



OPEN

DATA DESCRIPTOR

High temporal-resolution microclimate and photovoltaic power dataset for sustainable rooftop design in subtropical humid city

Ip Cheng¹, Liutao Chen¹, Haoran Chang², Zinan Lin¹, Mengying Li²✉ & Zhe Wang¹✉

This paper presents a high resolution microclimate monitoring dataset designed to explore rooftop microclimate responses to different rooftop configurations within a subtropical humid campus setting. The dataset comprises measurements from seven rooftops representing five roofing types: bare roof, bare roof with HVAC units, green roof, PV roof, and PV integrated green roof (PVIGR). The dataset is based on data collected over a one-and-a-half-year period on a university campus in Hong Kong. Microclimate data were collected using various sensors, including temperature and relative-humidity sensors, anemometers, and pyranometers. In addition, the dataset also provides PV power generation data, which supports studies of PV performance and the microclimate impacts of PVIGR designs. Lastly, weather data from a nearby weather station were also included in this dataset for future reference. This dataset addresses a significant gap in existing literature that primarily focuses on conventional roof types and supports the evaluation of alternative rooftop designs under comparable campus-scale conditions. This paper provides a comprehensive description of the rooftop sites, data collection methods, records, and validation processes. Potential applications encompass analyzing microclimate characteristics, supporting validation or benchmarking of computational fluid dynamics models under comparable rooftop conditions, optimizing photovoltaic conversion efficiency, and informing sustainable rooftop design studies.

Background & Summary

The urban heat island effect (UHI) has been documented in more than 400 major cities worldwide¹, with reported average temperature increases exceeding 4–5 °C and peak differences above 10 °C². UHI is associated with urban material properties, anthropogenic heat emissions, canyon geometry, reduced evaporative surfaces, and limited turbulent exchange in dense areas³. Elevated urban temperatures can increase cooling demand^{4,5}, peak electricity loads, local climate vulnerability, and heat-related morbidity and mortality⁶.

Buildings account for 20–40% of global energy consumption⁷ and related greenhouse gas emissions⁸. Rooftops can support several energy and climate related functions, including green roofs⁹, rainwater harvesting systems¹⁰, solar energy collectors¹¹, insulation layers¹², and cool roofs¹³. These systems can influence building energy use, renewable energy generation, usable rooftop space, and near surface microclimate conditions^{7,14}.

Existing open-source datasets on rooftop environments and urban microclimates generally fall into two categories. The first consists of site-specific monitoring datasets^{15,16}, such as rooftop weather observations under HVAC scenarios in Tennessee, USA¹⁵ and high-frequency thermal infrared imagery from two rooftops at the National University of Singapore¹⁶. The second consists of urban-scale meteorological networks^{17–19}, including

¹State Key Laboratory of Climate Resilience for Coastal Cities (SKLCC), Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region, Hong Kong, China. ²Department of Mechanical Engineering & Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, Hong Kong, China. ✉e-mail: mengying.li@polyu.edu.hk; cezhewang@ust.hk

rooftop or citywide monitoring datasets from Birmingham¹⁷, the UK¹⁸, and Sydney¹⁹, as well as meteorological data from UK monitoring stations²⁰. These datasets provide valuable meteorological and thermal observations, but most do not compare multiple rooftop configurations within the same monitored setting.

Current open-source rooftop meteorological datasets are mainly based on conventional roof types or urban-scale monitoring networks, leaving limited data for comparing alternative rooftop configurations under consistent site and climate conditions. To address this limitation, this study provides a high-temporal-resolution microclimate monitoring dataset from seven rooftops representing five configurations: bare roof (BR), BR with HVAC units, green roof (GR), PV roof (PVR), and PV integrated green roof (PVIGR). The measurements were collected on a university campus in Hong Kong and include rooftop microclimate data, PV power generation data for PVR and PVIGR systems, and weather station data for cross-reference. The PV power-generation and weather-station components extend a previously released campus PV and weather dataset²¹ by adding measurements across multiple rooftop configurations rather than focusing only on photovoltaic roofs.

The dataset has three main features:

- It combines high-temporal-resolution microclimate, PV power-generation, and weather data from seven rooftops representing five rooftop types over a one-and-a-half-year monitoring period.
 - The rooftops include temperature and relative-humidity sensors, anemometers, pyranometers, and soil moisture sensors positioned according to each rooftop configuration.
 - The PVIGR sites include three PV tilt angles (3°, 12°, and 22°), three module heights above ground (0.6 m, 0.75 m, and 0.9 m), and three vegetation types (*Wedelia trilobata*, *Sedum lineare*, and *Zoysia*), resulting in 12 configurations for comparing PVIGR design parameters under the monitored campus conditions.
- The potential applications for the dataset include, but are not limited to, the following:
- Analyzing the effects of different rooftop configurations on rooftop microclimate conditions^{16,22–24}.
 - Providing a high-temporal-resolution case-study dataset for benchmarking PV conversion-efficiency optimization²⁵, power prediction, and fault detection diagnosis²⁶.
 - Assisting in the development, benchmarking, or validation of Computational Fluid Dynamics (CFD) models^{27–29} and urban meteorological simulations, including the Weather Research and Forecasting (WRF) model, under comparable rooftop and climatic settings^{30,31}.
 - Informing sustainable rooftop design studies, climate adaptation assessments, and future urban sustainability research under comparable subtropical rooftop conditions^{32–35}.
 - Supporting studies on Internet of Things (IoT) communication and instrument selection for rooftop microclimate monitoring stations³⁶ and building energy load and PV generation estimation³⁷.

The spatial representativeness of this dataset should be interpreted considering the monitoring design. All rooftop measurements were collected within a single university campus in Hong Kong, providing a consistent subtropical humid coastal climate context for comparing multiple rooftop configurations. However, the monitored rooftops reflect the specific spatial, morphological, operational, and exposure characteristics of this campus, including building elevations, rooftop materials, surrounding terrain, coastal setting, and institutional land use. Therefore, the dataset should not be interpreted as fully representative of all rooftop environments or urban morphologies in subtropical humid cities. Instead, it provides a high-temporal-resolution case-study dataset for evaluating rooftop microclimate and PV performance responses under a campus-scale setting. Users should consider local morphology, building density, surface materials, shading, ventilation, and climatic exposure when transferring findings to other urban contexts.

To improve usability for different analytical needs, the dataset is organized into raw and processed versions with documented quality-control, down sampling, interpolation, and extraction settings. Users can therefore select the most appropriate version for their objectives: the baseline records preserve the original temporal structure for event-scale or sensor-level analyses; the down sampled records provide regular hourly data for model comparison and statistical assessment; and the downsampled + interpolated records support visualization, exploratory analysis, and workflows requiring continuous hourly inputs. The accompanying metadata, naming conventions, missing-rate summaries, and configurable curation scripts are intended to make the dataset transparent, reproducible, and adaptable for applications such as rooftop microclimate comparison, PV-performance assessment, model validation, and sustainable rooftop-design evaluation³⁸.

Methods

Sites Description. The dataset was collected from seven building rooftops and an automatic weather station on the campus of the Hong Kong University of Science and Technology. The university is located in the rural coastal area of Sai Kung District, Hong Kong (22.3363°N 114.2634°E), covers approximately 60 hectares, and is situated in a subtropical climate with average humidity levels exceeding 75% and temperatures ranging from about 10°C in winter to over 30°C in summer³⁹. The monitoring project, initiated in August 2023, compares rooftop microclimate conditions across different roof configurations, quantifies selected PVIGR design parameters, and supports numerical model evaluation under the monitored campus conditions. Figure 1 shows the rooftop locations and configurations, and Table 1 summarizes their surface materials, vegetation, mounted equipment, building type, elevation, and data-transmission method.

Please note that by mid-November 2024, *Wedelia trilobata* showed signs of mortality due to a malfunction in the irrigation system that lasted several days. The horticulture team removed it and replaced it with *Sedum lineare* on December 1, 2024.

Bare roof (BR). #1 BR is located on top of the campus library. It has a light-colored ceramic tile surface, no rooftop equipment, and limited routine access, providing a reference site without mounted systems or vegetation.

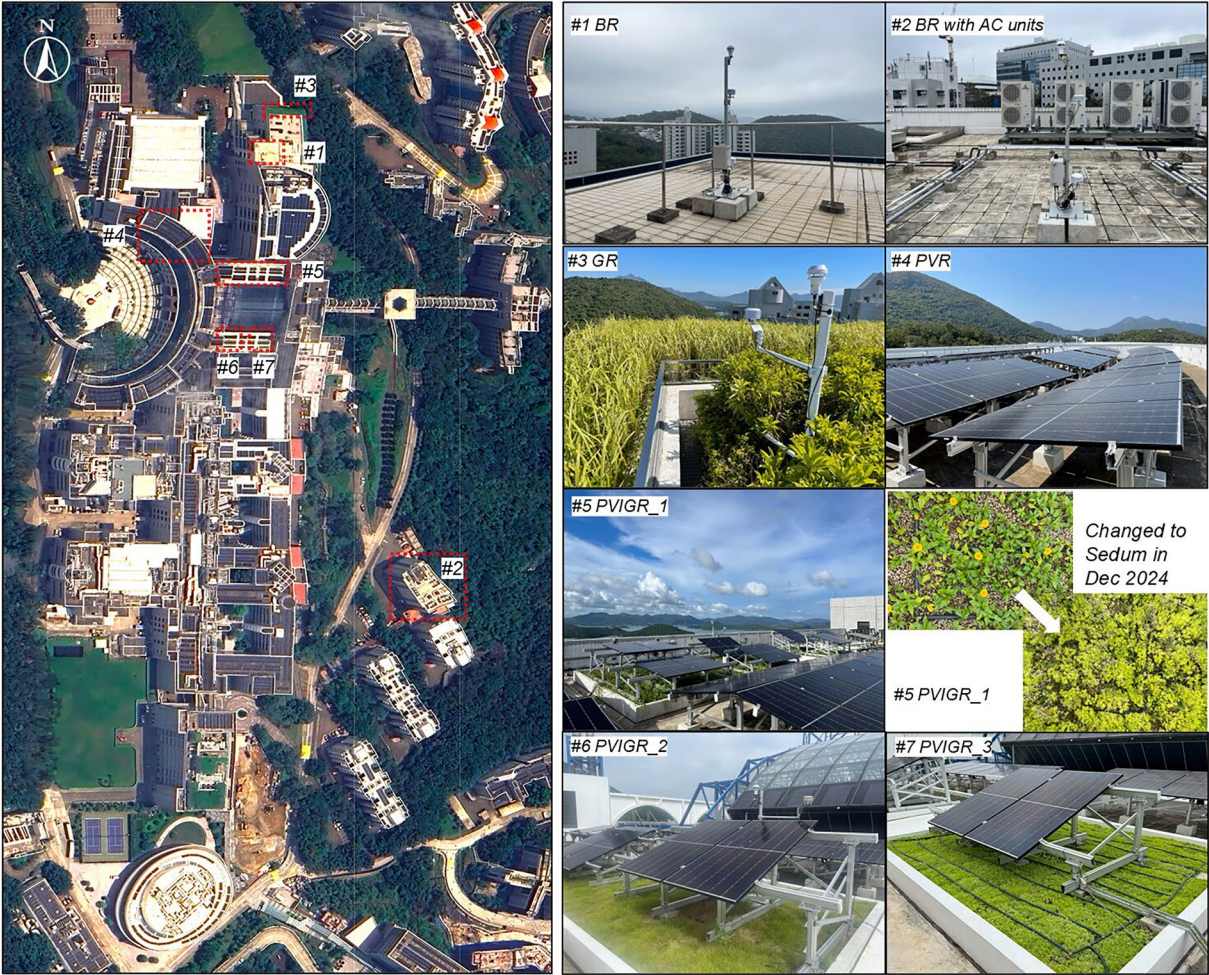


Fig. 1 Overview of rooftop numbering, locations, and configurations across the campus.

Rooftop ID	Rooftop Type	Surface Materials	Vegetation	Mounted Equipment	Building Type and Elevation	Method of Data Transmission
#1 BR	Bare roof	Light-colored ceramic tile	/	/	Library (97 m)	4G
#2 BR with HVAC units	BR with HVAC units	Concrete	/	AC outdoor units	Residence hall (107 m)	4G
#3 GR	Green roof	Soil	Cogongrass	/	Library (93 m)	LoRaWAN
#4 PVR	PV roof	Concrete	/	PV modules	Academic building (101 m)	LoRaWAN
#5 PVIGR_1	PV integrated green roof	Pumice stone	Wedelia trilobata/ Sedum lineare	PV modules	Academic building (101 m)	4G
#6 PVIGR_2	PV integrated green roof	Pumice stone	Sedum lineare	PV modules	Academic building (101 m)	4G
#7 PVIGR_3	PV integrated green roof	Pumice stone	Zoysia	PV modules	Academic building (101 m)	4G

Table 1. Overview of rooftop configurations and characteristics.

BR with HVAC units. #2 BR with HVAC units is located on Tower C of the faculty apartment dormitory. The rooftop has a concrete surface and contains outdoor air conditioning units serving the building, which may affect rooftop thermal conditions and local airflow through waste heat release and equipment induced obstruction.

Green roof (GR). #3 GR is an outdoor garden located on the ground floor of the library. It is planted with Cogongrass and is watered every morning by maintenance staff. The vegetation and substrate may contribute to local microclimate modification through shading, evapotranspiration, and surface moisture retention.

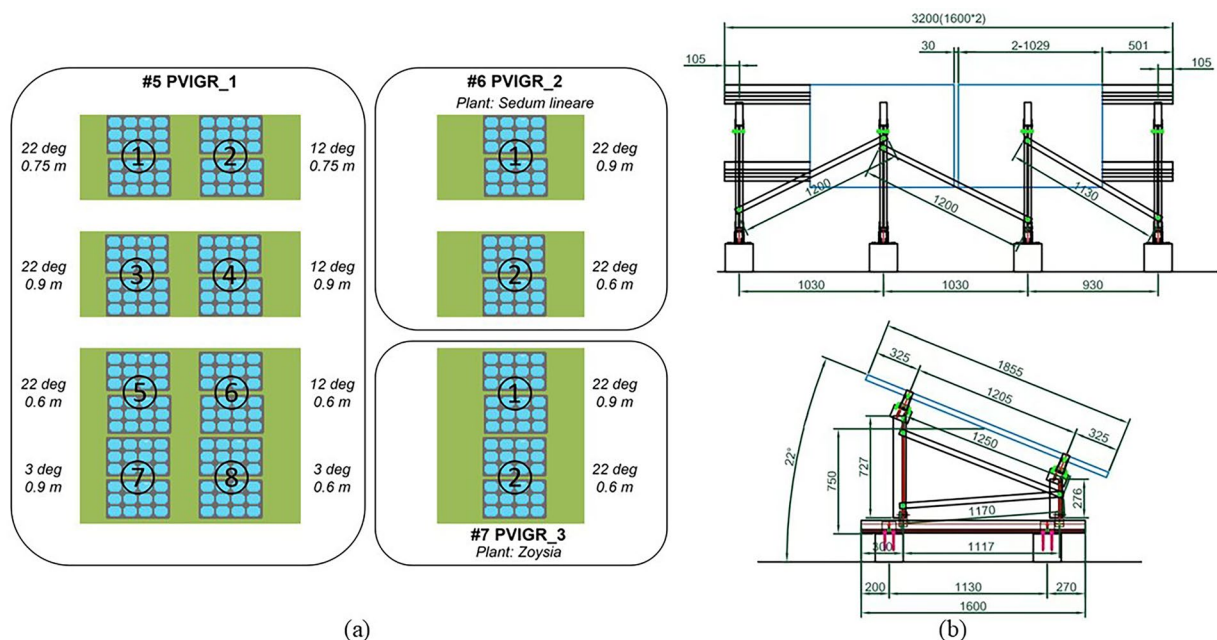


Fig. 2 (a) Configurations and identifiers of PVIGR systems for rooftops #5–7 and (b) PV framework design example drawing for #1 PV module in #5 PVIGR_1, which has the tilt angle of 22° and separation height of 0.75 m.

PV roof (PVR). #4 PVR is a concrete rooftop equipped with grid-connected photovoltaic systems. The PV modules cover approximately 50% of the rooftop area, and electricity generated by the system is used by the building, with surplus power fed into the local grid. The system includes 66 silicon monocrystalline PV modules, each measuring 1855 mm × 1029 mm × 30 mm and rated at 415 W. Each module is anchored by concrete plinths measuring 300 mm × 300 mm × 124 mm and weighing 20 kg, connected to stainless steel brackets. The monitored PV module is installed at a 10° tilt angle, oriented due southwest and northeast, with its lowest edge 20 cm above the rooftop surface.

PV integrated green roof (PVIGR). Rooftops #5 to #7 are PVIGR experimental sites on an academic building. The setup was designed to quantify how PV tilt angle, module height, and vegetation type influence PVIGR performance and rooftop microclimate conditions. Across the three sites, 24 silicon monocrystalline PV modules rated at 415 W were installed facing due south. The experimental variables include three tilt angles (3°, 12°, and 22°), three module heights above ground (0.6 m, 0.75 m, and 0.9 m), and three ground-cover plants: *Wedelia trilobata*, *Sedum lineare*, and *Zoysia*. These variables result in 12 combinations. Each pair of PV modules representing one configuration is connected to an optimizer that records power output and related electrical parameters. The identifiers and configurations of the PVIGR systems are shown in Fig. 2.

For all PVIGR setups, pumice stone was used as the substrate (Fig. 3d) because it retains water and nutrients while remaining relatively lightweight, with a dry density of 420 kg/m³ and a wet density of 620 kg/m³. The substrate also limits weed and insect presence, contains minimal soil content, and reduces soil erosion and settling. Each PV module is secured with four concrete plinths measuring 300 mm × 300 mm × 124 mm and weighing 20 kg (Fig. 3e), anchored in the substrate and connected to stainless steel brackets. Irrigation is managed by an automated system (Fig. 3a) with a control panel (Fig. 3c), and drip irrigation pipes are placed on the soil surface (Fig. 3b) to distribute water across the planted areas.

Automated Weather Station. The automatic weather station is located southeast of the campus and records wind speed, wind direction, air temperature, relative humidity, air pressure, visibility, and Global Horizontal Irradiance (GHI) hourly. Figure 4 shows the weather station's location and equipment.

Sensor Specifications. The dataset includes four microclimate monitoring sensor types, five weather monitoring sensor types, and optimizer-level PV power monitoring. Table 2 summarizes the measured variables, ranges, resolutions, and precision of these instruments.

Figure 5 shows the microclimate sensor configurations used on the rooftops. Each sensor tree consists of a 1.7 m metal pole anchored to a 300 mm × 300 mm × 250 mm concrete plinth weighing 30 kg, with attachment points for the data logger and sensors. Ultrasonic anemometers are installed 2 m above the rooftop surface and, where applicable, between the PV modules and rooftop surface to measure wind speed and wind direction. Temperature and relative humidity sensors are housed in white radiation shields and positioned 0.5 m and 1.8 m above the rooftop surface. Two south-facing plane-of-array irradiance (POA) pyranometers are mounted at the edges of PV modules tilted at 3° and 22°. At the PVIGR sites, three soil moisture sensors are installed to measure



Fig. 3 (a) Automated integration system, (b) Drip irrigation pipes placed on soil, (c) Control panel of the irrigation system, (d) Concrete plinths and (e) Connected PV mounting frame.

moisture conditions for the three vegetation types. All sensors are connected to data loggers and wired to waterproof sockets on rooftop walls, with power supplied by the local 220 V electrical system.

Data Transmission. Data transmission methods differ by data category and rooftop site.

Among the seven rooftops, five use 4 G wireless transmission and two use LoRaWAN, as listed in Table 1. For the 4 G sites, meteorological sensors transmit data to the data logger via RS-485 communication. The data logger is connected to a 4 G communication module with a SIM card, which uploads data to a cloud database through the 4 G network. Users can download the data from a web interface that also displays the connection status of the 4 G module. For the LoRaWAN sites, the connection between the meteorological sensors and data logger is the same, but the data logger output is connected to a LoRaWAN module via RS-485. The module connects to nearby LoRaWAN gateways, which relay the data to the database through wired connections. Users can download the data through an Application Programming Interface (API).

Each pair of PV modules is equipped with a module optimizer that records optimizer-level power generation data and applies maximum power point tracking (MPPT). The data are processed by the corresponding inverter and uploaded to SolarEdge Monitoring via Wi-Fi for online access.

Weather-station data are stored in a local data logger and transmitted to a centralized database via RS-232 protocol.



Fig. 4 (a) The location of the automatic weather station, (b) Equipment used at the automatic weather station⁴⁰.

Category	Sensor	Data Type	Range	Resolution	Precision
Microclimate	Anemometer	Wind speed and direction	0–70 m/s; 0–359°	0.1 m/s; 1°	±3%; ±3°
	Temperature and relative-humidity sensor	Air temperature and relative humidity	–40–60 °C; 0–100%	0.1 °C; 0.5%	±0.5 °C; ±3%
	Soil moisture sensor	Soil moisture	0–100%	/	±2%
	Pyranometer	Plane-of-array irradiance (POA)	0–1600 W/m ²	/	<7 W/m ²
Weather ¹⁹	Anemometer	Wind speed and direction	0–70 m/s; 0–359°	0.1 m/s; 1°	0 to 30 m/s: ±2% 30 to 70 m/s: ±3%
	Temperature and relative-humidity sensor	Air temperature and relative humidity	–40–60 °C; 0–100%	/	(Only precision for relative humidity is available) At 0–40 °C and 0–90% RH: ±2% At 0–40 °C and 90–100% RH: ±5% At 40–60 °C and 0–90% RH: ±5% At 40–60 °C and 90–100% RH: ±7%
	Barometer	Pressure	500–1100 hPa	/	/
	Visibility Sensor	Visibility	6 m to 80 km	/	/
	Pyranometer	Global Horizontal Irradiance (GHI)	/	/	±3% to ±5%
Power	Optimizer	PV power generation	0–1000 W	/	/

Table 2. Specifications of microclimate and weather monitoring sensors.

Data Curation. Data curation was implemented using Python and Jupyter Notebook. Modules were developed for each stage, and the overall pipeline can be executed through a Jupyter Notebook file with configurable parameters to generate dataset versions for different analysis requirements. The five core data curation stages are merge, quality control, down sampling, interpolation, and extraction. Figure 6 illustrates the workflow, and Table 3 summarizes each stage. Missing rate analysis is provided as an additional validation and diagnostic step to help users evaluate data completeness across dataset versions.

Stage 1: Merge. The raw microclimate, weather and power data from dataloggers and databases were downloaded in multiple time ranges due to limitations of the platforms, which need to be merged after collection. Each datalogger data file after this stage will contain the data that covers a whole year. Special functions were developed in Python to merge each datalogger’s data from multiple files into one file for further process.

Stage 2: Quality control. Removing invalid data from a dataset is critical to its application by ensuring validity. A three-step procedure was adopted for quality control. First, the procedure checks whether the standard deviation of all values within a user-defined time window (e.g., one day) is lower than the user-defined tolerance (e.g., standard deviation under 10^{-6}). This ensures that the values did not remain constant throughout the entire window, which is highly unlikely in real-life scenarios. If such a case is detected, the procedure sets all values within this window to NaN (Not a Number), indicating invalid values. Second, the procedure sets all values within known sensor failure periods for a specific data logger to NaN. Third, the procedure validates whether the values are within physical boundaries and sets the values to NaN if they are not within the boundaries. It is worth

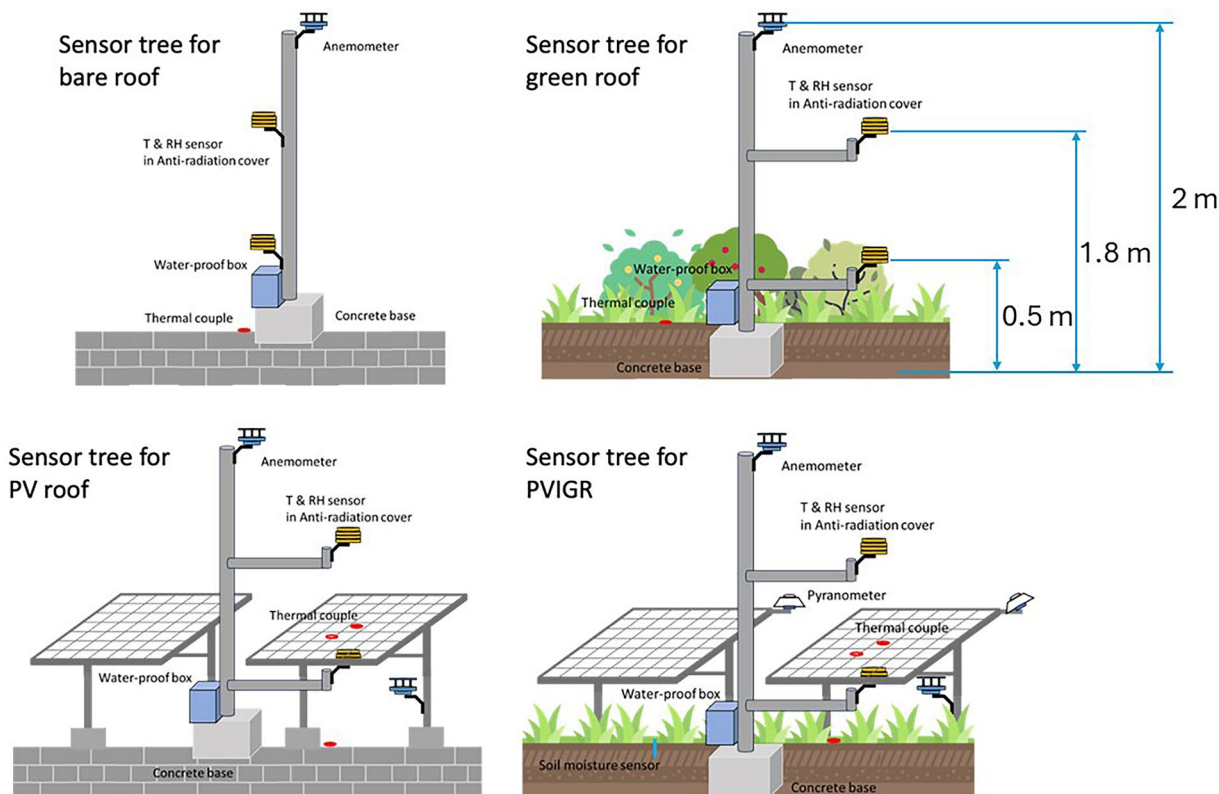


Fig. 5 Illustration of sensors configuration for different rooftops.

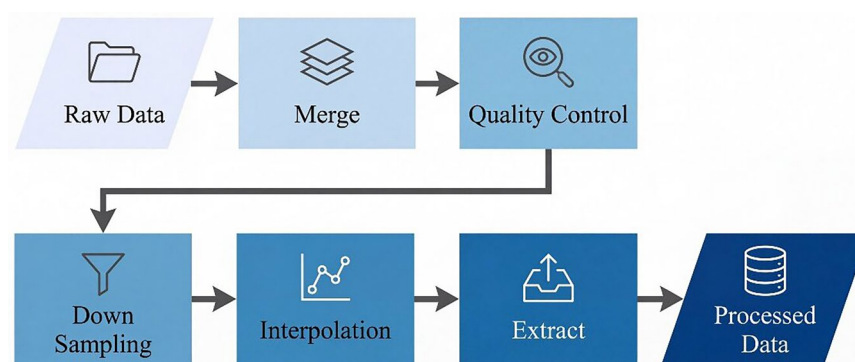


Fig. 6 The overall workflow of the data curation.

Stage	Purpose
Stage 1: Merge	Merge multiple files from different time ranges into one file
Stage 2: Quality Control	Filtering out invalid data from the dataset
Stage 3: Down sampling	Down sample the dataset with a specific sampling rate
Stage 4: Interpolation	Linear interpolation for the dataset on missing values
Stage 5: Extraction	Extract and organize the data into category-specific files

Table 3. Description of each stage in the data curation.

noting that irradiance and PV power generation are expected to be non-zero during daytime (i.e., between 7:00 and 18:00), as some solar radiation exists even under cloud shading. Hence, the lower boundaries for irradiance and PV power generation during daytime were set to 10 W/m^2 and 1 W , respectively. All steps of this procedure were implemented as individual functions and can be reconfigured in the related JSON files mentioned in Table 4.

Stage	Related Jupyter Notebook	Related Python Module	Related Configuration Files
Stage 1: Merge	01_merge_raw_data.ipynb	merge_data.py	N/A
Stage 2: Quality Control	02_quality_control.ipynb	quality_control.py	constant_value_detection.json sensor_failures.json qc_boundaries.json
Stage 3: Down sampling	03_downsampling.ipynb	downsampling.py	downsampling_config.json
Stage 4: Interpolation	04_interpolation.ipynb	interpolation.py	interpolation_config.json
Stage 5: Extraction	05_extract_data.ipynb	extract_data.py	extraction_config.json
Stage 6: Missing Rate Analysis	06_missing_rate_analysis.ipynb	missing_rate_analysis.py	N/A

Table 4. Description of related files for each stage of the data curation.

The QC procedure was designed as a transparent and reproducible first-pass screening process rather than a comprehensive anomaly-detection framework for all possible environmental-data artefacts. The near-zero-variance criterion is intended to identify possible sensor or data-logger freezing, repeated transmission values, or communication-related errors over unusually long user-defined time windows. However, low variance may also occur under valid environmental conditions, such as stable nocturnal periods, weak wind conditions, sustained overcast weather, or periods with limited solar radiation. Therefore, this criterion should be interpreted as a conservative screening rule, and both the time-window length and variance tolerance can be adjusted by users through the configuration files. Similarly, the physical-boundary checks are intended to remove values outside the sensor-supported range or locally plausible limits, but they cannot detect all anomalous yet physically plausible records, such as gradual sensor drift, calibration bias, local shading effects, partial sensor obstruction, or short-term microclimatic disturbances. Known sensor-failure periods were removed based on documented field observations and maintenance records. These QC procedures may influence downstream applications, particularly model validation, PV performance assessment, and fault-detection studies, where retaining anomalous values or removing valid stable periods could affect calculated performance metrics. For this reason, the raw and processed dataset versions are both provided, allowing users to apply additional application-specific QC criteria, compare results across dataset versions, and assess the sensitivity of their analyses to the selected QC settings. Please refer to Table 5 for the summary of each QC step.

The known sensor failure periods were listed in Table 6, and the physical boundaries for each sensor type were listed in Table 7.

Stage 3: Down sampling. Down sampling is a stage aimed at reducing data missing rate by averaging all valid values within a longer sampling period, to ensure the dataset's integrity and consistency. The default down sampling rate is one hour and can be user-defined in the configuration JSON file mentioned in Table 4. This stage is optional and can be skipped in the overall pipeline.

Stage 4: Interpolation. Interpolation is a process designed to fill missing data within a reasonable data gap, particularly when gaps in the data persist after down sampling. The current user-defined setting for the longest duration of missing data is 60 minutes, and only applicable to temperature, relative humidity and soil moisture by default. The setting prevents interpolation for gaps that are too long, as such gaps may result in inaccurate interpolation, and limits its application to variables that do not exhibit large fluctuations within a short period of time. These restrictions can be adjusted in the configuration JSON file mentioned in Table 4. Additionally, this stage is also optional and only executable if Stage 3: Down sampling is enabled. This ensures that the dataset is interpolated with a certain degree of consistency. Please refer to Table 9 regarding the setting for all sensor types for this dataset.

The interpolated dataset is provided as an optional version rather than the default dataset for all applications. Although interpolation can improve temporal continuity and reduce short gaps after down sampling, it may also alter the temporal dynamics of high-frequency microclimate data. In particular, linear interpolation can smooth short-term fluctuations, reduce local extremes, and obscure abrupt transitions caused by cloud cover, shading, rainfall, irrigation, wind gusts, HVAC operation, or rapid changes in rooftop surface conditions. For this reason, interpolation is limited by default to gaps of no more than 60 minutes and is enabled only for variables that are expected to vary relatively smoothly over short periods, namely air temperature, relative humidity, and soil moisture. Users should prefer the baseline or down sampled non-interpolated datasets when temporal integrity is critical, such as for CFD or WRF model validation, event-scale analysis, wind-field assessment, irradiance and PV-power analysis, and studies focusing on extremes or transient processes. The downsampled + interpolated dataset is more suitable for exploratory analysis, visualization, long-term trend assessment, and workflows requiring continuous hourly input, provided that users account for the potential smoothing effects introduced by interpolation. Please refer to Table 8 for the recommended use of dataset versions in relation to interpolation and temporal data integrity.

Angular interpolation was used for wind direction because wind direction is a circular variable rather than a linear one. In a circular coordinate system, 0° and 360° represent the same direction, so conventional linear interpolation can produce misleading values when the angle crosses this boundary. For example, if wind direction changes from 359° to 0° over a short interval, direct linear interpolation would pass through values near 180°, incorrectly implying a reversal in wind direction. In contrast, angular interpolation treats each direction as

QC step	Criterion applied	Rationale	Limitations	Downstream implications
Near-zero-variance detection	Values within a user-defined time window are flagged when their standard deviation is below a user-defined tolerance.	Designed to identify possible sensor freezing, data-logger freezing, repeated transmission values, or communication-related errors over extended periods.	Low variance may also occur under valid environmental conditions, including stable nocturnal periods, weak wind, sustained overcast weather, or low-radiation periods.	May remove valid stable periods if parameters are too strict; users should adjust the window and tolerance for application-specific studies.
Known sensor-failure removal	Values recorded during documented sensor-failure periods are set to NaN.	Uses field observations, maintenance records, and documented malfunction periods to remove records known to be unreliable.	Only documented failures can be removed; undocumented drift, partial obstruction, or intermittent faults may remain.	Improves reliability for affected periods but may leave some physically plausible anomalous values in the dataset.
Physical-boundary filtering	Values outside sensor-supported or locally plausible boundaries are set to NaN.	Provides conservative sanity checks using sensor specifications and dataset-specific thresholds appropriate for the monitored rooftop environment.	Cannot detect all anomalous but physically plausible values, such as sensor drift, calibration bias, shading effects, or partial obstruction.	Users conducting model validation, PV performance analysis, or fault detection should consider additional application-specific QC and sensitivity checks.

Table 5. Summary of quality-control criteria, rationale, limitations, and implications for downstream applications.

Site	Sensors Affected	Date Range
#1 BR	Temperature and relative-humidity sensor, and anemometer	July 31, 2024 – February 10, 2025
#2 BR with HVAC units	Temperature and relative-humidity sensor, and anemometer	July 28, 2024 – October 16, 2024

Table 6. Known sensor failure periods.

Sensor Type	Physical Boundary	Justification	Fixed or user defined
Air temperature	0–60 °C	Lower limit: it is rare to have air temperature lower than 0 °C in Hong Kong Upper limit: Hong Kong’s air temperature rarely exceeds 40 °C, setting the upper limit to 60 °C would provide a sufficient buffer for error	User defined
Relative humidity	0–100%	Definition of relative humidity	Fixed
Wind speed	0–70 m/s	Defined by sensor manufacturers	Fixed
Wind direction	0–360°	Defined by sensor manufacturers	Fixed
Irradiance	0–1600 W/m ²	Defined by sensor manufacturers	Fixed
Soil moisture	0–100%	Definition of soil moisture	Fixed
Power	0–1000 W	The rated power for a PV module used on all PV roofs is 415 W, and an optimizer is connected to two modules. Hence, the total rated power for an optimizer is 830 W. 1000 W was selected as a higher threshold in case the PV modules perform better than expectation.	User defined
Air pressure	500–1100 hPa	Defined by sensor manufacturers	Fixed
Visibility	6–80000 m	Defined by sensor manufacturers	Fixed

Table 7. Physical boundaries for each type of sensor.

a point on the unit circle and interpolates its sine and cosine components. This preserves the circular continuity of the data and avoids artificial jumps at the 0°/360° boundary. The procedure used in this dataset is as follows.

For each observed wind direction θ_i in degrees, convert the angle to radians and decompose it into unit-circle components:

$$\theta_{i-rad} = \theta_i^* \frac{\pi}{180}$$

$$x_i = \cos(\theta_{i-rad})$$

$$y_i = \sin(\theta_{i-rad})$$

For a missing value at time t between two valid observations at t_0 and t_1 , interpolate the two components separately:

$$x(t) = x_0 + (x_1 - x_0) \frac{(t - t_0)}{(t_1 - t_0)}$$

Dataset version	Interpolation status	Recommended use	Caution
Baseline	No down sampling and no interpolation	Recommended when the highest temporal fidelity is required, including event-scale analysis, sensor-level diagnostics, and studies of rapid microclimatic transitions.	May contain irregular timestamps and missing values; users should account for missing records explicitly.
Downsampled	Hourly down sampling; no interpolation	Recommended for CFD or WRF model validation, hourly model comparison, and analyses where regular timestamps are needed but estimated values should be avoided.	Missing values remain and should be handled explicitly in validation metrics or statistical analyses.
Downsampled + Interpolated	Hourly down sampling with interpolation for selected variables and gaps no longer than 60 minutes	Suitable for exploratory analysis, visualization, long-term trend assessment, and workflows requiring continuous hourly input for smoothly varying variables.	May smooth short-term dynamics and reduce extremes; not recommended for event-scale validation, wind, irradiance, PV-power, or transient-process analyses.

Table 8. Recommended use of dataset versions in relation to interpolation and temporal data integrity.

Sensor Type	Interpolation Method	Enabled by default?	Justification
Temperature	Linear	True	Low fluctuations within 60 minutes
Relative humidity	Linear	True	Low fluctuations within 60 minutes
Wind speed	Linear	False	High fluctuation may occur within 60 minutes
Wind direction	Angular	False	High fluctuation may occur within 60 minutes
Irradiance	Linear	False	High fluctuation may occur within 60 minutes due to cloud cover
Soil moisture	Linear	True	Low fluctuations within 60 minutes
Power	Linear	False	High fluctuation may occur within 60 minutes due to cloud cover
Air pressure	Linear	False	High fluctuation may occur within 60 minutes
Visibility	Linear	False	High fluctuation may occur within 60 minutes

Table 9. Interpolation setting for each type of sensor.

Categories	Description	Number of Files
PV Power	PV power-generation data from all three PVIGR sites, with WT, Sed, or Zoy as prefixes, and from the nearby PV module, with PV as the prefix	1
Microclimate	Microclimate data from all sites, including air temperature, relative humidity, POA irradiance, soil moisture, wind speed, and wind direction	14
Weather	Weather data from the automatic weather station, including wind speed, wind direction, air temperature, relative humidity, air pressure, visibility, and Global Horizontal Irradiance (GHI)	1

Table 10. Description of all categories in the extracted dataset.

$$y(t) = y_0 + (y_1 - y_0) \frac{(t - t_0)}{(t_1 - t_0)}$$

where the gap length is within the permitted interpolation window.

Recover the interpolated wind direction using the two-argument arctangent and convert it back to degrees:

$$\theta(t) = \text{atan2}(y(t), x(t)) * \frac{180}{\pi}$$

$$\theta_i = \begin{cases} \theta(t) + 360^\circ & \text{if } \theta(t) < 0^\circ \\ \theta(t) & \text{otherwise} \end{cases}$$

Stage 5: Extraction

The final stage of data curation is extraction, which extracts and organizes the data from April 1, 2024, to October 1, 2025, into CSV files stored in three categories. Please note that soil moisture data cover only the period from April 1, 2024, to August 31, 2024, due to sensor failure in September 2024. The details of each category are presented in Table 10.

Data Records

The data were deposited into Zenodo⁴¹, which contains three versions of processed data and the raw data.

Group	Abbreviation or pattern	Definition and usage
File/category prefix	PV	Photovoltaic; in file names, indicates the non-PVIGR PV roof/site.
File/category prefix	nrXXX	Near rooftop/site XXX, where XXX is the rooftop or site identifier.
PV configuration	SenTree; Land; Vert; Shd	PV module next to a sensor tree; PV module with landscape layout; PV module with vertical layout; PV module with relatively more shading.
Vegetation/configuration label	WedTri or WT; WT-to-Sed	PVIGR_1 before the vegetation change from Wedelia trilobata to Sedum lineare; WT is used for records before December 1, 2024. WT-to-Sed is used for records after the change on December 1, 2024.
Vegetation/configuration label	Sedum or Sed; Zoysia or Zoy	PVIGR with Sedum lineare (PVIGR_2); PVIGR with Zoysia (PVIGR_3). The shorter labels Sed and Zoy are used where compact file or column names are needed.
PV geometry	PVAzXXX; PVTiltXXX	PV module azimuth angle of XXX°, where 0° indicates north-facing and 90° indicates east-facing; PV module tilt angle of XXX°, where 0° indicates a horizontal module.
PV geometry	HgtXXX or PVHgt_XXXcm	Separation height between the center of the PV module and the rooftop surface, expressed in cm. PVHgt_XXXcm is the explicit height pattern; HgtXXX is a compact equivalent used where shorter labels are needed.
Sensor height/variable	SHXXX; Ta_XXXcm; RH_XXXcm	Sensor installation height of XXX cm; air temperature or relative humidity measured by a sensor installed at XXX cm above the rooftop surface.
Sensor height/variable	Ws_XXXcm; Wdir_XXXcm	Wind speed or wind direction measured by a sensor installed at XXX cm above the rooftop surface; for wind direction, 0° indicates wind from the north and 90° indicates wind from the east.
Sensor height/variable	Rad_Xdeg; SM	Pyranometer installed at X° tilt and facing south; soil moisture sensor.

Table 11. Abbreviation and corresponding definitions for columns and file names used in the extracted dataset.

Version name	Down sampling enabled?	Interpolation enabled?
Baseline	No	No
Downsampled	Yes (downsampling rate = 1 hour)	No
Downsampled + Interpolated	Yes (down sampling rate = 1 hour)	Yes (max gap = 1 hour)

Table 12. List of dataset versions.

Categories	Missing Rate (%)		
	Baseline	Downsampled	Downsampled + Interpolated
Power	30.99%	7.16%	7.16%
Microclimate	24.01%	18.44%	15.23%
Weather	0.27%	1.68%	1.36%

Table 13. General missing rate for each category.

Columns and files related to Wedelia trilobata power generation are divided into two parts because PVIGR_1 was replanted with Sedum lineare on December 1, 2024. The first part contains records before December 1, 2024, with later records marked as NaN; this part uses the suffix “WT”. The second part contains records from December 1, 2024, onward, with earlier records marked as NaN; this part uses the suffix “WT-to-Sed”. File and column names use abbreviated prefixes or suffixes to encode rooftop type, vegetation type, PV geometry, sensor height, and variable type. In the naming patterns below, “XXX” denotes a numeric value, and units such as “cm” and “deg” are included where applicable. Please refer to Table 11 on the abbreviations for the column names in the extracted dataset.

Three versions of the dataset were extracted with the data curation workflow, and they are shown in Table 12.

Technical Validation

A missing rate analysis was conducted for all three versions of the dataset to evaluate the quality of the dataset. The general missing rate for each category is shown in Table 13.

The missing-rate values reported for the three dataset versions should be interpreted within the context of each version and should not be treated as directly comparable indicators of overall data quality. In the Baseline dataset, missing rates are calculated only over timestamps that were recorded in the original data files; therefore, missing timestamps are not represented in the denominator. In contrast, the Downsampled and Downsampled + Interpolated datasets are evaluated on a regular hourly timestamp grid. Down sampling can introduce additional timestamp rows where no original observation is available, and these cells are represented as NaN, changing both the numerator and denominator of the missing-rate calculation. As a result, a higher missing rate in a processed version does not necessarily indicate deterioration in data quality, and a lower missing rate after interpolation reflects reduced missingness within the regular hourly grid rather than recovery of originally

	Baseline	Downsampled	Downsampled + Interpolated
Total number of cells	89,922	91,210	91,210
Number of missing cells	242	1,530	1,237
Missing Rate (%)	0.27%	1.68%	1.36%

Table 14. Total number of cells, number of missing cells and missing rate for weather data in each version of dataset.

observed values. In practice, the Baseline missing rate describes completeness among recorded observations, the Downsampled missing rate describes completeness relative to the hourly analysis grid, and the Downsampled + Interpolated missing rate describes remaining missingness after short-gap interpolation. Users should therefore select the dataset version based on the temporal structure required by their analysis and should interpret missing-rate metrics accordingly.

Please note that missing rate for PV power generation data is only computed for data between 6:00 and 18:00 because this is the common time range that SolarEdge would record the power generation data.

On the other hand, the “Baseline” dataset ignored missing timestamps when computing the missing rate (i.e., it only counts the number of cells with missing values and divides by the total number of cells) due to the lack of down sampling in the process. Down sampling generates missing timestamps, fills the cells in those rows with NaN, and then computes the missing rate. This may cause the missing rate for the “Baseline” dataset to be lower than those of the other two versions. For example, 184 timestamps were added in the weather data after down sampling, which added 1,288 cells with NaN to the data (i.e., 7 variables × 184 timestamps = 1,288 cells). Given that the weather data already had a one-hour sampling rate, no cells were removed during down sampling. As a result, the total number of cells and number of missing cells both increased by 1,288, increasing the missing rate because the relative change in the numerator was larger than that in the denominator, as shown in Table 14. Users should be cautious about missing timestamps when using the “Baseline” dataset.

Monthly heatmaps of missing data rates for all sensors in the “Downsampled + Interpolated” dataset are presented in Fig. 9 of the Appendix. In addition, the monthly missing rate heatmaps are available in the subfolder “missing_rate” of every version of the dataset.

Data availability

All data in this article are available at Zenodo (<https://doi.org/10.5281/zenodo.19123202>).

Code availability

All Python module files, configuration JSON files, and Jupyter notebooks are available at <https://github.com/donald323/HK-Rooftop-Microclimate-PV>. Users can modify the configuration JSON files to adjust the parameters for any stage of data curation to suit their needs and then run the Jupyter notebook “00_main_pipeline.ipynb” for data curation to process the raw data. In addition, a TXT file named “Set up environment.txt” is also included for users to install all Python dependencies required for the notebook.

Received: 11 March 2026; Accepted: 11 September 2026;

Published: xx xx xxxx

References

1. Santamouris, M. Minimizing Energy Consumption, Energy Poverty and Global and Local Climate Change in the Built Environment: Innovating to Zero: Causalities and Impacts in a Zero Concept World (Elsevier, 2018).

2. Santamouris, M. Recent progress on urban overheating and heat island research. Integrated assessment of the energy, environmental, vulnerability and health impact. Synergies with the global climate change. *Energy Build.* **207**, 109482 (2020).

3. Oke, T. R., Johnson, G. T., Steyn, D. G. & Watson, I. D. Simulation of surface urban heat islands under ‘ideal’ conditions at night part 2: Diagnosis of causation. *Bound.-Layer Meteorol.* **56**, 339–358 (1991).

4. Li, X. C., Zhao, L., Qin, Y., Oleson, K. & Zhang, Y. Elevated urban energy risks due to climate-driven biophysical feedbacks. *Nat. Clim. Change.* **14**, 1056–1063 (2024).

5. Chen, L. *et al.* Mitigating urban climate-energy feedback with citywide building-integrated photovoltaics implementation. *Nexus.* **3**, 100114 (2026).

6. Santamouris, M. Regulating the damaged thermostat of the cities—Status, impacts and mitigation challenges. *Energy Build.* **91**, 43–56 (2015).

7. Pérez-Lombard, L., Ortiz, J. & Pout, C. A review on buildings energy consumption information. *Energy Build.* **40**, 394–398 (2008).

8. Varbanov, P. S., Klemeš, J. J. & Wang, X. Methods optimisation, Process Integration and modelling for energy saving and pollution reduction. *Energy* **146**, 1–3 (2018).

9. Wolch, J. R., Byrne, J. & Newell, J. P. Urban green space, public health, and environmental justice: The challenge of making cities ‘just green enough’. *Landsc. URBAN Plan.* **125**, 234–244 (2014).

10. Campisano, A. *et al.* Urban rainwater harvesting systems: Research, implementation and future perspectives. *WATER Res.* **115**, 195–209 (2017).

11. Li, G. *et al.* Building integrated solar concentrating systems: A review. *Appl. ENERGY* **260**, 114288 (2020).

12. Ramamurthy, P., Sun, T., Rule, K. & Bou-Zeid, E. The joint influence of albedo and insulation on roof performance: An observational study. *ENERGY Build.* **93**, 249–258 (2015).

13. Androutsopoulos, A. V., Stavrakakis, G. M. & Damasiotis, M. Cool Roof Impacts on a School-Building Thermal and Energy Performance in Athens, Greece. in Sustainable Synergies From Buildings to the urban scale (eds. Bikas, D., Theodosiou, T. & Katerina, T.) vol. 38 178–186 (Elsevier Science Bv, Amsterdam, 2017).

14. Castleton, H. F., Stovin, V., Beck, S. B. M. & Davison, J. B. Green roofs; building energy savings and the potential for retrofit. *ENERGY Build.* **42**, 1582–1591 (2010).

15. Yoon, Y., Jung, S., Im, P. & Gehl, A. Datasets of a Multizone Office Building under Different HVAC System Operation Scenarios. *Sci. Data* **9**, 775 (2022).
16. Lin, S. *et al.* District-scale surface temperatures generated from high-resolution longitudinal thermal infrared images. *Sci. DATA* **10**, 859 (2023).
17. Warren, E. L. *et al.* The Birmingham Urban Climate Laboratory—A high density, urban meteorological dataset, from 2012–2014. *Sci. Data* **3**, 160038 (2016).
18. Reani, M., Lowe, D., Gledson, A., Topping, D. & Jay, C. UK daily meteorology, air quality, and pollen measurements for 2016–2019, with estimates for missing data. *Sci. Data* **9**, 43 (2022).
19. Ulpiani, G. *et al.* A citizen centred urban network for weather and air quality in Australian schools. *Sci. Data* **9**, 129 (2022).
20. Weather and climate change. Met Office <https://www.metoffice.gov.uk/> (2024).
21. Lin, Z. *et al.* A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics. *Sci. DATA* **12**, 63 (2025).
22. Santamouris, M. Cooling the cities - A review of reflective and green roof mitigation technologies to fight heat island and improve comfort in urban environments. *Sol. ENERGY* **103**, 682–703 (2014).
23. Pisello, A. L. & Cotana, F. The thermal effect of an innovative cool roof on residential buildings in Italy: Results from two years of continuous monitoring. *ENERGY Build.* **69**, 154–164 (2014).
24. Baik, J.-J., Kwak, K.-H., Park, S.-B. & Ryu, Y.-H. Effects of building roof greening on air quality in street canyons. *Atmos. Environ.* **61**, 48–55 (2012).
25. Liu, H., Li, Y., Zhang, Y., Zhang, J. & Zhao, X. Spectral Splitting for Full-Spectrum Solar Utilization: From Efficiency Improvement to Spectral Intelligence. *Energy Use.* **1**, 100021 (2025).
26. Lazzaretti, A. E. D. *et al.* A Monitoring System for Online Fault Detection and Classification in Photovoltaic Plants. *Sensors*. **20**, 4688 (2020).
27. Peren, J. I., van Hooff, T., Leite, B. C. C. & Blocken, B. CFD analysis of cross-ventilation of a generic isolated building with asymmetric opening positions: Impact of roof angle and opening location. *Build. Environ.* **85**, 263–276 (2015).
28. Tominaga, Y., Akabayashi, S., Kitahara, T. & Arinami, Y. Air flow around isolated gable-roof buildings with different roof pitches: Wind tunnel experiments and CFD simulations. *Build. Environ.* **84**, 204–213 (2015).
29. Zhou, Q., Dong, P., Li, M. & Wang, Z. Analyzing the interactions between photovoltaic system and its ambient environment using CFD techniques: A review. *Energy Build.* **296**, 113394 (2023).
30. Zonato, A. *et al.* Exploring the effects of rooftop mitigation strategies on urban temperatures and energy consumption. *J. Geophys. Res.* **126** (2021).
31. Sun, Y., Takane, Y., Yamaguchi, K. & Ihara, T. The photovoltaic heat island effect in Tokyo: Morphological dependence and spatiotemporal dynamics of PV-induced urban warming. *Build. Environ.* **288**, 113941 (2026).
32. Chen, L. *et al.* Impacts of photovoltaics and integrated green roofs on urban climate: Experimental insights for urban land surface modelling. *Renewable Sustainable Energy Rev.* **217**, 115709 (2025).
33. Chen, L. *et al.* The Resilience Paradox of Rooftop PV: Building Cooling Penalties and Heat Risks. *Build. Environ.* **282**, 113233 (2025).
34. Dassen, T., Kunseler, E. & van Kessenich, L. M. The Sustainable City: An Analytical-Deliberative Approach to Assess Policy in the Context of Sustainable Urban Development. *Sustain. Dev.* **21**, 193–205 (2013).
35. Degirmenci, K., Desouza, K. C., Fieuw, W., Watson, R. T. & Yigitcanlar, T. Understanding policy and technology responses in mitigating urban heat islands: A literature review and directions for future research. *Sustain. CITIES Soc.* **70**, 102873 (2021).
36. Faïd, A., Sadik, M. & Sabir, E. An Agile AI and IoT-Augmented Smart Farming: A Cost-Effective Cognitive Weather Station. *Agric. - BASEL* **12**, 35 (2022).
37. Fan, C., Lei, Y., Piscitelli, M. S. & Mo, J. Generalizable Data-Driven Building Operation Control Methods for Scalable Cross-Building Applications. *Energy Use.* **1**, 100008.
38. Chen, L. *et al.* Comparing designs for photovoltaic-green roofs: A year-long field study in a subtropical climate. *Renew. Energ.* **268**, 125805.
39. Siu, L. W. & Hart, M. A. Quantifying urban heat island intensity in Hong Kong SAR, China. *Environ. Monit. Assess.* **185**, 4383–4398 (2013).
40. Environmental Central Facility. The Hong Kong University of Science and Technology <https://envf-ienv.hkust.edu.hk/>.
41. Cheng, I. *et al.* High temporal-resolution microclimate and photovoltaic power dataset for sustainable rooftop design in subtropical humid city. *Zenodo* <https://doi.org/10.5281/zenodo.19123202> (2026).

Acknowledgements

The authors acknowledge the support from the State Key Laboratory of Climate Resilience for Coastal Cities (ITC-SKLCRCC26EG01), and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. C6003-22Y).

Author contributions

I.C. and L.C. led the conceptualization and methodology, carried out the experiments and investigation, and prepared the figures on the manuscript. They also contributed to revisions. I.C. also drafted the manuscript and was responsible for curating the data. H.C. wrote the rationale for the sensor tree and contributed to revisions. Z.L. contributed to the experiments, investigation, and drafting of the manuscript. M.L. provided supervision and manuscript review, and secured funding. Z.W. oversaw the project, provided manuscript review, and secured funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-08340-7>.

Correspondence and requests for materials should be addressed to M.L. or Z.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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.