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.

A homogenization-free air-temperature urban heat-island dataset for 1,117 cities from GHCN-Daily, 1941–2020

In Chang Hwang

ichwang@si.re.kr

The Seoul Institute

data-descriptor

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11282626/v1

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

Additional Declarations: No competing interests reported.

Abstract

Station-based urban heat island (UHI) studies generally use homogenized records, and homogenization removes the gradual local warming a heat island produces, erasing the signal under study: it removes about 60% of the global UHI intensification since 1971–90 and 83% of it in North America. This record therefore builds UHI from GHCN-Daily air temperatures that passed daily quality control but no homogenization. A city's UHI is its urban station minus the median of rural stations 40–150 km away, corrected for elevation, and reported for the daily mean, night and day. It covers 1,117 cities over 2001–2020 and 719 with long station histories, from an annual panel of 8,637 stations, with within-city panels at five-year epochs from 1975 and 20-year windows to 1941. Every product is issued with and without a 12 km rural-reference screen. Coverage is concentrated in North America, Europe and Asia; the other continents hold about 13%. Validation covers the reference radius, lapse rate, station siting, per-city uncertainty and replication from an independent archive.

Background & Summary

Homogenizing a station network removes most of the urban heat island (UHI) signal it carries. On the same stations released here, homogenization removes about 60% of the global UHI intensification since 1971–90 and 83% of it in North America (Technical Validation). This is a direct property of the procedure, not an incidental cost, since homogenization is designed to remove exactly the kind of gradual, local, non-climatic change a heat island itself produces. The UHI effect, in which city centres run warmer than the surrounding countryside, is an important factor governing heat exposure, energy demand and health impacts, and research on it has followed two paths. The first uses satellite data: satellites provide high-resolution coverage of nearly the entire globe, but they measure land surface temperature (LST) rather than the near-surface air temperature to which people are actually exposed. The second uses homogenized station temperature records, such as the GHCN-M adjusted products or Berkeley Earth, which run long but, for the reason just given, are ill-suited to isolating the effect this release measures. Some studies[1–3] have extracted a UHI signal from raw records instead, but each rests on a single national network and none of them distributes the city-level values it derived. Hausfather et al.[2] come closest: the network they work from is public and a supplementary document describes the method, but neither the derived series nor the code behind them is released. To our knowledge, no publicly available multi-region, raw air-temperature UHI dataset exists.

That gap matters for four concrete uses. It lets a user compute the observed UHI under one standardized procedure instead of reimplementing Karl et al.[1] or Hausfather et al.[2] independently, which until now was the only way to obtain a city-level UHI estimate from raw records. It supports city-temperature scaling and energy-demand or heat-exposure models, which need a city's own urban-rural decomposition rather than a satellite proxy calibrated against it. It supports heatwave exposure modeling that needs the night and day signals as separate channels, which this record's three-channel construction provides (Methods). And it supports climate-economy and social-cost-of-carbon modeling, where the urban contribution to a city's observed warming has to be separated from the global and

regional terms before it can be priced on its own. This release ships that separation as a derived product (decomposition_cities.csv), and the descriptor itself makes no claim from it.

This paper builds that dataset from publicly available sources, with a transparent processing workflow and accompanying analysis code so that every step can be reproduced. Measuring the UHI as an urban station minus a rural reference drawn from raw records goes back to Karl et al.[1], who established the log-population relationship on United States data, and Hausfather et al.[2] is its closest modern descendant. What we add, against those studies and the gridded products that followed them (Table 1), is the same construction applied across regions from 1941 to 2020, with the resulting values, the intermediate inputs behind them and the analysis code released together as a citable dataset.

The coverage this buys is uneven by construction, not by oversight. A station enters this record only if it clears the daily-quality and multi-decade completeness rules in Methods without any homogenizing adjustment. Dense, long-running, non-homogenized networks of that kind exist disproportionately in Europe, North America and parts of Asia. They account for 40%, 31% and 17% of the 1,117 cities, respectively, with the remaining 13% spread across the rest of the world (Fig. 1a). That is the same constraint, read from the other side, that confines the three prior studies above to a single country each. A raw, non-homogenized record demands a station density that most of the world does not yet carry at the required length (Table 2). Technical Validation reports the resulting imbalance in full and names what it rules out (tropical, developing and low-income cities, for which satellite or crowdsourced temperature records remain the right tool[4]), instead of treating the gap as incidental.

Table 1
This release against the prior literature

Study	Region	Station data	Output
Karl et al.[1]	United States	raw	findings
Peterson[5]	contiguous US	homogenized	findings
Hansen et al.[6]	global	homogenized, night-light adjusted	findings + adjusted product
Muller et al.[7]; Wickham et al.[8]	global	homogenized	findings + gridded product
Hausfather et al.[2]	United States	raw *and* homogenized	findings
Sun et al.[3]	China	national network	findings
This release	multi-region, 1,117 cities	raw, non-homogenized	dataset + codes

Methods

Source data and station selection

GHCN-Daily[9] was extracted using a NOAA AWS public data mirror. The inventory retrieved for this release lists 132,501 stations, of which 40,800 reported maximum, minimum, or mean temperature at least once. NOAA updates the inventory periodically, so this count drifts slightly with the retrieval date. Stations were selected based on the following criteria:

- Daily values are used only when the GHCN-Daily quality flag is empty (i.e., when all approximately 19 automated quality checks are passed).
- Monthly averages are generated only when there are at least 20 valid days.
- Annual averages are generated only when there are at least 11 valid months.
- Homogenization is not applied.

These criteria are strict. Out of the 40,800 stations, most fail to meet these criteria for even a single year because their records are too short or incomplete to fill 11 months, or because daily values fail quality control. Stations that pass these criteria for at least one year total 8,637, and these comprise the compiled data (annual_by_elem.csv), the station-level record from which all city-level values are derived.

Table 2
How the station count falls at each step, and why. GHCN-Daily is predominantly a precipitation network, and roughly seven stations in ten carry no thermometer at all. Only the second step discards temperature data, and it discards most of it, because the completeness rules below are strict.

Step	Stations	Reason
GHCN-Daily inventory	132,501	—
Report TMAX/TMIN/TAVG at any point	40,800	91,701 report no temperature; 99.9% of those are precipitation-only
Clear the daily-quality and completeness rules for ≥ 1 year	8,637	short records (20% of temperature stations span < 10 years and 48% span < 30), and months in which too many daily values fail the quality flag
Enter a city match (broad set, 2000–2020)	7,794	no city within range, or the station falls outside the 25 km / 40–150 km geometry
Enter a city match (long-record set)	4,135	as above, plus the ≤ 1955-to-≥2010 span requirement

Station data aggregation

The aggregation of daily values into monthly and annual means was performed directly from the raw .dly files, so that no intermediate adjustment could enter. The annual mean temperature is computed month by month as

TAVG = (monthly maximum temperature TMAX + monthly minimum temperature TMIN) / 2, (1)

where TMAX and TMIN are the monthly maximum and minimum. In months where those two are missing but a station-reported TAVG exists, the reported value is used instead. Nighttime records are represented by TMIN and daytime records by TMAX, and both are aggregated under the same rules. The code that does this is parse_dly_elem.py and its output is annual_by_elem.csv.

The 20-day and 11-month rules follow common climatological practice, under which a monthly mean is not computed when more than about ten daily values are missing. Two points follow. First, the 20-day rule is applied regardless of month length, so it demands 20 of 28 days in February against 20 of 31 in January. Second, no rule is applied to consecutive missing days: ten scattered gaps and one ten-day block are treated alike.

The annual rule has the larger effect. When a year is built from eleven months, the result differs from the true twelve-month mean by

bias = (annual average − missing month's average) / 11, (2)

so both the size and the direction of the departure depend on which month is absent. For a mid-latitude station whose coldest and warmest months lie 20°C apart, a missing January makes the year about 0.9°C too warm and a missing July about 0.9°C too cold. Years filled from only eleven months are a small minority in this record, and a 20-year mean spreads them out so that opposite-signed departures cancel. The UHI is also a difference between an urban and a rural station, and outages tend to be regional. When the urban station and its rural reference lose the same month, what survives is far smaller than the 0.9°C carried by either station alone. The rural side is a median over at least three stations, which further limits what survives. A city whose urban station repeatedly loses the same season, however (winter gaps at a high-latitude site, for instance) carries a systematic rather than a random departure, and averaging over years does not remove it.

The direct test, which we did not run for this release, is to rebuild the record using only twelve-month years and compare. The released annual_by_elem.csv carries annual values without the underlying month counts, so this restriction cannot be applied to it directly; a user who needs it can regenerate the annual panel from the raw GHCN-Daily archive with parse_dly_elem.py, which applies the rules above and records which months contributed.

Working sets

Two working sets are defined from the GHCN-Daily station list. They are not disjoint: 704 of the 818 long-record cities also appear in the broad set, because a station with a long record usually satisfies the broad span requirement too. What separates them is the span each demands, not the cities each reaches.

- **Long-record subset.** Stations where both TMAX and TMIN begin before 1955 and continue past 2010, or where TAVG satisfies those conditions. This supports the 1941–2020 products and a 2001–2020 climatology on long-record stations.
- **Broad subset.** Stations where temperature observation begins before 2001 and continues past 2015. This supports six-continent analyses for 2000–2020.

These limits are applied to the GHCN-Daily inventory, which records only a station's first and last reporting year, so they can screen for span but not for usability. The long-record limits of 1955 and 2010 leave about fourteen years of slack at each end of the 1941–2020 period, and the broad limits of 2001 and 2015 allow a similar tolerance around 2000–2020. Sufficiency is enforced later, against the data themselves, by the requirement of at least eight valid years inside a twenty-year window, set out under UHI construction below. The broad subset buys geographic coverage at the price of record length, so values carrying the same name in the two sets are not directly comparable. The two subsets are distributed as separate files, city_station_match_broad.csv and city_station_match_longrecord.csv.

City–station matching

Cities are the urban centres of GHS-UCDB R2024A[10]. Distances are computed with an approximation that takes account of each city's latitude, Eq. (3):

$$d = \sqrt{[(\Delta lat \cdot 111)^2 + (\Delta lon \cdot 111 \cdot \cos \varphi)^2]} \text{ km} \qquad (3)$$

This is the standard equirectangular approximation; against a haversine great-circle distance it errs by at most 1.1% at the 150 km outer radius and by 0.25% at the 12 km contamination screen, well below the spacing at which any boundary binds.

Cities are matched in two steps.

1. **Urban station.** The nearest station within 25 km of the urban centre, within the relevant working set. If no such station exists, the city is excluded.

2. **Rural baseline.** All stations between 40 km and 150 km from the urban centre, within the same subset. If fewer than three qualify, the city is excluded.

The three distances answer three different questions.

• **The 25 km urban radius** asks whether a station can represent the city at all. GHS cities are continuous built-up areas defined by population density, so their footprints vary by orders of magnitude: the median urban centre has an equivalent radius of 2.7 km, but 43 of the 11,422 exceed 20 km and the largest reaches 46 km. The 25 km limit therefore accommodates all but a few dozen of the largest centres, and in practice it rarely binds. Only the nearest qualifying station is used, which fixes the choice as a reproducible rule; Technical Validation reports what that choice costs.

• **The 40 km inner radius** asks how far the city's thermal influence extends. If it is set too close, the reference is drawn from suburbs that already carry part of the heat island, and the urban–rural difference then understates the true contrast. If it is set too far, usable cities are lost: a 15–50 km ring reduces the sample from 1,054 cities to 469.

• **The 150 km outer radius** asks how many cities remain usable against how far the reference has moved. Regime contrasts are not a simple function of distance, so this radius is a practical limit rather than a physical boundary. Discordant individual stations are handled instead by the median over at least three references and by the elevation correction.

• **The requirement of at least three rural stations** follows from using a median. With two stations the median equals the mean and gives no protection against an outlier; with one, a single faulty station becomes the city's UHI. Three is the smallest number at which the median does any work.

Rural contamination screen. A rural station is flagged and excluded from the reference if it lies within 12 km of any GHS urban centre, not only the city being measured, because a station near a second city carries that city's heat island into the reference. Values quoted in this paper are from the unscreened series, which is the broader sample and the conservative one: applying the screen raises the pooled median from + 0.18 to + 0.19°C. The 12 km radius is a working choice, set at roughly half the 25 km limit for accepting an urban station, so that a station treated as rural for one city cannot sit closer to a second city than an urban station may sit from its own. It matters less than it might appear: rebuilding the record at screen radii of 8, 12, 16, 20 and 25 km leaves the pooled median between + 0.19 and + 0.23°C and the negative-UHI share between 36% and 37%. Applying no screen at all is the one setting that stands out, at + 0.17°C and 40% negative. Where the 40 km inner radius guards against contamination from the city being measured, the 12 km screen guards against contamination from every other city.

Why a distance ring. Satellite studies define the reference geometrically, as a ring outside the built-up boundary, whereas station studies in the tradition of Karl et al.[1] and Hausfather et al.[2] classify each station by the settlement it sits in. The graded form of that classification cannot be carried over to this record. GHS-UCDB admits only centres above 50,000 inhabitants, so Karl's three-way split places 6,782 of 7,794 stations in the lowest class and just 14 in the middle one; filling the middle class would require population for settlements below 50,000, which no global urban database supplies. The satellite classifiers Hausfather et al. use, night-time lights and impervious surface, begin in the 1990s and the 2000s respectively, against a record that starts in 1941.

One classification the database does support is a binary test: whether a station falls inside a mapped urban centre at all. We therefore follow the geometric convention, with radii far larger than a satellite ring, because thermometers are where they are and the ring has to reach until it finds three of them, and we apply that binary test alongside it in the form of the 12 km screen. The ring controls regional climate by construction; the screen supplies the population criterion.

Applying the same binary test at its strictest (treating every station inside any urban-centre polygon as urban, not only those within 12 km of a centre point) moves the pooled median from + 0.18 to + 0.23°C. Pushing the population criterion further therefore raises the measured heat island rather than lowering it, and the values released here are a lower bound with respect to this choice, not a maximal one.

Distances are measured from the fixed urban-centre centroid, not from the built-up boundary, which is no fixed object over a record of this length. Mean urban-centre area in this sample grows from 58 km² in 2000 to 107 km² in 2020, and only 8% of cities hold still to within 1%. A centroid keeps the rule defining the reference the same in 1941 as in 2020.

Duplicate urban-station matches. A subset of centres resolve to the *same* nearest station inside the 25 km radius. GHS-UCDB defines urban centres far more finely than GHCN-Daily's station network is spaced, so a dense conurbation can carry several distinct centres too close together, relative to actual station spacing, for each to hold its own thermometer. The Ruhr valley and the Fergana valley are two examples. Left unaddressed, this double-counts a single physical station's temperature record as two or more separate "cities" in any aggregate statistic. Across the released matching tables this affected 271 of 1,314 stations (40% of rows before the fix) in the broad set and 153 of 818 (36%) in the long-record set. The fix keeps, for each shared station, only the nearer of the competing centroids and drops the other; duplicate_station_matches_dropped.csv, deposited with this release, documents every dropped city, its distance to the shared station, and which city it lost to. All city counts in this descriptor are post-fix.

UHI construction

The UHI for each city and period is computed as Eq. (4),

$$UHI = T_{urban} - \text{median}(T_{rural}) + \Gamma \cdot (z_{urban} - \text{median}(z_{rural})), \qquad (4)$$

where the median is taken over the clean-rural stations reporting in that period and $\Gamma = 6.5$°C/km is the standard atmospheric lapse rate, a decrease of 6.5°C per kilometer of elevation.

The elevation term adjusts the urban station temperature to the elevation of the rural baseline. Every published UHI value therefore already incorporates the elevation correction, and elevation should not be re-introduced as an explanatory variable in subsequent analyses.

The urban station is held fixed across all periods, so a difference between two periods is a difference in the same thermometer. The rural set is not fixed, because the median is retaken over whichever clean-rural stations report in the period being measured. A station entering or leaving therefore moves the rural reference, and a difference between two periods can read that movement as a change in temperature.

For a level over a single period this is harmless and even helpful, since more reporting stations sample the background better. For a change between periods it is neither, and the problem grows with the length

of the interval being differenced.

Over a twenty-year difference the median count of references usable at both ends falls only from 16 to 15, and the two constructions of the change agree at r = 0.81. Over forty-five years the count falls from 10 to 7 and the agreement drops to r = 0.61. About two thirds of the variance in the change is then a property of the construction rather than of the city.

All three constructions are distributed side by side in city_uhi_epoch_panel.csv, so a user working with long differences can take a composition-free column directly. Data Records names the columns.

The composition-free construction replaces the retaken median with an estimated reference. Fitting T(s,t) = αs *+ τ(c,t) by least squares across all of a city's clean-rural stations gives a station term α*s that absorbs elevation, exposure and siting, and a period term τ(c,t) identified from the overlaps between stations rather than from whichever of them happens to report. The reference is then R(c,t) = median_s(α_s) + τ(c,t). The median of α is used rather than the mean, so that levels stay on the same scale as the median-of-stations values above; a mean inflates them by about 0.5°C.

One further rule applies when more than one temperature element is wanted. Applying Eq. (4) separately to each element does not preserve the arithmetic relation between them: a user may expect UHI_day = 2·UHI_mean − UHI_night to hold, and it does not, because the rural median for each element need not fall on the same station. On this record the two differ by a median of 0.11°C and by as much as 5.8°C. Two of the released files satisfy the identity and one does not. The city files are built for levels, one element at a time, except for uhi_tmax in longrecord_city_uhi.csv, which is derived from the identity rather than computed and therefore carries the combined uncertainty of the two terms behind it. The day and night epoch panel is built for the identity throughout: coverage is aligned first, the mean is formed as (TMIN + TMAX)/2, and the period term is fitted separately for each element, so the relation holds to machine precision at every city-epoch.

The calculation sequence is also strictly defined. When computing period values, the urban temperature and the median rural temperature must each be averaged over the period first and differenced afterwards; differences must not be taken year by year and then averaged. The two orders differ by a median of 0.05°C across cities, which is small next to the per-city uncertainty and shows no systematic offset over the warming period. Residual sensitivity in cities with sparse coverage is indicated by the bootstrap standard error uhi_se (Technical Validation). The 20-year average requires at least 8 valid years.

That floor is permissive (40% coverage against the roughly 90% expected of a WMO climatological normal), and it is set there because the difference removes most of what year-to-year variation would otherwise cost. Such variation is largely regional, so it moves a city and its rural reference alike: an urban station's annual mean varies with a median standard deviation of 0.62°C against 0.27°C for the annual UHI, a reduction of about 56%. Tightening the rule therefore buys precision that is invisible next to the per-city uncertainty in Technical Validation, ± 0.31°C from the rural side alone. Recomputing the 2001–

2020 UHI at minimum coverages of 8, 12, 15, 18 and 20 valid years moves the pooled median only between + 0.181 and + 0.232°C and the negative-UHI share between 36% and 39%, while the usable sample falls from 1,117 cities to 506. Users who prefer a stricter rule can impose one from annual_by_elem.csv.

Temporal products

The same formula produces values at three temporal resolutions: annual, the five-year GHS epochs of 1975–2020, and 20-year windows on the long-record subset. Data Records lists the file each resolution is distributed in.

Both city files are climatologies over the same window, 2001–2020, and a station enters one only if it reports in at least eight of those twenty years. They differ in which stations were eligible to form the pairing, not in the period averaged. longrecord_city_uhi.csv is built from the long-record station set, so its cities are those whose thermometers hold up across the whole record; broad_city_uhi.csv is built from the 2000–2020 set and reaches more cities on shorter station histories. A user wanting a heat island averaged over the whole record has to build it from the annual panel. The two files are not interchangeable, because the samples differ (Technical Validation).

Data Records

The complete dataset is publicly available on Zenodo under a CC-BY-4.0 license (DOI: 10.5281/zenodo.22689038) and consists of the components listed in Table 3:

Table 3

The files this release comprises. Sizes and scopes are those of version v3, the version this paper describes, in which the duplicate urban-station matches were removed (Methods); the data dictionary below gives the columns of each file.

#	File	Contents	Scope
1	annual_by_elem.csv	station-level annual TMIN, TMAX and TAVG; the aggregated raw panel	8,637 stations, 1815–2026
2	city_station_match_broad.csv, city_station_match_longrecord.csv	urban and rural station lists per city, with distances, rural counts and country	1,314 and 818 cities
3	broad_city_uhi.csv	daily-mean and nighttime UHI with bootstrap uncertainty uhi_se (Technical Validation), urban-station distance and rural count	1,117 cities, 2001–2020
4	balanced_uhi_cities.csv	city UHI by 20-year window, long-record subset	1941–2020
5	decomposition_cities.csv, decomposition_cities_1971_90.csv	warming decomposition into global, regional and urban terms; the second is city-level on a 1971–90 baseline and is the one to reuse	derived product
6	station_overview.csv	station metadata: continent, income group, Köppen zone[12], record duration, operational status	8,637 stations
7	need_broad_meta.csv, need_stations_meta.csv	station id, latitude, longitude, elevation, name	7,794 stations
8	longrecord_city_uhi.csv	climatological UHI over 2001–2020 for daily mean, night and day, on the long-record station set	719 cities, 2001–2020
9	city_uhi_epoch_panel.csv	UHI at ten five-year epochs under all three rural-reference constructions; the within-city panel for trends	1,216 cities, 8,880 city-epochs; 910 and 6,736 screened
10	city_uhi_epoch_panel_daynight.csv	the same panel split into its nocturnal and daytime channels, all three constructions; the mean is formed from the extremes so that uhi_day = 2·uhi_mean − uhi_night holds exactly	887 cities, 5,893 city-epochs; 716 and 4,989 screened
11	city_population.csv	population and sample classification for stations-	1,352 cities

#	File	Contents	Scope
		linked cities	
12	city_centroids.csv	coordinates of the urban centres the contamination screen was built from	11,422 centres
13	homogenization_sensitivity.csv	raw against homogenized UHI enhancement, global and North American, behind Fig. 2 and Technical Validation	city counts per product
14	ghcnd-stations.txt	original GHCN-Daily station metadata used for the siting analysis in Technical Validation	132,501 stations
15	ghsl_population_2020.csv	2020 population per GHS urban centre on GHS boundary criteria	used only in Technical Validation
16	ghcnm_qcu_validation.csv	published daily-mean UHI against the same construction on GHCN-M v4 QCU (Technical Validation)	574 cities
17	refined_siting_cities.csv	city-level results of the station-siting check (Technical Validation)	city-level
18	city_area.csv	surface area of every GHS urban centre; the equivalent radius sqrt(area/pi) is the built-up edge the boundary rings of Table 5 and the clearance figures in Technical Validation are measured from	11,422 centres
19	duplicate_station_matches_dropped.csv	cities dropped where two or more urban centres resolved to the same nearest station, with each dropped city's distance to that station and the city it lost to (Methods)	655 dropped rows

Most of these can be used as they are. Four carry a caution that is not visible from the column headers, and getting one of them wrong changes the answer rather than the presentation.

The two matching files are not interchangeable (2). They select different stations even for the same city, agreeing on the urban station for 80% of the cities they share. Use the file matching the period being analyzed.

The daytime column is not independent information (8). uhi_tmax is the identity 2 × daily mean − nighttime, not a separate computation, and holds to machine precision across all 540 cities carrying all

three values. The column is supplied so that users need not form it themselves. Its uncertainty is the combined uncertainty of the two terms behind it, and an independent daytime estimate has to be recalculated directly from the station data.

The epoch panel offers three rural references (9). Users differencing across more than about two decades should take a composition-free column, not the varying one. Methods gives the argument and the data dictionary names the columns.

Population comes in two incompatible extractions (11, 15). In city_population.csv, pop is the population used for the baseline calculations in Technical Validation. ghsl_population_2020.csv is a separate extraction on GHS boundary criteria, running about 21% higher at the median, and substituting one for the other is not an update but a change in the target of estimation: it substantially alters the scaling coefficients (Technical Validation). File 15 should not be used outside Technical Validation.

Three notes on scope. city_centroids.csv carries coordinates only and no city identifier, so which centres it holds cannot be read from it directly. homogenization_sensitivity.csv cannot be regenerated from the release at all, because it needs the original GHCN-M v4 QCU and QCF monthly files; the same holds for the identically-processed correlation quoted in Technical Validation, which needs both archives rebuilt side by side. Those, like the raw daily files, are large and are not redistributed here; both can be obtained from their original source.

Data dictionary

All files are UTF-8 encoded CSVs with a header row. Temperatures are in degrees Celsius (∘C), coordinates are in decimal degrees (WGS84), and elevations are in meters. Missing values are left blank. Table 4 gives the columns file by file.

Table 4
Data dictionary. Every column of every deposited file, with its units and the section that defines it.

File	Column	Description
annual_by_elem.csv	id, year	GHCN-Daily station ID; calendar year
	tmin, tmax, tavg	station annual mean of daily minimum / maximum / average temperature (°C); tavg = (tmax+tmin)/2, reported TAVG as fallback
city_station_match_broad.csv	city_id, country	GHS-UCDB urban-centre ID; ISO-3 country
	lat, lon	urban-centre centroid
	urban, urban_km	matched in-city station ID; its distance from the centroid (km)
	n_rural, rural	number of rural-reference stations; ;-separated rural station IDs
broad_city_uhi.csv	city_id, country, continent, koppen	city keys and climate group (Köppen main class A/B/C/D/E)
	uhi_tavg, uhi_tmin, uhi_tmax	mean, nocturnal and daytime UHI over 2001–2020 (°C), elevation-corrected; uhi_tmax follows from the identity 2·uhi_tavg − uhi_tmin
	uhi_*_screened, n_rural_screened, uhi_se_screened	the same quantities with the 12 km rural contamination screen applied (Methods); the unscreened columns are the primary series
	uhi_se	per-city bootstrap standard error of UHI from rural-reference resampling (°C; Technical Validation)
	urban_km, n_rural	urban-station distance; rural-reference count (for user screening)
balanced_uhi_cities.csv	region, city_id	continent; city ID (balanced 1941–2020 subset)
	L0–L3	UHI **level** (°C) in 20-yr windows 1941–60, 1961–80, 1981–2000, 2001–20
	C0–C3	cumulative UHI **change** relative to the 1941–60 window (C_i = L_i − L_0; C0 = 0)
decomposition_cities.csv	region, country	continent; ISO-3 country (per-city rows with city_id are in decomposition_cities_1971_90.csv)
	dTurb, dTrur, dGMT	Δ urban-station warming; Δ rural warming; Δ global-mean warming (°C)

File	Column	Description
	amp, dUHI	regional amplification (dTrur − dGMT); urban excess Δ(urban−rural) (°C)
station_overview.csv	id, iso3, continent, income, koppen	station ID and attributes (World Bank income group, FY2020 classification (https://datahelpdesk.worldbank.org); Köppen)
	first, last, span, active2020, longrecord	first/last year with data; record length; active-in-2020 flag; long-record-set flag
need_broad_meta.csv	id, lat, lon, elev, name	station ID, coordinates, elevation (m), name
longrecord_city_uhi.csv	city_id, CityID	city keys (CityID is the GHS-UCDB urban-centre ID)
	uhi_tavg, uhi_tmin, uhi_tmax	mean, nocturnal and daytime UHI over 2001–2020 (°C), elevation-corrected; uhi_tmax follows from the identity 2·uhi_tavg − uhi_tmin
	uhi_*_screened, n_rural_screened	the same quantities with the 12 km screen applied (Methods)
	country, continent, koppen, urban_km, n_rural	city attributes and matching metadata
city_uhi_epoch_panel.csv	CityID, year	city key; five-year GHS epoch, 1975–2020
	uhi_obs, n_ref	city UHI (°C) under the composition-free anomaly reference, unscreened; the primary series (Methods)
	uhi_obs_varying, uhi_obs_fixed	the same epoch under the two alternative references: the median over whichever stations report, and the set valid at every epoch
	*_screened	every column above also in a screened variant, with its own n_ref
city_uhi_epoch_panel_daynight.csv	CityID, year	city key; five-year GHS epoch, 1975–2020
	uhi_mean, uhi_night, uhi_day, n_ref	the same panel split by channel, on coverage aligned so that uhi_day = 2·uhi_mean − uhi_night holds exactly (Methods)
	*_varying, *_fixed, *_screened	the same three constructions and two series as the mean panel

File	Column	Description
city_population.csv	CityID, city_id, pop, pop_sum	city keys; population used in the scaling analysis, and the GHS-UCDB total
	abslat, income, latband, kop, sampled	absolute latitude, World Bank income group, latitude band, Köppen class, in-sample flag
city_centroids.csv	lon, lat	coordinates of all 11,422 urban centres of GHS-UCDB R2024A, the list the 12 km rural-contamination screen is applied against (Methods)
city_area.csv	city_id, area_km2	GHS-UCDB urban-centre ID; surface area of the urban centre in square kilometers (GC_UCA_KM2_2025 of GHS-UCDB R2024A)

Technical Validation

Sensitivity to rural reference radii

On the unscreened series, we recalculated the 2001–2020 UHI by varying the rural reference range (Table 5). Varying the range from 25–75 km to 100–300 km keeps the overall median between + 0.19 and + 0.25°C and the North American median between + 0.28 and + 0.56°C, confirming that the results are not an artifact of the 40–150 km selection. North American values rise as the radius expands because the reference is drawn from cooler, less urbanized land.

A narrow ring costs coverage above all: at 15–50 km the usable sample falls from 1,054 cities to 469. What narrowing does to the level depends on whether the screen is applied. On the screened series the median falls from + 0.215 to + 0.203°C, which is the suburban contamination a narrow inner radius is expected to admit. On the unscreened series it moves only slightly, from + 0.192 to + 0.172°C, because a reference that already carries contaminated stations at every radius gains little more from being drawn closer. Users who vary these ranges should anticipate this magnitude of variation and state both the range and the series they used.

The reference point matters as little as the radii. All distances in this release are measured from the urban-centre centroid, so the clearance between the built-up edge and the inner rural limit narrows as city size grows: 36 km at the median city and 18 km at the 99th percentile. Rebuilding the record against the city boundary instead removes that dependence and changes little: across eight ring definitions spanning both reference points, the pooled median stays between + 0.19 and + 0.25°C and the negative share between 36% and 39%.

A boundary-referenced ring applies the same clearance to every city and costs nothing in coverage, at 1,080 cities against 1,054 here. The boundary ring is reported as a sensitivity rather than adopted,

because the GHS boundary is re-delineated at every epoch whereas the centroid is fixed for the life of the record (Methods).

Table 5
The 2001–2020 UHI under eight rural-reference rings. "Centroid" rings are measured from the urban-centre point, "boundary" rings from the built-up edge, taken as the equivalent radius of the GHS urban centre. All rows are computed on the unscreened series with the three-station minimum held fixed; the city count varies because a ring that reaches less far finds fewer qualifying references.

Reference ring	Cities	Median UHI (°C)	Negative share
centroid 40–150 km (this release)	1,054	+ 0.192	38%
centroid 40–100 km	727	+ 0.231	37%
centroid 40–250 km	1,171	+ 0.248	36%
boundary + 5 to + 105 km	841	+ 0.187	36%
boundary + 5 to + 145 km	1,087	+ 0.193	38%
boundary + 20 to + 150 km	1,080	+ 0.195	38%
boundary + 40 to + 105 km	775	+ 0.193	38%
boundary + 40 to + 150 km	1,067	+ 0.191	39%

Station siting environment

Categorized using network codes, WMO numbers, and GSN/CRN flags, airport/WBAN stations do not underestimate the UHI. When matched for city scale, they show values equal to or higher than cooperative stations, which are volunteer-run and typically sit within the town rather than at its airfield. Therefore, the approximately 34% of long-record cities carrying a negative daily-mean value are not caused by the siting environment.

Proximity of observation points

Urban stations lie a median of 5.5 km from the city centre, and 14.5 km at the 90th percentile. As distance increases, the measured UHI declines by only − 0.009°C/km ($p = 0.20$) for the daily mean and shows no trend at night (− 0.001°C/km, $p = 0.96$). Restricting the analysis to the cities whose urban station lies within 10 km moves the pooled median of uhi_tavg only from + 0.23 to + 0.25°C. Proximity is therefore not a material source of distortion, and the slight decline contributed by outlying stations is consistent with the record acting as a conservative lower bound.

Independent raw-data replication

Separate from the homogenization sensitivity test, which uses adjusted QCF data, we recalculated the UHI from an entirely different raw dataset, the unadjusted GHCN-M v4 monthly record (QCU). That archive shares neither GHCN-Daily's daily quality control nor its annual aggregation. Computed on the

same cities and with the same elevation correction, it reproduces the published values. Across the 574 cities both archives resolve, the correlation is r = 0.72, and r = 0.98 once both sides pass through identical simple processing. An independent reconstruction of that procedure, window-free, screen-free and minimally filtered, obtains r = 0.86 both before and after the duplicate-match correction. It does not reproduce the published coefficient exactly, but it shows that the correction leaves this comparison materially unchanged. The remaining difference stems from the public release's use of a fixed 2001–2020 window, its contamination removal and its daily aggregation. Aggregate levels agree, at + 0.30°C here against + 0.26°C for QCU, as do the negative shares, at 32% against 32%, with the two agreeing on the sign for 82% of cities. The published values are therefore not an artifact of GHCN-Daily or of its particular quality-control scheme. Because GHCN-M v4 distributes mean temperatures only, the nocturnal values are not cross-validated by this route.

City-level uncertainty and single-station representativeness

We provide a bootstrap standard error uhi_se for every city, from resampling the rural reference stations 200 times with replacement. The median city-level uncertainty from rural sampling alone is ± 0.31°C, with an interquartile range of 0.19 to 0.57.

The urban side contributes about as much. Of all cities, 42% have two or more candidate stations within 25 km, and those deviate from one another by a median of ± 0.32°C. For a city that offers such a choice the two sources are independent and combine to about ± 0.44°C; for the remaining 58%, which have a single candidate, only the rural term is measurable and the urban contribution is unquantified rather than absent. Either way the per-city figure is of order ± 0.3°C or larger, which is the scale against which the sensitivities elsewhere in this section should be read.

Among the 42% that offer a choice, the released nearest-station value and the average over all in-city candidates differ by a median of 0.23°C and by 0.71°C at the 90th percentile. The two constructions correlate at 0.90 on those cities and at 0.97 over all of them.

Taken over all cities the median difference is 0.000°C, but that is an artifact of the 58% with a single candidate, for which the two constructions are the same number, and it should not be read as evidence that the choice does not matter. The single-station rule makes city values reproducible rather than precise.

One quantity sits outside this budget. The displacement an eleven-month year introduces when the absent month is seasonally extreme (Methods) is a structural bias on particular station-years rather than a sampling spread.

Homogenization sensitivity

Compared against the homogenized counterpart of the same station network (GHCN-M v4 QCF)[13], the intensification of the urban heat island from 1971–1990 to 2001–2020 is reduced by about 60% globally

(Fig. 2) (+ 0.08 to + 0.03°C, n = 633/619) and by about 83% in North America (+ 0.21 to + 0.04°C, n = 244/234). An independent reconstruction of this comparison, run identically before and after the duplicate-station correction, confirms the correction changes neither the qualitative pattern nor its magnitude (global damping unchanged; North American damping if anything slightly less severe), though it does not reproduce the published city counts exactly. This reproduces in air temperature what Hausfather et al.[2] documented for the United States: pairwise homogenization[14] removes the urban signal directly from the temperature record. It also justifies the decision to omit homogenization here.

Standard adjusted datasets erase most of the signal this record is built to measure. That does not make them flawed; they remain well suited to their original purpose of providing unbiased climate records. That trade-off is precisely what makes a separate raw-data record necessary.

Replication verification from raw data

This release has two tiers: the raw observation record with its city-station matching, and the derived city tables built from it. The script verify_release.py recomputes a fixed set of summary quantities from the deposited files alone and exits with an error on any mismatch, so a reader holding the download can confirm that the second tier is the one described here. The reference values are given in Table 6.

Table 6
Reference values for the rebuild check. Each is a property of the released files rather than a scientific result, and between them they touch every stage of the pipeline, so a fault in the station panel, the matching, the rural median or the elevation correction moves at least one. verify_release.py recomputes them from the released files and exits non-zero on any mismatch.

Quantity	Reference value
annual_by_elem.csv: stations; station-years	8,637; 484,193
longrecord_city_uhi.csv: cities with daily mean / night / day	719 / 542 / 540
longrecord_city_uhi.csv: median UHI, daily mean / night	+ 0.277 / +0.554°C
longrecord_city_uhi.csv: negative share; median rural count	34.0%; 23
broad_city_uhi.csv: cities; median UHI; negative share	1,117; +0.182°C; 39.5%
city_uhi_epoch_panel.csv: cities unscreened / screened; city-epochs	1,216 / 910; 8,880
uhi_tmax = 2·uhi_tavg − uhi_tmin, largest residual	below 1×10^{-9} °C

Reproducing these values confirms that the analysis-ready tier follows from the raw one. They are checks on the files, not findings, and nothing in them is borrowed from an analysis.

Two common reuse pitfalls.

First, level analyses should use longrecord_city_uhi.csv rather than broad_city_uhi.csv. Both average the same 2001–2020 window, so the difference is not one of period but of which stations were eligible to

form the pairing: the broad file reaches 1,117 cities against 719, on urban stations whose records run a median of 76 years against 84. The 493 cities it adds are not a random extension of the sample. On the 624 cities the two files share, the broad file's median UHI is + 0.290°C, essentially the long-record value; on the cities it adds alone the median is + 0.045°C and 48% of them read negative. Its pooled median of + 0.182°C is that mixture, so an analysis run on the broad file is answering a question about a different set of cities.

Second, population variables must use pop from city_population.csv rather than a fresh extraction from GHS-UCDB. ghsl_population_2020.csv is such an extraction on GHS boundary criteria, and it is not a rescaling of pop. The ratio between the two has a median of 1.21, an interquartile range of 1.01 to 1.50 and a 5th-to-95th-percentile range of 0.82 to 2.32, and for 24% of cities the GHSL value is the lower of the two. The two therefore reorder cities by size, with 7% crossing a size-class boundary, and any analysis conditioning on city size gives a different answer depending on which is used. File 15 should not be used outside this section.

Sensitivity to the lapse rate

The elevation term uses a fixed Γ = 6.5°C/km, and it carries little, for a reason visible in the data: the urban station and its rural median sit at nearly the same height. Across the long-record cities the urban-minus-rural elevation difference has a median of − 13 m, a 5th-to-95th-percentile range of − 278 to + 223 m, and exceeds 100 m in absolute value for only 27% of cities, so Γ multiplies a small number.

Rebuilding the daily-mean record from the released station panel and station metadata with Γ swept from 0 to 9.8°C/km moves the pooled median UHI from + 0.340°C, with no correction at all, to + 0.214°C, and the negative-UHI share from 32.4% to 39.0%. Across the physically defensible range alone (5.0°C/km, a moist adiabat, to 9.8°C/km, the dry adiabat), the median moves only from + 0.286 to + 0.214°C and the negative share from 31.4% to 39.0%. Relative to the released Γ = 6.5°C/km the median shifts by at most 0.051°C in either direction, an order of magnitude below the per-city uncertainty reported above, which is ± 0.31°C from the rural side alone. The lapse rate is therefore not a material driver of the released values. A user who prefers a different Γ can recompute the correction from need_stations_meta.csv and need_broad_meta.csv, which carry the station elevations. These figures come from a rebuild that reproduces the released series exactly at Γ = 6.5°C/km, so the sweep is the released record recomputed, not an approximation to it.

Usage Notes

Getting started. Readers new to the record should open longrecord_city_uhi.csv first, the smallest of the city-level files. It carries all three temperature elements (daily mean, day and night) for 719 cities and is the basis of the reference values in Technical Validation, so it gives the quickest picture of what a UHI value in this record looks like. broad_city_uhi.csv then extends the coverage to 1,117 cities for 2001–2020, and city_uhi_epoch_panel.csv adds the within-city time dimension. The station-level panel annual_by_elem.csv is needed only by users who intend to rebuild the UHI from scratch. Data Records gives the full inventory. Three files in it (the balanced panel and the two decomposition tables) are

derived products computed on the record rather than part of it, and no claim in this descriptor rests on them.

Screened and unscreened values, and which to use. The record is distributed with and without the 12 km rural contamination screen, because the screen is not equally justified everywhere and the choice between the two columns is not neutral with respect to geography. Screening retains 94% of North American and 80% of European cities but 44% of Asian, 20% of South American and 9% of African ones, and 19% of the tropical sample. Whether that loss buys anything can be measured directly: within each city's own set of rural candidates, the stations the screen removes should run warmer than those it keeps, since removing a warm reference is the whole point. Tables 7 and 8 report that comparison, elevation-adjusted, city by city.

Table 7
Does the screen remove warmer stations? By region. A positive value means the screen is removing the contamination it exists to remove. The African row rests on six cities and the tropical row on thirteen; both should be read as indicative only.

Region	Cities	Warm excess of removed stations (°C)	Share above zero
Oceania	24	+ 0.15	58%
North America	304	+ 0.15	57%
Europe	303	+ 0.14	67%
Asia	88	−0.06	43%
Africa	6	−1.33	17%
all cities	731	+ 0.13	59%

Table 8
The same comparison by climate zone. The arid row is the informative one: there the screen removes stations that are 0.74°C cooler than those it keeps, and − 1.09°C in arid North America.

Köppen group	Cities	Warm excess of removed stations (°C)	Share above zero
D continental	279	+ 0.31	72%
C temperate	356	+ 0.04	55%
A tropical	13	+ 0.07	69%
B arid	83	−0.74	34%
all cities	731	+ 0.13	59%

Two failures are visible, and they have different causes. In Asia and Africa the removed stations are no warmer than the retained ones. A GHS urban centre is defined by population while a heat island is

produced by built fabric, and a settlement of fifty thousand in sub-Saharan Africa is not the same physical object as one in Germany. There the rule discards stations without removing contamination. In arid zones the sign is wrong altogether, because irrigation and urban vegetation can make the settlement cooler than the desert around it (the cool-island condition behind the negative values reported in Technical Validation), so screening biases the reference toward pure desert and inflates the measured heat island. The screen is therefore justified where it was designed: temperate and continental cities in the developed world, where it removes a genuine + 0.1 to + 0.3°C of contamination. Analyses confined to those settings should prefer the screened column. Analyses that turn on regional or climatic contrast, or that reach into the tropics, the drylands or the developing world, should use the unscreened column, where the screen costs coverage without buying cleanliness, and should note that unscreened values carry a downward bias of about 0.03°C at the pooled median.

Uneven geographic distribution; weight or stratify, and report the sensitivity. About 77% of the network is in North America and about 79% in high-income regions, so a global average over this record is effectively an average over developed nations. The imbalance is sharper still in the long-record subset. Sensitivity to it has to be checked per statistic: removing the United States or balancing by continent moves aggregates in either direction, so an unweighted mean is not automatically the conservative choice. Continent and country fields ship in the city files and station_overview.csv supports network stratification. Relationships for tropical, developing and low-income cities cannot be recovered from this record alone, and require satellite or crowdsourced temperature data instead [4].

Station surroundings are not stationary over the record. A station's immediate environment can change across two decades through road construction, building work or altered vegetation, and because no homogenization is applied such changes enter the measured UHI rather than being removed from it. Gradual urban change is what the record exists to capture, by design, but a single station-year should not be read as a clean standalone measurement of its city. Aggregate over cities, or use the 20-year windows provided, instead of working at that granularity.

"Levels" and "trends" target different phenomena. Use the cross-sectional city files to ask why cities differ from one another, and the epoch panel with city fixed effects to ask how a city has changed. The two serve different purposes, and the difference is not one of precision.

Daytime and nighttime. Nighttime (TMIN) carries the strongest and cleanest urban heat island signal, and the daily mean (TAVG) is a conservative lower bound on it. The daytime channel is distributed in both the city files and the epoch panel, but it is derived from an identity rather than observed directly, so it carries the combined uncertainty of the two terms behind it (Methods).

Data availability

The dataset is publicly available on Zenodo under a CC-BY-4.0 licence, and two DOIs point to it for different purposes.

• **Concept DOI (**10.5281/zenodo.22006932**)** represents all versions and always resolves to the most recent one. Users who want the current data should bookmark this one.

• **Version DOI (**10.5281/zenodo.22689038**)**[15] is the fixed identifier for the version this paper describes. Cite this one to reproduce the numbers reported here.

Every file can be downloaded directly through either DOI without registration or log-in. The processing code is deposited separately under an MIT licence in its own Zenodo record (below), linked to this one through related-identifier metadata, so that data and code can be versioned independently. Because GHCN-Daily continues to be updated by NOAA, this record is a snapshot rather than a live service; future revisions will be registered as new versions under the same concept DOI.

Input data and licensing. GHCN-Daily is in the public domain as a U.S. government work (NOAA/NCEI). GHS-UCDB R2024A is released under CC-BY-4.0 by the European Commission Joint Research Centre, and the unadjusted GHCN-M v4 (QCU) data used for the validation in Technical Validation is likewise in the public domain (NOAA/NCEI). The global term of the decomposition is the Berkeley Earth land and ocean record[16]. Berkeley Earth licenses its data under CC BY-NC 4.0, for non-commercial use only, which CC-BY-4.0 does not permit redistributing. The series itself is therefore not included in this release; obtain it from www.berkeleyearth.org under their terms. What remains is dGMT, a single global-mean warming value per reference period, carried as a constant column in the decomposition tables. A scalar read off a published series is a fact rather than a substantial part of a database, and it is attributed to Berkeley Earth throughout. The coastline and land polygons in Fig. 1 are from Natural Earth[11], public domain. The large source files (the GHCN-Daily daily files and the GHCN-M v4 QCU and QCF monthly files) are not redistributed here; a one-line command to download the former is provided, and the latter enter only through pre-computed tables.

Code availability

All processing and analysis code (Python) is developed openly at https://github.com/ichwang79/UHI_AIR and archived under the Zenodo concept DOI 10.5281/zenodo.22006819, which always links to the latest release. The version described in this paper is v1.2.3, DOI 10.5281/zenodo.23190822. All code is available under an MIT licence without registration.

The repository covers raw-data parsing, city-station matching, UHI construction and every check in Technical Validation that the deposited files support, which is all of them but the two quantities Data Records names as needing the GHCN-M v4 monthly archives. It includes verify_release.py, which recomputes the reference values in Technical Validation from the released files alone and exits with an error on any mismatch, and reproduce_manuscript.py, which does the same for every quantitative claim in this paper's text and tables, together with fig_coverage.py and make_hero_figure.py, which redraw Figs. 1 and 2.

One script needs an input this release does not carry. build_decomposition.py rebuilds the warming decomposition from the global-mean series, and that series is Berkeley Earth's, withheld for the licence reason given above. Pass it with --gmt, having downloaded it from www.berkeleyearth.org; the script requires the option and does not default to a file that is no longer here. Nothing else in the repository depends on it, and the deposited decomposition tables already carry dGMT as a column, so they can be used without rerunning that step.

Declarations

Author contributions

In Chang Hwang is the sole author and is responsible for all CRediT roles: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, and writing of the original draft.

Competing interests

The author declares that he has no conflict of interest.

Acknowledgements

I thank NOAA/NCEI for maintaining and openly distributing GHCN-Daily and GHCN-M v4, and the European Commission Joint Research Centre for the GHS Urban Centre Database. During the preparation of this study, the author used Claude (Anthropic) to assist with writing data-processing and validation code, generating figures, and editing the manuscript. All analysis code was reviewed by the author before execution, and all data records, technical validation results, tables, figures and references were verified by the author, who takes full responsibility for their content. The scientific interpretation and all substantive claims are the author's own.

Funding

This work was supported by the Korea Environmental Industry and Technology Institute (KEITI) through the R&D Program for Response to a New Climate Regime, funded by the Ministry of Environment (RS-2023-00218794). The Seoul Institute supported the project *Sectoral Climate Change Damage Costs and Policy Benefits for Seoul* (2026-BR-17).

References

1. Karl, T. R., Diaz, H. F. & Kukla, G. Urbanization: its detection and effect in the United States climate record. J. Clim. 1, 1099–1123 (1988).

2. Hausfather, Z. et al. Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records. J. Geophys. Res. Atmos. 118, 481–494 (2013).
3. Sun, Y., Zhang, X., Ren, G., Zwiers, F. W. & Hu, T. Contribution of urbanization to warming in China. Nat. Clim. Change 6, 706–709 (2016).
4. Venter, Z. S., Chakraborty, T. & Lee, X. Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. Sci. Adv. 7, eabb9569 (2021).
5. Peterson, T. C. Assessment of urban versus rural in situ surface temperatures in the contiguous United States: no difference found. J. Clim. 16, 2941–2959 (2003).
6. Hansen, J., Ruedy, R., Sato, M. & Lo, K. Global surface temperature change. Rev. Geophys. 48, RG4004 (2010).
7. Muller, R. A. et al. A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinfor. Geostat. Overv. 1:1 (2013).
8. Wickham, C. et al. Influence of urban heating on the global temperature land average using rural sites identified from MODIS classifications. Geoinfor. Geostat. Overv. 1:2 (2013).
9. Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. & Houston, T. G. An overview of the Global Historical Climatology Network-Daily database. J. Atmos. Ocean. Technol. 29, 897–910 (2012).
10. Mari Rivero, I. et al. GHS-UCDB R2024A — GHS Urban Centre Database 2025. European Commission, Joint Research Centre [data set] https://doi.org/10.2905/1a338be6-7eaf-480c-9664-3a8ade88cbcd (2024).
11. Natural Earth. Free vector and raster map data. naturalearthdata.com (public domain, 2024).
12. Beck, H. E. et al. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5, 180214 (2018).
13. 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).
14. Menne, M. J. & Williams, C. N. Homogenization of temperature series via pairwise comparisons. J. Clim. 22, 1700–1717 (2009).
15. Hwang, I. C. A raw, homogenization-free air-temperature urban heat-island dataset for 1,117 cities from GHCN-Daily, 1941–2020. Zenodo https://doi.org/10.5281/zenodo.22689038 (2026).
16. Rohde, R. A. & Hausfather, Z. The Berkeley Earth Land/Ocean Temperature Record. Earth Syst. Sci. Data 12, 3469–3479 (2020).

Figures

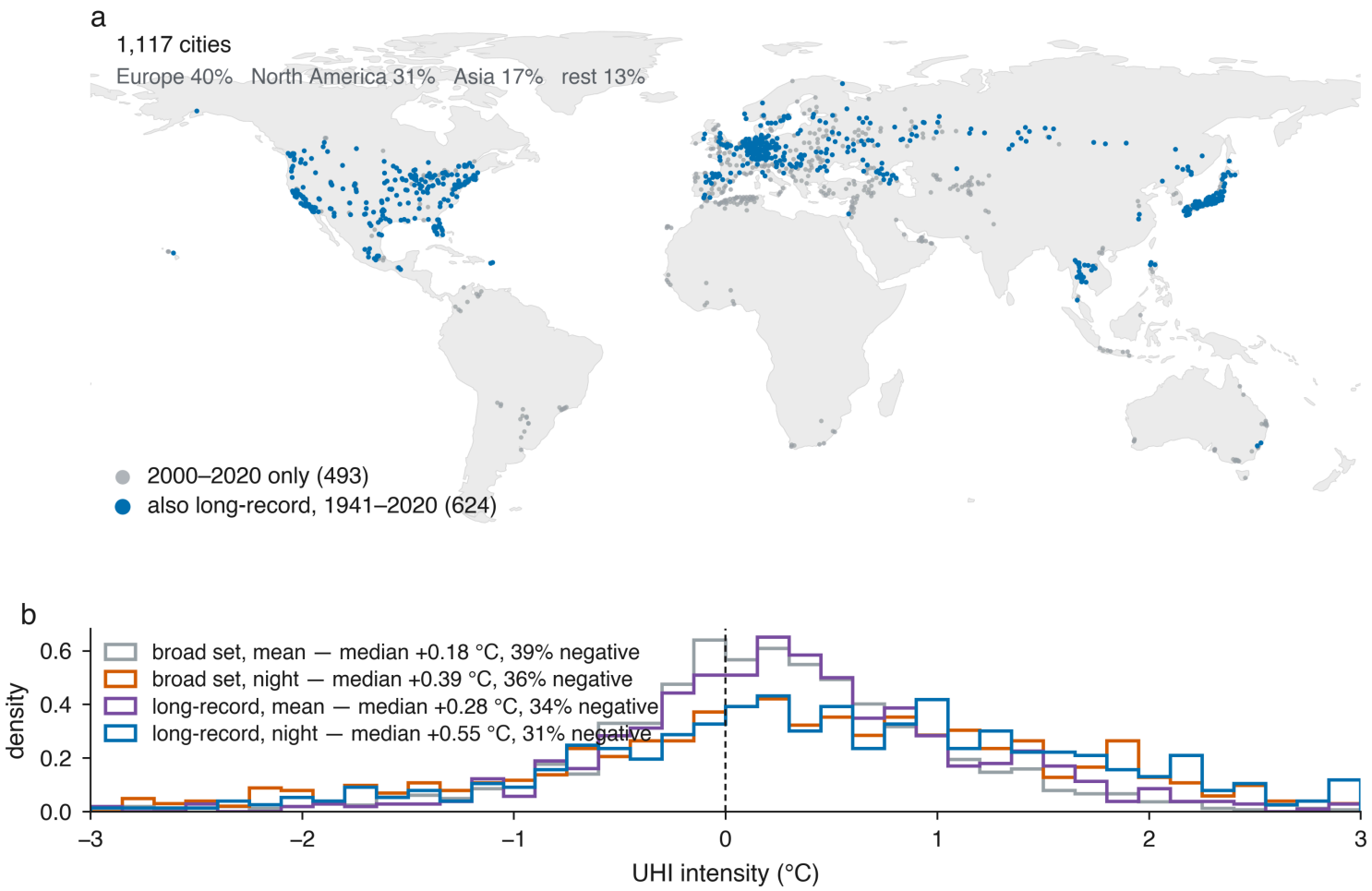


Figure 1

What is in the record, and where. (a) All 1,117 cities of the 2001–2020 broad set, with the 624 that also carry a long record (station spans ≤1955 to ≥2010) marked in blue; those are the cities the 1941–2020 analyses run on. The developed-world concentration the Usage Notes warn about (Europe 40 %, North America 31 %, Asia 17 %, everywhere else 13 %) is the first thing the map shows, and it bounds every statistic computed from this record. (b) The UHI distribution on both working sets, for the annual mean and for the night. A large minority of cities read negative, and how large depends on which element and which set is used: 39.5 % on the broad set's annual mean against 31.0 % on the long-record night. Medians run +0.18 to +0.55 °C. Users should choose the set and element that match their question before computing anything (broad_city_uhi.csv, longrecord_city_uhi.csv; scripts/fig_coverage.py; land polygons from Natural Earth[11], public domain).

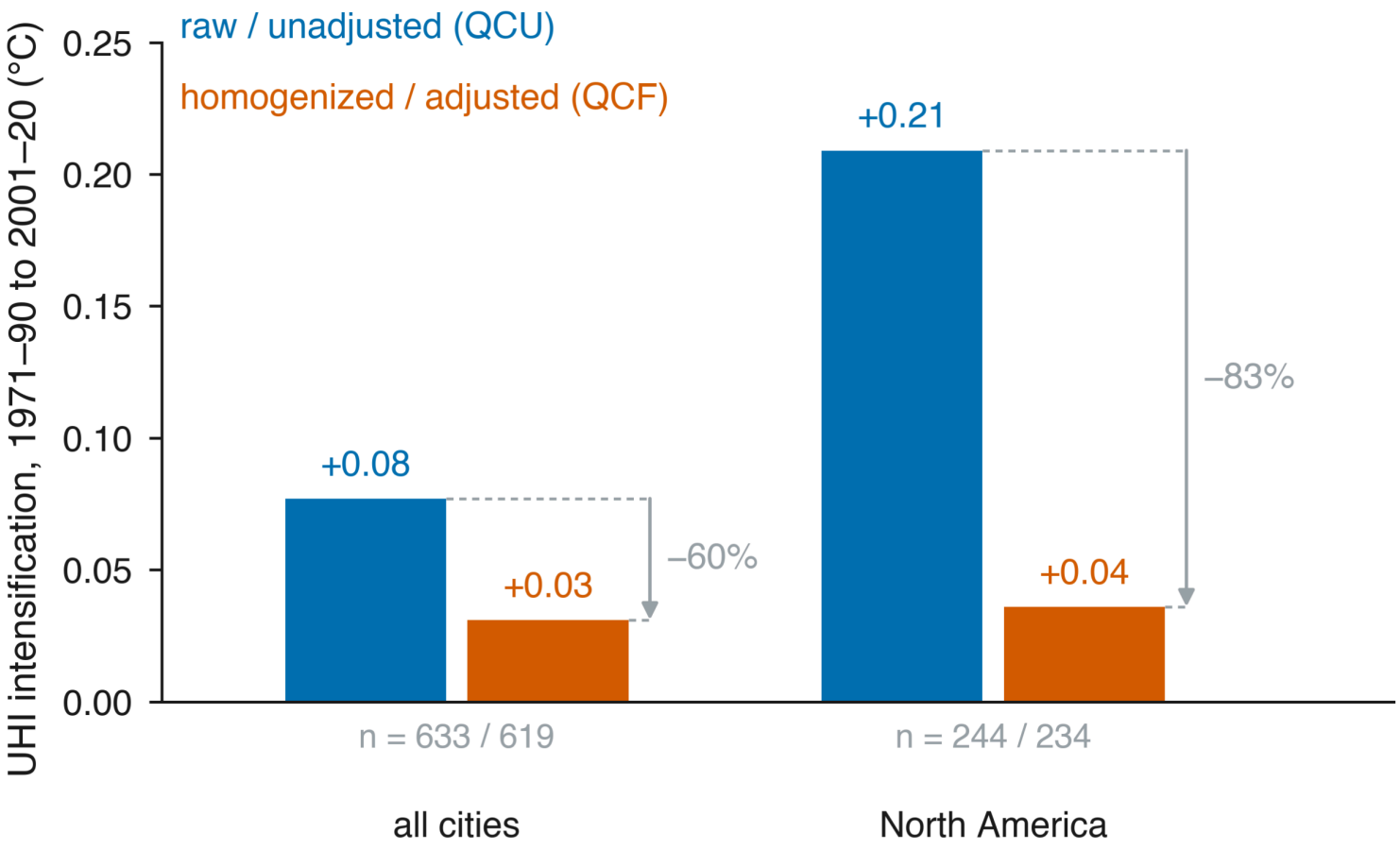


Figure 2

Why the release is homogenization-free. UHI intensification from a 1971–90 baseline to 2001–20, computed on the raw/unadjusted GHCN-M v4 monthly record (QCU, blue) versus the homogenized/adjusted version of the same network (QCF, orange). Homogenization removes most of the signal this dataset is built to measure: North America's +0.21 °C intensification falls to +0.04 °C (−83 %) and the global aggregate falls by 60 % (+0.08 → +0.03 °C). This happens because a gradual urban trend relative to rural neighbors is exactly the local inhomogeneity pairwise algorithms target. Distributing a quality-controlled but non-homogenized UHI record follows from that trade-off; sample sizes are annotated (homogenization_sensitivity.csv).