scientific data

OPEN

DATA DESCRIPTOR

CHUWD-H v1.0: a comprehensive historical hourly weather database for U.S. urban energy system modeling

Chenghao Wang[1,2]✉, Chengbin Deng[2,3], Henry Horsey[4], Janet L. Reyna[4], Di Liu[2,3], Sarah Feron[5,6], Raúl R. Cordero[5], Jiyun Song[7] & Robert B. Jackson[8,9,10]

Reliable and continuous meteorological data are crucial for modeling the responses of energy systems and their components to weather and climate conditions, particularly in densely populated urban areas. However, existing long-term datasets often suffer from spatial and temporal gaps and inconsistencies, posing great challenges for detailed urban energy system modeling and cross-city comparison under realistic weather conditions. Here we introduce the Historical Comprehensive Hourly Urban Weather Database (CHUWD-H) v1.0, a 23-year (1998–2020) gap-free and quality-controlled hourly weather dataset covering 550 weather station locations across all urban areas in the contiguous United States. CHUWD-H v1.0 synthesizes hourly weather observations from stations with outputs from a physics-based solar radiation model and a reanalysis dataset through a multi-step gap filling approach. A 10-fold Monte Carlo cross-validation suggests that the accuracy of this gap filling approach surpasses that of conventional gap filling methods. Designed primarily for urban energy system modeling, CHUWD-H v1.0 should also support historical urban meteorological and climate studies, including the validation and evaluation of urban climate modeling.

Background & Summary

Energy systems are essential in the global effort to reduce energy use and greenhouse gas emissions for climate change mitigation[1,2]. Achieving the ambitious regional and national emission reduction targets relies on the rapid transformation and decarbonization of current energy systems, particularly in densely populated urban areas[3]. As home to more than half of the global population, urban areas are responsible for around 75% of global primary energy consumption and 70% of anthropogenic greenhouse gas emissions[4,5]. With ongoing urbanization amid climate change, cities are expected to face evolving challenges brought by changes in energy demand, energy supply, and their interactions[3]. Addressing these challenges requires robust and reliable urban energy system modeling. Such modeling also plays a crucial role in supporting the planning and development of new energy infrastructures, guiding energy policy decisions, and enhancing the resilience of urban energy systems to extreme weather events and climate change.

Reliable urban energy system modeling should be able to integrate energy demand and supply sectors with weather and climate conditions[6–8]. This integration is particularly important for components of the energy system that are susceptible to weather fluctuations. For example, the outputs of variable renewable energy systems (e.g., solar and wind power), as well as the cooling and space heating demands of buildings, are heavily influenced by weather conditions[9–12]. To assess the performance of energy system components under typical

[1]School of Meteorology, University of Oklahoma, Norman, OK, USA. [2]Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, OK, USA. [3]Center for Spatial Analysis, University of Oklahoma, Norman, OK, USA. [4]National Renewable Energy Laboratory, Golden, CO, USA. [5]Knowledge Infrastructure, University of Groningen, Wirdumerdijk 34, 8911, CE, Leeuwarden, Netherlands. [6]Universidad de Santiago de Chile. Av. Bernardo O'Higgins, 3363, Santiago, Chile. [7]State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan, China. [8]Department of Earth System Science, Stanford University, Stanford, CA, USA. [9]Woods Institute for the Environment, Stanford University, Stanford, CA, USA. [10]Precourt Institute for Energy, Stanford University, Stanford, CA, USA. ✉e-mail: chenghao.wang@ou.edu

weather conditions, typical meteorological year (TMY) data have been widely used in existing modeling efforts, especially for the building sector[13–18]. TMY data comprise yearlong, hourly meteorological and solar radiation data that represent typical weather conditions over a long period of time for a specific location. They essentially consist of 12 typical months of data selected from different calendar years. As one of the standard input weather datasets for energy system modeling, TMY data have also been modified with climate projection data using methods such as morphing for simulations under future scenarios[19,20].

However, TMY data are insufficient for capturing interannual variability and the frequency and intensity of peak loads and demands under extreme meteorological conditions, as suggested in previous studies[7,21,22]. Moreover, TMY data are becoming increasingly outdated; for example, the latest official TMY collection, TMY3, was developed based on data only up to 2005[23]. Relying on TMY data for current urban energy system modeling can potentially lead to large uncertainties, particularly during peak periods. Alternatives aiming to capture extreme conditions, such as the extreme meteorological year, typical hot year, and typical cold year datasets, have been proposed using similar concepts of selecting representative months or years[24–26]. Although these datasets provide improved insight into extreme weather conditions, they still fall short of capturing the full spectrum of variability. In comparison, actual meteorological year (AMY) data, which include actual hourly weather data for specific locations and years, are indispensable for assessing the long-term trends, interannual variability, and historical extremes of energy systems[22,27]. Unfortunately, largely due to spatial and temporal data gaps in weather observations, long-term AMY data available for energy system modeling are still very rare.

Recent studies have attempted to address this data scarcity with gap-free, gridded AMY data derived from the results of numerical weather and climate simulations[25,28]. Nevertheless, this approach cannot fully substitute for observation-based AMY data for urban energy system modeling due to the potential biases and uncertainties inherent in these simulations. One of the notable limitations is the inadequate representation of urban climates (e.g., the urban heat island effect) in most regional and global climate models and reanalysis datasets[29,30]. To comprehensively understand the dynamics of urban energy systems and their dependence on local meteorological conditions and broader background climate, there is a pressing need for a long-term, gap-free, hourly weather dataset that covers a diverse array of urban areas.

Here we introduce the Historical Comprehensive Hourly Urban Weather Database (CHUWD-H) v1.0, a long-term (1998–2020), gap-free, and quality-controlled hourly weather dataset designed for urban energy system modeling in the United States (U.S.). CHUWD-H v1.0 includes data from 550 weather station locations, covering all 481 urban areas across the entire contiguous U.S. (CONUS). This database is primarily constructed from observations at ground-based weather stations, complemented by outputs from a physics-based solar radiation model and reanalysis data. The current version (v1.0) features 14 gap-free meteorological variables at hourly intervals. The accuracy of the gap-filled hourly data in CHUWD-H v1.0 exceeds that of other commonly used gap filling methods, as evidenced by a 10-fold Monte Carlo cross-validation. CHUWD-H v1.0 is publicly accessible through an online data repository[31] and an interactive platform[32]. This database is expected to facilitate a broad spectrum of applications that extend well beyond urban energy system modeling.

Methods

Selection of weather stations. To develop a historical hourly weather database based on station observations for urban energy system modeling, the first step is to select representative stations. We selected a subset of weather stations from the TMY3 dataset, a widely used dataset developed by the National Renewable Energy Laboratory (NREL)[23]. This selection can facilitate comparisons of energy system modeling efforts based on the official TMY3 dataset and CHUWD-H v1.0. As the latest version of the TMY datasets, the TMY3 dataset encompasses 925 weather stations across the CONUS, constructed using measured and modeled data spanning either 30 years (1976–2005) or 15 years (1991–2005). Weather stations in the TMY3 dataset are categorized into three classes based on the quality of the source data: Class I stations have the lowest uncertainty, Class II stations have moderate uncertainty, and Class III stations have the most data gaps. Within the CONUS, there are 217 Class I, 564 Class II, and 144 Class III stations. This classification also indicates the general reliability and completeness of the weather observations at each station. To create a continuous, gap-free dataset, we prioritized Classes I and II stations in the station selection process. However, in the absence of Class I or II stations within a target urban area, Class III stations were also considered.

We then used the U.S. Census Bureau's 2010 Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line Shapefiles data[33] to identify representative stations for urban areas. The U.S. Census Bureau delineates the boundaries of 481 densely developed urban areas (or "urbanized areas"), each with a population of at least 50,000. The urban boundary shapefile can be retrieved from the Census Bureau's official website: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2010.html. We identified 392 weather stations located directly within these urban boundaries, and an additional 162 stations in close proximity to at least one urban area. We further inspected the operational status of individual stations using weather observations (see "Ground-based hourly weather observations" section), which is critical to the future updates of CHUWD-H. Four stations that discontinued weather observations after ~2010 were then excluded from the database. Consequently, CHUWD-H v1.0 covers 550 stations across the CONUS, including 181 Class I stations (141 within urban boundaries), 323 Class II stations (223 within urban boundaries), and 46 Class III stations (26 within urban boundaries).

When retrieving the observational records from the selected 550 weather stations, we found that several stations in the official TMY3 dataset[23] have inaccurate latitude, longitude, and/or time zone. These inaccuracies could influence source data retrieval processes. To address this, we leveraged an updated version of the TMY3 dataset developed by Climate.OneBuilding.Org (https://climate.onebuilding.org/WMO_Region_4_North_and_Central_America/USA_United_States_of_America/index.html), which has undergone extensive verification[34]. Additionally, we cross-referenced station information from the Integrated Surface Database (ISD)[35] and the

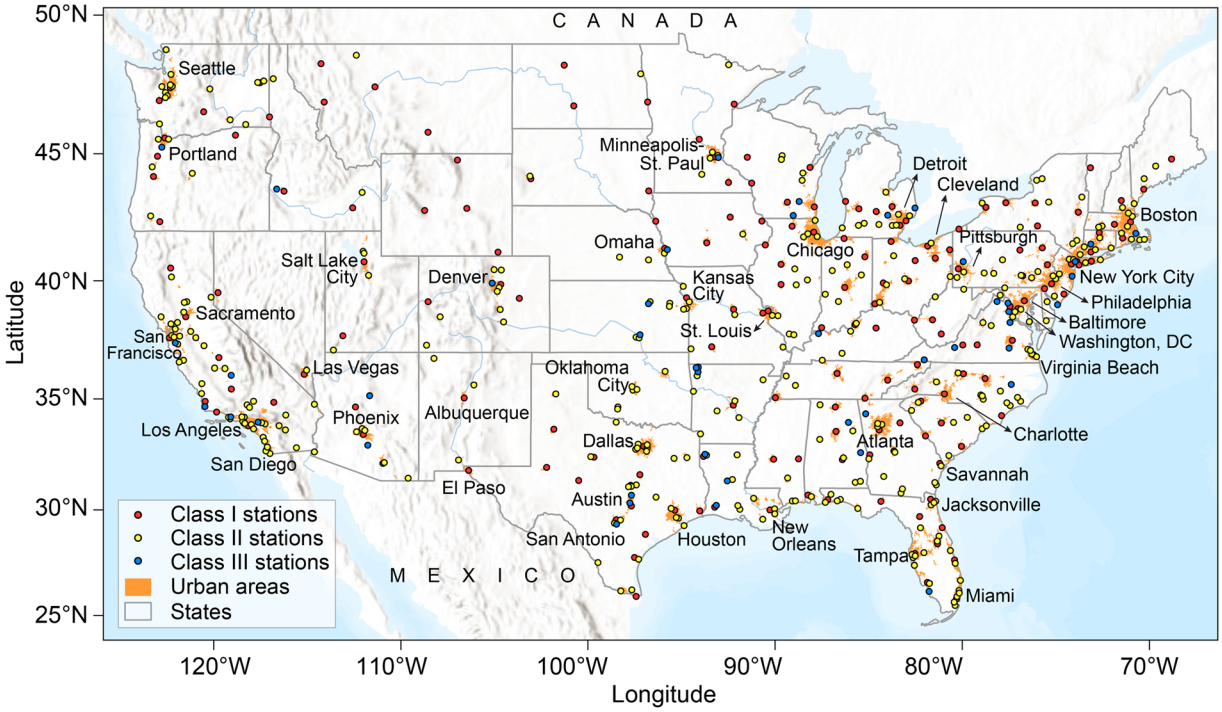

Fig. 1 Spatial distribution of the 550 representative weather stations in CHUWD-H v1.0, color coded by classification according to the official TMY3 dataset. Class I stations have the lowest uncertainty, Class II stations have moderate uncertainty, and Class III stations have the most data gaps. Shaded areas in orange are urban areas with populations of at least 50,000. The basemap is World Terrain data from ArcGIS Pro.

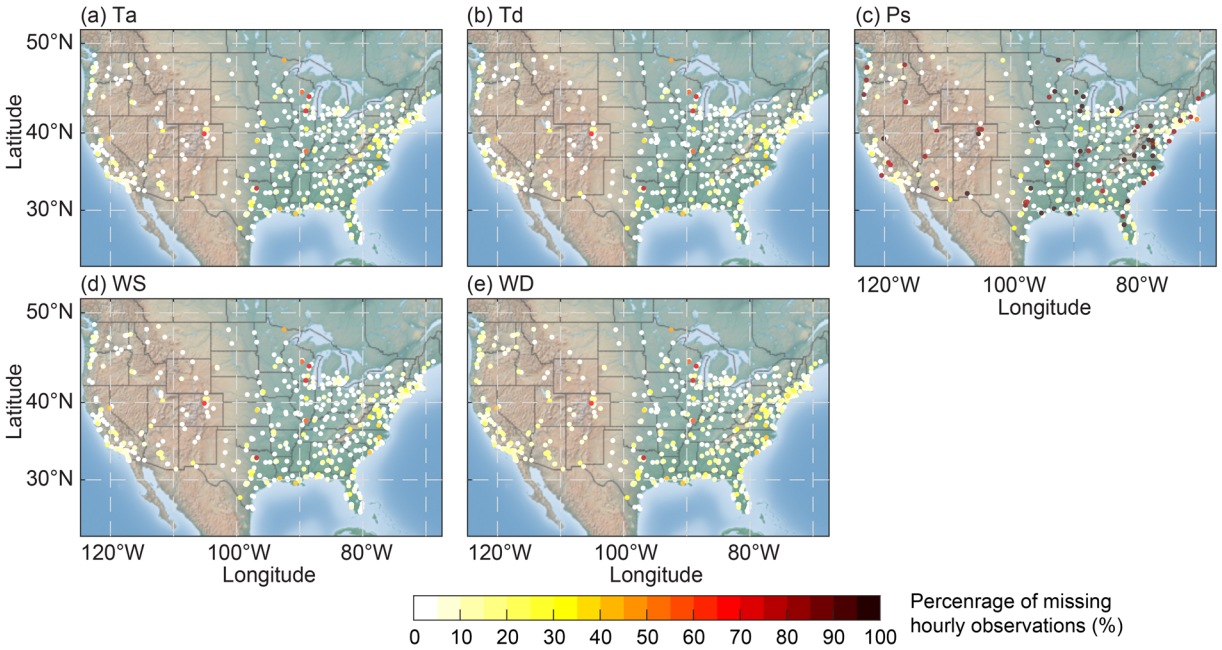

Fig. 2 Percentage of missing hourly observations over the 23-year period (1998–2020) for (**a**) near-surface air temperature (Ta), (**b**) dew point temperature (Td), (**c**) surface pressure (Ps), (**d**) wind speed (WS), and (**e**) wind direction (WD) for all stations in CHUWD-H v1.0. Each dot represents an individual station. The basemap is the shaded relief map blended with a land cover palette from MATLAB.

National Centers for Environmental Information (NCEI)'s Climate Data Online (CDO)[36]. This allowed us to identify discrepancies in station locations, time zones, and elevations. In total, we corrected geographical locations and/or time zones for 43 weather stations. Figure 1 shows the spatial distribution of all 550 stations in CHUWD-H v1.0.

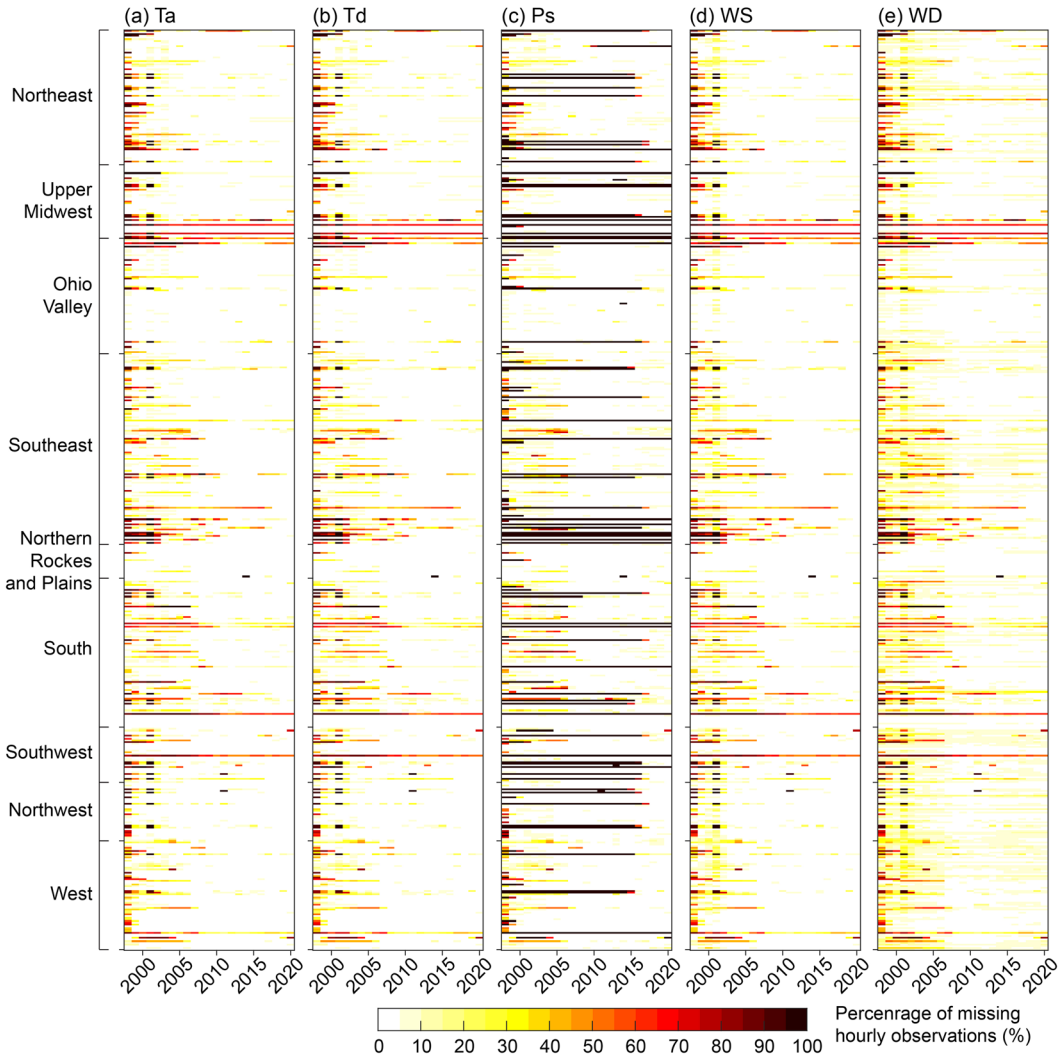

Fig. 3 Percentage of missing hourly observations by year and station for (**a**) near-surface air temperature (Ta), (**b**) dew point temperature (Td), (**c**) surface pressure (Ps), (**d**) wind speed (WS), and (**e**) wind direction (WD) for all stations in CHUWD-H v1.0. Note that weather stations are grouped according to the nine climate regions defined by the NCEI[51]. In each subplot, each row shows the percentage of missing hourly observations for a specific station across different years.

Ground-based hourly weather observations. Hourly weather observations for representative stations were retrieved from the Integrated Surface Database (ISD) (https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database). ISD is a global database that includes hourly and synoptic surface observations from more than 100 original data sources[35]. As one of the flagship climate data products of the NCEI, ISD currently covers over 14,000 active stations. To ensure the high quality of data in ISD, various quality control measures were carried out, including validity checks, extreme value checks, internal consistency checks, and external continuity checks. These measures extend beyond the internal quality controls already present in source datasets (e.g., the Automated Surface/Weather Observing Systems; ASOS/AWOS). To further filter out duplicate values and sub-hourly data, we retrieved hourly air temperature, dew point temperature, sea level pressure, wind direction, wind speed, one-hour accumulated liquid precipitation, and six-hour accumulated liquid precipitation data from ISD-Lite for 1998–2021.

It is noteworthy that the same weather station can be listed under different station IDs over time. To minimize gaps in observational data, we compared the geographical locations, names, and elevations of stations in proximity to each target station to identify those that have undergone changes in name, ID, and/or location. For stations missing entire years of observations, we identified and used nearby stations with similar elevations ($<$100 m difference) and geographical conditions to fill in data gaps. The final dataset includes 312 stations merged from multiple station records, with nine of these stations partially supplemented with data from nearby stations. Note that data retrieved from ISD directly reflect raw observations, which may not be homogenized for some stations.

Fig. 4 The multi-step gap filling (MSGF) approach to fill a missing air temperature data point at the target weather station. MLR and SLR models are multiple linear regression and simple linear regression models, respectively. For clarity, this flowchart omits connections from statistically insignificant results in the MLR or SLR models based on 30-day, 90-day, and one-year windows (i.e., "No" decision in the light pink box "Statistically significant") to the data gap decision boxes for subsequent longer windows (light green boxes on the left for MLR and right for SLR).

Given that most energy system models require meteorological data in local time for model input, we converted all hourly observations from Coordinated Universal Time (UTC) to local time. To maintain consistency with TMY3 data, we also removed February 29 in leap years from our database. We further converted the sea level pressure to surface pressure for each station using the hypsometric equation and the station's elevation.

Radiation data from NSRDB. Solar radiation data were retrieved from the National Solar Radiation Data Base (NSRDB) (https://nsrdb.nrel.gov/), developed by the NREL[37]. This database provides 4-km resolution solar irradiation data at 30-min intervals covering the entire CONUS from 1998 to the present. NSRDB uses the two-step Physical Solar Model (PSM), which computes solar radiation from satellite data under both clear sky and cloudy conditions with the Fast All-sky Radiation Model for Solar Applications (FARMS)[38]. More specifically, FARMS integrates cloud properties from satellite data, aerosol optical depth from Moderate Resolution Imaging Spectroradiometer (MODIS) and Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) data, along with additional atmospheric and land surface data from multiple sources to calculate Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI). Evaluation against concurrent ground-based measurements suggests a mean bias error of $\pm 5\%$ for the estimated hourly GHI and $\pm 10\%$ for DNI[39]. We retrieved 30-min GHI, DNI, DHI, clear-sky GHI, clear-sky DNI, clear-sky DHI, and zenith angle from NSRDB for each station location, and converted these instantaneous values to hourly average (during the preceding hour) using the trapezoidal rule.

Reanalysis data from MERRA-2. We further used the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) dataset as an additional resource to fill gaps in weather observations. As a reanalysis dataset developed by NASA's Global Modeling and Assimilation Office, MERRA-2 is generated with the Goddard Earth Observing System (GEOS) atmospheric data assimilation system[40]. It provides hourly data from 1980 to the present at a spatial resolution of 0.5° latitude × 0.625° longitude. The current version includes several major improvements over its predecessor (MERRA), such as the assimilation of aerosol observations, improved characterization of stratospheric processes, and better representation of glaciated land surface processes. However, the coarse native resolution of MERRA-2 is not suitable for direct station-level gap filling. Here we used hourly air temperature, dew point temperature, surface pressure, wind speed, and wind direction from downscaled 4-km MERRA-2 data developed by NREL (https://nsrdb.nrel.gov/). The high-resolution temperature and pressure data were downscaled from the native resolution of MERRA-2 using an elevation scaling, while the

Variable	Unit	Definition
Year	—	Year of the data.
Month	—	Month of the data.
Day	—	Day of the data.
Hour	—	Hour of the data (1–24).
Minute	—	Minute of the data.
Ta	°C	Air temperature at the time indicated.
Ta_flag	—	Gap filling flag for air temperature: "0" is based on observation, "1" is gap filled.
Td	°C	Dew point temperature at the time indicated.
Td_flag	—	Gap filling flag for dew point temperature: "0" is based on observation, "1" is gap filled.
RH	%	Relative humidity at the time indicated.
Pa	Pa	Atmospheric pressure at the time indicated.
Pa_flag	—	Gap filling flag for atmospheric pressure: "0" is based on observation, "1" is gap filled.
GHI	W h m^{-2}	Global horizontal radiation during the hour preceding the time indicated.
DNI	W h m^{-2}	Direct normal radiation during the hour preceding the time indicated.
DHI	W h m^{-2}	Diffuse horizontal radiation during the hour preceding the time indicated.
CS_GHI	W h m^{-2}	Global horizontal radiation assuming clear sky condition during the hour preceding the time indicated.
CS_DNI	W h m^{-2}	Direct normal radiation assuming clear sky condition during the hour preceding the time indicated.
CS_DHI	W h m^{-2}	Diffuse horizontal radiation assuming clear sky condition during the hour preceding the time indicated.
WD	°	Wind direction at the time indicated, following the convention North = 0.0, East = 90.0, South = 180.0, West = 270.0. If calm, direction equals zero. (The angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing.)
WD_flag	—	Gap filling flag for wind direction: "0" is based on observation, "1" is gap filled.
WS	m s^{-1}	Wind speed at the time indicated.
WS_flag	—	Gap filling flag for wind speed: "0" is based on observation, "1" is gap filled.
PCW	cm	Total precipitable water contained in a column of unit cross section extending from the earth's surface to the top of the atmosphere at the time indicated.
PC_1hr	mm	Liquid precipitation depth measured over a one-hour accumulation period. Trace precipitation is coded as "−1", while missing data are "999".
PC_6hr	mm	Liquid precipitation depth measured over a six-hour accumulation period. Trace precipitation is coded as "−1", while missing data are "999".
Zenith	°	Zenith angle at the time indicated.

Table 1. Summary of variables in CHUWD-H v1.0.

Variable	Definition
CHUWD_ID	Station ID used in CHUWD-H v1.0.
CHUWD_Name	Station name used in CHUWD-H v1.0.
ISD_USAF	United States Air Force (USAF) Station ID in ISD.
ISD_WBAN	Weather Bureau Army Navy (WBAN) Station ID in ISD.
ISD_Name	Station name in ISD.
TMY3_USAF	USAF Station ID in the official TMY3 dataset.
TMY3_Name	Station name in the official TMY3 dataset.
OB_USAF	USAF Station ID in the updated TMY3 dataset from Climate.OneBuilding.Org.
OB_Name	Station name in the updated TMY3 dataset from Climate.OneBuilding.Org.
Class	Station class (I, II, or III).
State	Station state (abbreviations).
Lat	Latitude of the station (unit: °).
Lon	Longitude of the station (unit: °).
Elev	Elevation of the station (unit: m).
Time_zone	Number of hours from UTC (Standard Time).
Within_UA	Whether the station is located within the boundary of an urban area (population ≥ 50,000).

Table 2. Definitions of variables in "CHUWD-H v1.0 stations.xlsx".

high-resolution humidity and wind data were downscaled with a nearest-neighbor approach[37]. It is noteworthy that the elevations of the downscaled 4-km MERRA-2 grids may slightly differ from those of station locations. Although the elevation discrepancy for most stations is less than 100 m, which is likely to have only marginal

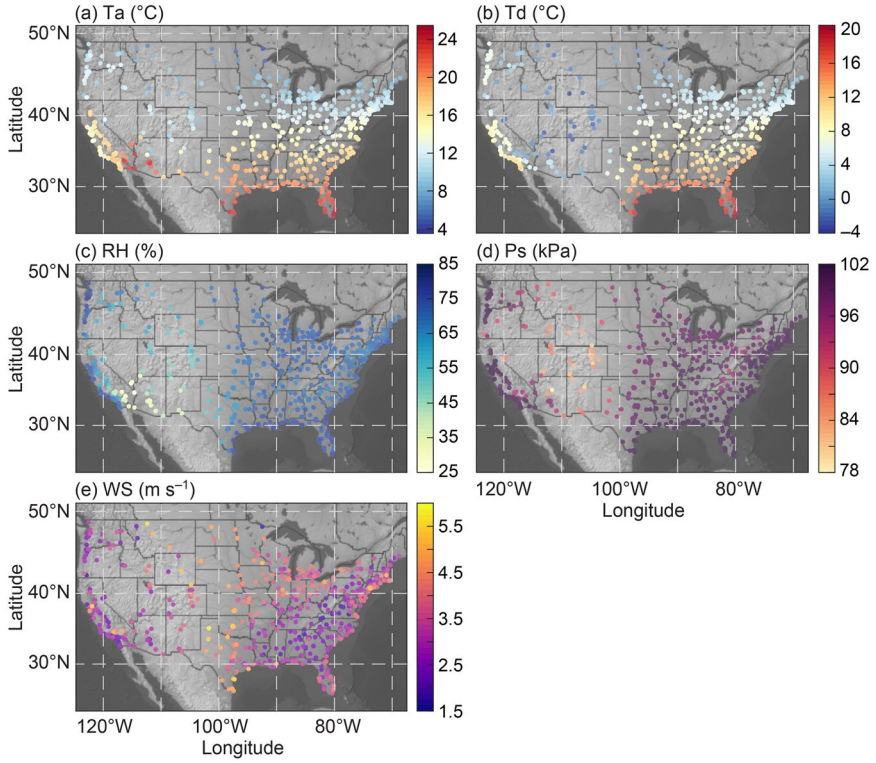

Fig. 5 Hourly (**a**) air temperature (Ta; °C), (**b**) dew point temperature (Td; °C), (**c**) relative humidity (RH; %), (**d**) atmospheric pressure (Ps; kPa), and (**e**) wind speed (WS; m s^{-1}) averaged over the 23-year period (1998–2020) for all stations in CHUWD-H v1.0.

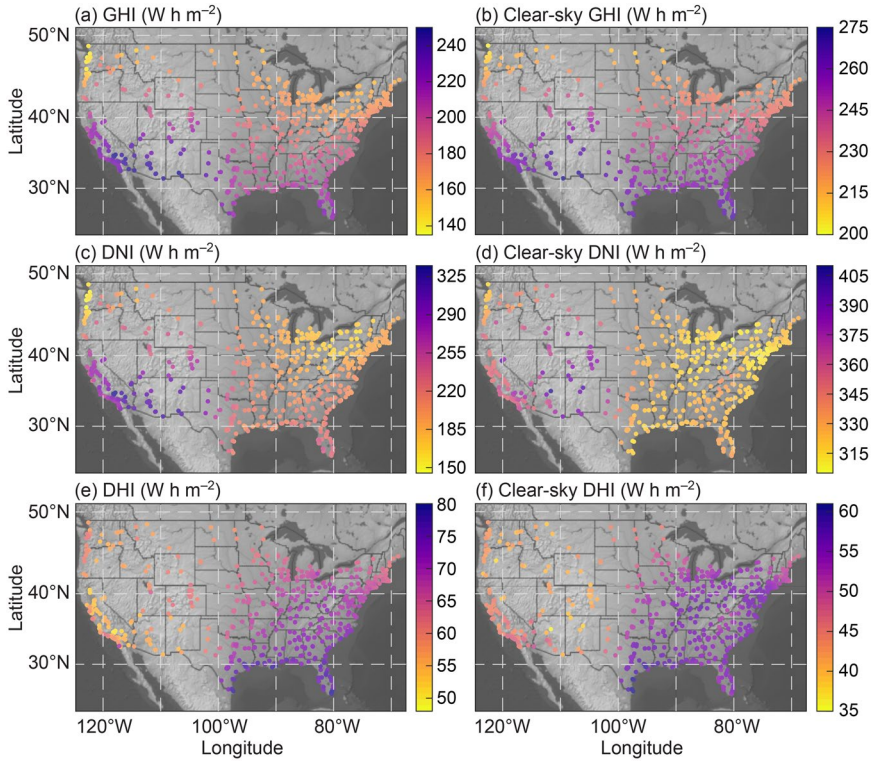

Fig. 6 Hourly (**a**) global horizontal irradiance (GHI), (**b**) clear-sky GHI, (**c**) direct normal irradiance (DNI), (**d**) clear-sky DNI, (**e**) diffuse horizontal irradiance (DHI), and (**f**) clear-sky DHI averaged over the 23-year period (1998–2020) for all stations in CHUWD-H v1.0. The unit is W h m^{-2}.

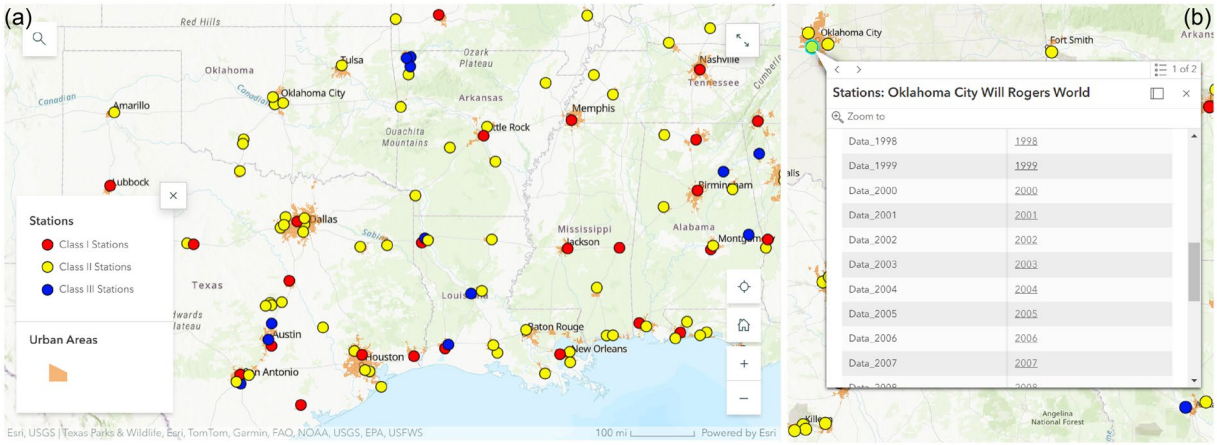

Fig. 7 Data inspection and downloading platform for CHUWD-H v1.0, featuring (a) interactive station inspection, and (b) downloading of individual annual hourly weather files.

effects on meteorological variables, we specifically corrected the pressure data for two stations where the elevation difference from the nearest MERRA-2 grid exceeds 100 m.

Quality control and gap filling methods. While the hourly irradiance variables from NSRDB are gap free, substantial gaps are prevalent in the raw observations from weather stations. Figures 2 and 3 illustrate the distributions of data gaps in time series of air temperature, dew point temperature, surface pressure, wind speed, and wind direction. Most data gaps occurred in the years prior to ~2005. On average, missing observations account for $5.46 \pm 9.81\%$ (mean $\pm$ 1 standard deviation or SD) for air temperature, $5.60 \pm 9.89\%$ for dew point temperature, $14.23 \pm 26.84\%$ for surface pressure, $5.72 \pm 9.70\%$ for wind speed, and $9.01 \pm 9.50\%$ for wind direction over the entire 23-year time series. No clear variations in data gaps were observed among different climate regions. While all stations recorded some measurements for air temperature, dew point temperature, wind speed, and wind direction over the 23-year period, there are 28 stations where surface pressure data were almost entirely missing (data gaps >99.99%).

Despite the internal quality controls conducted in the original data sources and ISD, we identified several potential erroneous data points (outliers) in the observations, which could introduce uncertainties into the subsequent gap filling process. Therefore, we carried out additional controls to remove these outliers based on the variations in raw data from both ISD and MERRA-2. Specifically, the acceptable upper and lower limits for temperature and pressure data were established using the mean $\pm$ 5 SD of ISD and MERRA-2 data and the range of MERRA-2 data over a 3-month moving window[41,42]. Any temperature or pressure data points falling outside these limits were considered outliers and subsequently removed. A similar procedure was applied to raw wind speed data, although a broader 10 SD threshold was used[43]. Following this quality control procedure, up to 0.09% of the ISD raw data were removed for individual stations. Additionally, all negative wind speed observations were removed.

We developed a multi-step gap filling (MSGF) approach to fill observational data gaps for all stations. This approach leverages data from both ISD and MERRA-2, aiming to retain as much observational data from ISD as possible. We constructed locally adaptive regression models that integrate available data from the target weather station, its nearest station, and the closest MERRA-2 grid[44–46]. For each missing data point, we first trained a multiple linear regression (MLR) model using available data at the target station and its nearest station as well as data from the nearest MERRA-2 grid within a specified moving window centered around the missing data point. The size of the moving window is 30 days (~ a month), 90 days (~ a season), 1 year, or 3 years for air temperature, dew point temperature, and wind speed, but 1 year or 3 years for pressure due to larger data gaps. The selection of the window size was determined by two key criteria: the statistical significance of the MLR model, which must achieve a p-value < 0.05 in a two-tailed t-test, and the number of non-missing observational data pairs from both stations within the window, which must be $\geq 10\%$. If the trained MLR model was statistically insignificant or if over 90% of observational data pairs were missing in the longest window (3 years), we switched to a simple linear regression (SLR) model trained using only the target station and its nearest MERRA-2 grid data, again selecting the moving window size based on statistical significance and data availability. If neither MLR nor SLR models sufficed, we filled the gap with data directly from the nearest MERRA-2 grid. To further enhance the robustness of the regression models, we excluded calm wind observations prior to constructing regression models. As an example, Figure 4 illustrates the MSGF approach for gap filling air temperature data.

Unlike air temperature, dew point temperature, surface pressure, and wind speed, missing wind direction data for each weather station were directly filled with data from the nearest MERRA-2 grid. Additional quality controls were applied to the gap filled data, such as adjusting dew point temperature higher than air temperature and correcting negative wind speed data. Furthermore, we derived hourly relative humidity from the gap filled air temperature and dew point temperature data[47].

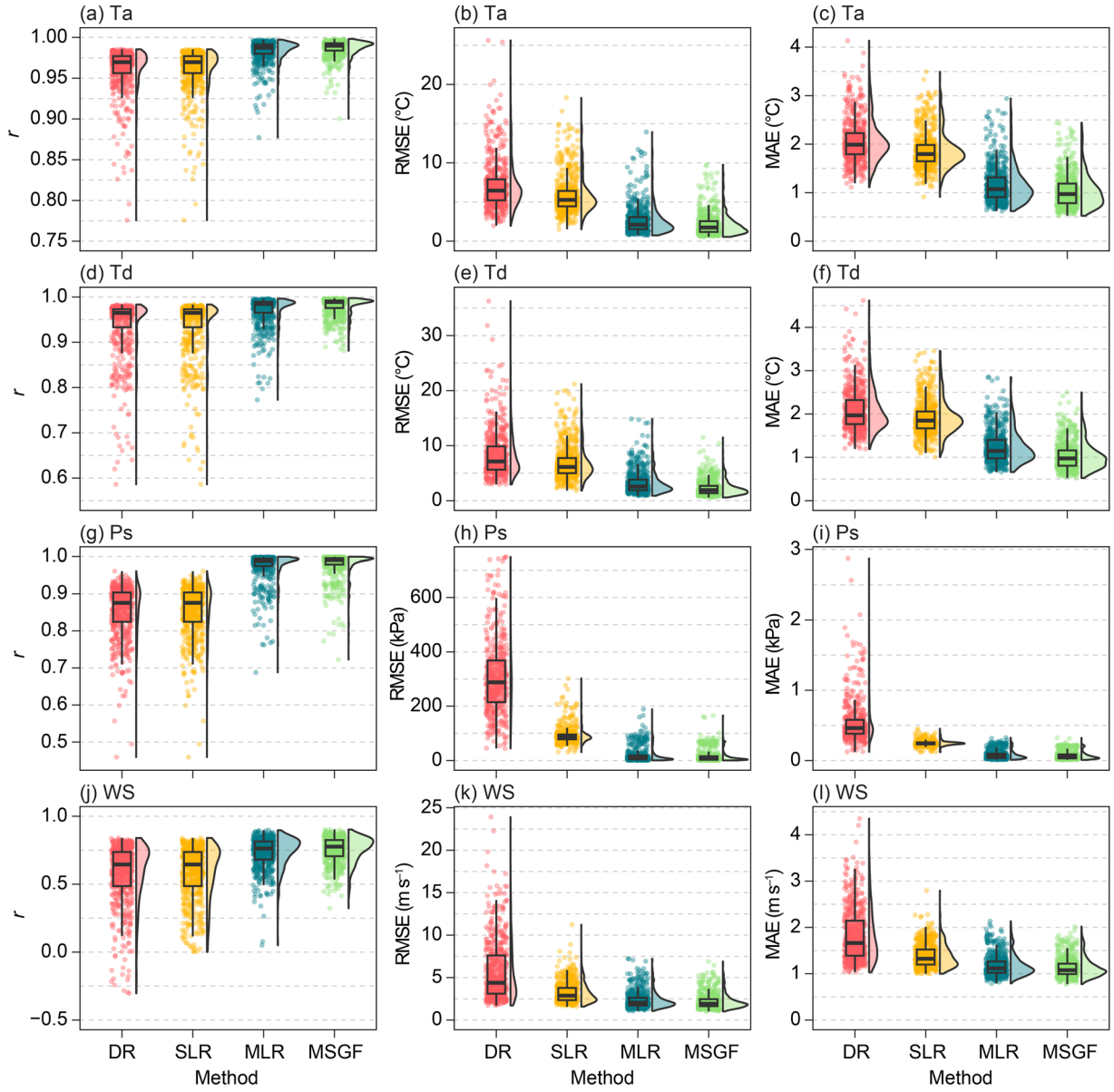

Fig. 8 Results of 10-fold Monte Carlo cross-validation, evaluated by the (**a,d,g**, and **j**) Pearson correlation coefficient (*r*), (**b,e,h**, and **k**) Root Mean Square Error (RMSE), and (**c,f,i**, and **l**) Mean Absolute Error (MAE), for (**a**–**c**) air temperature (Ta), (**d**–**f**) dew point temperature (Td), (**g**–**i**) surface pressure (Ps), and (**j**–**l**) wind speed (WS) using four gap filling methods: Direct Replacement (DR) with data from the nearest MERRA-2 grid, Simple Linear Regression (SLR) model constructed using 23-year data, Multiple Linear Regression (MLR) model constructed using 23-year data, and the multi-step gap filling (MSGF) approach proposed in this study. The box bounds the interquartile range divided by the median, with whiskers extending to ±1.5 times the interquartile range beyond the box. Circles are sample data points, and their distributions are represented by halved violin plots. Sample size *N* is 550 for air temperature, dew point temperature, and wind speed but 522 for surface pressure.

Data Records

The CHUWD-H v1.0 is publicly available through the Open Science Framework[31]. Hourly weather data from 1998 to 2020 for each station are organized into individual csv files for each year under individual project components, resulting in a total of 12,650 csv files following the naming convention "S******_year_Lat_***_Lon_***_State_Class.csv". For example, the file "S690150_2020_Lat_34.29_Lon_-116.15_CA_II" stores hourly data in 2020 for Station (S) ID 690150 (Twentynine Palms), a Class II station located in California (CA) at latitude 34.29° (34.29°N) and longitude −116.15° (116.15°W). Each csv file contains an annual time series of 8,760 hourly data points across 26 variables. These include five date and time variables, 16 meteorological variables, and five gap filling flag variables that indicate whether data points for air temperature, dew point temperature, pressure, wind speed, and wind direction were gap filled.

Variable	Correlation coefficient (r)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)
Air temperature (°C)	0.99 ± 0.01	2.14 ± 1.45	1.04 ± 0.34
Dew point temperature (°C)	0.98 ± 0.02	2.32 ± 1.45	1.04 ± 0.32
Surface pressure (kPa)	0.98 ± 0.04	15.99 ± 23.40	0.07 ± 0.06
Wind speed (m s^{-1})	0.76 ± 0.09	2.21 ± 0.92	1.13 ± 0.21

Table 3. Summary of cross-validation results (mean ± 1 standard deviation) for the multi-step gap filling (MSGF) approach. Note that units are for RMSE and MAE.

Note that CHUWD-H v1.0 also includes three auxiliary precipitation-related variables: total precipitable water from MERRA-2, and observed one-hour and six-hour liquid precipitation depths, sourced from ISD with existing data gaps. Due to the absence of reliable hourly precipitation sources for gap filling station-scale data, the current version does not include gap-free hourly precipitation data. However, these variables are included to support potential future applications such as model evaluation and the further development of CHUWD-H.

Details of all variables, their units, and definitions are summarized in Table 1. Information on all 550 weather stations in CHUWD-H v1.0 is provided in "CHUWD-H v1.0 stations.xlsx", accessible via the same data repository[31], with variables detailed in Table 2.

Figures 5 and 6 show major hourly variables averaged over the 23-year period (1998–2020) for 550 stations in CHUWD-H v1.0. To make this database more accessible to both researchers and the general public, we developed an interactive webpage[32] using ArcGIS Online (Fig. 7). This platform features clickable weather station locations, each linked to a pop-up box that provides detailed information about the weather station and download links for annual hourly weather data files.

Technical Validation

In addition to thorough quality controls on both raw data and gap filled data, we carried out a 10-fold Monte Carlo cross-validation (MCCV)[48,49] to evaluate the performance of the proposed MSGF approach against three commonly used gap filling methods. The alternative gap filling methods are: (1) Direct Replacement (DR), which fills missing data points with data from the nearest MERRA-2 grid; (2) an SLR model that uses the entire 23-year time series of data from both the target station and the nearest MERRA-2 grid; and (3) an MLR model that uses the entire 23-year time series of data from the target station, the nearest station (when data are non-missing), and the nearest MERRA-2 grid. Figure 8 shows the results of these four gap filling methods through 10-fold MCCV, evaluated by the Pearson correlation coefficient (r), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The MSGF approach consistently outperforms the other methods, achieving higher r values and much lower RMSE and MAE values. Specifically, the average MAE value is 1.04 °C for air temperature, 1.04 °C for dew point temperature, 0.07 kPa for surface pressure, and 1.13 m s^{-1} for wind speed (Table 3).

Usage Notes

The national coverage of CHUWD-H v1.0 will facilitate detailed hourly urban energy system modeling and enable cross-regional comparisons across cities under various realistic weather conditions, which will advance our understanding of how urban energy systems respond to extreme weather events and climate change. Originally developed for energy system modeling in urban areas, CHUWD-H v1.0 is also expected to support a wide range of applications in historical urban meteorological and climate studies, including the validation and evaluation of urban climate models. Moreover, it will serve as a valuable resource for future dates of CHUWD-H. Notably, parts of CHUWD-H v1.0 have been used in the first long-term, city-scale building energy consumption modeling across the entire U.S. to assess the impacts of climate change, population dynamics, and power sector decarbonization on urban building energy use[12]. Additionally, this database has supported analyses of casual interactions among U.S. cities during historical heat waves[50].

Code availability

The MATLAB code to perform the proposed multi-step gap filling approach for missing data points is available through the Open Science Framework[31].

Received: 22 June 2024; Accepted: 4 December 2024;
Published online: 18 December 2024

References

1. Kang, J.-N. *et al*. Energy systems for climate change mitigation: A systematic review. *Appl. Energy* **263**, 114602, https://doi.org/10.1016/j.apenergy.2020.114602 (2020).
2. DeAngelo, J. *et al*. Energy systems in scenarios at net-zero CO_2 emissions. *Nat. Commun.* **12**, 6096, https://doi.org/10.1038/s41467-021-26356-y (2021).
3. Nik, V. M., Perera, A. T. D. & Chen, D. Towards climate resilient urban energy systems: a review. *Natl. Sci. Rev.* **8**, nwaa134, https://doi.org/10.1093/nsr/nwaa134 (2021).
4. United Nations. World Urbanization Prospects: The 2018 Revision. Report No. ST/ESA/SER.A/420 (2019).
5. IEA. *Empowering Urban Energy Transitions* (International Energy Agency, 2024).
6. Craig, M. T. *et al*. Overcoming the disconnect between energy system and climate modeling. *Joule* **6**, 1405–1417, https://doi.org/10.1016/j.joule.2022.05.010 (2022).
7. Bloomfield, H. C. *et al*. The importance of weather and climate to energy systems: A workshop on next generation challenges in energy–climate modeling. *Bull. Am. Meteorol. Soc.* **102**, E159–E167, https://doi.org/10.1175/BAMS-D-20-0256.1 (2021).

8. Pfenninger, S., Hawkes, A. & Keirstead, J. Energy systems modeling for twenty-first century energy challenges. *Renew. Sustain. Energy Rev.* **33**, 74–86, https://doi.org/10.1016/j.rser.2014.02.003 (2014).
9. van der Wiel, K. *et al*. Meteorological conditions leading to extreme low variable renewable energy production and extreme high energy shortfall. *Renew. Sustain. Energy Rev.* **111**, 261–275, https://doi.org/10.1016/j.rser.2019.04.065 (2019).
10. Perera, A. T. D., Nik, V. M., Chen, D., Scartezzini, J.-L. & Hong, T. Quantifying the impacts of climate change and extreme climate events on energy systems. *Nat. Energy* **5**, 150–159, https://doi.org/10.1038/s41560-020-0558-0 (2020).
11. Berardi, U. & Jafarpur, P. Assessing the impact of climate change on building heating and cooling energy demand in Canada. *Renew. Sustain. Energy Rev.* **121**, 109681, https://doi.org/10.1016/j.rser.2019.109681 (2020).
12. Wang, C. *et al*. Impacts of climate change, population growth, and power sector decarbonization on urban building energy use. *Nat. Commun.* **14**, 6434, https://doi.org/10.1038/s41467-023-41458-5 (2023).
13. Al-Mofeez, I. A., Numan, M. Y., Alshaibani, K. A. & Al-Maziad, F. A. Review of typical vs. synthesized energy modeling weather files. *J. Renew. Sustain. Energy* **4**, 012702, https://doi.org/10.1063/1.3672191 (2012).
14. Chan, A. L. S. Generation of typical meteorological years using genetic algorithm for different energy systems. *Renew. Energy* **90**, 1–13, https://doi.org/10.1016/j.renene.2015.12.052 (2016).
15. Herrera, M. *et al*. A review of current and future weather data for building simulation. *Build. Serv. Eng. Res. Technol.* **38**, 602–627, https://doi.org/10.1177/0143624417705937 (2017).
16. Berrill, P., Wilson, E. J. H., Reyna, J. L., Fontanini, A. D. & Hertwich, E. G. Decarbonization pathways for the residential sector in the United States. *Nat. Clim. Change* **12**, 712–718, https://doi.org/10.1038/s41558-022-01429-y (2022).
17. Wei, W., Ramakrishnan, S., Needell, Z. A. & Trancik, J. E. Personal vehicle electrification and charging solutions for high-energy days. *Nat. Energy* **6**, 105–114, https://doi.org/10.1038/s41560-020-00752-y (2021).
18. Sweeney, J. F., Pate, M. B. & Choi, W. Life cycle production and costs of a residential solar hot water and grid-connected photovoltaic system in humid subtropical Texas. *J. Renew. Sustain. Energy* **8**, 053702, https://doi.org/10.1063/1.4963238 (2016).
19. Jentsch, M. F., James, P. A. B., Bourikas, L. & Bahaj, A. S. Transforming existing weather data for worldwide locations to enable energy and building performance simulation under future climates. *Renew. Energy* **55**, 514–524, https://doi.org/10.1016/j.renene.2012.12.049 (2013).
20. Shen, P. Impacts of climate change on U.S. building energy use by using downscaled hourly future weather data. *Energy Build* **134**, 61–70, https://doi.org/10.1016/j.enbuild.2016.09.028 (2017).
21. Bryce, R. *et al*. Consequences of neglecting the interannual variability of the solar resource: A case study of photovoltaic power among the Hawaiian Islands. *Sol. Energy* **167**, 61–75, https://doi.org/10.1016/j.solener.2018.03.085 (2018).
22. Hong, T., Chang, W.-K. & Lin, H.-W. A fresh look at weather impact on peak electricity demand and energy use of buildings using 30-year actual weather data. *Appl. Energy* **111**, 333–350, https://doi.org/10.1016/j.apenergy.2013.05.019 (2013).
23. Wilcox, S. & Marion, W. *Users Manual for TMY3 Data Sets*. Report No. NREL/TP-581-43156 https://doi.org/10.2172/928611 (National Renewable Energy Laboratory, 2008).
24. Li, H. *et al*. A new method of generating extreme building energy year and its application. *Energy* **278**, 128020, https://doi.org/10.1016/j.energy.2023.128020 (2023).
25. Doutreloup, S. *et al*. Historical and future weather data for dynamic building simulations in Belgium using the regional climate model MAR: typical and extreme meteorological year and heatwaves. *Earth Syst. Sci. Data* **14**, 3039–3051, https://doi.org/10.5194/essd-14-3039-2022 (2022).
26. Machard, A. *et al*. Typical and extreme weather datasets for studying the resilience of buildings to climate change and heatwaves. *Sci. Data* **11**, 531, https://doi.org/10.1038/s41597-024-03319-8 (2024).
27. White, P. R., Rhodes, J. D., Wilson, E. J. H. & Webber, M. E. Quantifying the impact of residential space heating electrification on the Texas electric grid. *Appl. Energy* **298**, 117113, https://doi.org/10.1016/j.apenergy.2021.117113 (2021).
28. Bloomfield, H. C., Brayshaw, D. J., Deakin, M. & Greenwood, D. Hourly historical and near-future weather and climate variables for energy system modelling. *Earth Syst. Sci. Data* **14**, 2749–2766, https://doi.org/10.5194/essd-14-2749-2022 (2022).
29. Zhao, L. *et al*. Global multi-model projections of local urban climates. *Nat. Clim. Change* **11**, 152–157, https://doi.org/10.1038/s41558-020-00958-8 (2021).
30. Zou, J. *et al*. Performance of air temperature from ERA5-Land reanalysis in coastal urban agglomeration of Southeast China. *Sci. Total Environ.* **828**, 154459, https://doi.org/10.1016/j.scitotenv.2022.154459 (2022).
31. Wang, C. & Deng, C. Historical Comprehensive Hourly Urban Weather Database (CHUWD-H) v1.0. *Open Science Framework (OSF)* https://doi.org/10.17605/OSF.IO/5DP8E (2024).
32. Liu, D., Deng, C. & Wang, C. *Historical Comprehensive Hourly Urban Weather Database v1.0* https://arcg.is/COWWe (2024).
33. U.S. Census Bureau. 2010 *TIGER/Line Shapefiles: Technical Document* (U.S. Census Bureau, 2012).
34. Climate.OneBuilding. Climate.OneBuilding.Org https://climate.onebuilding.org/default.html (2024).
35. Smith, A., Lott, N. & Vose, R. The Integrated Surface Database: Recent developments and partnerships. *Bull. Am. Meteorol. Soc.* **92**, 704–708, https://doi.org/10.1175/2011BAMS3015.1 (2011).
36. National Centers for Environmental Information. *Climate Data Online (CDO)* https://www.ncei.noaa.gov/cdo-web/ (2023).
37. Sengupta, M. *et al*. The National Solar Radiation Data Base (NSRDB). *Renew. Sustain. Energy Rev.* **89**, 51–60, https://doi.org/10.1016/j.rser.2018.03.003 (2018).
38. Xie, Y., Sengupta, M. & Dudhia, J. A Fast All-sky Radiation Model for Solar applications (FARMS): Algorithm and performance evaluation. *Sol. Energy* **135**, 435–445, https://doi.org/10.1016/j.solener.2016.06.003 (2016).
39. Habte, A., Sengupta, M. & Lopez, A. *Evaluation of the National Solar Radiation Database (NSRDB Version 2): 1998–2015*. Report No. NREL/TP-5D00-67722 https://doi.org/10.2172/1351859 (National Renewable Energy Laboratory, 2017).
40. Gelaro, R. *et al*. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1 (2017).
41. Osborn, T. J. *et al*. Land surface air temperature variations across the globe updated to 2019: The CRUTEM5 data set. *J. Geophys. Res. Atmospheres* **126**, e2019JD032352, https://doi.org/10.1029/2019JD032352 (2021).
42. Peterson, T. C., Vose, R., Schmoyer, R. & Razuvaëv, V. Global historical climatology network (GHCN) quality control of monthly temperature data. *Int. J. Climatol.* **18**, 1169–1179 https://doi.org/10.1002/(SICI)1097-0088(199809)18:11<1169::AID-JOC309>3.0.CO;2-U (1998).
43. Zona, D. *et al*. Characterization of the carbon fluxes of a vegetated drained lake basin chronosequence on the Alaskan Arctic Coastal Plain. *Glob. Change Biol.* **16**, 1870–1882, https://doi.org/10.1111/j.1365-2486.2009.02107.x (2010).
44. Luzio, M. D., Johnson, G. L., Daly, C., Eischeid, J. K. & Arnold, J. G. Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. *J. Appl. Meteorol. Climatol.* **47**, 475–497, https://doi.org/10.1175/2007JAMC1356.1 (2008).
45. Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S. & Lott, N. J. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteorol. Climatol.* **39**, 1580–1591 https://doi.org/10.1175/1520-0450(2000)039<1580:CASCND>2.0.CO;2 (2000).
46. Wang, C. & Wang, Z.-H. A statistical view ofs the Phoenix urban heat island during the past 86 years (1933–2018). (Central Arizona–Phoenix Long-Term Ecological Research 21st Annual All Scientists Meeting and Poster Symposium, 2019).
47. Shuttleworth, W. J. *Terrestrial Hydrometeorology*. https://doi.org/10.1002/9781119951933 (John Wiley & Sons, 2012).
48. Zhang, T., Zhou, Y., Wang, L., Zhao, K. & Zhu, Z. Estimating 1 km gridded daily air temperature using a spatially varying coefficient model with sign preservation. *Remote Sens. Environ.* **277**, 113072, https://doi.org/10.1016/j.rse.2022.113072 (2022).

49. Xiao, Q. *et al*. Evaluation of gap-filling approaches in satellite-based daily $PM_{2.5}$ prediction models. *Atmos. Environ.* **244**, 117921, https://doi.org/10.1016/j.atmosenv.2020.117921 (2021).
50. Yang, X., Wang, Z.-H., Wang, C. & Lai, Y.-C. Megacities are causal pacemakers of extreme heatwaves. *npj Urban Sustain* **4**, 8, https://doi.org/10.1038/s42949-024-00148-x (2024).
51. Karl, T. R. & Koss, W. J. *Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, 1895-1983* (National Climatic Data Center, 1984).

Acknowledgements

This research was supported by the U.S. National Science Foundation under grant Nos. OIA-2327435 and CNS-2301858, the National Oceanic and Atmospheric Administration under grant No. NA21OAR4590361, and the Seed Funding Grant funded by the Data Institute for Societal Challenges (DISC) at the University of Oklahoma. Financial support for publication was provided by the University of Oklahoma Libraries' Open Access Fund. R.R.C. and S.F. acknowledge support of ANID ACT210046. R.B.J. acknowledges support from the Stanford Natural Gas Initiative. We would like to acknowledge Carlo Bianchi (National Renewable Energy Laboratory, USA), Linda Lawrie (Climate.OneBuilding.Org), and Dru Crawley (Climate.OneBuilding.Org) for their help with weather data processing. This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding was provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Author contributions

Conceptualization: C.W.; Methodology: C.W., H.H. and J.L.R.; Data Curation: C.W., H.H., J.L.R. and J.S.; Formal Analysis: C.W.; Software: C.W., C.D. and D.L.; Validation: C.W., S.F., and R.R.C.; Visualization: C.W., C.D., D.L., S.F., R.R.C., and R.B.J.; Funding Acquisition: C.W.; Resources: C.W., C.D., and R.B.J.; Writing – original draft: C.W. and C.D.; Writing – review & editing: H.H., J.L.R., D.L., S.F., R.R.C., J.S. and R.B.J.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.