in °C per log unit, the built-up share in °C per unit of share, NDVI terms in °C per unit of NDVI; number of cities in italics. The night channel is the day/night station set, the daily mean the primary panel.

specification	population term	added series
primary window, daily mean, 2000–2020	+0.27 (+0.12 to +0.43) <i>885</i>	
+ building volume	+0.26 (+0.10 to +0.41) <i>884</i>	volume +0.02 (−0.02 to +0.06)
within-city standard deviation, volume / population; within-city correlation	<i>0.20 / 0.09; −0.04</i>	
volume by night and day, by region, by Köppen group, and without possibly truncated polygons	population unchanged in every split	volume null in every split (Table 1 of the main text; Supplementary Table S2)
night, 2000–2020, constant 574- city sample, population alone	+0.35 (+0.03 to +0.66) <i>574</i>	
+ built-up share	+0.31 (−0.03 to +0.64)	share +0.64 (−2.76 to +4.04)
+ vegetation contrast, core minus 0–10 km ring	+0.34 (+0.03 to +0.66)	contrast −0.14 (−1.15 to +0.87)

+ core vegetation	+0.35 (+0.03 to +0.67)	core +0.25 (−0.37 to +0.88)
+ built-up share and vegetation contrast	+0.31 (−0.03 to +0.64)	share +0.63 (−2.77 to +4.03); contrast −0.13 (−1.14 to +0.88)
night, North America, alone / + both series	+0.71 (+0.09 to +1.34) / +0.74 (+0.13 to +1.35) 308	share −0.44 (−4.62 to +3.75); contrast −0.14 (−1.75 to +1.46)
night, Europe, alone / + both series	−0.12 (−0.58 to +0.34) / −0.17 (−0.63 to +0.30) 231	share +1.83 (−5.40 to +9.05); contrast +0.16 (−0.97 to +1.28)
night, 1975–2020, alone / + built-up share	+0.32 (+0.16 to +0.47) / +0.25 (+0.10 to +0.41) 746	share +1.68 (−0.28 to +3.63)
daily mean, 2000–2020, alone / + built-up share	+0.27 (+0.12 to +0.43) / +0.23 (+0.07 to +0.39) 885	share +1.19 (−0.68 to +3.06)
daily mean, + core vegetation / + vegetation contrast	+0.32 (+0.13 to +0.51) / +0.32 (+0.13 to +0.51) 868	core −0.05 (−0.47 to +0.37); contrast −0.41 (−1.04 to +0.21)
within-city standard deviation, built-up share / core NDVI (574-city sample)	0.011 / 0.027	

S1.7 The regional gap by season

The seasonal split uses the same construction as the annual panel, applied to summer (June to August) and winter (December to February) means of each element built from the daily

record, with the composition-free rural reference refitted for each season and element (Methods of the main text). Table 2 of the main text gives every estimate. On the primary 2000–2020 window the seasonal series are too thin to separate the regions: pooled, the night-time term is present in summer and not in winter, but the North American intervals span more than a degree and neither season is distinguishable from the other. The full 1975–2020 record resolves it. In North America the night-time term is present in both seasons and of similar size, $+0.44\text{ }^{\circ}\text{C}$ per log unit in summer (95% CI $+0.20$ to $+0.67$) and $+0.33\text{ }^{\circ}\text{C}$ in winter ($+0.11$ to $+0.56$), and the summer-minus-winter difference, $+0.10\text{ }^{\circ}\text{C}$ (-0.07 to $+0.28$), does not exclude zero. The daily mean shows the same pattern at about half the size, and the daytime maximum carries no positive term in either season. Outside North America both seasons are null. Within North America the arid zone carries the largest term in both seasons and the temperate and continental zones smaller ones in both, so the seasonal shape does not depend on climate class. The night-time level of the heat island is itself larger in summer than in winter. Over 2000–2020 the city-mean night-time UHI is $+0.66\text{ }^{\circ}\text{C}$ in summer (95% CI $+0.48$ to $+0.84$) against $+0.41\text{ }^{\circ}\text{C}$ in winter ($+0.23$ to $+0.58$) in North America, and $+0.68\text{ }^{\circ}\text{C}$ ($+0.47$ to $+0.89$) against $+0.13\text{ }^{\circ}\text{C}$ (-0.08 to $+0.33$) in Europe. The within-city term therefore acts on a level that is highest in the warm season.

A year-round night-time term narrows the candidates. Cooling waste heat alone is unlikely to produce it, since that flux is absent in a North American winter, and vegetation loss is an unlikely main channel, since its evaporative cooling is a summer quantity. What remains is heat that is released in every season and mostly at night: the anthropogenic heat of heating, cooling and traffic taken together, which scales with per-capita energy use, and the heat a building stock stores by day and releases after dark. Both are plausibly larger per resident in North America than in Europe, the first through higher per-capita energy use and the second wherever the building stock is larger per person, and our station panel cannot yet separate

them from each other. That is left to a follow-up built on city-level energy and building-stock data, and on the full seasonal and diurnal structure of the station record, which this paper opens but does not exhaust. What the seasonal split establishes here is that the regional gap is a property of the night in every season, not of the summer.

Supplementary Note 2. What each statistical test asks, and what our result means

Robustness tables report numbers; they rarely say what a number would have to look like for the claim to fail. This note states, for every test used, the question it answers, what a failure would have looked like, what we actually found, and what follows.

S2.1 Is the relationship real, or an artifact of how it was specified?

Extreme-bounds analysis (384 specifications). *Question:* does the coefficient survive every reasonable combination of controls, or only the combination we happened to choose? *Failure would look like:* a driver significant in some specifications and not others, with the sign flipping across them. *Result:* city size is significant in 100% of the 256 specifications in which it appears (of 384 total) and never changes sign; density, built-up fraction, vegetation, wind, coastal distance and latitude all fail this bar. *Meaning:* the size law is not an artifact of control selection, and the drivers that fail should not be reported as drivers at all.

Incremental covariate ladder (nine steps). *Question:* does the coefficient drift as confounders are added one at a time? *Failure would look like:* the slope decaying toward zero as climate, latitude, income and urban form are absorbed. *Result:* the night size-slope strengthens, from $+0.65$ °C per tenfold population (95% CI $+0.30$ to $+0.99$) to $+0.89$ °C ($+0.64$ to $+1.13$). *Meaning:* the relationship is not a proxy for climate or wealth; if anything those covariates were suppressing it.

Restricted cubic spline on GDP. *Question:* is a driver absent, or merely non-linear? *Failure would look like:* the spline adding nothing over a straight line. *Result:* the spline is jointly significant ($p < 0.001$ for the annual mean, $p = 0.03$ at night) and beats the straight line on fit, while the linear term survives in only 44% of the extreme-bounds specifications. *Meaning:* income acts, but only through a high-income upturn that a linear specification averages away.

S2.2 Is the significance real, or an artifact of how the data are correlated?

Country-clustered standard errors. *Question:* cities in the same country share weather, station networks and reporting practice; does the result survive treating them as one unit rather than as independent observations? *Result:* both relationships remain significant, though the within-city trend's primary-window significance is markedly weaker than the full record's. *Meaning:* the result is not produced by counting correlated cities as independent evidence, though the primary window's inference is genuinely thinner.

Naïve versus clustered pooling. *Question:* how much would the conventional shortcut inflate confidence? *Result:* treating each city-year as independent roughly halves the standard errors. *Meaning:* reported significance in panels of this kind should be read with the clustering scheme in view; we report the conservative version throughout.

S2.3 Could the method find a signal in noise?

Permutation of the predictor (500×) and within-city placebo shuffle. *Question:* if the relationship were absent, would this procedure manufacture one anyway? *Result:* the permuted size-law slope is 0.00 ± 0.13 °C per tenfold population against $+0.66$ °C observed; the within-city shuffle returns 0.00 ± 0.06 °C per log unit on 2000–2020 and 0.00 ± 0.02 on the full record, against $+0.28$ and $+0.22$ °C observed. *Meaning:* the estimation procedure does not generate slopes from noise on either window. This is the strongest single guard against a spurious finding, because it tests the whole procedure rather than one assumption.

S2.4 Do the two dimensions really differ, or is that an artifact of estimating them separately?

Mundlak/Hausman test. *Question:* can one coefficient describe both the between- and within-city relationships? *Result:* it rejects on both windows. *Meaning:* the opposite signs are a joint feature of the data, not an artifact of running two regressions, which is what licenses the paper's central claim.

S2.5 Is the model the right shape, and is it driven by a few cities?

RESET functional-form test. *Question:* is a straight line in log population adequate, or is curvature being missed? *Result:* $F = 0.17$, $p = 0.68$, no misspecification detected. *Meaning:* the log-linear Oke form is appropriate for these data.

Breusch–Pagan and HC3 standard errors. *Question:* does non-constant error variance distort inference? *Result:* $p = 0.19$; the HC3 slope is $+0.66$ °C per tenfold population (95% CI $+0.43$ to $+0.88$ under HC3, against $+0.32$ to $+0.99$ country-clustered), unchanged. *Meaning:* heteroskedasticity does not change the inference on the size law.

Cook's distance (drop the most influential 1%). *Question:* do a handful of cities carry the result? *Result:* the slope moves from $+0.66$ to $+0.69$ °C per tenfold population (95% CI $+0.38$ to $+1.00$, 501 cities). *Meaning:* the slope does not rest on a handful of influential cities.

S2.6 Does it generalize?

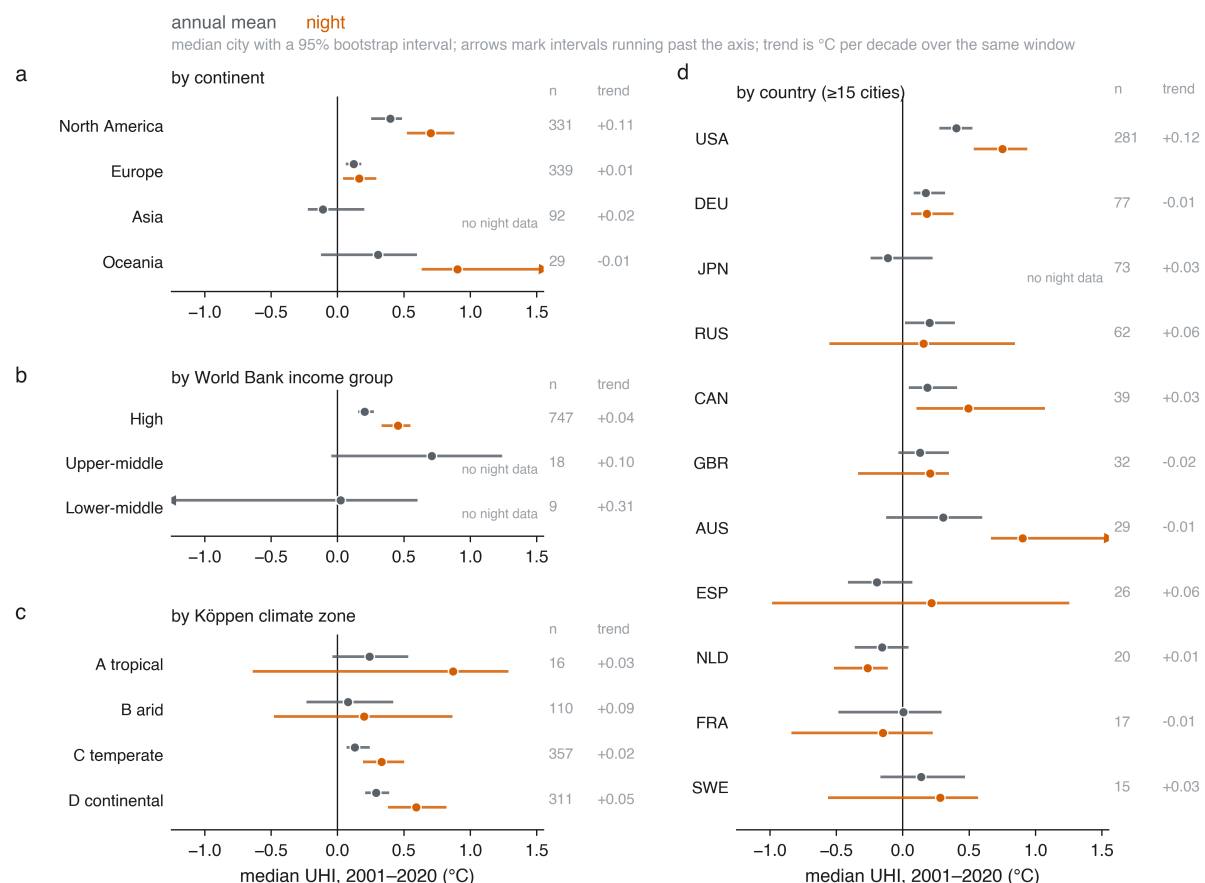
Leave-one-region-out. *Question:* is the finding an artifact of the dominant region? *Result:* it collapses for the within-city trend, more sharply on the primary window than the full record; the size law survives every single-region exclusion. *Meaning:* the level relationship generalizes across the sampled regions; the trend is a North American regularity on this record.

Independent temperature product (GHCN-M v4 QCU). *Question:* is the result a property of one archive's quality-control choices? *Result:* the size law is positive and significant on GHCN-M QCU and agrees closely with a matched GHCN-Daily reconstruction ($r = 0.98$).

The GHCN-M panel was matched before the duplicate-urban-station rule was applied; removing the cities that rule dropped moves the within-city term on GHCN-M from +0.18 to +0.21 °C per log unit on 539 cities (Supplementary Table S1). Meaning: neither relationship's sign or significance is an artifact of the specific dataset.

Supplementary Figure S1

Supplementary Fig. S1. Median UHI by continent, income group, climate zone and country. Median 2001–2020 UHI for the annual mean (gray) and the night (orange), by **a**, continent, **b**, World Bank income group, **c**, Köppen climate zone and **d**, country, for the 798 cities of the broad clean-rural sample; groups below eight cities (fifteen for countries) are not shown. Bars are 95% confidence intervals from 2,000 bootstrap resamples of the cities in each group, and the right-hand columns give the group's city count and its median trend in °C per decade. Four groups have too few TMIN cities to show a night estimate.



Supplementary references

Chakraborty, T. & Lee, X. A simplified urban-extent algorithm to characterize surface urban heat islands on a global scale and examine vegetation control on their spatiotemporal variability. *Int. J. Appl. Earth Obs. Geoinf.* 74, 269–280 (2019).

Chakraborty, T. & Lee, X. Yale Center for Earth Observation (YCEO) Surface Urban Heat Islands, Version 4, 2003–2018. NASA Socioeconomic Data and Applications Center (SEDAC) [data set] <https://doi.org/10.7927/S5M5-ZK14> (2023).

Driscoll, J. C. & Kraay, A. C. Consistent covariance matrix estimation with spatially dependent panel data. *Rev. Econ. Stat.* 80, 549–560 (1998).

Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J. & Lawrimore, J. H. The Global Historical Climatology Network Monthly Temperature Dataset, Version 4. *J. Clim.* 31, 9835–9854 (2018).

Pesaran, M. H. General diagnostic tests for cross-sectional dependence in panels. *Empir. Econ.* 60, 13–50 (2021).

Yang, Q. et al. A global urban heat island intensity dataset: generation, comparison, and analysis. *Remote Sens. Environ.* 312, 114343 (2024).