

scientific data



OPEN

DATA DESCRIPTOR

A two-year dataset of energy, environment, and system operations for an ultra-low energy office building

Jung Min Han^{1,2,3}, Ali Malkawi^{1,2}, Xu Han^{1,2,4}✉, Sunghwan Lim^{1,2}, Elence Xinzhu Chen^{1,2}, Sang Won Kang^{1,2}, Yiwei Lyu^{1,2} & Peter Howard²

This paper describes a two-year high-fidelity dataset for an ultra-low energy office building and living laboratory called *HouseZero®*. The building integrates multiple low-energy technologies, such as natural ventilation with automatic windows, ground source heat pump, and thermally activated building systems. The building's performance is continuously monitored with an extensive sensor network. The dataset consists of breakdown energy end uses, photovoltaic (PV) production, zone-level indoor environment including indoor air temperature, CO₂ concentration, and relative humidity, micro-climatic conditions, building façade temperature, and detailed system operations including zone-level BTU meter, valve status, slab temperature, window/skylight opening status, heat pump, and geothermal well operations. The data can be used to support data analytics of ultra-low-energy building operations, and data-driven modeling of low-energy building systems.

Background & Summary

The building sector, including residential and commercial buildings, accounts for approximately one-third of global energy consumption and carbon dioxide (CO₂) emissions¹. To improve the energy efficiency of building operations, data-driven approaches have been widely used for building load forecasting², occupant behavior modeling³, machine-learning based control of building systems^{4–6}, building analytics⁷, and energy management⁸. To support the development of the data-driven approaches, high-fidelity data with detailed building operational information becomes essential.

With the equipment of building management systems (BMS) and smart meters in buildings, there are various open-source datasets available with different levels of fidelity. Most datasets are focused on building energy consumption^{9–12}. Several datasets have been released with a focus on occupancy data^{13–16}. More comprehensive datasets have also been proposed that consist of energy consumption, indoor environment, occupancy, weather conditions, and HVAC operations^{17–19}. However, very limited number of datasets are reported for ultra-low energy buildings with low-energy and passive technologies, such as natural ventilation, ground source heat pumps, and thermally active building systems (TABS), which play a crucial role in achieving the carbon-neutral goal for the building sector. Agee, Nikdel and Roberts^{20,21} proposed a dataset for a zero-energy building that consists of energy uses, photovoltaic (PV) production, and building air leakage data, but doesn't include detailed heating, ventilation, and air conditioning (HVAC) system operational data. Schweiker, Kleber and Wagner^{22,23} introduced a dataset for a naturally ventilated office building. However, the operation status of the manually operated windows is recorded as closed or open without specific information about the window opening percentage, which may be required to develop natural ventilation prediction and smart window controls. Therefore, a dataset containing detailed system operational information of ultra-low energy buildings with low-energy and passive technologies is needed.

¹Graduate School of Design, Harvard University, Cambridge, MA, 02138, USA. ²Harvard Center for Green Buildings and Cities, Harvard University, Cambridge, MA, 02138, USA. ³Department of Architecture and Architectural Engineering, Yonsei University, Seoul, 03722, Republic of Korea. ⁴School of Engineering, University of Kansas, Lawrence, KS, 66045, USA. ✉e-mail: xuhan@ku.edu

Sensors	Methods
Weather	Air temperature
	Relative humidity
	Wind speed
	Wind direction
	Rain
	Solar radiation
Building façade	Façade temperature
Load	Heating/ Cooling
	Domestic Hot Water
	Lighting
	Plug
	IT
	Control
	Others
Indoor sensors	Air temperature
	Relative humidity
	CO ₂ concentration
	Slab temperature
Sensors for systems	BTU meters
	Window openings
	Valve status

Table 1. Summary of abnormality detection and filtering methods for different sensors.

This paper describes a high-fidelity dataset which provides granular insights into the performance and operations of an ultra-low energy building. With its intricate sensor network, the building captures diverse performance parameters as listed in Table 3. The dataset includes the following unique aspects compared to existing datasets in this area:

- It provides data of integrated low-energy building systems, such as natural ventilation with automatic windows combined with geothermal powered TABS for heating and cooling, automatic operable skylights, and PV systems. The data of such a low-energy naturally ventilated building combined with geo-powered TABS has not been reported in existing datasets to our best knowledge.
- It provides data from an extensive sensor network, including not only energy uses and indoor environment data as reported in existing datasets, but also detailed system operational data. Examples include window openings, temperature and flowrate of both source side and load side water loops of the heat pump, as well as outdoor sensors, such as localized weather stations and building façade temperatures that provide the boundaries of the microclimate.
- It provides data of zone-level BTU meters for the TABS that was rarely reported in existing datasets, which helps understand the zone-level thermal load and detailed operations, e.g., water temperature, flowrate and valve status, of the TABS for each zone in response to the disturbances.

In summary, this dataset provides high-fidelity data regarding micro-climatic conditions, façade temperature, zone-level TABS, and thermal information with loads for a naturally ventilated building that utilizes geothermal heating and cooling. With this dataset, the users will be able to have a better understanding of the operations of such integrated low-energy and passive building technologies in a real building and thereby develop advanced methods/algorithms to better design and operate such systems. For example, similar to the analysis in⁶, the user can investigate the operational performance of the coupled NV and TABS and identify operation issues and potential improvement strategies, which may provide valuable information for researchers or operators of other similar buildings.

Methods

The building called *HouseZero*[®] (See Fig. 1) was retrofitted from a pre-1940s house into an office building and living-laboratory that functions as a prototype for an ultra-efficiency building. It has a total floor area of 356 m².

HouseZero[®] consists of four floors: basement (level 0), first floor (level 1), second floor (level 2), and third floor (level 3), and all zones are identified in Fig. 2. The basement is located at the underground lower level and has a large conference room, a server room, and a kitchenette. The first floor has direct access to the main entrance with semi-open spaces, as well as an open lounge which is designed for five occupants. The second floor is an open workspace, designed for 16 occupants. Lastly, the third floor is connected to the open lounge, with one laboratory ('Live Lab') and other workspaces designed for two occupants. The 'Live Lab' is designed to conduct room-scale experiments with functionality that represents the systems and operation of the building. In addition, it has the capability for experimentation with different façade systems, as the entire window system is designed to be removed and replaced with other experimental systems.



Fig. 1 The office building HouseZero® in Cambridge, Massachusetts, USA.

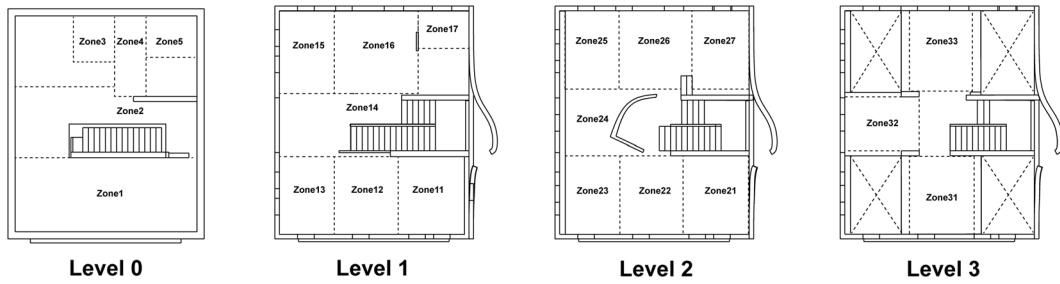


Fig. 2 Layout of each floor and thermal zones.

The building integrates multiple low-energy technologies, including natural ventilation with automatic windows and operable skylights, an automatic light system, TABS, a heat pump, a geothermal system, solar PV, and a hot water system, as well as a sensor network and advanced controls²⁴. Figure 3 describes the system configurations at *HouseZero*®. For a detailed description of the network, systems and controls, please refer to^{7,8,24}.

The workflow of data collection and processing is shown in Fig. 4. *HouseZero*® has dedicated Building Automation System (BAS) server networks for controls and performance optimization. The raw dataset was first downloaded from the BAS servers and then processed with four steps.

The first step is outlier filtering. Outliers were identified using four methods: Absolute Difference²⁵, Z-score^{26,27}, DBSCAN and Mahalanobis distance²⁸. Identifying values with high absolute difference from preceding data points helps pinpoint distinct peaks in the initial dataset. Z-score identifies abnormal data; points with a Z-score exceeding a certain threshold (typically, standard deviation = 3) were excluded. DBSCAN identifies dense clusters by grouping data points that are closely packed together and labels those not belonging to a cluster with at least 15 neighbors as outliers. The Mahalanobis distance measures a data point's distance from the distribution's center. Points with a distance exceeding a certain limit are deemed outliers. This limit is derived from the Chi-squared distribution, considering the dataset's significance level (0.05 - 0.01) and variable count. This method is especially effective for datasets with correlated variables, and thus applied to zone temperature, slab temperature, and other BTU-related data.

The second step is data aggregation. This step is performed to resample data to hourly intervals for several reasons. The main reason is that, in practice and research studies, hourly interval data is often used, such as typical meteorological year weather files and operational data, for the building energy simulation. Moreover, the

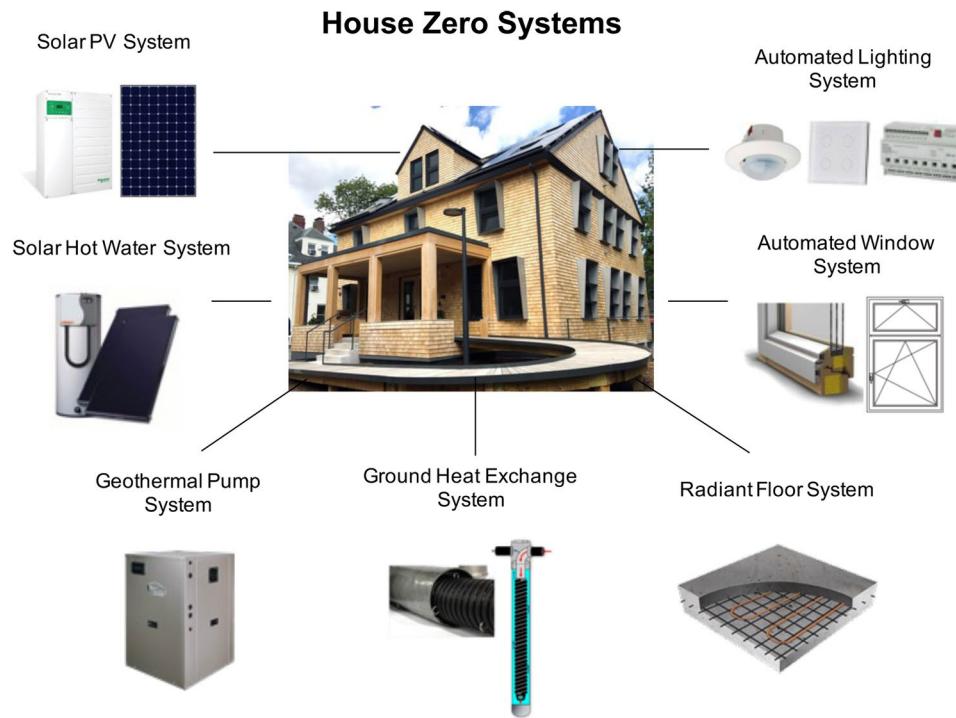


Fig. 3 Schematics of the building systems in HouseZero®.

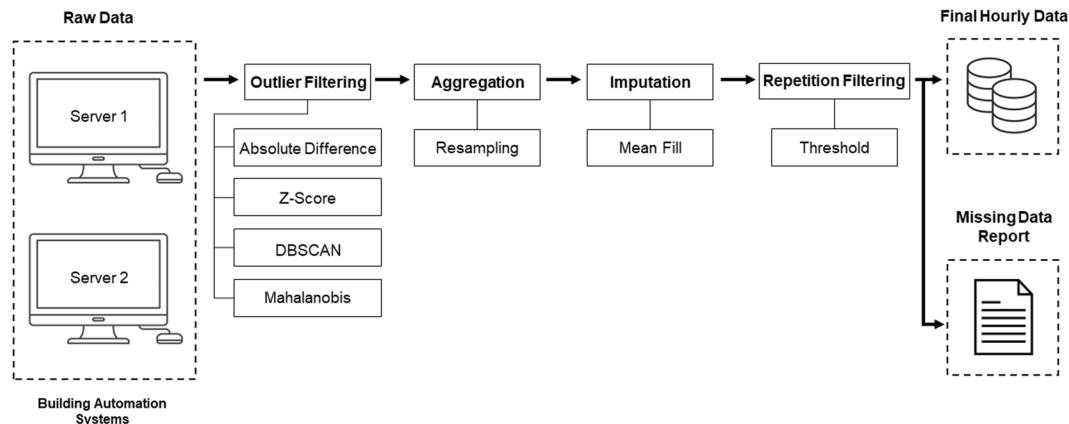


Fig. 4 Workflow of the data collection and processing.

size of the dataset is reasonable and easier to utilize with hourly intervals. However, it is noteworthy that this may lose the fidelity of short-term control dynamics, such as the winter pulse ventilation that occurred in an hourly basis with a 30 seconds duration of window opening.

The third step is data imputation. During data resampling to hourly intervals, imputation with a mean fill technique is employed to handle short-term missing data such as removed values for outliers, and enable the completion of the dataset. This technique involves identifying missing intervals in the data and calculating the average between the last non-missing value before the gap and the first non-missing value following the gap. This calculated average is then used to fill the missing values. For the long-term missing data when the missing period is longer than a threshold, the values of those data points are left blank without imputation in the dataset and the missing data periods are documented in the data report. The thresholds to identify the long-term missing data are different for different sensor data, which can be found in detail in the data processing Python code provided in the Section “Code Availability”.

Finally, repetition filtering eliminates values in the dataset when they appear consecutively repeated over a set number of hours, based on thresholds tailored to assumptions about sensor errors and data characteristics. Such repetition is assumed to indicate a sensor malfunction or servers offline, which was recorded as missing data, ensuring data integrity. The repetition filtering was conducted at the last step instead of the first step to avoid the filtered repetition data points being filled again in the third step. Table 1 summarizes all of the methodologies for different sensor data described in this section.

Folder	Subfolder	File	Column Description	Unit
Outdoor Sensors		Local_weather_hourly	Air temperature, Relative humidity, Wind speed, Wind direction, Rain, Solar radiation	—
		Facade_temp_hourly	Façade temperature - High, mid, low location (9 sensors)	°C
Indoor Sensors		Zone_temp_hourly	Zone air temperature (17 zones)	°C
		Slab_temp_hourly	Slab temperature (20 zones)	°C
Sensors for Systems		CO ₂ _hourly	CO ₂ concentration (15 zones)	ppm
		RH_hourly	Relative humidity (15 zones)	%
Sensors for Systems	Year 1: June 2022 - May 2023 Year 2: June 2023 - May 2024	Load_hourly	Breakdown loads - Heat pump, Heat pump Electric, Cooling, DHW (Solar hot water/ Electric hot water), Lighting, Controls, Others (Exhaust fan/ Elevator/ Elevator light/Fire alarm/Sump pump/ Battery cabinet/Solar rapid shutdown), Plug load (Basement/1 st /2 nd /3 rd) and IT (18 sensors)	kWh
		BTU_GEO_hourly	Flowrate, Supply temperature, Return temperature, Energy (calculated) (3 zones for Geothermal 8 zones for Tabs)	°C, GPM, BTUs
		BTU_TABS_hourly		8 zones °C, GPM, BTUs
		Window_opening_hourly	Window openings (33 windows)	%
		Zone_valve_hourly	Valve status (19 valves)	%

Table 2. Folder Structure of the Dataset.

Sensors/Meters			Number of Sensors/Meters	Missing data (%)	
				Year 1	Year 2
Outdoor sensors	Localized weather stations	Air temperature	1	0%	0%
		Relative humidity	1	0%	0%
		Wind speed	2	0%	0%
		Wind direction	2	0%	0%
		Rain	1	0%	1.02%
		Solar radiation	1	0%	0%
Indoor sensors	Indoor environment	Building façade	9	6.37%	12.73%
		Air temperature	17	6.92%	0.19%
		Relative humidity	15	1.95%	0%
	Building structure	CO ₂ concentration	15	2.04%	0%
Sensors for systems	Meters	Building structure	Slab temperature	20	3.13% 0.57%
		PV production	2	1.96%	0%
		Breakdown loads: Heat pump, Heat pump Electric, Cooling, Domestic hot water (Solar hot water/ Electric hot water), Emergency Lighting, Lighting, Controls, Others (Exhaust fan/ Elevator/ Fire alarm/Sump pump/ Battery cabinet/Solar rapid shutdown), Plug load (Basement/1 st /2 nd /3 rd) and IT	19	0%	0%
		Net meter	1	0%	0%
		Import and export (from Net meter)	—	0%	0%
	BTU meters	Building structure	Flowrate	10	0% 0%
		Building structure	Supply water temperature	10	0% 0%
		Building structure	Return water temperature	10	0% 0%
		Building structure	Energy (calculated)	—	0% 0%
	System status	Building structure	Window openings	34	0.06% 0%
		Building structure	Valve status	19	0.39% 0%

Table 3. Summary of released data and missing data percentages.

As shown in Table 2, sensor data is organized within the database into three main categories: outdoor sensors, indoor sensors, and sensors for systems, each tracking various environmental and system-related parameters.

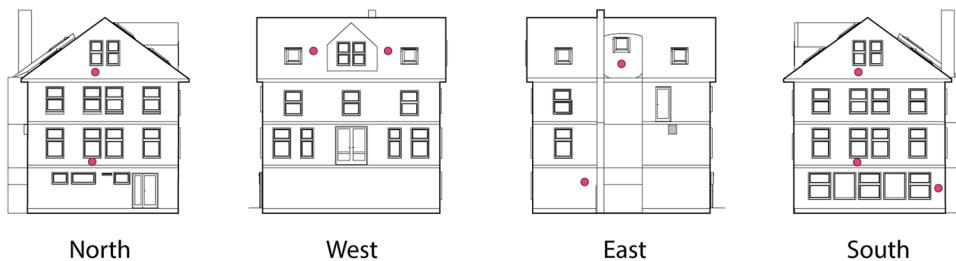


Fig. 5 Locations of the building façade temperature sensors.

Starting Date	Ending Date	Issues
07/02/2022	07/03/2022	PV meter 2 connection issue
07/19/2022	07/21/2022	PV meter 1 and meter 2 connection issues
08/06/2022	08/07/2022	PV meter 2 connection issue
09/29/2022	09/30/2022	PV meter 2 connection issue
01/16/2023	01/16/2023	PV snow covering
02/07/2023	03/11/2023	PV meter #2 was off. To optimize the system for winter operations, the system was changed to one inverter.
3/14/2023	8/16/2023	The solar hot water was offline.
8/16/2023	9/5/2023	The solar hot water was temporarily powered from the basement circuit, this led to 3 kWh total increase on the plug loads which was not included in the basement plug loads.
12/04/2023	03/15/2024	PV meter #2 was off. To optimize the system for winter operations, the system was changed to one inverter.
Multiple missing periods, refer to the report		Façade sensor issue

Table 4. Summary of the issues causing missing data.

Outdoor sensors provide hourly data on local weather conditions and façade temperatures, while indoor sensors measure zone and slab temperatures, CO₂ concentrations, and relative humidity across different zones. Sensors for systems offer detailed insights into the building systems operations, including heat pumps, domestic hot water systems, lighting, and other loads, as well as operational data from geothermal systems, Thermal Active Building Systems (TABS), window openings, and valve statuses.

To ensure transparency and reproducibility, the missing periods were documented in a data report. The missing periods were identified during data processing, which may be due to sensor malfunctions, server offline or other technical issues. The report included the start and end dates of each missing period, as well as the names of data points. This allows for greater transparency and ensures that the processed data can be used reliably and accurately in future analysis.

Data Records

The time-series data with one-hour intervals are in CSV format. The data is hosted at figShare²⁹. A data report documenting the detailed description of the dataset is available in the same repository²⁹.

This section provides a description of the data and special events that occurred within the data collection period. Table 3 includes all available sensors and missing data percentages. The available periods for the data collected are from June 2022 to the end of May 2024. (Year 1: June 2022 to May 2023; Year 2: June 2023 to May 2024). The last two columns indicate the percentage of data missing across Year 1 and Year 2 for different sensor data. In this paper, the outliers and missing data are reported in a separate file.

While the building has more than 300 sensors and meters, this paper releases data from 189 sensors and meters that are closely related to the main operational performance of the building. Outdoor sensors include two localized weather stations and nine façade temperature sensors on the building's façades (See Fig. 5 for the locations of the façade temperature sensors). There are two localized weather stations with one installed on the roof of HouseZero® building and the other installed on the roof of a nearby building. These weather stations independently measure outdoor weather conditions, and their readings are cross-checked against each other for consistency.

Indoor sensors monitor zone-level air temperature, CO₂, and relative humidity, as well as slab temperature. Sensors for systems are related to operational status of the integrated building systems shown in Fig. 3. Meters have been installed to monitor all load breakdowns from individual breakers, including PV, loads and net meters. Multiple meters are installed in the building for cross-validation of the electric loads to enhance accuracy and ensure data fidelity. For example, in addition to the electrical provider utility meter, a net meter was installed to validate both the PV production and loads as well as to provide the export and import electrical data. BTU meters are included for the radiant floor system in each zone, as well as the heat pump and geothermal systems. The status of each individual window and valve is also monitored.

Over the course of the data collection period, there were some times when certain meters and sensors had temporary interruptions and data loss. Table 4 summarizes issues in operation over the two years. This table is

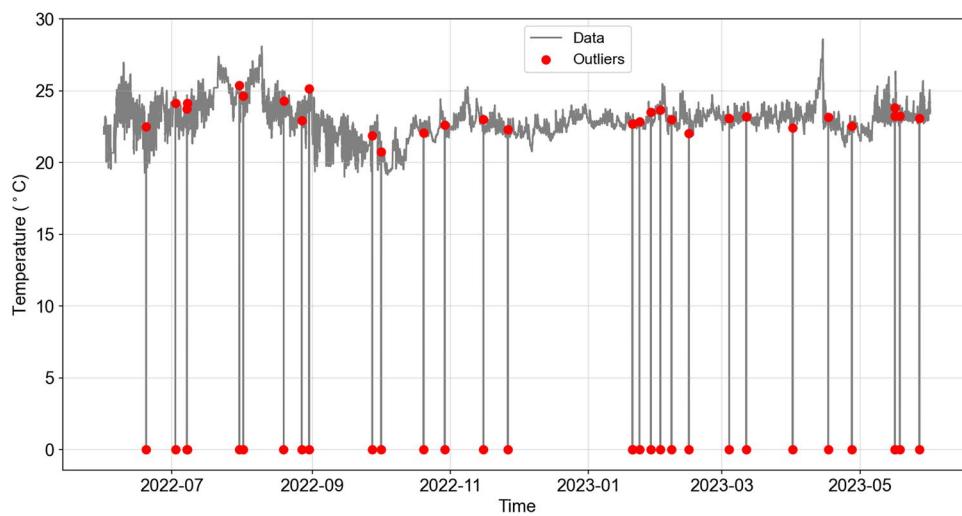


Fig. 6 An example of outlier filtering using Z-score.

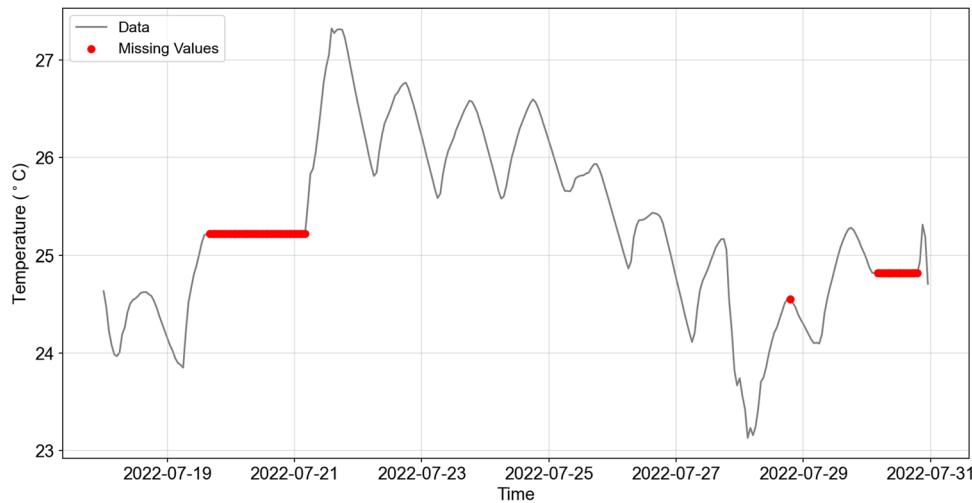


Fig. 7 An example of data repetition filtering.

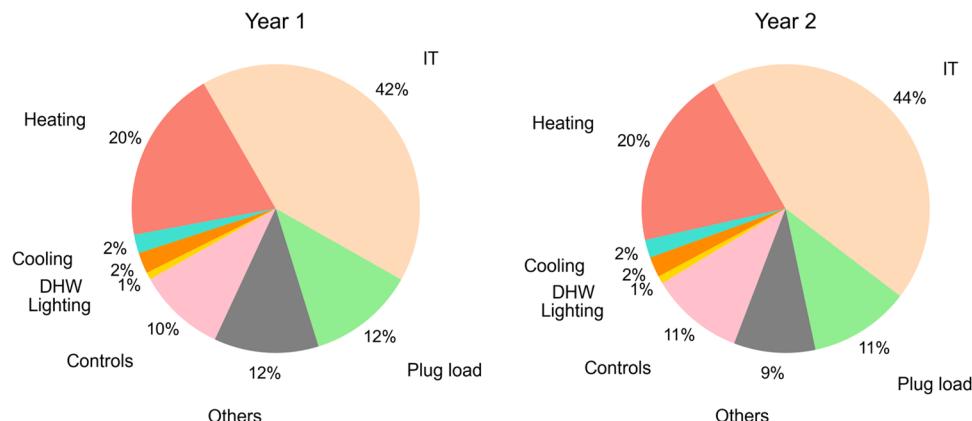


Fig. 8 Pie chart of the breakdown energy end uses for two years.

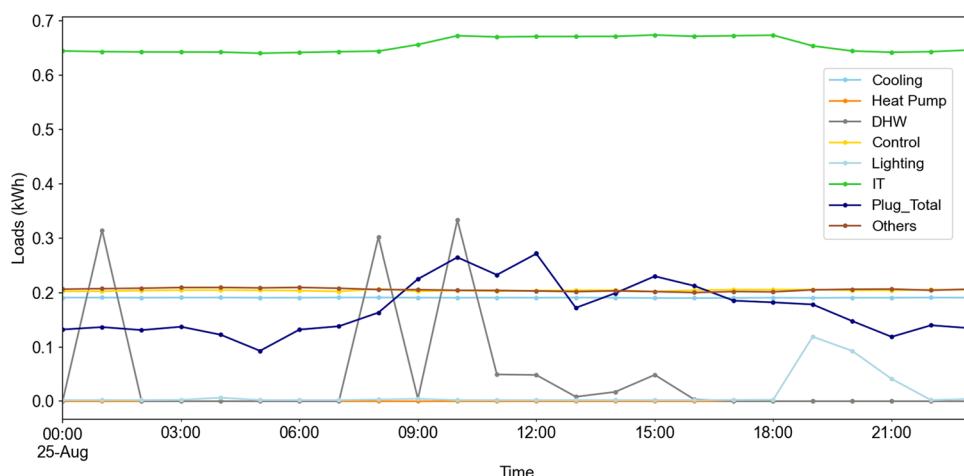


Fig. 9 Daily pattern of electricity end uses in a sample summer day.

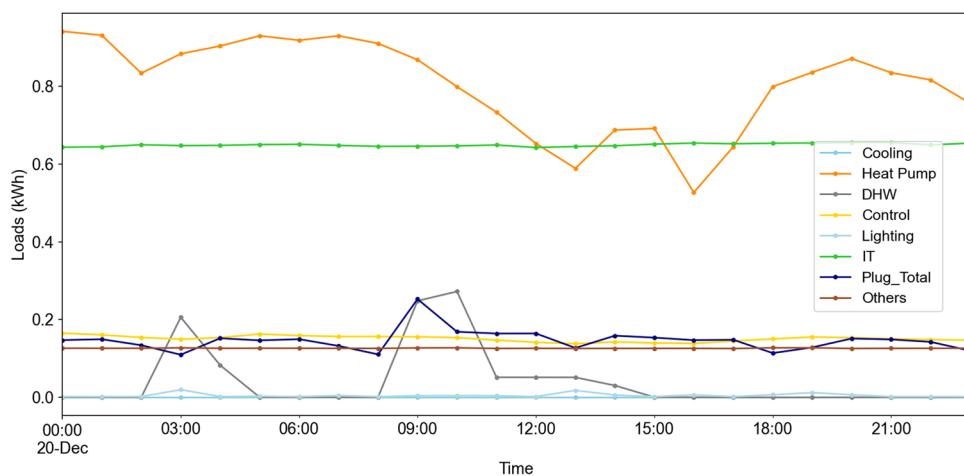


Fig. 10 Daily pattern of electricity end uses in a sample winter day.

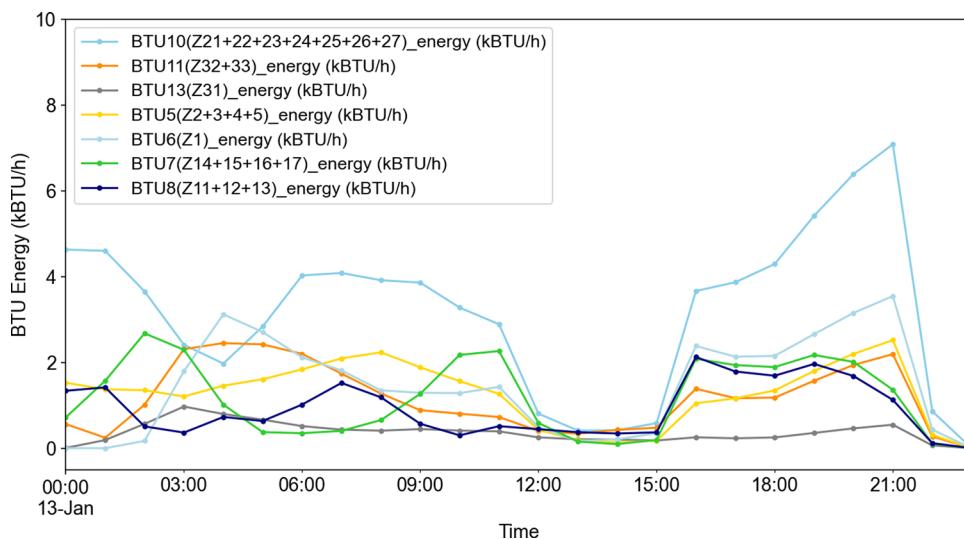


Fig. 11 Daily pattern of energy rate data from a BTU meter in a sample winter day.

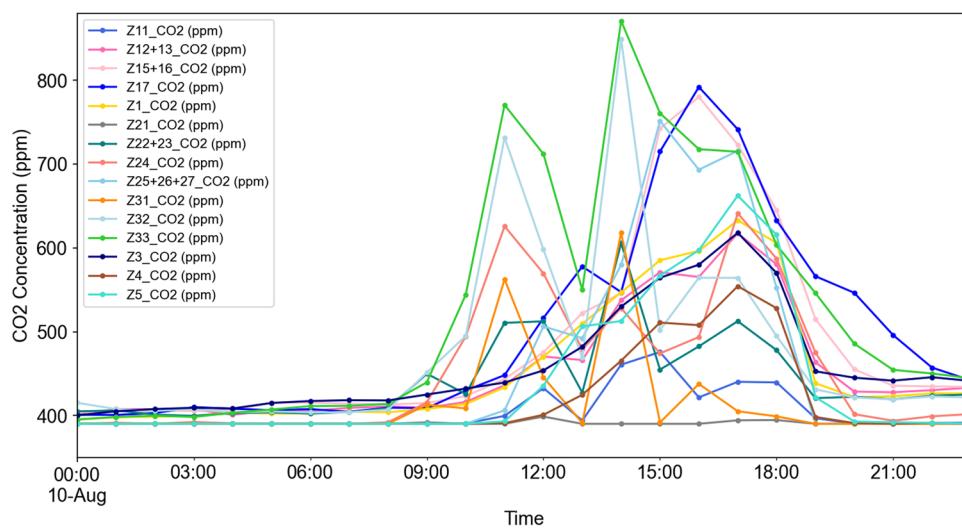


Fig. 12 Daily pattern of the indoor CO₂ concentrations in a sample summer day.

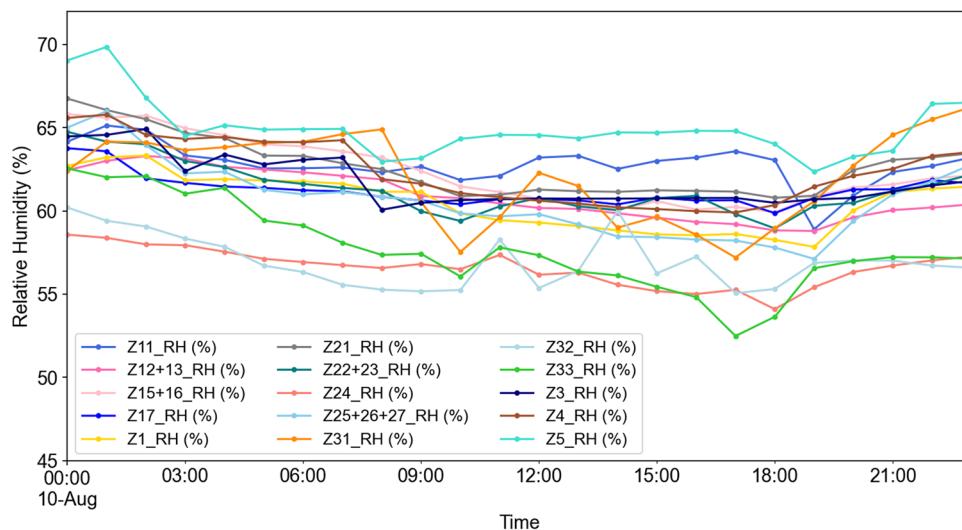


Fig. 13 Daily pattern of relative humidity in a sample summer day.

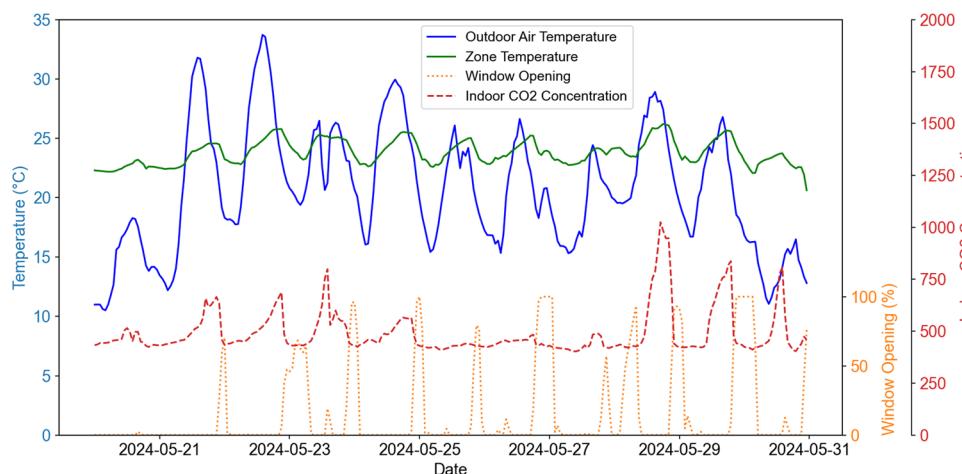


Fig. 14 Indoor temperature and CO₂ with natural ventilation in the passive mode.

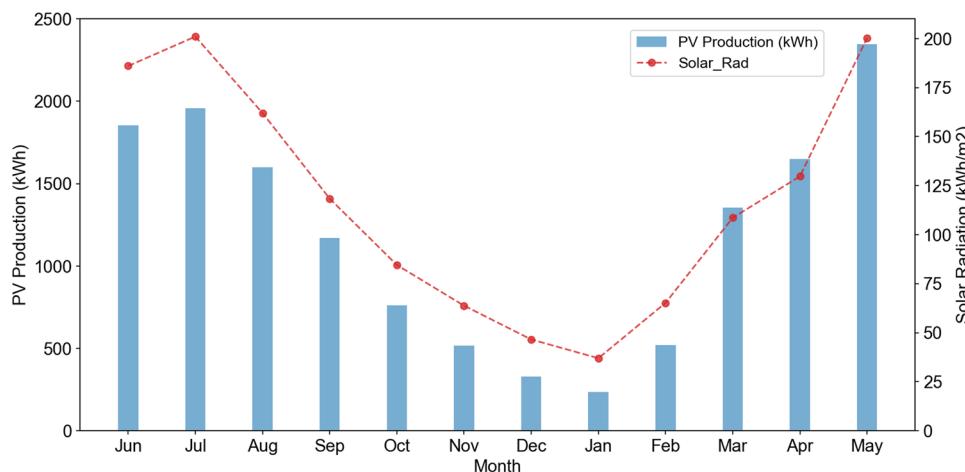


Fig. 15 Monthly PV production of Year 1.

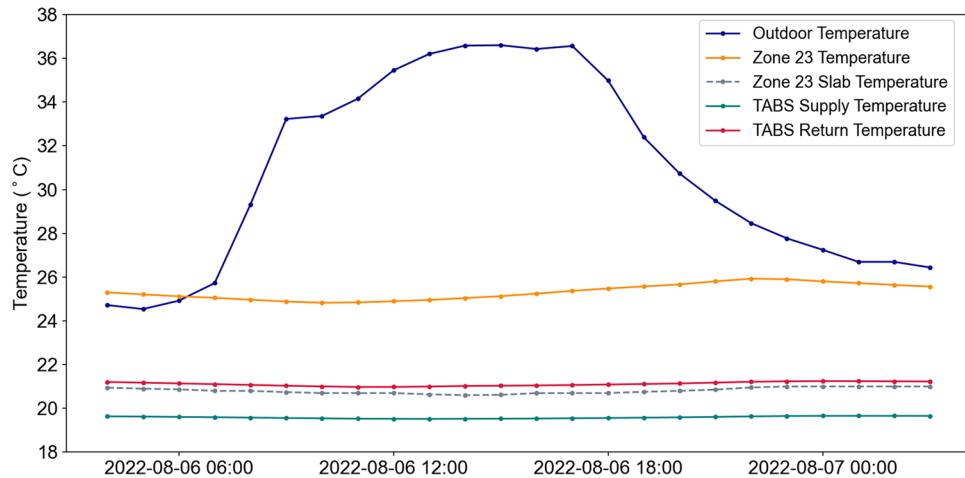


Fig. 16 Operation of the TABS in a sample summer day of Zone 23.

derived from the building operation log that provides additional information to help understand the causes of the missing data. The events that may cause missing data include system updates, snow covering, sensor issues, server communication issues, database maintenance issues, and other issues as listed in Table 4.

Technical Validation

The datasets ensure completeness by minimizing temporal gaps in the two-year data collection and the necessary sensors are employed to capture critical data of the building such as temperature, weather, system operations, and window controls.

In this section, examples of data processing results and data samples from different sensors and meters are presented to demonstrate the data quality and coverage of the dataset. To ensure the validity and soundness of the dataset, the raw data has been processed using the methods as described in the Methods section. Figures 6–7 show the examples of detection of outliers and data repetition from the raw data, which demonstrate the efficacy of our data processing methods, underscoring the dataset's reliability.

The pie charts in Fig. 8 depict the breakdown of energy end uses over the course of two years. The annual energy consumptions are 38.5 kWh/m² and 36.3 kWh/m² for Years 1 and 2, respectively. In both years, the IT load emerged as the most dominant, with heating and plug load following suit consistently throughout the two years.

Figures 9 and 10 present the daily load trends observed during sample summer and winter days in Year 1. In Year 1, the cooling loads were consistently maintained at an average of approximately 0.2 kWh. The largest proportion of the total loads during summer was attributed to IT and plug loads. Meanwhile, cooling, control, and others exhibited a constant energy demand. During the winter season, heating energy consumption emerged as the highest-demand load and IT loads represented the second-largest contributor to the total load during winter. The other loads exhibited stability and remained predominantly similar during both typical summer and winter days.

BTU meters measured the water flowrates (in gallons per minute, GPM), energy rates (in kilo British Thermal Units per hour, kBTU/h), and supply and return water temperatures. Figure 11 illustrates the energy rate data from BTU meters in different zones in a sample winter day as an example. The energy rate data reflected

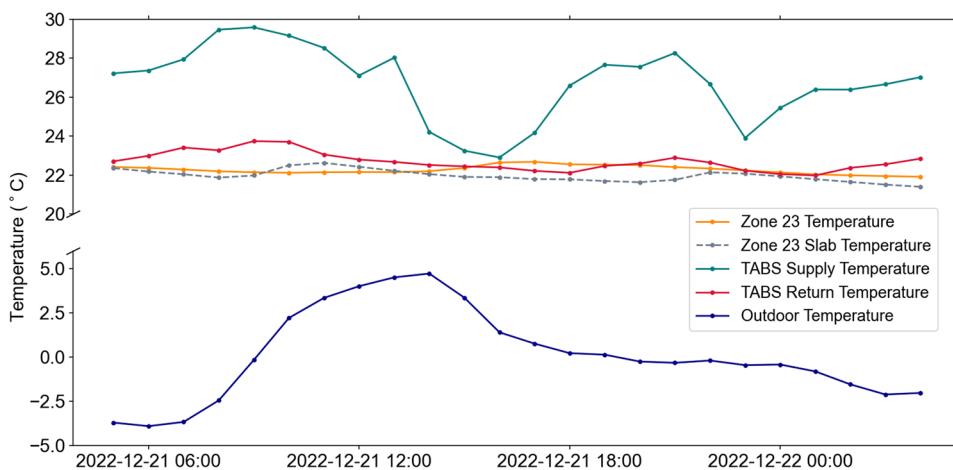


Fig. 17 Operation of the TABS in a sample winter day of Zone 23.

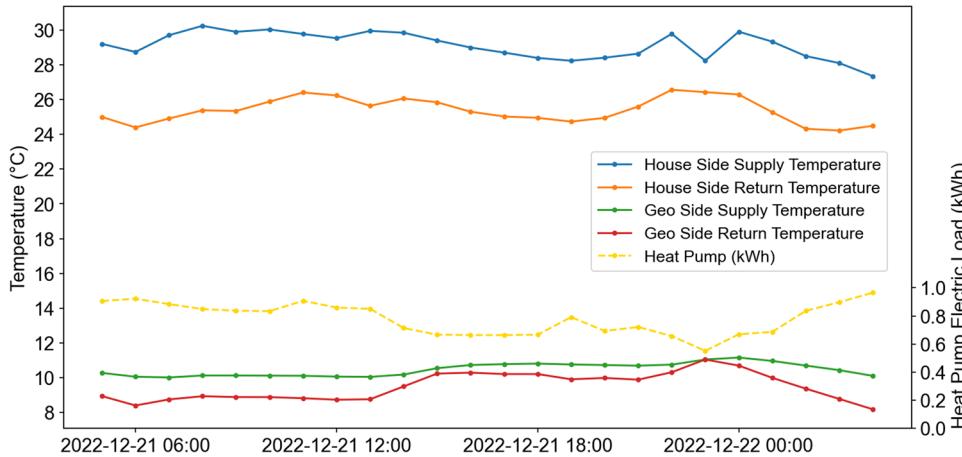


Fig. 18 Operation of the heat pump in a sample winter day.

the heating or cooling demand of different zones. In winter, the variation of the heating demand depends on the slab temperature and the slab temperature setpoint.

As shown in Fig. 12, the daily pattern of the indoor CO₂ concentration exhibits a diurnal cycle with varying concentration levels throughout the day across multiple data series. For most of the zones, the CO₂ level rose in the morning as occupants entered the office, reached peak in the afternoon, and decreased from the evening. The pattern indicates a potential correlation with daily occupancy profile in the building.

Figure 13 presents the relative humidity (RH, %) across various zones in a sample summer day. The RH values remained relatively constant, with slight fluctuations during the day. The pattern showed a potential correlation with the occupant schedule.

Figure 14 illustrates the temperature trends and CO₂ variation with natural ventilation during the passive mode. The windows were controlled based on indoor and outdoor air temperatures as well as indoor CO₂ concentration to maintain the indoor air temperature and CO₂ concentration within the comfortable range or acceptable level. During nighttime, the windows were operated for free cooling with night flushing to further improve the energy efficiency.

Figure 15 shows the monthly PV production and solar radiation in Year 1. Overall, the PV production and solar radiation follow the same trend throughout the year. A new inverter with higher efficiency was installed in early 2023, which contributed to the improved PV efficiency from March 2023.

Figure 16 shows how TABS operated on a sample summer day. The cold water with an average temperature of 19.6°C drawn from the geothermal well was directly supplied to the building and circulated through the piping systems in the slab. The mean temperature of the return water that carried the heat away from indoors was about 21.1°C. The slab temperature is slightly lower than the return water temperature. Thus, the indoor temperature was maintained between 24–26°C.

Figure 17 shows how TABS operated on a sample winter day. The hot water with an average temperature of 26.7°C generated from the ground source heat pump was supplied to the slab-embedded piping systems. The supply water temperature was controlled as a function of the outdoor temperature. The water returned at a temperature of around 22.7°C. The slab temperature and indoor temperature remained at approximately 22°C and stayed within the comfort zone.

Figure 18 shows how the heat pump operated on a sample winter day. Detailed operational performance of the heat pump was monitored, including the power consumption, supply and return water temperature as well as flowrates for both house side and geo side, which supports the analysis of heat pump operations and development of prediction models and advanced controls. Part of the data was illustrated in this figure as an example.

This dataset contributes to the development of sustainable building practices from the following aspects. Firstly, the dataset provides a building energy benchmark for an ultra-efficient building with integrated low-energy and passive building technologies. Secondly, the dataset provides a better understanding of the operations of an ultra-efficient building with integrated natural ventilation and geo-powered TABS systems with detailed operational data. This may include load shape analysis, energy prediction, system operational pattern analysis as well as data-driven modeling. In addition, it can also be used for validation of building simulation models and development of learning-based control algorithms. The operation data of the building demonstrates the effectiveness of NV combined with TABS for maintaining a comfortable indoor environment while achieving high energy efficiency, which help promote the wider adoption of such low-energy technologies.

There are some usage restrictions and limitations for the usage of this dataset due to the specific boundaries of the building. First, the dataset is dependent on the configurations and characteristics, such as occupancy, design, and envelope/materials, of the subject building. Second, the dataset is dependent on local weather conditions under the 5A climate zone. Third, the dataset is dependent on sensor settings, such as locations and resolution/accuracy.

Code availability

The Python code for data processing is available at <https://github.com/Harvard-CGBC-at-GSD/HouseZero-two-year-dataset-June-2022-May-2024-Processing-Code>.

Received: 11 September 2023; Accepted: 14 August 2024;

Published online: 28 August 2024

References

1. Agency, I. E. Transition to sustainable buildings: strategies and opportunities to 2050. (2013).
2. Zhang, L. *et al.* A review of machine learning in building load prediction. *Appl. Energy* **285**, 116452 (2021).
3. Yan, D. *et al.* Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy Build.* **107**, 264–278 (2015).
4. Chen, E. X., Han, X., Malkawi, A. & Li, N. Ensembled Deep Learning-based Model Predictive Control for Automatic Window Operations in Winter. in *2023 ASHRAE Winter Conference* (Atlanta, Georgia, 2023).
5. Chen, E. X., Han, X., Malkawi, A., Zhang, R. & Li, N. Adaptive model predictive control with ensembled multi-time scale deep-learning models for smart control of natural ventilation. *Build Environ.* **110**, 519 (2023).
6. Han, X. & Malkawi, A. Model-Free Reinforcement Learning-Based Control for Radiant Floor Heating Systems. in *the 5th International Conference on Building Energy and Environment (COBEE 2022)* (Montreal, CA, 2022).
7. Yan, B. *et al.* Comprehensive Assessment of Operational Performance of Coupled Natural Ventilation and Thermally Active Building System via an Extensive Sensor Network. *Energy Build.* **111**, 921 (2022).
8. Han, J. M. *et al.* Data-informed building energy management (DiBEM) towards ultra-low energy buildings. *Energy and Buildings* **281**, 112761 (2023).
9. Schlemminger, M., Ohrdes, T., Schneider, E. & Knoop, M. Dataset on electrical single-family house and heat pump load profiles in Germany. *Scientific Data* **9**, 1–11 (2022).
10. Kriechbaumer, T. & Jacobsen, H.-A. BLOND, a building-level office environment dataset of typical electrical appliances. *Scientific data* **5**, 1–14 (2018).
11. Schlemminger, M., Ohrdes, T., Schneider, E. & Knoop, M. Dataset on electrical single-family house and heat pump load profiles in Germany. *Zenodo* <https://doi.org/10.1038/s41597-022-01156-1> (2022).
12. Kriechbaumer, T. & Jacobsen, H.-A. BLOND: Building-Level Office eNvironment Dataset [Dataset]. *mediaTUM* <https://doi.org/10.14459/2017mp1375836> (2017).
13. Jacoby, M., Tan, S. Y., Henze, G. & Sarkar, S. A high-fidelity residential building occupancy detection dataset. *Scientific Data* **8**, 1–14 (2021).
14. Dong, B. *et al.* A global building occupant behavior database. *Scientific data* **9**, 1–15 (2022).
15. Jacoby, M., Tan, S. Y., Henze, G. & Sarkar, S. A high-fidelity residential building occupancy detection dataset. *Figshare* <https://doi.org/10.6084/m9.figshare.c.5364449> (2021).
16. Dong, B. *et al.* A global building occupant behavior database. *Figshare* <https://doi.org/10.6084/m9.figshare.16920118.v6> (2022).
17. Tekler, Z. D. *et al.* in *Building Simulation*. 2127–2137 (Springer).
18. Hong, T., Luo, N., Blum, D. & Wang, Z. A three-year dataset supporting research on building energy management and occupancy analytics [Dataset]. *Dryad* <https://doi.org/10.7941/D1N33Q> (2022).
19. Luo, N. *et al.* A three-year dataset supporting research on building energy management and occupancy analytics. *Scientific Data* **9**, 156 (2022).
20. Agee, P., Nikdel, L. & Roberts, S. A measured energy use, solar production, and building air leakage dataset for a zero energy commercial building. *Scientific Data* **8**, 1–8 (2021).
21. Agee, P. & Nikdel, L. An energy use, energy production, and building air leakage dataset for a zero energy commercial building. *Open Science Framework* <https://doi.org/10.17605/OSF.IO/3KDHQ> (2021).
22. Schweiker, M., Kleber, M. & Wagner, A. Long-term monitoring data from a naturally ventilated office building. *Scientific data* **6**, 1–6 (2019).
23. Schweiker, M., Kleber, M. & Wagner, A. Long-term monitoring data from a naturally ventilated office building. *Open Science Framework* <https://doi.org/10.17605/OSF.IO/2YDZG> (2019).
24. Malkawi, A. *et al.* Design and Applications of an IoT Architecture for Data-Driven Smart Building Operations and Experimentation. *Energy Build.* (2023).
25. Hawkins, D. M. *Identification of outliers*. Vol. 11 (Springer, 1980).
26. Saleem, S., Aslam, M. & Shaukat, M. R. A review and empirical comparison of univariate outlier detection methods. *Pakistan Journal of Statistics* **37** (2021).
27. Kannan, K. S., Manoj, K. & Arumugam, S. Labeling methods for identifying outliers. *International Journal of Statistics and Systems* **10**, 231–238 (2015).

28. Ghorbani, H. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583–595 (2019).
29. Han, J. *et al.* HouseZero® two year dataset (June 2022–May 2024). *figshare*. Journal contribution. <https://doi.org/10.6084/m9.figshare.26499595.v2> (2024).

Acknowledgements

The authors would like to thank Mayuri Rajput for the help with cross-checking the data.

Author contributions

Jung Min Han led data processing and the development of the dataset, managed data curation of the dataset, and wrote and edited the manuscript. Ali Malkawi supervised the research effort, contributed to conceptualization, design of the architecture of the dataset, as well as writing and editing the manuscript. Xu Han led the initial draft of the manuscript, wrote and edited the manuscript. Sunghwan Lim managed raw data collection, participated in development of the dataset, wrote and edited the manuscript. Elence Xinzu Chen participated in development of the dataset, wrote and edited the manuscript. Sang Won Kang participated in development of the dataset and edited the manuscript. Yiwei Lyu participated in development of the dataset and edited the manuscript. Peter Howard participated in the development of the dataset, and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.H.

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) 2024