Nature Communications

https://doi.org/10.1038/s41467-026-73716-7

Article in Press

Transfer learning reveals large discrepancies between air and land surface temperatures in cities

Received: 7 June 2025

Accepted: 18 May 2026

Cite this article as: Zhang, Y., Zhao, L., Chakraborty, T. *et al.* Transfer learning reveals large discrepancies between air and land surface temperatures in cities. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-73716-7

Yiwen Zhang, Lei Zhao, TC Chakraborty, Priyam Mazumdar, Keer Zhang & Pierre Gentine

Transfer learning reveals large discrepancies between air and land surface temperatures in cities

Authors: Yiwen Zhang[1], Lei Zhao[1,2,3,4*], TC Chakraborty[5], Priyam Mazumdar[6], Keer Zhang[7], Pierre Gentine[8*]

Affiliations:

[1]Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign; Urbana, IL, USA

[2]National Center for Supercomputing Applications, University of Illinois Urbana-Champaign; Urbana, IL, USA

[3]Institute for Sustainability, Energy, and Environment (iSEE), University of Illinois Urbana-Champaign; Urbana, IL, USA

[4]Department of Climate, Meteorology & Atmospheric Sciences, University of Illinois Urbana-Champaign; Urbana, IL, USA

[5]Pacific Northwest National Laboratory; Richland, WA, USA

[6]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign; Urbana, IL, USA

[7]High Meadows Environmental Institute, Princeton University; Princeton, NJ, USA

[8]Department of Earth and Environmental Engineering, Columbia University; New York, NY, USA

*Corresponding author. Email: leizhao@illinois.edu, pg2328@columbia.edu

Abstract: Understanding of urban weather and climate is severely limited by data poverty resulting from a dearth of true urban weather stations. As a result, land surface temperature (T_s), obtained from remote sensing platforms, has been widely used as a stand-in for near-surface air temperature (T_a) despite their fundamental differences, especially in urban areas. Although T_s provides important scientific insights and practical utility for urban climate studies, this substitution risks introducing large uncertainties and biased characterization of urban heat stress and urban climate impacts. Here we develop an urban transfer learning framework (U-TL) to address this critical gap and to provide urban high-resolution air temperature (U-HAT) data at large scales across the contiguous United States (CONUS). U-TL demonstrates high accuracy and strong robustness in predicting urban T_a, even with limited training data. The resulting U-HAT is a high-resolution urban T_a dataset capable of accurately reproducing observed and well-established urban climatology. U-HAT reveals substantial T_s–T_a discrepancies and therefore cautions the use of T_s to characterize urban heat. We show that satellite-measured T_s substantially overestimates both urban heat stress magnitude and intra-city spatial variability, which have consequential implications for urban heat exposure, vulnerability, and adaptation policy making.

Introduction

Cities, despite being widely recognized as global centers of climate hazards and exposure[1–3], remain severely underrepresented in terms of in-situ meteorological observations. This gap stems primarily from the scarcity of research-grade weather stations in urban areas[4,5] (Fig. 1). Standard weather stations are disproportionately placed in non-urban landscapes (Fig. 1a,b). Even those stations that are classified as "urban" (see Fig. 1c–g) are most often located at airports or adjacent open fields, thereby failing to capture the true meteorological and climatological characteristics of the built environments[6–8]. A key reason for this misrepresentation lies in the World Meteorological Organization (WMO) siting guidelines, which stipulate weather stations to be installed on open, grass-covered ground and far away from built obstructions[9]. While the WMO guidelines aim to ensure consistent, uncontaminated measurements for capturing broader weather and climate signals, adherence to these standards effectively precludes the placement of conventional weather stations in true built environments, leaving only peripheral locations such as airports or city outskirts as apparent viable options[10]. Consequently, very few stations are situated in core urban areas worldwide. Even in the rare occasions where stations do exist in true built environments, their sparse presence (e.g., just one or two stations per city) fails to capture the inherent large intra-city heterogeneity of urban landscapes[11]. This pervasive urban data poverty has been a major obstacle to urban climate research for decades, hindering observation-based studies, compromising validation efforts for climate and weather models, and limiting data assimilation in historical reanalysis products and weather forecasting systems, particularly for urban areas[12].

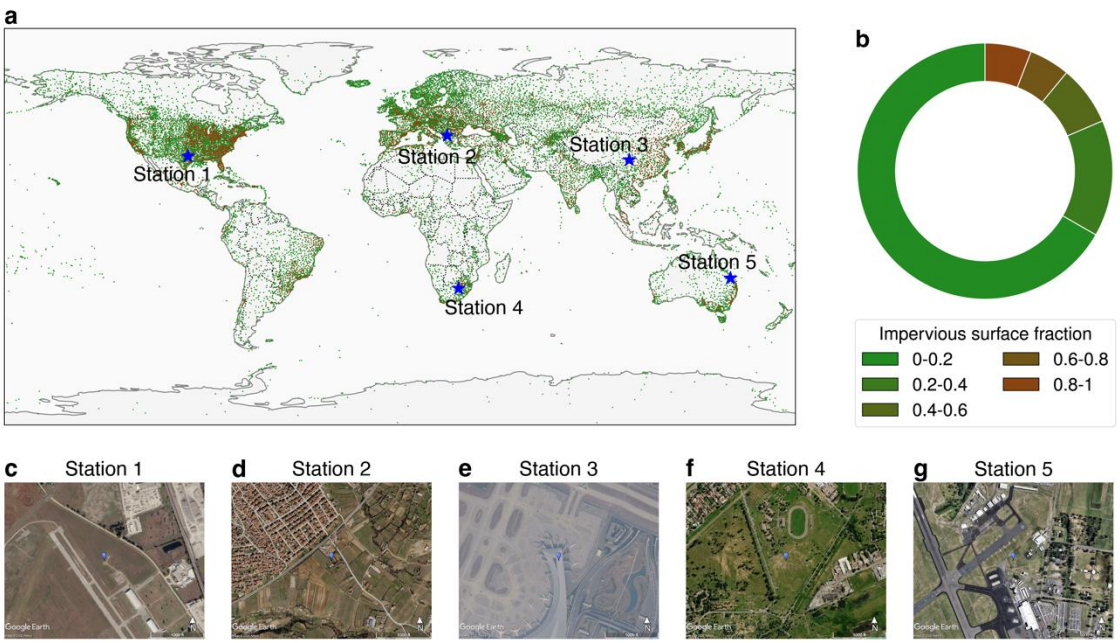

Fig. 1: Limited true urban representation in global weather stations. a, Meteorological stations in the Global Historical Climatology Network Hourly[13]. Stations located within urban boundaries as defined by the Global Urban Boundaries dataset[14] are shown as brown dots while those outside are in green. Blue stars indicate selected stations that are considered "urban" based on this broad classification. **b,** Proportion of meteorological stations categorized by impervious surface fraction within a 200-meter buffer. Impervious surface fraction is based on the "built-up" category in the ESA WorldCover dataset[15]. **c-g,** Satellite images from Google Earth (http://earth.google.com) showing the surrounding environments of selected stations. Basemap from Natural Earth (https://www.naturalearthdata.com/).

Land surface temperature (T_s), a bulk radiative skin temperature of objects on the ground as estimated from a satellite view, has therefore been commonly used as a stand-in for near-surface air temperature at the screen height (T_a) to study the urban thermal environment in recent

years[16–20]. In many cases, findings based on T_s have been used to discuss implications for an amorphous "urban heat", which can be often confused with those from T_a[21]. Satellite-measured T_s offers several advantages over sparse weather stations, including global coverage, consistency through measurements from the same satellite sensor, and spatial continuity[22,23], making it particularly suitable for large-scale studies and detailed intra-urban analyses. Recent advances in drone sensing can also provide on-demand, meter- to submeter-level T_s observations that capture very fine-scale urban heterogeneity at both high spatial and temporal resolutions[24]. These T_s-based applications provide important scientific insights and practical utility into urban climate studies[25,26]. For example, T_s is directly related to urban radiative thermal environment, provides the lower boundary condition for the atmosphere above, and directly modulates the surface energy balance and land-atmosphere interactions. However, unlike T_a which is strongly correlated with human-perceived heat stress, T_s does not directly reflect physiologically-relevant heat hazard and are thus subject to critical limitations when drawing impact- or risk-related implications such as those related to public health or urban adaptation policies[27].

Indeed, T_s itself does not inherently represent human-perceived thermal conditions or other impact-relevant variables (e.g., heat-related health risks or energy use). Therefore, studies relying on T_s to infer urban climate impacts and adaptation strategies have to assume – explicitly or implicitly – that either (i) T_s has a strong spatial correlation with T_a within cities, or (ii) urban-rural difference in T_s (surface urban heat island or UHI_s) is strongly correlated with urban-rural difference in T_a (canopy urban heat island or UHI_a). However, multiple recent observational studies in diverse cities or regions have demonstrated that neither of the two assumptions is justified[6,27,28]. T_s and T_a, while related, are conceptually and physically distinct[29]. Their

relationship is influenced by the complex interplay of many factors, including land surface properties (e.g., vegetation cover, imperviousness, and aerodynamic roughness), sky conditions (e.g., water vapor and cloud cover), and boundary layer processes[29–32], especially in highly heterogeneous urban landscapes. This raises significant concerns about the validity of using T_s as a proxy for T_a, as this substitution can lead to substantial uncertainties and systematic errors in the assessments of urban heat stress and climate impacts. For example, T_s generally shows stronger modifications due to urbanization than T_a or more comprehensive heat stress indices, thereby leading to quantitative, and sometimes qualitative, errors in such assessments.

Recent observational-based attempts to address this critical urban data gap have primarily focused on two approaches: crowdsourced citizen weather station (CWS) measurements[33,34] and empirical data-driven methods to infer urban T_a[35,36]. The former has emerged as a promising avenue for local and regional urban observational studies due to the recent proliferation of CWS networks[27,37–39]. Their measurements, however, are subject to two critical limitations. First, CWS measurements are prone to substantial errors, biases, and uncertainties due to improper station placement (e.g., on rooftops or near HVAC [Heating, ventilation, and air conditioning] exhausts), lack of calibration and maintenance, instrument or internet malfunctions, and inconsistent metadata[34,40,41]. Second, CWSs are subject to uneven distribution both within cities and across regions[42,43]. They are, by design, primarily installed in more conveniently accessible locations (e.g., backyards of residential buildings), and often in more affluent neighborhoods or regions, leading to sampling biases that undermine the representativeness of their measurements. These factors significantly limit the reliability of CWS data and complicate their use for global or large-scale urban climate studies.

On the other hand, data-driven approaches aim to empirically infer urban T_a using standard observations and/or auxiliary parameters. Broadly, these methods fall into two categories: spatial interpolation from standard weather stations[44–46] and statistical/machine learning techniques to predict urban T_a[47–54]. Both, however, are fundamentally constrained by the overall dearth of stations in true urban environments. The former often produces interpolated T_a datasets substantially biased toward non-urban landscapes due to their much higher station density, resulting in substantial inaccuracy in urban areas[35]. The latter, particularly data-intensive deep learning techniques, struggle to establish robust relationships between predictors and urban T_a due to insufficient representative training samples from urban cores. Many of such methods have to rely heavily on T_s as a predominant input feature. Consequently, their T_a products often fail to reproduce observed urban climate signals, for example, the well-established reversal of diurnal patterns for UHI_a and UHI_s, preventing their applications in urban climate research[55].

To address this critical data poverty issue, we present a framework based on the idea of transfer learning, termed U-TL (Supplementary Fig. S1a), which provides urban high-resolution air temperature (U-HAT) data at large scales. U-HAT is developed at 1-km resolution, covering 384 largest cities in the contiguous United States (CONUS) from 04-01-2013 to 12-31-2023 twice a day, at 1:30 AM and 1:30 PM local time. Comprehensive validation of the U-HAT data demonstrates strong performance in capturing urban-specific climate signals and gradients with high fidelity. Unlike existing methods, our framework effectively reproduces observed urban climatology both within and across cities, addressing a longstanding barrier in urban climate

research. Future releases of U-HAT will extend coverage to all urban areas globally. Combining U-HAT with T_s observations, we provide a comprehensive assessment of urban T_s–T_a differences on a pixel-to-pixel level at continental scale. We find large discrepancies between air and land surface temperatures in cities in both the magnitude and intra-city spatial variability in urban heat exposure.

Results

U-TL: an urban transfer learning framework for air temperature reconstruction

We overcome the fundamental barrier of sparse true urban weather stations by leveraging a transfer learning approach, a machine learning technique which addresses data scarcity by adapting knowledge from data-rich, pretrained models to data-limited tasks through "fine-tuning"[56]. The U-TL framework consists of two main steps (Supplementary Fig. S1a; Methods). In the first step, we pretrain a deep neural network (DNN) model to predict urban T_s, a data-abundant target due to extensive satellite observations, as a function of meteorological forcings and surface properties. This pretrained model is then fine-tuned with representative, ground-based true urban T_a observations to estimate urban T_a in the second step. Both T_a and T_s are essentially governed by urban land-atmosphere interactions. U-TL leverages this shared physics by initially learning a generalized representation of the land-atmosphere interactions driving urban T_s in the first step, and then transferring that knowledge to a different, yet related, task of estimating T_a via fine-tuning in the second step. This physical basis distinguishes U-TL from previous statistical models that directly map T_s to T_a based solely on empirical correlations, which often break down in data-scarce and spatially heterogeneous urban environments. Because

U-TL acquires most of its "understanding" of urban land-atmosphere interactions in a data-rich environment (i.e., using urban T_s labels), only a small amount of true urban T_a data is required during the fine-tuning step for accurate prediction, making the approach both efficient and scalable. In other words, the purpose for predicting T_s in the first step is to pretrain the model with a data-rich proxy so that the knowledge gained can be transferred to the target task of estimating T_a.

Urban land-atmosphere interactions are primarily characterized by meteorological forcings and surface properties (Supplementary Fig. S1b). To emulate these interactions, the DNN in the pretraining step ingests the atmospheric forcing fields and urban surface properties as predictors (Supplementary Table S1) and outputs the corresponding urban T_s (Methods). For the atmospheric forcings, we include all the fields commonly required to drive an urban land model in climate or weather models[57,58]. The urban surface properties are represented using satellite imagery, surface elevation, and building height. This design allows the DNN to predict urban T_s in a manner reminiscent of climate or weather modeling, but "solving" for T_s "statistically" rather than "numerically", as the model takes the same input variables as a numerical model would require for solving a system of equations[57]. In the second step, the pretrained DNN is further fine-tuned using urban T_a observations from 52 weather stations that were carefully selected based on our site-selection criteria (Methods). These stations are sourced from 18 cities across diverse climate regimes (Supplementary Fig. S2).

We conduct a thorough validation of U-TL against unseen observations and established urban temperature climatology (Methods). U-TL exhibits high accuracy in predicting urban T_a across scales (Supplementary Fig. S3), and this strong overall predictive power is achieved with very limited urban training labels. It is able to reproduce the observed diurnal, daily, and spatial variabilities of urban T_a (Supplementary Fig. S4). Importantly, such performance remains highly robust under increased data scarcity (Supplementary Fig. S5) – a challenge mirroring the real-world extreme scarcity of true urban observations.

U-HAT reproduces observed urban climatology

As discussed above, inconsistencies with observed urban climatology have long hindered the widespread adoption of existing T_a datasets in urban climate and weather studies. Many such products, including most reanalysis datasets[59,60], fail to capture urban signals such as the UHI effect. Even those specifically designed to incorporate urban signals often fail to reproduce observed spatiotemporal patterns of UHI_a. For instance, a majority of observational studies have concluded that UHI_a is typically stronger at night than during the day, whereas UHI_s often shows the opposite[6,23,61,62]. However, many existing urban T_a datasets reveal a contradictory diurnal relationship, featuring a stronger UHI_a during daytime rather than at night[55]. This discrepancy largely arises from the fact that those datasets derive urban T_a directly from urban T_s, effectively making them "magnitude-adjusted T_s" products.

Here we compared the diurnal and seasonal UHI_a patterns in U-HAT against two recent T_a datasets[48,49] as illustrative examples. We employed a Simplified Urban Extent (SUE) method[63]

(Methods), which defines UHI as the difference in temperatures between the urban and non-urban pixels within each urban extent. We acknowledge that using non-urban pixels inside urban boundaries might introduce exaggerated "rural" reference temperatures. To ensure the robustness of our findings, we repeated these analyses using the PRISM data[64] and ERA5 reanalysis[65] as rural reference temperatures (Methods).

Our results show that U-HAT consistently captures the observed UHI_a trends remarkably well across calculation methods. With U-HAT, the nighttime UHI_a intensity is consistently more pronounced than during the day (Fig. 2a, Supplementary Fig. S6). For example, the average summer (winter) daytime UHI_a across CONUS cities shown by SUE is 0.15 K (0.14 K), which is 0.71 K (0.31 K) lower than nighttime values. In contrast, the average summer (winter) UHI_s, measured by T_s is 1.55 K (0.72 K) higher during the day than at night. Observational studies have also noted that UHI_s tends to be higher in summer and lower in winter, whereas UHI_a does not necessarily[66]. These diurnal and seasonal differences are however misrepresented in the two existing T_a datasets—Yao et al.[48] and Zhang et al.[49]. Using SUE, both datasets show no significant diurnal differences, and consistently report higher summer UHI_a than winter. With PRISM and ERA5, they both demonstrate stronger UHI_a during daytime than at night and a greater summer peak than winter, a diurnal and seasonal pattern largely reflecting that of UHI_s. For instance, Yao et al.[48] and Zhang et al.[49] show a 0.95 K (0.66 K) and 0.85 K (0.60 K) higher average summer (winter) daytime UHI_a than nighttime using PRISM, respectively. In other words, UHI_a patterns demonstrated by U-HAT, which align with long-term documented observations, are not captured by these existing T_a datasets.

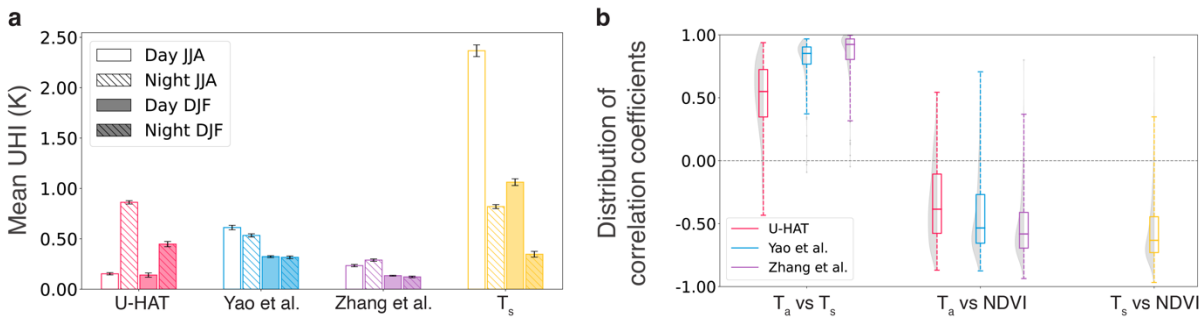

Fig. 2: U-HAT accurately reproduces observed urban climatology compared to existing T_a (near-surface air temperature) datasets. a, Mean UHI (urban heat island) intensities across cities in the CONUS (contiguous United States) during daytime and nighttime in JJA (June–August) and DJF (December–February) from 2013–2020 calculated using U-HAT, Yao et al.[48], Zhang et al.[49] and T_s. Error bars indicate the standard error of the mean. **b,** Distributions of correlation coefficients (r) from linear regressions between T_a, T_s (land surface temperature) and NDVI (Normalized Difference Vegetation Index) across CONUS cities. T_a estimates are from U-HAT, Yao et al.[48] and Zhang et al.[49]. Each data point is computed using JJA mean values of variables from 2013–2020, based on all available pixels within one cluster. Center bars represent the median, box edges the 25th and 75th percentiles, and error bars extend to 3×IQR (interquartile range) from the quartiles. Shaded violin plots indicate the underlying distribution.

Recent studies using CWSs also observed relatively weak and diverse T_s–T_a correlations across cities, cautioning the use of T_s for urban heat assessment and mitigation recommendations[27]. These studies have further indicated that the Normalized Difference Vegetation Index (NDVI) had significantly weaker associations with T_a than with T_s. U-HAT reproduces these relationships well (Fig. 2b). Across CONUS cities, U-HAT shows relatively small yet variable T_s–T_a correlation coefficients (r) of 0.51 ± 0.28 (mean ± standard deviation), in broad agreement

with recent studies over European cities using CWS observations[27]. In contrast, Yao et al.[48] and Zhang et al.[49] exhibit substantially high and more spatially consistent correlations (r = 0.81 ± 0.15 and 0.86 ± 0.16, respectively). Similarly, the T_a–NDVI correlation from U-HAT (r = -0.33 ± 0.31) is notably weaker than in the Yao et al.[48] and Zhang et al.[49] datasets (r = -0.43 ± 0.31 and -0.52 ± 0.24, respectively). Collectively, these results suggest that existing T_a datasets fail to capture true urban T_a signals, and instead largely reflect T_s-driven patterns. This shortcoming likely stems from the heavy reliance on urban T_s as a primary predictor in those T_a data products.

Large skin to air temperature discrepancies

U-HAT provides an unprecedented opportunity for pixel-to-pixel level comparisons with satellite-derived T_s, enabling characterization of urban human-perceived thermal variations both within and across cities. Previous T_s–T_a comparisons were primarily based on station observations on non-urban surfaces, revealing substantial differences between these two temperatures[30,67]. However, similar analyses in urban environments remain limited. Even among the few urban case studies, inconsistencies arise due to measurement biases and representativeness issues[28]. Here by combining U-HAT and T_s estimated from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor onboard NASA's Aqua satellite (MYD11A1), we examined the 9-year average urban vertical temperature gradient ($\Delta T = T_s - T_a$) at 1 km resolution, as well as the intra-city variability of urban T_a vs. T_s ($\Delta\sigma_T = \sigma_{T_s} - \sigma_{T_a}$) across CONUS cities. Our results reveal substantial spatiotemporal discrepancies between T_s and T_a in urban areas, reinforcing that T_s alone can be a highly biased proxy for T_a when assessing urban heat hazard.

We observe strong diurnal and climatic signals in the urban vertical temperature gradient (Fig. 3a,b). Across all available pixels in 384 CNOUS urban clusters, ΔT is, on average, positive during the day (5.01 $\pm$ 3.10 K) and negative at night (-1.58 $\pm$ 0.96 K). This means that using T_s in place of T_a can substantially overestimate the urban diurnal temperature range (DTR), an important metric of climate change. Hence, caution is warranted when drawing implications of urban effects on DTR using satellite T_s. These city-average ΔT signals exhibit a distinct climatic and geographic pattern: substantially larger (positive or negative) ΔT magnitudes occur in dry climates (desert and semiarid regions of CONUS) than in wetter temperate climate zones (e.g., eastern CONUS). This spatial pattern complies with previous observations over non-urban landscapes[30], with cities showing more intra-city variation in ΔT (Fig. 3a,b). For example, in Seattle (Phoenix, Dallas, New York), the range of daytime ΔT spans approximately 20 K (10 K, 13 K, 17 K respectively), and nighttime ΔT about 11 K (4 K, 4 K, 6 K).

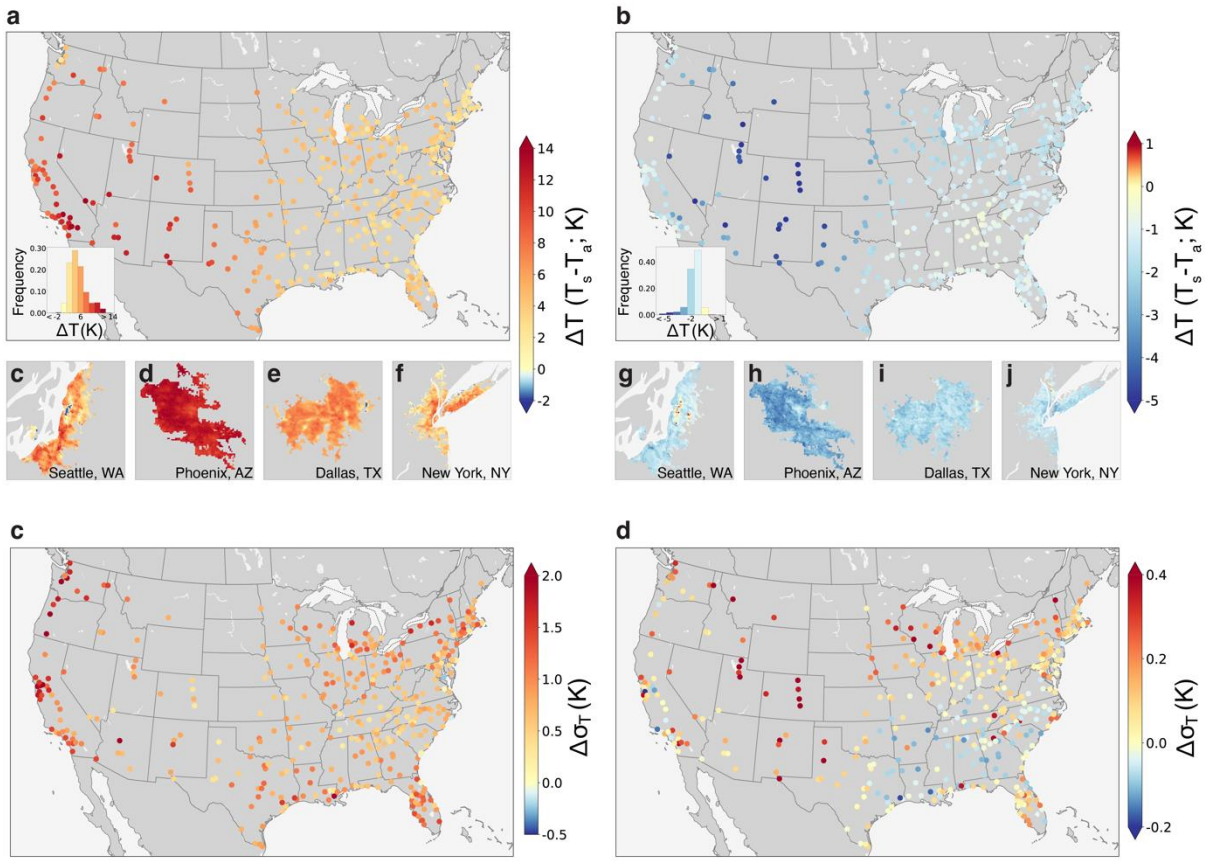

Fig. 3: U-HAT demonstrates substantial differences in magnitude and spatial variability between T_s (land surface temperature) and T_a (near-surface air temperature) over 384 cities in the CONUS (contiguous United States). a,b, Map of the city-wise mean vertical temperature gradient ($\Delta T = T_s - T_a$) averaged across 9 years at daytime (**a**) and nighttime (**b**). For each map, pixel-wise ΔT in Seattle, Phoenix, Dallas, and New York are shown below. The frequency distribution of pixel-wise ΔT is also shown. **c,d,** Map of the difference between the city-wise standard deviation of 9-year mean of T_s and T_a ($\Delta\sigma_T$) at daytime (**c**) and nighttime (**d**). Basemap from Natural Earth (https://www.naturalearthdata.com/).

These intracity and intercity ΔT variabilities can be primarily attributed to two factors: background climate and urbanization. In dry regions, with less water available for evaporation,

more net radiation is partitioned into sensible heat flux rather than latent heat, leading to reduced evaporative cooling and weaker land-atmosphere coupling, thereby increased ΔT. In addition, lower moisture availability of vegetated surfaces further reduces thermal inertia and heat capacity, causing the land surface to heat up more quickly during the day and cool more quickly at night compared to the air above[30]. However, these climatic signals are diminished at more "urbanized" landscapes (Supplementary Fig. S7a,b).

When considering only pixels with impervious surface fraction (ISF) greater than 0.5 instead of all available pixels, ΔT increases more significantly in wet climates but remains largely unchanged in dry climates, narrowing the difference between them. To further examine the effects of urbanization on ΔT, we categorized pixels into 10 equally spaced bins based on their ISF and compared the average ΔT within each bin. We find that T_s and T_a are strongly decoupled over more urbanized surfaces. Specifically, highly impervious surfaces (ISF>0.9) show 9.10 K higher daytime ΔT and 1.25 K lower nighttime ΔT than highly pervious surfaces (ISF<0.1) (Supplementary Fig. S7c,d). This is primarily due to the thermal properties and energy partitioning towards sensible heat flux of impervious surfaces, which resemble those of dry natural surfaces as described above. These similarities between dry and impervious surfaces also explain why ΔT changes less with increasing ISF in dry climates. In these regions, transitioning from dry to impervious surfaces has little impact, whereas in wet climates, the conversion from moist soils or vegetated surfaces to impervious materials results in a much greater reduction in evaporative cooling and surface heat capacity and thus lead to a large increase in ΔT. This

further cautions the use of T_s as a proxy of T_a over highly urbanized surfaces where T_s and T_a are largely decoupled regardless of the background climate zones, especially during daytime.

We also find a large discrepancy in intra-city variability between T_a and T_s. Across the CONUS, a majority of cities (99.2% at daytime and 69.8% at nighttime) show larger σ_{T_s} than σ_{T_a} (Fig. 3c,d). The smaller variability of T_a within cities reflects the role of air advection and horizontal mixing in redistributing heat citywide, an effect that minimally influences T_s. Such an effect is especially evident during the daytime. At night, σ_{T_s} and σ_{T_a} become closer, likely due to weaker air mixing. These differences suggest that using T_s might misrepresent human-perceived heat hotspots in urban areas.

Large skin to air temperature discrepancies in heat exposure

This large T_s–T_a discrepancy in intra-city variability is further amplified when examining population-scale exposure to urban heat extremes. To illustrate this, we compared daytime extreme heat exposure measured by T_a (HE_a) and by T_s (HE_s) across each city. We define extreme heat exposure as the total number of days exceeding the city's 99th percentile of temperature over the current entire U-HAT timespan (2013–2023), multiplied by the exposed population[68] at each 1 km pixel. We find that HE_s shows much wider intra-city distributions than HE_a across CONUS cities (Fig. 4, Supplementary Fig. S8a). Specifically, 96.9% of the cities examined show a larger coefficient of variation for HE_s than for HE_a. This discrepancy is often more pronounced in major, populated U.S. metropolitan areas, as illustrated in Fig. 4a,b.

Numerous recent studies have utilized high-resolution satellite data to examine urban environmental disparities or inequities[19,20,69,70]. However, our results indicate that these T_s-based estimates of intra-city heat spatial variability are substantially overestimated, consistent with patterns previously identified using process-based modeling[71]. Insights derived from these measures therefore may not translate directly to effective equitable policy. We also find that HE_s distributions are commonly more skewed toward higher ends than those of HE_a, as evidenced by a greater difference between their mean and median. 87.5% of the cities show higher skewness for HE_s than HE_a (Supplementary Fig. S8b), suggesting that not only intra-city spatial variability but also the overall magnitude of extreme heat exposure are likely overestimated when relying on satellite observed T_s.

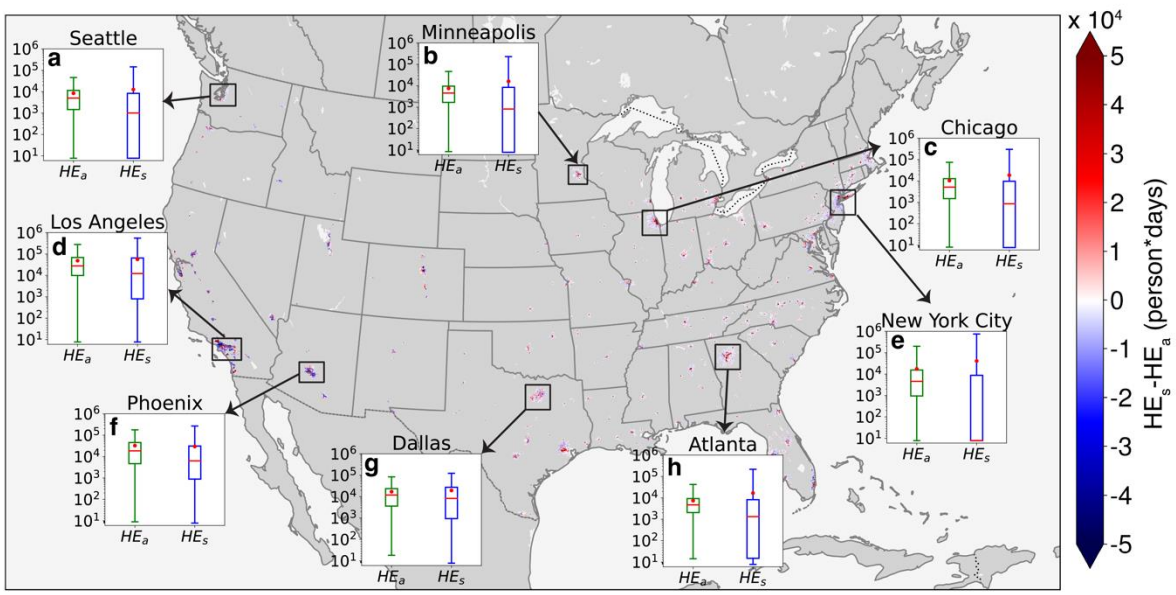

Fig. 4 U-HAT demonstrates substantial differences between daytime extreme heat exposure measured by T_s (HE_s) and T_a (HE_a) in their magnitude and spatial variability over 384 cities in the CONUS (contiguous United States). The map shows pixel-wise differences between daytime HE_s and HE_a. **a–h**, Distributions of HE_a and HE_s in selected cities on log scale. Center bars represent the median,

box edges the 25th and 75th percentiles, and error bars the 1st and 99th percentiles. Basemap from Natural Earth (https://www.naturalearthdata.com/).

This overestimation also displays an important intra-city pattern: T_s tends to overestimate heat exposure in urban cores while underestimating it in peripheral neighborhoods. Both HE_a and HE_s indicate that heat exposure is often higher in city centers (Supplementary Fig. S9), likely due to the combined effects of greater population density and elevated temperatures. However, their spatial distributions differ significantly. HE_a typically shows a less steep spatial gradient than HE_s within a city, encompassing both central and suburban areas – a pattern evident in Chicago, Atlanta, Dallas, and other major cities. The more uneven distribution of HE_s – from city center to its periphery – is shaped by a multiplicative effect of T_s and population. T_s has a more localized spatial pattern than T_a, as its redistribution is much less affected by air advection and horizontal mixing. This effect is further amplified by the uneven distribution of population, resulting in a much more intensified pattern of HE_s than HE_a in densely populated, high-temperature areas. These results suggest that relying on T_s alone risks biased policy interventions that focus disproportionately on city centers while overlooking broader, dispersed populations in outlying residential areas who are also dangerously exposed to extreme heat stress. In the U.S., these often include retired and elderly individuals who are highly vulnerable to heat[20]. T_a-based analyses are therefore essential in ensuring effective and equitable resource allocation, strategy development, and decision making for urban heat mitigation efforts and climate actions. Nevertheless, we note that U-HAT is not uncertainty-free. While more impact- and exposure-relevant than T_s, U-HAT remains subject to uncertainties tied to limited training domains and

spatial transferability. These uncertainties should be considered when interpreting equity and policy implications.

Discussion

In this work, we developed a transfer learning framework, U-TL, to reconstruct urban high-resolution air temperature. Accurate urban T_a estimation has long been challenging due to scarce true urban observations and strong urban heterogeneity. U-TL addresses these fundamental barriers and provides the urban T_a dataset capable of accurately reproducing observed urban climatology. Notably, U-TL achieves high predictive power as a "global" model, meaning it forgoes any local empirical tuning or spatially varying parameters.

Importantly, U-TL demonstrates the possibility of using data-rich proxies to support accurate prediction of data-poor variables in a highly heterogenous environment. This framework is not restricted to air temperature prediction or to urban systems. With appropriate pretraining targets and urban observations, the framework can be trained to predict other sparsely measured urban variables such as humidity, precipitation, and radiative or turbulent fluxes. Nor is data poverty an issue unique to urban landscapes; many other regions around the world, such as the Global South, also significantly lack quality observations. Similar transfer learning models can be adapted for those contexts or systems such as rivers and lakes, wildlands, croplands, or other vegetated ecosystems, where in situ data are scarce.

The 1-km U-HAT dataset produced by U-TL fills a critical urban data gap that has hindered urban climate and weather research for decades, and provides an impact-relevant thermal characterization both within and across cities at a large scale, thus advancing our understanding of the urban thermal environment. Due to the absence of high-fidelity, high-resolution, and spatial continuous urban T_a data, satellite-derived T_s has been widely used, often informing impact-, decision-, or policy-oriented implications under the assumption that T_s (or UHI_s) strongly correlates with T_a (UHI_a). U-HAT enables direct evaluations of these assumptions at the pixel-to-pixel level. Our findings reveal substantial discrepancies between T_a and T_s both within and across cities. These discrepancies are further exacerbated in urban heat exposure. These results underscore the criticality of T_a-based measures in urban studies, and have consequential implications for urban heat mitigation. For instance, although numerous remote sensing studies confirm the effectiveness of vegetation-based strategies in reducing urban T_s, such cooling efficacy does not necessarily translate to air temperature. Compared with T_s, the much weaker correlation (r =-0.33 vs. -0.56) (Fig. 2b) and lower sensitivity (-1.28 K vs. -13.26 K) of urban T_a to NDVI in U-HAT (Supplementary Fig. S10) suggest substantially lower cooling potential of urban greenery for air temperature. This cooling benefit for human-perceived heat stress is likely further dampened when air humidity is factored in[27,72–74]. That said, T_s estimates in urban climate studies are still important as they provide globally consistent metrics for mechanistic understanding of urban thermal environments and how it relates to the surface energy balance[26,75,76]. Moreover, as also seen from our U-TL workflow, it remains a critical input for state-of-the-art algorithms that can more accurately capture physiologically relevant urban climate signals.

U-HAT opens promising new opportunities for long-standing challenges stemming from the shortage of spatiotemporal urban-resolving T_a data. For example, weather forecasting models and data assimilation systems are known to integrate little-to-no urban ground observations, and this dataset could potentially improve weather prediction and reanalysis products over urban areas significantly. In large-scale climate and Earth system modeling, U-HAT can provide spatiotemporally consistent urban T_a data to evaluate, benchmark, and constrain urban climate simulations. Furthermore, many urban epidemiological and socioeconomic studies have long been constrained by the lack of high-quality, high-resolution urban air temperature data[19,77]. U-HAT can provide a crucial input feature for applications in these fields.

We note three main limitations of this study. First, the resolution and accuracy of our final data are limited by the uncertainty and resolution of satellite data. The spatial resolution of current U-HAT is 1 km as it is constrained by the resolution of MODIS T_s. However, one could easily train U-TL to produce higher resolution T_a when higher-resolution satellite T_s is used. Part of the U-HAT uncertainty is inherited from the uncertainty of T_s, which is less than 2 K (Methods). However, because the transfer learning approach extracts generalizable features from the pretrained task rather than relying on exact pixel values, the impact of the T_s uncertainty on U-TL's performance is likely mitigated, especially when our pretrained task is sufficiently general, i.e., trained on monthly mean T_s instead of daily values (Methods). Second, the model currently provides estimations only for non-cloudy days, as the availability and accuracy of satellite-derived data under cloudy conditions remain limited. However, the large variability of the training data encompasses a wide range of atmospheric and surface conditions, presumably enabling the model to generalize beyond clear-sky scenarios. This suggests that the framework

has the potential to predict T_a under cloudy conditions, although larger biases may be introduced. As more advanced methods are being developed to enhance spatial and temporal resolutions, improve retrieval accuracy and reconstruct missing observations of satellite products, these challenges can likely be addressed by leveraging higher-resolution, higher-accuracy, and gap-filled datasets as they become available. Third, the physical relationship between T_s and T_a is variable and context-dependent across diverse urban morphologies and climates. Our validation results demonstrate strong generalizability and spatial transferability of U-TL within CONUS. U-TL nevertheless provides an approximation of the underlying complex, variable T_s–T_a mapping, and thus its accuracy may be sensitive to input uncertainties.

Methods

The U-TL framework

The U-TL framework (Supplementary Fig. S1a) leverages a transfer learning approach that adapts knowledge learned from data-rich tasks to improve performance on data-limited tasks[56,78,79]. We achieve this by initially training a deep neural network (DNN) to predict urban land surface temperature (T_s), and then fine-tuning it to estimate urban near-surface air temperature at the screen height (T_a) in the second step.

Step 1: Pretraining to estimate urban T_s

In the first step, we trained the model to estimate T_s at each urban pixel using its corresponding surface properties (i.e., satellite imagery, building height and elevation), atmospheric forcing fields and month indicators as predictors. The backbone of the model architecture is a DNN modified from a 2-layer ResNet[80] concatenated with a separate set of linear layers. The DNN ingests satellite imagery and elevation as inputs, and its output is average-pooled to generate a feature vector. The input satellite imagery includes blue (SR_B2), green (SR_B3), red (SR_B4), near infrared (SR_B5), shortwave infrared 1 (SR_B6), normalized difference vegetation index (NDVI) and normalized difference built-up index (NDBI) bands. In addition, we applied three linear layers to extract another feature vector from forcing fields, building height and month indicators. The atmospheric forcing fields include surface solar radiation downwards (ssrd), surface thermal radiation downwards (strd), surface pressure (sp), total precipitation (tp), u-component of wind (u), v-component of wind (v), temperature (t) and relative humidity (r) at a

blending height. The two feature vectors are concatenated and passed through three linear layers to mimic the land-atmosphere interactions for the final T_s prediction.

The T_s data is obtained from the Aqua MODIS daily land surface temperature product (MYD11A1) at 1 km resolution, retrieved twice a day at approximately 1:30 PM and 1:30 AM local time for daytime and nighttime, respectively[81]. Note that the U-TL framework itself is not limited to Aqua samples and can be applied to Terra data seamlessly. Here Aqua is chosen over Terra to generate urban T_a data because its overpass times better correspond to daily air temperature extremes.

We only retained pixels that are labeled with an estimated T_s error of less than 2 K according to the quality control bands and included only days with at least 90% available pixels to minimize cloud contamination effects. For each sample, satellite imagery and elevation at 30 m resolution are cropped into 33×33-pixel patches, centered on the corresponding T_s grid cell. All other input variables are represented as scalars. In transfer learning, coarse graining is often applied to the pretraining task to improve its generalizability for the subsequent target task[78]. Therefore, we used monthly averaged T_s and forcing fields instead of daily values for model training. Two separate models were trained to predict daytime and nighttime T_s. The nighttime model takes in one fewer input from the forcing fields than the daytime model, as downward surface solar radiation is absent at night.

Step 2: Transfer learning to estimate urban T_a

In the second step of U-TL, we fine-tune the pretrained T_s model using urban ground T_a observations. To optimize the knowledge transfer, we conduct several experiments to determine the best configuration. While model performance remains stable across structural adjustments, we selected the overall best-performing model.

First, while the initial layers of the T_s model can resolve generalized representation of the land-atmosphere interactions, the final layers are more task-specific and may be less useful for T_a prediction due to T_s–T_a differences. We therefore experimented with progressively unfreezing or discarding layers of the pretrained T_s model from the final layer upward. Here “Unfreezing” refers to allowing the weights of pretrained layers to be updated during fine-tuning, while "discarding" removes them entirely. Optimal configurations were determined separately for daytime and nighttime models. For both, unfreezing all layers proved most effective, likely due to the need to adjust for temperature distribution differences between T_s and T_a. In the daytime model, the last three linear layers of the T_s model were discarded and replaced with three new layers containing fewer parameters, initialized with random weights. For the nighttime model, only the last layer was discarded and replaced with a randomly initialized one.

Second, we found that larger T_s–T_a differences were associated with higher residual errors during fine-tuning, particularly in the nighttime model, suggesting that greater temperature differences may hinder effective knowledge transfer. To mitigate this issue, we introduced a set of linear

layers as skip connections[82] using T_s as input to learn the residuals from the transfer learning step (Supplementary Fig. S1a). We found that this implementation improved performance, particularly for the nighttime model.

Third, we incorporated a curriculum learning-inspired loss function to improve training stability and the learning of hard samples for both the pretraining and fine-tuning steps. Curriculum learning prioritizes easier samples at the beginning of training and gradually increases the influence of harder ones as the model improves[83]. We implemented this by dynamically adjusting sample importance based on their loss[84]. By initially downweighing high-loss samples, the model first captures fundamental patterns, leading to better feature extraction for complex cases later. This prevents the model from being overwhelmed early and ensures that harder samples are learned more effectively over time.

Data preprocessing

Urban extent

Urban boundaries are defined using the TIGER/Line Files and Shapefiles provided by the United States Census Bureau. Cities are ranked by their sizes, and the largest 400 are initially selected. After excluding 16 cities located outside the Contiguous United States (CONUS), the final sample includes 384 cities.

Station data preparation

For the fine-tuning step, we carefully screened 52 stations from 18 cities with hourly T_a records (Supplementary Fig. S2). We applied a set of selection criteria to ensure their representativeness of true urban environments, though some subjectivity remains. Each meteorological network was scrutinized to verify that the screen height is at 1.3–3 m above the surface. Generally, stations were selected if they were located on surfaces with an impervious surface fraction (ISF) greater than 0.5 within a 200 m buffer, and we also visually inspected station locations and surrounding environments using Google Earth (http://earth.google.com). Airport urban stations were only included when they were within or close enough to cities. The selected stations cover diverse landscapes and climate zones. To match the sampling time of T_s, we calculated mean T_a values between 1 PM and 2 PM local time as daytime T_a, and mean T_a values between 1 AM and 2 AM as nighttime T_a. Stations came from Synoptic data, Integrated Surface Database (ISD) and local sources.

Auxiliary datasets

All auxiliary datasets are summarized in Supplementary Table S1. Satellite imagery is from the Landsat 8 surface reflectance product[85]. Although the model can infer vegetation and built-up features from the available bands, we explicitly calculated and included NDVI and NDBI (**Eq. 1** and **2**) as inputs, as prior testing showed that their inclusion enhances model performance.

$$NDVI = \frac{SR\_B5 - SR\_B4}{SR\_B5 + SR\_B4} \quad (1)$$

$$NDBI = \frac{SR\_B6 - SR\_B5}{SR\_B6 + SR\_B5} \quad (2)$$

Where SR_B4, SR_B5 and SR_B6 are red, near infrared and shortwave infrared 1 bands, respectively.

Clouds and cloud shadows are masked using the quality control bands. We converted raw values into surface reflectance, and masked pixels exceeding the radiometric saturation range (0–1) to remove physically unrealistic values, which can result from Landsat's retrieval issues over water bodies and snow surfaces. As urban surface characteristics (except vegetation phenology) are generally static over a short period, for input satellite imagery we use the monthly median composites over the corresponding five-year period (2013–2018 or 2019–2023) for each training sample. This choice is considered reasonable because urban expansion, particularly in U.S. cities, typically occurs on decadal timescales[86]. However, this assumption may not hold universally. In regions experiencing rapid urban development over short timescales, the use of multi-year composites could obscure temporal evolution in urban characteristics and potentially introduce bias into T_a predictions. In such cases, yearly composites may be more appropriate, albeit at the cost of insufficient cloud-free surface imagery availability.

To test the effect of the composite window length on model performance, we conducted a sensitivity analysis comparing model performance when using 3-year versus 5-year monthly median composite surface imagery in both the T_s pretraining and T_a finetuning steps. For this analysis, we trained both models using data from the 18 cities with available station observations to ensure a fair comparison. We find negligible differences in model performance: the daytime MAE is 0.93 K for the 3-year composite compared to 0.90 K for the 5-year composite

experiment, while the nighttime MAE is 1.22 K for both cases (Supplementary Fig. S11). To further assess the sensitivity of the resulting T_a estimates to shorter composite windows, we examined the temporal characteristics of estimates derived from models using 1-year, 3-year, and 5-year monthly median composite surface imagery as inputs. The results show highly consistent temporal patterns across all three composite windows (Supplementary Fig. S12). The mean bias differences between the 1-year and 3-year composites and between the 3-year and 5-year composites are -0.3 K and -0.5 K, respectively. Together, these results indicate that the choice of composite window length (up to 5 years) has minimal influence on the resulting T_a estimates, demonstrating the robustness of the model to this design choice over the CONUS region.

We used only images without masked pixels for model training and data production, which results in an average of 84.4% available samples across the entire dataset spanning 384 cities. For building height[87], we applied gap-filling with the mean height of neighboring 1.5 pixels. We reprojected Landsat imagery, elevation[88] and building height data to the T_s projection, and then preprocessed them using the Google Earth Engine platform[89].

The atmospheric forcing fields are from the ERA5 hourly reanalysis dataset. Specifically, the variables u, v, t and r are at the 1000 hPa level[90], and ssrd, strd, sp and tp are at the surface level[91]. Precipitation values were aggregated daily, whereas all other variables were extracted at Aqua's overpass times in the local time zone. At the forcing height (i.e., blending height in the atmospheric boundary layer), the air is well blended, and thus the influence of fine-scale urban structure on the forcing meteorology at this height is largely diminished. The rationale of using

coarse large-scale atmospheric forcings as inputs in U-TL aligns with standard methodological structure in climate modeling, where coarse-resolution atmospheric forcings (such as ERA5 reanalysis or GCM outputs) are routinely used to drive high-resolution land surface models or WRF[57,58]. Therefore, the fine-scale variability captured in U-HAT arises from the interaction between large-scale atmospheric forcing and fine-scale surface characteristics. To ensure that these forcing variables are compatible with the 1-km output temperatures and can be consistently applied into the model, we re-gridded all forcing fields from 0.25° to 1 km using a "patch" method provided by xESMF[92].

Validation of U-TL

Validation of the T_s model

We randomly selected approximately 35 million samples from 70% of each city's data for training the T_s model. We drew an equal number of samples from each city to ensure sample balance. To compensate for scarcer data in smaller cities, we applied a bootstrap stratified sampling method in which existing samples from data-scarce cities were stochastically resampled during training as needed. An additional 3.5 million samples were randomly selected from the remaining 30% for testing. Model validation on the test data demonstrates high accuracy in predicting urban T_s (Supplementary Fig. S13), with mean absolute errors (MAEs) of 0.61 K and 0.37 K, and R^2 scores of 0.99 and 1 for daytime and nighttime, respectively. The model also shows very low spatial biases: 99.5% (day) and 99.8% (night) of pixels have a mean error between -2 to 2 K (i.e., within the quality control threshold of the T_s data), with errors randomly distributed within cities (Supplementary Fig. S14). Our result underscores the

pretrained model's ability to capture the mapping from land-atmosphere interactions to urban surface temperatures, laying a strong foundation for the subsequent transfer learning step.

Validation of the T_a model

To validate the generalizability of U-TL to unseen locations, we randomly divided stations into 10 folds, each containing 5–6 ones. In each iteration, 1 fold was held out for testing, while a subset of 3, 6, or 9 folds from the remaining ones was used for training. This process was repeated 10 times, ensuring that each fold was used for testing once. Note that entire data from held-out stations, rather than individual samples, were excluded in the training. This design ensures that the testing results better reflect the model's performance and generalizability to unseen locations. We repeated the whole process 5 times to ensure the robustness of the results.

The U-TL framework exhibits high accuracy and strong robustness in predicting urban T_a across scales (Supplementary Figs. S3–5). We validate U-TL's predicted urban T_a against unseen observations via the leave-station-out 10-fold cross-validation across 18 cities in the CONUS (Supplementary Fig. S3). The results indicate excellent accuracy of our T_a predictions, with out-of-sample mean absolute errors (MAEs) of 0.91 K during daytime and 1.24 K at night (Supplementary Fig. S3). The model accurately captures daily T_a variabilities across space and time, with R^2 of 0.98 and 0.97 for daytime and nighttime, respectively. This high predictive accuracy is observed both across and within cities. When evaluated at each individual city, R^2 remains high with values ranging from 0.88 to 0.99 at daytime and from 0.86 to 0.98 at nighttime (Supplementary Figs. S15–S16). Notably, this high overall predictive performance is achieved

with very limited urban T_a labels, illustrating the power of our transfer learning approach. Daytime results slightly outperform nighttime ones, likely because the model does not explicitly account for the heat storage term, which plays a more significant role at night.

We further demonstrate U-HAT's high accuracy by comparison against the observationally based PRISM dataset[64]. Ten metropolitan areas across a range of background climate zones are selected for assessment. We acknowledge that PRISM data is spatially interpolated from standard weather stations and thus in general subject to rural biases. However, previous studies have showed that it could partially capture urban signals in certain large metropolitan areas due to a relatively higher density of urban stations within these areas[57,93]. Our results show that U-HAT accurately reproduces the observed daily distributions of diurnal average, daytime, and nighttime urban T_a within each city and the spatial variability across the ten cities for the whole time span as well as summer and winter months from 2013–2023 (Supplementary Fig. S4). The higher median of U-HAT's diurnal average T_a compared to PRISM is likely due to PRISM's limited incorporation of true urban stations, which leads to its underestimation of urban temperature. Another caveat is the difference in sampling times between these two datasets (PRISM only reports daily maximum and minimum T_a), which might cause differences in their median values across times of day. Nevertheless, these results show strong evidence for U-TL's ability to reproduce the observed diurnal, daily, and spatial variabilities of urban T_a.

To evaluate whether our approach improves upon direct prediction of urban T_a from scratch, we compared U-TL with a model that directly predicts urban T_a using the same predictors without

skip connections – hereinafter referred to as the "scratch model". We divided the selected 52 stations into 10 folds and tested both models' performance under fewer training samples by progressively reducing the number of training folds from 9 to 6 to 3.

The results reveal that U-TL significantly enhances the accuracy, generalizability, and robustness in urban T_a prediction compared to the scratch model, particularly under data-scarce conditions (Supplementary Fig. S5). When relatively abundant training samples are available (i.e., training with 9 folds), U-TL reduces mean absolute error (MAE) by 13.1% and 26.2% for daytime and nighttime, respectively. U-TL also shows narrower error distributions than the scratch model, with standard deviation reductions of 12.6% during the day and 26.0% at night, demonstrating higher robustness of our approach.

This advantage becomes more pronounced when data are even scarcer. The performance of U-TL remains remarkably stable as the number of training samples is largely reduced especially at nighttime, while the scratch model's errors worsen substantially (Supplementary Fig. S5). From 9 to 3 training folds, U-TL's MAE increases by only 8.5% during daytime and 7.1% at nighttime, compared to 13.4% and 18.0% for the scratch model. Similarly, the error standard deviation increases by only 8.2% during daytime and 6.9% at nighttime for U-TL, compared to 13.7% and 15.6% for the scratch model.

Generalizability of U-TL across climates, surface types and urban forms. To validate U-TL's generalizability to unseen climates, we conducted a leave-climate-zone-out cross-validation experiment. In addition, we assessed U-TL's generalizability to unseen climates and urban forms using leave-climate-zone-out and leave-city-out cross-validation, respectively. First, for the leave-climate-zone-out experiment, we grouped all stations by their Köppen-Geiger climate zone and, in each training round, withheld one entire climate-zone group for testing while training on stations from all other zones. Overall, U-TL demonstrates strong spatial generalization ability: across all held-out climate zones, the mean absolute error (MAE) is 1.04 K and 1.41 K, with corresponding R^2 values of 0.97 and 0.96 for daytime and nighttime, respectively (Supplementary Fig. S17). Similarly, the leave-city-out cross-validation experiment withholds all stations from each city for testing while training on stations from all other cities in each training round. U-TL remains highly robust under this setting, with MAE of 1.07 K during daytime and 1.32 K during nighttime, and R^2 values of 0.97 for both times (Supplementary Figs. S18–19). Performance is generally consistent across climate zones and cities, although we do observe slightly reduced skill at coastal sites (e.g., Csa: hot-summer Mediterranean climate; Csb: warm-summer Mediterranean climate; and Am: tropical monsoon climate). This reduction is likely due to local processes, such as sea-breeze circulations, which might only be partially represented in the atmospheric forcing inputs but not fully resolved by our current model. However, when stations with similar coastal characteristics are included in the training set, as in our standard 9-fold cross-validation, the performance improves substantially (Supplementary Figs. S15–16). For example, MAE and R^2 in Csa (hot-summer Mediterranean climate) zones improve to 1.2 K and 0.92 at daytime and 1.22 K and 0.89 at nighttime, respectively. This suggests that U-TL is able to transfer knowledge from stations with similar local dynamics when such information is

represented in the training data, rather than being fundamentally limited in representing these processes.

We further assess whether specific surface types or urban geometries challenge the framework by examining the relationship between MAE and indicators of local surface characteristics, for which we use both National Land Cover Dataset (NLCD)[94] and Local Climate Zone (LCZ)[95] classes to ensure the robustness of the results. For each station, we identified the dominant NLCD or LCZ class within the 1-km buffer and then computed the average MAE from the 9-fold cross-validation experiment across all stations in that class (Supplementary Tables S2–S3). Results show that MAE vary modestly across classes. Among NLCD classes (Supplementary Table S2), values span 0.84 K (developed medium intensity) to 1.04 K (developed low intensity) at daytime and 1.15 K (developed medium intensity) to 1.56 K (shrub/scrub) at nighttime. Among LCZ classes (Supplementary Table S3), errors range from 0.77 K (sparsely built) to 1.23 K (bare rock or paved) at daytime and 1.05 K (sparsely built) to 1.78 K (bare rock or paved) at nighttime. These results suggest that model performance shows little bias with respect to surface type or urban form.

Comparison with existing T_a models and datasets. To evaluate U-TL against previous models, we compiled a summary of major gridded air temperature datasets covering the contiguous United States and their reported error metrics (Supplementary Table S4). Note that because most products report overall accuracy based on predominantly rural stations, those statistics are not urban-specific MAE and therefore likely provide a lower bound on their potential errors in urban

areas. In addition, the magnitude of MAE for different products can be influenced by their temporal scales; for example, monthly T_a is likely to show lower errors than daily T_a, as evidenced by the comparison between the MAE of daily values and monthly means from U-HAT. Nevertheless, among datasets at similar spatial and temporal scales (i.e., ~1 km and daily scale), U-TL generally achieves lower errors, particularly at daytime. We further computed the urban-specific MAE for each product against the 52 urban stations used in this study over a common period (04-01-2013 to 12-31-2016). Unlike our leave-station-out cross-validation, these evaluations are not necessarily out-of-sample, since some of these datasets may have trained on these stations. Even so, U-HAT shows lower MAE than most previous datasets, particularly those derived from statistical or traditional machine learning techniques (i.e., Zhang et al.). Some interpolation-based products (e.g., PRISM) exhibit lower errors against selected stations because of a relatively higher density of urban stations incorporated in these areas. However, their accuracy is likely to degrade where stations are sparse, which is the very problem U-TL is intended to address. To provide a more independent validation of model performance, we conducted an additional evaluation using only stations from local observational networks that are not used in the training of other T_a products. When evaluated against this independent subset of local stations, we find an even larger performance advantage of U-HAT relative to existing datasets (Supplementary Table S4). Most models exhibit degraded performance, particularly those relying on spatial interpolation methods (e.g., PRISM, TopoWx, MacDonald et al., and Daymet). In contrast, U-HAT shows better performance, with lower MAE by 25.3% during daytime and 8.9% during nighttime. These comparisons further confirm the high accuracy of U-HAT and its value for urban areas with limited in-situ observations.

To provide a more robust evaluation, we further added a regression-based baseline model – specifically multiple linear regression model using the same predictors as U-TL (hereafter referred to as Baseline-MLR) – and conducted additional analyses, as the majority of the existing traditional methods reported in the literature is regression-based approaches. We then repeated the same 10-fold cross-validation applied to the main model using 3 folds of training data. Our results show that U-TL substantially outperforms this linear regression baseline, where mean absolute error (MAE) is reduced by 11.4 % during daytime and 14.6 % during nighttime (Supplementary Fig. S20). In addition, U-TL yields narrower error distributions, with standard deviation reductions of 8.1 % (day) and 11.6 % (night), indicating improved robustness relative to the linear model.

More importantly, we evaluated whether these datasets reproduce observed urban climatology by conducting a same urban heat island (UHI) analysis using the same Simplified Urban Extent (SUE) method as Fig. 2a for consistency. As shown in Supplementary Fig. S21, none of the alternative datasets captures the observed UHI patterns due to inherent limitations in their approaches. For example, the Baseline-MLR model shows cooler urban temperatures than rural at daytime, contradicting widely observed daytime UHI phenomenon. Daymet and PRISM show near-zero UHI across seasons and times of day because their estimates are produced through spatial interpolation from stations that are mostly located in rural or open areas, essentially missing the urban signal. HUMID, while generated with physics-based model and bias-corrected approach, produces much stronger daytime than nighttime UHI, which is again inconsistent with widely observed and well established UHI climatology. This is possibly because HUMID is produced from land-only simulations, therefore missing the two-way land-atmosphere

interactions and lateral advection[71]. These results further suggest that lower errors (e.g., for PRISM) do not equal faithful urban signal representation, particularly where station coverage is sparse. Collectively, these findings underscore the criticality of TL principle and U-TL's advantage over traditional methods.

These findings highlight U-TL's strong performance in estimating high-resolution urban T_a with high generalizability to unseen locations and robustness under limited training data – a challenge mirroring the real-world extreme scarcity of true urban observations. This advance largely stems from the transfer learning approach, in which abundant T_s data from the pretraining step helps the model learn a generalized representation of land-atmosphere interactions, thereby enabling its adaptability across diverse urban environments and under data-scarce conditions. Notably, U-TL achieves this high predictive performance through a unified global model and not a locally refined one. Unlike previous statistical or machine learning methods that are often only locally applicable and may fail in regions with limited training data, U-TL's reliance on underlying physical principles makes it universally applicable. Therefore, although the number of selected stations is limited because of the extreme scarcity of true urban weather stations, such data limitation is overcome by the transfer learning approach. That said, U-TL will benefit when more true and representative urban weather stations become available in the future.

While T_a variability is primarily influenced by atmospheric forcing fields, surface properties also play a role through land-atmosphere interactions. To assess the significance of surface properties in T_a prediction, we conducted an experiment where training samples were assigned randomly

shuffled images and the correct forcings. We compared four scenarios: (1) correct images in both pretraining and fine-tuning steps, (2) correct images in the pretraining step but mismatched images in the fine-tuning step, (3) mismatched images in the pretraining step but correct images in the fine-tuning step, and (4) mismatched images in both pretraining and fine-tuning steps. We followed the same 10-fold cross-validation procedure with 5 repetitions. We found that scenario (1) outperformed all other scenarios in both accuracy and stability, with a reduction of 7.1%–14.5% in MAE and 8.4–30.0% in error standard deviation for daytime and a reduction of 6.3–8.7% in MAE and 5.9–8.3% in error standard deviation for nighttime predictions (Supplementary Fig. S22). The result highlights the critical role of surface properties in predicting urban T_a.

Urban heat island calculations

UHI is defined as the temperature differences between an urban area and its surrounding rural area. Here we compare UHI intensities derived from U-HAT, Yao et al.[48], Zhang et al.[49] and T_s at daytime and nighttime in summer and winter months of 2013–2020. Both Yao et al.[48] and Zhang et al.[49] provide seamless 1 km air temperature datasets, with the former offering monthly-averaged values of daily maximum and minimum temperatures, and the latter providing daily values. Since U-HAT provides urban temperature only, we first adopted the simplified urban extent (SUE) method[63] to ensure fair comparisons. SUE defines UHI as the temperature difference between urban and non-urban pixels within each urban cluster, and shows UHI characteristics in line with results from other algorithms[63]. To classify pixels into “urban” and “non-urban”, we used MODIS Land Cover Type dataset (MCD12Q1), where urban pixels

correspond to urban land use, and non-urban pixels include all other non-urban and non-water land uses. To minimize elevation-related effects, we further filtered non-urban pixels to retain only those with an elevation difference of less than 50 m from the median urban elevation of each cluster. Elevation data were obtained from the Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010)[96], and all datasets were reprojected to match its projection. We then computed the mean temperatures over 2013–2020 for the urban and non-urban subsets within each cluster, and their difference yields the UHI intensity. Urban areas lacking available non-urban pixels were excluded, resulting in a final dataset of 335 urban clusters.

To ensure the robustness of our findings, we repeated these analyses using an equal-area rural buffer method following a previous study[97]. We use both PRISM[64] and ERA5[65] as the rural reference temperatures for the canopy UHI (UHI_a) calculations. For surface UHI (UHI_s) we use the MODIS land surface temperature (LST) data for both urban and rural T_s. The urban extent is defined using the Global Human Settlement Index for 2015[98]. For the rural reference of each urban cluster, we generated an equal-area buffer around the urban region using an iterative 300 m step size. To prevent overlap with nearby urban clusters, pixels classified as urban land use within the rural buffer were excluded. This results in a final dataset of 183 urban clusters.

Given the differences between 1:30 AM/PM overpass times and timing of daily extremes, to test the robustness of the UHI results by U-HAT and ensure comparability with other datasets, we trained U-TL using T_{min} and T_{max} observations and recomputed the UHI results using the same SUE method. We found close agreement between these two results (Supplementary Fig. S23).

This is consistent with previous studies showing similar UHI_a intensity between 1:30 AM/PM and daily T_a extremes[6] further supporting the validity of our results. We conducted all UHI calculations using the Google Earth Engine platform[89].

Definitions of all the abbreviations used in this study are listed in Supplementary Table S5.

Data availability

The U-HAT data are publicly available at https://doi.org/10.5281/zenodo.20057651 (ref[99]). Source data underlying the main figures are provided as Source Data files. Source data underlying the Supplementary Figures are publicly available at https://doi.org/10.5281/zenodo.20082808. The Global Historical Climatology Network Hourly data are available at https://www.ncei.noaa.gov/products/global-historical-climatology-network-hourly. The ESA WorldCover data are available at https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v200. The Aqua MODIS LST data are available at https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MYD11A1. The Landsat 8 Surface Reflectance data are available at https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2. The ERA5 hourly data on single levels are available at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview. The ERA5 hourly data on pressure levels are available at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview. The building height data are available at https://doi.org/10.34894/4QAGYL. The NASA STRM elevation data are available at https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003. The ISD data are available at https://www.ncei.noaa.gov/access/search/data-search/global-hourly. The Synoptic data are available at https://download.synopticdata.com/. The station observation data from Arizona Mesonet are available at https://cals.arizona.edu/AZMET/az-data.htm. The station observation data for Madison are available at https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-ntl.324.24. The station observation data from Oklahoma Mesonet and Harrisburg are available upon request. The ERA5-Land daily data are available at https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR. The PRISM data are available at https://prism.oregonstate.edu/. The LandScan population data is available at https://landscan.ornl.gov/. The TIGER/Line Files and Shapefiles are available at https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. The MODIS Land Cover Type dataset is available at https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MCD12Q1. The Global Multi-resolution Terrain Elevation Data 2010 are available at https://developers.google.com/earth-engine/datasets/catalog/USGS_GMTED2010_FULL. The local climate zone map is available at https://developers.google.com/earth-engine/datasets/catalog/RUB_RUBCLIM_LCZ_global_lcz_map_latest.

Code availability

The python code necessary to train and apply the U-TL model to reproduce the main results are publicly available from Github at https://github.com/Yiwen-Zhang97/U-TL.

References

1. Grimm, N. B. *et al.* Global Change and the Ecology of Cities. *Science* **319**, 756–760 (2008).
2. Bulkeley, H. *Cities and Climate Change*. (Routledge, London, 2013). doi:10.4324/9780203077207.
3. Mora, C. *et al.* Global risk of deadly heat. *Nat. Clim. Chang.* **7**, 501–506 (2017).
4. Muller, C. L., Chapman, L., Grimmond, C. S. B., Young, D. T. & Cai, X. Sensors and the city: a review of urban meteorological networks. *Int. J. Climatol.* **33**, 1585–1600 (2013).
5. Li, J. *et al.* Satellite-Based Ranking of the World's Hottest and Coldest Cities Reveals Inequitable Distribution of Temperature Extremes. *Bull. Am. Meteorol. Soc.* **104**, E1268–E1281 (2023).
6. 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).
7. Gough, W. A. Thermal Signatures of Peri-Urban Landscapes. *J. Appl. Meteorol. Climatol.* **59**, 1443–1452 (2020).
8. Oke, T. R. Siting and Exposure of Meteorological Instruments at Urban Sites. in *Air Pollution Modeling and Its Application XVII* (eds Borrego, C. & Norman, A.-L.) 615–631 (Springer US, Boston, MA, 2007). doi:10.1007/978-0-387-68854-1_66.
9. World Meteorological Organization. *Guide to Instruments and Methods of Observation, Volume V – Quality Assurance and Management of Observing Systems*. (WMO, Geneva, 2023).
10. Chakraborty, T. C. & Qian, Y. Urbanization exacerbates continental- to regional-scale warming. *One Earth* **7**, 1387–1401 (2024).

11. Muller, C. L., Chapman, L., Grimmond, C. S. B., Young, D. T. & Cai, X.-M. Toward a Standardized Metadata Protocol for Urban Meteorological Networks. *Bull. Am. Meteorol. Soc.* **94**, 1161–1185 (2013).
12. Kalnay, E. & Cai, M. Impact of urbanization and land-use change on climate. *Nature* **423**, 528–531 (2003).
13. Dunn, R. J. H., Willett, K. M., Parker, D. E. & Mitchell, L. Expanding HadISD: quality-controlled, sub-daily station data from 1931. *Geosci. Instrum. Method. Data Syst.* **5**, 473–491 (2016).
14. Li, X. *et al.* Mapping global urban boundaries from the global artificial impervious area (GAIA) data. *Environ. Res. Lett.* **15**, 094044 (2020).
15. Zanaga, D. *et al.* ESA WorldCover 10 m 2021 v200. Zenodo https://doi.org/10.5281/zenodo.7254221 (2022).
16. Manoli, G. *et al.* Magnitude of urban heat islands largely explained by climate and population. *Nature* **573**, 55–60 (2019).
17. Schwaab, J. *et al.* The role of urban trees in reducing land surface temperatures in European cities. *Nat. Commun.* **12**, 6763 (2021).
18. Massaro, E. *et al.* Spatially-optimized urban greening for reduction of population exposure to land surface temperature extremes. *Nat. Commun.* **14**, 2903 (2023).
19. Chakraborty, T., Hsu, A., Manya, D. & Sheriff, G. Disproportionately higher exposure to urban heat in lower-income neighborhoods: a multi-city perspective. *Environ. Res. Lett.* **14**, 105003 (2019).
20. Hsu, A., Sheriff, G., Chakraborty, T. & Manya, D. Disproportionate exposure to urban heat island intensity across major US cities. *Nat. Commun.* **12**, 2721 (2021).

21. Du, M. *et al.* Daytime cooling efficiencies of urban trees derived from land surface temperature are much higher than those for air temperature. *Environ. Res. Lett.* **19**, 044037 (2024).
22. Weng, Q. Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends. *ISPRS J. Photogramm. Remote Sens.* **64**, 335–344 (2009).
23. Deilami, K., Kamruzzaman, Md. & Liu, Y. Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures. *Int. J. Appl. Earth Obs. Geoinf.* **67**, 30–42 (2018).
24. Desai, A. R. *et al.* Multi-Sensor Approach for High Space and Time Resolution Land Surface Temperature. *Earth Space Sci.* **8**, e2021EA001842 (2021).
25. Voogt, J. A. & Oke, T. R. Thermal remote sensing of urban climates. *Remote Sens. Environ.* **86**, 370–384 (2003).
26. Zhao, L., Lee, X., Smith, R. B. & Oleson, K. Strong contributions of local background climate to urban heat islands. *Nature* **511**, 216–219 (2014).
27. Chakraborty, T., Venter, Z. S., Qian, Y. & Lee, X. Lower Urban Humidity Moderates Outdoor Heat Stress. *AGU Adv.* **3**, e2022AV000729 (2022).
28. Naserikia, M. *et al.* Land surface and air temperature dynamics: The role of urban form and seasonality. *Sci. Total Environ.* **905**, 167306 (2023).
29. Jin, M. & Dickinson, R. E. Land surface skin temperature climatology: benefitting from the strengths of satellite observations. *Environ. Res. Lett.* **5**, 044004 (2010).
30. Lian, X. *et al.* Spatiotemporal variations in the difference between satellite-observed daily maximum land surface temperature and station-based daily maximum near-surface air temperature. *J. Geophys. Res. Atmos.* **122**, 2254–2268 (2017).

31. Benali, A., Carvalho, A. C., Nunes, J. P., Carvalhais, N. & Santos, A. Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sens. Environ.* **124**, 108–121 (2012).

32. Good, E. J., Ghent, D. J., Bulgin, C. E. & Remedios, J. J. A spatiotemporal analysis of the relationship between near-surface air temperature and satellite land surface temperatures using 17 years of data from the ATSR series. *J. Geophys. Res. Atmos.* **122**, 9185–9210 (2017).

33. Muller, C. L. *et al.* Crowdsourcing for climate and atmospheric sciences: current status and future potential. *Int. J. Climatol.* **35**, 3185–3203 (2015).

34. Meier, F., Fenner, D., Grassmann, T., Otto, M. & Scherer, D. Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Clim.* **19**, 170–191 (2017).

35. Stahl, K., Moore, R. D., Floyer, J. A., Asplin, M. G. & McKendry, I. G. Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. *Agric. For. Meteorol.* **139**, 224–236 (2006).

36. Vedrí, J. *et al.* Empirical methods to determine surface air temperature from satellite-retrieved data. *Int. J. Appl. Earth Obs. Geoinf.* **136**, 104380 (2025).

37. Fenner, D., Meier, F., Bechtel, B., Otto, M. & Scherer, D. Intra and inter 'local climate zone' variability of air temperature as observed by crowdsourced citizen weather stations in Berlin, Germany. *Meteorol. Z.* **26**, 525–547 (2017).

38. Brousse, O. *et al.* Evidence of horizontal urban heat advection in London using six years of data from a citizen weather station network. *Environ. Res. Lett.* **17**, 044041 (2022).

39. Varentsov, M. I. *et al.* Urban heat island of the Moscow megacity: the long-term trends and new approaches for monitoring and research based on crowdsourcing data. *IOP Conf. Ser.: Earth Environ. Sci.* **606**, 012063 (2020).

40. Bell, S., Cornford, D. & Bastin, L. How good are citizen weather stations? Addressing a biased opinion. *Weather* **70**, 75–84 (2015).

41. Chapman, L., Bell, C. & Bell, S. Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. *Int. J. Climatol.* **37**, 3597–3605 (2017).

42. Brousse, O., Simpson, C. H., Poorthuis, A. & Heaviside, C. Unequal distributions of crowdsourced weather data in England and Wales. *Nat. Commun.* **15**, 4828 (2024).

43. Sakthivel, P. & Sengupta, R. Spatial bias in placement of citizen and conventional weather stations and their impact on urban climate research: A case study of the Urban Heat Island effect in Canada. *Urban Clim.* **59**, 102280 (2025).

44. Thornton, P. E. *et al.* Gridded daily weather data for North America with comprehensive uncertainty quantification. *Sci. Data* **8**, 190 (2021).

45. Shtiliyanova, A. *et al.* Kriging-based approach to predict missing air temperature data. *Comput. Electron. Agric.* **142**, 440–449 (2017).

46. DeGaetano, A. T. & Belcher, B. N. Spatial Interpolation of Daily Maximum and Minimum Air Temperature Based on Meteorological Model Analyses and Independent Observations. *J. Appl. Meteorol. Climatol.* **46**, 1981–1992 (2007).

47. Hooker, J., Duveiller, G. & Cescatti, A. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Sci. Data* **5**, 180246 (2018).

48. Yao, R. *et al.* Global seamless and high-resolution temperature dataset (GSHTD), 2001–2020. *Remote Sens. Environ.* **286**, 113422 (2023).

49. Zhang, T. *et al.* A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020). *Earth Syst. Sci. Data* **14**, 5637–5649 (2022).

50. Chen, Y. *et al.* An all-sky 1 km daily land surface air temperature product over mainland China for 2003–2019 from MODIS and ancillary data. *Earth Syst. Sci. Data* **13**, 4241–4261 (2021).

51. Oyler, J. W., Ballantyne, A., Jencso, K., Sweet, M. & Running, S. W. Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *Int. J. Climatol.* **35**, 2258–2279 (2015).

52. Li, X., Zhou, Y., Asrar, G. R. & Zhu, Z. Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sens. Environ.* **215**, 74–84 (2018).

53. Shen, H. *et al.* Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. *Remote Sens. Environ.* **240**, 111692 (2020).

54. dos Santos, R. S. Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. *Int. J. Appl. Earth Obs. Geoinf.* **88**, 102066 (2020).

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

56. Immorlano, F. *et al.* Transferring climate change physical knowledge. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2413503122 (2025).

57. Zhao, L. *et al.* Global multi-model projections of local urban climates. *Nat. Clim. Chang.* **11**, 152–157 (2021).

58. Zheng, Z., Zhao, L. & Oleson, K. W. Large model structural uncertainty in global projections of urban heat waves. *Nat. Commun.* **12**, 3736 (2021).

59. Lee, J. & Dessler, A. E. Improved Surface Urban Heat Impact Assessment Using GOES Satellite Data: A Comparative Study With ERA-5. *Geophys. Res. Lett.* **51**, e2023GL107364 (2024).

60. Chen, J., Qian, Y., Chakraborty, T. C. & Yang, Z. Complexities of urban impacts on long-term seasonal trends in a mid-sized arid city. *Environ. Res. Commun.* **6**, 021004 (2024).

61. Wang, K. *et al.* Comparing the diurnal and seasonal variabilities of atmospheric and surface urban heat islands based on the Beijing urban meteorological network. *J. Geophys. Res. Atmos.* **122**, 2131–2154 (2017).

62. Tzavali, A., Paravantis, J. P. & Mihalakakou, G. Urban Heat Island Intensity: A literature review. *Fresen. Environ. Bull.* **24**, (2015).

63. 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).

64. PRISM Climate Group, Oregon State University. https://prism.oregonstate.edu (2014).

65. Copernicus Climate Change Service. ERA5-Land post-processed daily statistics from 1950 to present. ECMWF https://doi.org/10.24381/CDS.E9C9C792 (2024).

66. Du, H. *et al.* Simultaneous investigation of surface and canopy urban heat islands over global cities. *ISPRS J. Photogramm. Remote Sens.* **181**, 67–83 (2021).

67. Mildrexler, D. J., Zhao, M. & Running, S. W. A global comparison between station air temperatures and MODIS land surface temperatures reveals the cooling role of forests. *J. Geophys. Res. Biogeosci.* **116**, (2011).

68. Dobson, J., Bright, E., Coleman, P., Durfee, R. & Worley, B. LandScan: A Global Population Database for Estimating Populations at Risk. *Photogramm. Eng. Remote Sens.* **66**, 849–857 (2000).

69. Li, Y. *et al.* Green spaces provide substantial but unequal urban cooling globally. *Nat. Commun.* **15**, 7108 (2024).

70. Mitchell, B. C. & Chakraborty, J. Landscapes of thermal inequity: disproportionate exposure to urban heat in the three largest US cities. *Environ. Res. Lett.* **10**, 115005 (2015).

71. Chakraborty, T., Newman, A. J., Qian, Y., Hsu, A. & Sheriff, G. Residential segregation and outdoor urban moist heat stress disparities in the United States. *One Earth* **6**, 738–750 (2023).

72. Meili, N. *et al.* An urban ecohydrological model to quantify the effect of vegetation on urban climate and hydrology (UT&C v1.0). *Geosci. Model Dev.* **13**, 335–362 (2020).

73. Zhang, K. *et al.* Increased heat risk in wet climate induced by urban humid heat. *Nature* **617**, 738–742 (2023).

74. Krayenhoff, E. S. *et al.* Cooling hot cities: a systematic and critical review of the numerical modelling literature. *Environ. Res. Lett.* **16**, 053007 (2021).

75. Zhao, L. *et al.* Interactions between urban heat islands and heat waves. *Environ. Res. Lett.* **13**, 034003 (2018).

76. Zhao, L., Lee, X. & Schultz, N. M. A wedge strategy for mitigation of urban warming in future climate scenarios. *Atmos. Chem. Phys.* **17**, 9067–9080 (2017).

77. Mistry, M. N. *et al.* Comparison of weather station and climate reanalysis data for modelling temperature-related mortality. *Sci. Rep.* **12**, 5178 (2022).

78. Jean, N. *et al.* Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).

79. Ma, Y., Chen, S., Ermon, S. & Lobell, D. B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **301**, 113924 (2024).

80. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). doi:10.1109/CVPR.2016.90.

81. Zhengming Wan & Dozier, J. A generalized split-window algorithm for retrieving land-surface temperature from space. *IEEE Trans. Geosci. Remote Sensing* **34**, 892–905 (1996).

82. Zheng, G. *et al.* Knowledge-based Residual Learning. in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* 1653–1659 (International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 2021). doi:10.24963/ijcai.2021/228.

83. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. in *Proceedings of the 26th Annual International Conference on Machine Learning* 41–48 (Association for Computing Machinery, New York, NY, USA, 2009). doi:10.1145/1553374.1553380.

84. Castells, T., Weinzaepfel, P. & Revaud, J. SuperLoss: A Generic Loss for Robust Curriculum Learning. in *Advances in Neural Information Processing Systems* vol. 33 4308–4319 (Curran Associates, Inc., 2020).

85. Earth Resources Observation and Science (EROS) Center. Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2. U.S. Geological Survey https://doi.org/10.5066/P9OGBGM6 (2020).

86. Li, X., Zhou, Y., Zhu, Z. & Cao, W. A national dataset of 30 m annual urban extent dynamics (1985–2015) in the conterminous United States. *Earth Syst. Sci. Data* **12**, 357–371 (2020).

87. Li, M., Wang, Y., Rosier, J. F., Verburg, P. H. & van Vliet, J. Global maps of 3D built-up patterns for urban morphological analysis. *Int. J. Appl. Earth Obs. Geoinf.* **114**, 103048 (2022).

88. Farr, T. G. *et al.* The Shuttle Radar Topography Mission. *Rev. Geophys.* **45**, (2007).

89. Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).

90. Copernicus Climate Change Service, Climate Data Store. ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) https://doi.org/10.24381/CDS.BD0915C6 (2023).

91. Copernicus Climate Change Service, Climate Data Store. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) https://doi.org/10.24381/CDS.ADBB2D47 (2023).

92. Jiawei Zhuang *et al.* pangeo-data/xESMF: v0.8.8. Zenodo https://doi.org/10.5281/ZENODO.4294774 (2024).

93. Krayenhoff, E. S., Moustaoui, M., Broadbent, A. M., Gupta, V. & Georgescu, M. Diurnal interaction between urban expansion, climate change and adaptation in US cities. *Nat. Clim. Chang.* **8**, 1097–1103 (2018).

94. Jon Dewitz. National Land Cover Database (NLCD) 2021 Products. U.S. Geological Survey https://doi.org/10.5066/P9JZ7AO3 (2023).

95. Stewart, I. D. & Oke, T. R. Local Climate Zones for Urban Temperature Studies. *Bull. Am. Meteorol. Soc.* **93**, 1879–1900 (2012).

96. Danielson, J. J. & Gesch, D. B. *Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010)*. (2011).

97. Chakraborty, T. C., Sarangi, C. & Lee, X. Reduction in human activity can enhance the urban heat island: insights from the COVID-19 lockdown. *Environ. Res. Lett.* **16**, 054060 (2021).

98. Pesaresi, M. *et al.* A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **6**, 2102–2131 (2013).

99. Zhang, Y. *et al.* U-HAT: Urban High-resolution Air Temperature Data. Zenodo https://doi.org/10.5281/ZENODO.20057651 (2026).

Acknowledgments
We acknowledge the high-performance computing support provided by the U.S. NSF National Center for Atmospheric Research (NCAR)'s Computational and Information Systems Laboratory, sponsored by the US National Science Foundation. We thank the cloud computing support provided by Google Earth Engine. We also thank the three reviewers for their valuable suggestions and feedback to this work.

Funding
L.Z. acknowledges the support by the U.S. National Science Foundation (CAREER Award Grant No. 2145362). L.Z. and T.C. acknowledge the U.S. National Aeronautics and Space Administration (NASA) through the AMT and LCLUC programs (Grant #80NSSC25K7322). T.C.'s contribution was also supported jointly by the U.S. Department of Energy's (DOE) Office of Science Biological and Environmental Research's Earth System Model Development and Regional and Global Model Analysis program Areas via a DOE Early Career Award and by an Interdisciplinary Research in Earth Science grant (number 80NSSC24K0505) funded by NASA. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830.

Author contributions
L.Z. and P.G. proposed and designed the study. Y.Z. developed the U-TL model with contributions from L.Z. and P.M. Y.Z. performed data collection with contributions from K.Z. Y.Z. performed the U-HAT data processing and analysis. L.Z. and T.C. contributed ideas to the data analysis. Y.Z. and L.Z. drafted the manuscript. All authors edited and revised the manuscript.

Competing interests
Authors declare that they have no competing interests.

Figure captions

Fig. 1: Limited true urban representation in global weather stations. a, Meteorological stations in the Global Historical Climatology Network Hourly[13]. Stations located within urban boundaries as defined by the Global Urban Boundaries dataset[14] are shown as brown dots while those outside are in green. Blue stars indicate selected stations that are considered "urban" based on this broad classification. **b,** Proportion of meteorological stations categorized by impervious surface fraction within a 200-meter buffer. Impervious surface fraction is based on the "built-up" category in the ESA WorldCover dataset[15]. **c-g,** Satellite images from Google Earth (http://earth.google.com) showing the surrounding environments of selected stations. Basemap from Natural Earth (https://www.naturalearthdata.com/).

Fig. 2: U-HAT accurately reproduces observed urban climatology compared to existing T_a (near-surface air temperature) datasets. a, Mean UHI (urban heat island) intensities across cities in the CONUS (contiguous United States) during daytime and nighttime in JJA (June–August) and DJF (December–February) from 2013–2020 calculated using U-HAT, Yao et al.[48], Zhang et al.[49] and T_s. Error

bars indicate the standard error of the mean. **b,** Distributions of correlation coefficients (r) from linear regressions between T_a, T_s (land surface temperature) and the Normalized Difference Vegetation Index (NDVI) across CONUS cities. T_a estimates are from U-HAT, Yao et al.[48] and Zhang et al.[49]. Each data point is computed using JJA mean values of variables from 2013–2020, based on all available pixels within one cluster. Center bars represent the median, box edges the 25th and 75th percentiles, and error bars extend to 3×IQR (interquartile range) from the quartiles. Shaded violin plots indicate the underlying distribution.

Fig. 3: U-HAT demonstrates substantial differences in magnitude and spatial variability between T_s (land surface temperature) and T_a (near-surface air temperature) over 384 cities in the CONUS (contiguous United States). a,b, Map of the city-wise mean vertical temperature gradient ($\Delta T = T_s - T_a$) averaged across 9 years at daytime (**a**) and nighttime (**b**). For each map, pixel-wise ΔT in Seattle, Phoenix, Dallas, and New York are shown below. The frequency distribution of pixel-wise ΔT is also shown. **c,d,** Map of the difference between the city-wise standard deviation of 9-year mean of T_s and T_a ($\Delta\sigma_T$) at daytime (**c**) and nighttime (**d**). Basemap from Natural Earth (https://www.naturalearthdata.com/).

Fig. 4 U-HAT demonstrates substantial differences between daytime extreme heat exposure measured by T_s (HE_s) and T_a (HE_a) in their magnitude and spatial variability over 384 cities in the CONUS (contiguous United States). The map shows pixel-wise differences between daytime HE_s and HE_a. **a–h**, Distributions of HE_a and HE_s in selected cities on log scale. Center bars represent the median, box edges the 25th and 75th percentiles, and error bars the 1st and 99th percentiles. Basemap from Natural Earth (https://www.naturalearthdata.com/).

Editorial Summary

This study develops a transfer-learning method to estimate high-resolution urban air temperature across the contiguous United States and shows that satellite land surface temperature substantially overestimates heat stress and temperature variability in those cities.

Peer review information: *Nature Communications* thanks Lahouari Bounoua and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.