SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition

Clayton Miller[1]✉, Anjukan Kathirgamanathan[2], Bianca Picchetti[3], Pandarasamy Arjunan[4], June Young Park[5], Zoltan Nagy[5], Paul Raftery[6], Brodie W. Hobson[7], Zixiao Shi[7] & Forrest Meggers[8]

This paper describes an open data set of 3,053 energy meters from 1,636 non-residential buildings with a range of two full years (2016 and 2017) at an hourly frequency (17,544 measurements per meter resulting in approximately 53.6 million measurements). These meters were collected from 19 sites across North America and Europe, with one or more meters per building measuring whole building electrical, heating and cooling water, steam, and solar energy as well as water and irrigation meters. Part of these data was used in the Great Energy Predictor III (GEPIII) competition hosted by the American Society of Heating, Refrigeration, and Air-Conditioning Engineers (ASHRAE) in October-December 2019. GEPIII was a machine learning competition for long-term prediction with an application to measurement and verification. This paper describes the process of data collection, cleaning, and convergence of time-series meter data, the meta-data about the buildings, and complementary weather data. This data set can be used for further prediction benchmarking and prototyping as well as anomaly detection, energy analysis, and building type classification.

Background & Summary

Building performance analytics and commissioning processes have significant opportunities to save energy, reduce carbon emissions of buildings, and reduce the operating costs of building owners world-wide[1]. Machine learning and prediction techniques are a vital component of many of the ways of finding savings opportunities and quantifying the risk and reward of undertaking such efforts. Despite the significant research body of knowledge developed, there is still a lack of understanding of how to scale techniques across the highly heterogeneous building stock[2]. When it comes to machine learning innovation in academia, one of the most significant assets can be large and open data sets that the community can use to prototype and quantitatively compare techniques in ways that show better value in terms of speed, accuracy, or implementation ease. This statement is supported by the significant efforts in time-series data classification[3], image recognition[4], and the larger machine learning community in general, both hardware and software[5].

The building energy analytics community has only just started to use open data sets towards the efforts of creating benchmarking data sets. Several prominent open building energy-related data sets have been released in

[1]Building and Urban Data Science (BUDS) Lab, School of Design and Environment (SDE), National University of Singapore (NUS), 4 Architecture Drive, Singapore, 117566, Singapore. [2]UCD Energy Institute, O'Brien Science Building, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland. [3]Gerencia del Ciclo de Combustible Nuclear, Comisión Nacional de Energía Atómica, Avenida General Paz 1499, Buenos Aires, 1650, Argentina. [4]Berkeley Education Alliance for Research in Singapore (BEARS), 1 Create Way, #11-01, CREATE Tower, Singapore, 138602, Singapore. [5]Intelligent Environments Lab (IEL), Department of Civil, Architectural and Environmental Engineering, Cockrell School of Engineering, The University of Texas at Austin, 301 E Dean Keeton Street St C1700, Austin, TX, 78712, USA. [6]Center for the Built Environment, University of California, 390 Wurster Hall, Berkeley, CA, 94720, USA. [7]Department of Civil and Environmental Engineering, Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada. [8]CHAOS Laboratory, School of Architecture, Princeton University, 86 Olden St, Princeton, NJ, 08540, USA. ✉e-mail: clayton@nus.edu.sg

Site	UID	Kaggle	Actual Site Name	Location	Climate	Buildings	Meters
Panther	1P4YFG	0	Univ. of Central Florida (UCF)	Orlando, FL	2A	136	299
Robin	1TKL5P	1	Univ. College London (UCL)	London, UK	4A	52	67
Fox	4QFLSM	2	Arizona State Univ. (ASU)	Tempe, AZ	2B	137	306
Rat	72SGIQ	3	Washington DC - City Buildings	Washington DC	4A	305	305
Bear	7E44IQ	4	Univ. of California - Berkeley	Berkeley, CA	3C	92	92
Lamb	9T5ZA2	5	Cardiff - City Buildings	Cardiff, UK	4A	147	265
Eagle	EQDHIP	6	Anonymous	N/A	4A	47	106
Moose	H7PNXU	7	Ottawa - City Buildings	Ottawa, Ontario	6A	15	43
Gator	I9U4WZ	8	Anonymous	N/A	2A	74	74
Bull	JG98YH	9	Univ. of Texas - Austin	Austin, TX	2A	124	308
Bobcat	JP4TNW	10	Anonymous	N/A	5B	36	116
Crow	JTM0LY	11	Carleton Univ.	Ottawa, Ontario	6A	5	15
Wolf	RFO3TV	12	Univ. College Dublin (UCD)	Dublin, Ireland	5A	36	66
Hog	SREOJG	13	Anonymous	Anonymous	6A	163	336
Peacock	WI83D6	14	Princeton University	Princeton, NJ	5A	106	298
Cockatoo	YYAFES	15	Cornell University	Cornell, NY	6A	124	282
Shrew	L2HJLD	—	UK Parliment	London, UK	4A	9	13
Swan	N950XM	—	Anonymous	N/A	3C	21	55
Mouse	ZVJUMW	—	Ormand Street Hospital	London, UK	4A	7	7

Table 1. Overview of the sites from which the building energy meter data was collected. Each site is given an animal-like site code name, a UID that corresponds to some of the data convergence processes, the Kaggle Site ID that was included in the competition, and the Actual Site Name, Location and Climate Zone. Several of the sites are to remain anonymous based on discussions with the data donors. The last two columns indicate the number of buildings and meters where two years of hourly, whole building meter data were collected from each site.

recent years including applications to building-level office[6] and residential[7] appliances, occupant behavior[8], heat pump[9] and natural ventilation systems[10], as well as commercial and residential energy meter data[11–13]. The use of open data sets in the built environment enables the analysis of large numbers of buildings in applications such as benchmarking[14]. From the machine learning perspective, there have also been efforts towards using large data sets to benchmark various machine learning techniques as applied to building energy performance analytics[15].

This paper focuses on the development of a data set that builds upon these motivations. The data set is part of the *Building Data Genome Project*, an international consortium of building energy-related academics and practitioners who seek to create large, open data sets that increase the understanding of the foundations of building behavior and energy use in buildings. The first phase of the project had a data set that was released in 2017 and included one year of hourly data from over 500 buildings[16].

The newest version of the data set is described in this publication as the Building Data Genome Project 2 data set. This open data repository has data from 1,636 non-residential buildings. It includes hourly whole-building data for two years, from different kinds of meters: electricity, chilled water, steam, hot water, gas, water, irrigation, and solar. The hourly frequency for the data set was targeted as it provides enough resolution to support analytics techniques targeting several scales, including daily, weekly, monthly, seasonal, and annual patterns of use. Each of the buildings has metadata such as area, weather, and primary use type collated. This data set can be used to benchmark various statistical learning algorithms and other data science techniques. It can also be used merely as a teaching or learning tool to practice dealing with measured performance data from large numbers of non-residential buildings. This data set was collected from 19 different locations from around the world. These locations, climates, and the number of buildings from each site are found in Table 1. This table also includes information about which buildings were used in the ASHRAE-sponsored Great Energy Predictor III (GEPIII) competition that was held on the Kaggle platform from October to December 2019 (https://www.kaggle.com/c/ashrae-energy-prediction)[17]. These buildings represent several different primary use type categories from several industries. Figure 1 illustrates the breakdown of the buildings according to the principal use category and subcategory, industry and sub-industry, timezone, and meter type. The remaining parts of this paper focus on how the data were collected, processed, and how users can find and use the data for several example applications.

Methods

Energy data sources overview and collection. The collection of the metadata and whole building meter data from the various sites outlined in Table 1 was done by the authors of this paper from September 2017 until May of 2019. Eight of the sites from this list are online data sources that are freely downloadable without the use of login credentials. These sites are considered open access data sources and are publicly available. Table 2 outlines these eight sites and the online link to the main interface for downloading the data. The remaining eleven sites did not have online, publicly available data feeds. In those situations, there were facilities management professionals involved in the process of data collection and organization for those subsets. Data collection from these sites was

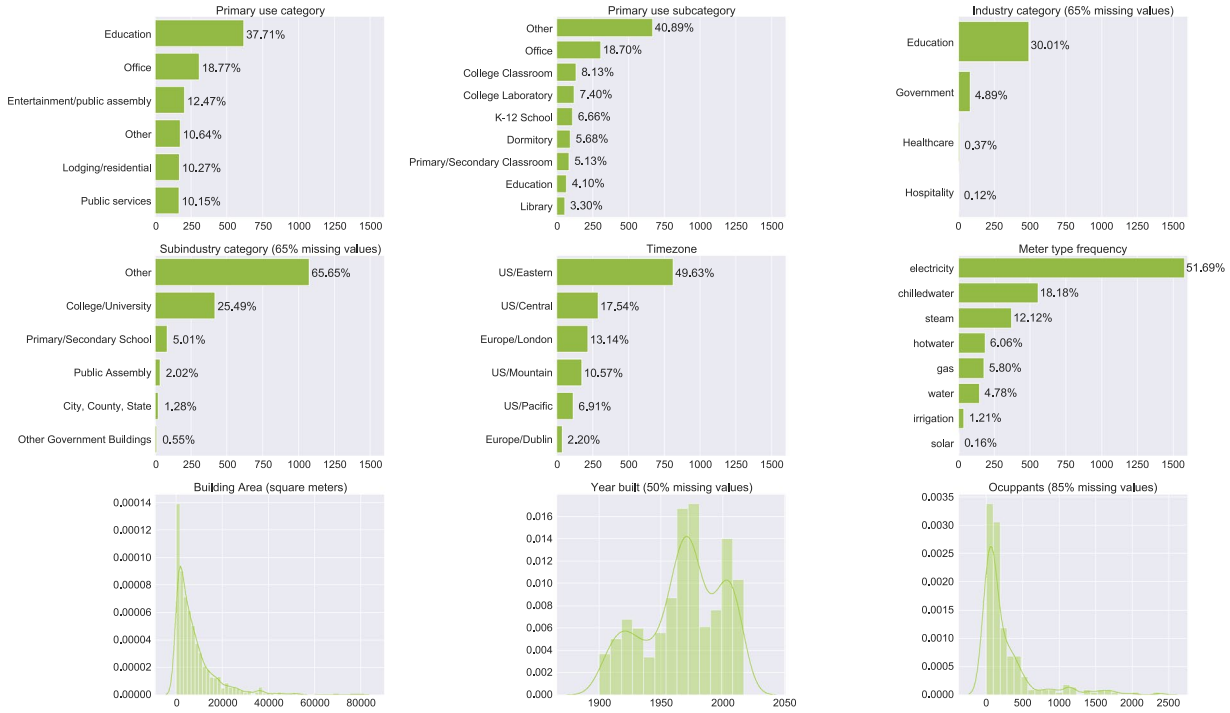

Fig. 1 Main features distribution in metadata file that describes the various buildings from which the meter data was collected. Several of the meta-data categories are available for all buildings including the Primary Use Category of the building (`primaryspaceusage`), the Sub-primary Use Category (`subprimaryspaceusage`), Gross Floor Area (`sqm`), Time Zone (`timezone`), Weather Data, and Meter Type.

Site	Actual Site Name	Online source
Panther (0)	Univ. of Central Florida (UCF)	http://oeis.ucf.edu/
Robin (1)	Univ. College London (UCL)	https://platform.carbonculture.net/communities/ucl/30/
Fox (2)	Arizona State Univ. (ASU)	https://cm.asu.edu/
Bear (4)	UC Berkeley (UCB)	https://engagementdashboard.com/ucb/ucb/
Lamb (5)	Cardiff - City Buildings	https://platform.carbonculture.net/communities/cardiff-council/19/
Cockatoo (15)	Cornell University	https://portal.emcs.cornell.edu/
Shrew	UK Parliment	https://platform.carbonculture.net/communities/uk-parliament/2/
Mouse	Ormand Street Hospital	https://platform.carbonculture.net/communities/great-ormond-street-hospital/4/

Table 2. Sites with data that are publicly available to download online. The site name includes the Kaggle ID in parentheses.

a manual process that included site visits, in-person meetings, and data collection workshops, numerous digital communications via video calls and emails. The raw meter data for these sites were downloaded and provided to the technical team, usually via emailing flat files. These raw data sources are not included in the data repository; however, the process of convergence, cleaning, and normalization is included in this paper's subsequent subsections.

Weather data overview and collection. One of the critical comparative data sources for building energy meter data is outside weather conditions, which are among the key influencing factors for energy consumption in buildings. Each of the building sites has a corresponding weather data file with hourly data related to the outdoor temperature, humidity, cloud cover, and other conditions that influence energy consumption. Hourly weather data for this data set were collected using the National Centers for Environmental Information (NCEI) National Oceanic and Atmospheric Administration (NOAA) Integrated Surface Database (ISD) (https://www.ncdc.noaa.gov/isd). The ISD-Lite version was used for easy hourly data capture. The closest station with available data for the period 2016–2017 was selected for each site, as outlined in Table 3. The ISD-Lite data set includes the eight climatological variables for each station with a modified timestamp, which corresponds to the nearest hour of actual observation. In the preparation step for this data set, scaling (where applied) was removed and missing

Site	ISD Station Code
Panther (0)	722050-12815
Robin (1)	037720-99999
Fox (2)	722780-23183
Rat (3)	724050-13743
Bear (4)	724930-23230
Lamb (5)	037150-99999
Moose (7)	710630-99999
Bull (9)	722544-13958
Crow (11)	710630-99999
Wolf (12)	039690-99999
Peacock (14)	724095-14792
Cockatoo (15)	725155-94761
Shrew	037720-99999
Mouse	037720-99999

Table 3. ISD weather station data sources for the non-anonymous sites. The site name includes the Kaggle ID in parentheses.

values were processed to be NaN instead of −9999 as per the raw data. The final processed weather data is summarised in Fig. 2.

Data cleaning and normalization. After collection of the raw data from the sites and weather sources, the data were transformed in ways that create consistency and uniformity across the data sets so that they could be converged into one large data set. These steps were completed in a private, non-public data repository as the preparation for the data was done in the Kaggle GEPIII competition context. These data and processes were kept secure as the premature release of the data would have compromised the competition's integrity. This subsection describes those steps used to create both the data set for the competition and this data repository.

The first step in this process was the normalization of measurement units for the various energy meter types. Table 4 summarises the original measurement units for the raw data collected from every site. A conversion process was undertaken to convert into standard units for every meter type, as outlined in Table 5. Following the standardization of the units, a few additional steps were undertaken to clean and process the data. All meters with only a single value were removed, duplicate meter data (if present) were removed, and negative meter values were replaced with `NaN`. Where there were more than 50% of negative meter readings, this meter was also removed. This step removes the possibility of including meters from net-zero energy buildings, although we are not aware that there were any of these buildings in the data set specifically. Meters with significant consecutive missing values (over 100 consecutive days) were excluded. There were still meters with very high-value outliers, and in this case, standard outlier removal techniques won't work as these outliers are large enough to skew measures such as the mean. A log conversion and pruning technique were used with any high outliers greater than three standard deviations from the mean on the log-transformed data converted to `NaN` values. Finally, all meter data was rounded to four decimal places.

For the metadata of the buildings, where necessary, floor area in `sqm` and `sqft` was converted from whichever floor area data was available. Latitude and longitude data were set to the central location of either the site or the city in which the site is located. In all cases, all buildings are within a 25-mile (40-kilometer) radius of the central location of the site or city. For the `year_built` attribute, a valid range was considered to be 1900 to 2018, and invalid or implausible years were filled as missing values. Primary space usage (`primary_use`) metadata for all buildings was mapped using the Energy Star scheme building description types. Based upon the meter and metadata as described above, a further filter was done to synchronize both sets of data and remove meter data for which the building metadata did not exist and likewise remove metadata for which meter data did not exist.

Data Records

This section documents the file types and structure for the data set (https://github.com/buds-lab/building-data-genome-project-2) that has a v1.0 release deposited in Zenodo[18]. The following subsections outline the data files that can be found in the repository to guide their use. Each building in the data set can be connected to this publication through its *Unique Site Identifier* that was created with the following structure: *animal name* (unique per site) + *primary space usage abbreviation* + *Human-like name* (unique per building). An example of a building name is *Raven_Education_Nina*.

Building metadata. The building meta data file (`data/metadata/metadata.csv`) contains information about the whole building characteristics that enable the analysis of the associated meter data with various aspects of the building such as floor area, weather, and primary use type. These data were collected either from the operations teams from which the data was collected or from descriptors from the online data portals if collected from a public data source. Only the attributes for building unique identifier (`building_id`), site identifier (`site_id`), floor area (`sqft` and `sqm`), and time zone (`timezone`) are found for all the buildings. The remaining meta data descriptors have missing value rates from 4–99%. A more detailed overview of these attributes can be found in the repository

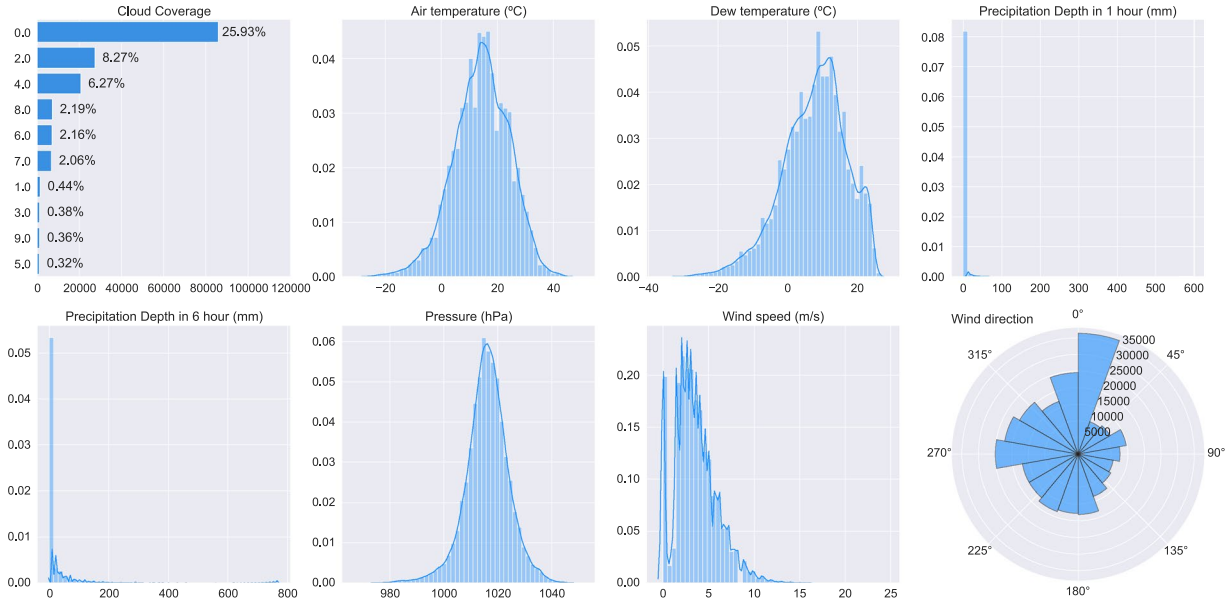

Fig. 2 Main feature distributions of the weather data set.

Site	Chilled water	Electricity	Gas	Hot water	Solar	Steam	Water	Irrigation
Panther (0)	kBTU*sum*	kBTU*sum*	$kBTU_{sum}$				gallons	gallons
Robin (1)		kWh_{sum}						
Fox (2)	$Tons_{avg}$	kW_{avg}		$mmBTU_{sum}$				
Rat (3)		kWh_{sum}						
Bear (4)		kW_{avg}						
Lamb (5)		kWh_{sum}	kWh_{sum}					
Eagle (6)	$mmBTU_{sum}$	kW_{avg}		$mmBTU_{sum}$		$lbs_{perhour}$		
Moose (7)	KJ	KJ						
Gator (8)		kWh_{avg}						
Bull (9)	$Tons_{sum}$	kWh_{sum}				$lbs_{perhour}$		
Bobcat (10)	kBTU	kWh_{sum}	kBTU	kBTU	kWh_{sum}		gallons	
Crow (11)	kWh_{sum}	kWh_{sum}	kWh_{sum}					
Wolf (12)		kWh_{sum}	m3				liters	
Hog (13)	$Tons_{avg}$	kWh_{avg}				$lbs_{perhour}$		
Peacock (14)	$Tons_{avg}$	kW_{avg}				$lbs_{perhour}$		
Cockatoo (15)	$Tons_{avg}$	kW_{avg}		$Tons_{avg}$		$lbs_{perhour}$		
Shrew		kWh_{sum}	kWh_{sum}					
Swan	$Tons_{avg}$	kWh_{sum}		$kBTU_{sum}$				
Mouse		kWh_{sum}						

Table 4. Overview of original measurement units for the raw data collected from each site. All data were subsequently converted to kWh_{sun} or liters. The site name includes the Kaggle ID in parentheses.

documentation (https://github.com/buds-lab/building-data-genome-project-2/wiki/Metadata-description). The following are the attributes or column headings and the description of the data found in the file:

- `building_id`: building code-name with the structure - *_UniqueSiteID _primaryspaceusage _UniqueFirstName*.
- `site_id`: animal-code-name for the site.
- `primaryspaceusage`: Primary space usage of all buildings is mapped using the Energy Star scheme building description types.
- `sqft`: Floor area of building in square feet (ft^2).
- `lat`: Latitude of building location to city level. This attribute is available for all non-anonymous locations.
- `lng`: Longitude of building location to city level.This attribute is available for all non-anonymous locations.
- `electricity`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `hotwater`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `chilledwater`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.

Unit	Conversion Factor
kW_{avg}	1 kWh_{sum} = kW_{avg} * 1
tons	1 kWh_{sum} = tons * 3.51685
kBTU	1 kWh_{sum} = kBTU * 0.293071
MJ	1 kWh_{sum} = MJ * 3.6
mmBTU	1 kWh_{sum} = mmBTU * 293.071
therm	1 kWh_{sum} = therm * 29.3071
cubic meter gas	1 kWh_{sum} = cubic meter * 11.4772
lb/hour steam	1 kWh_{sum} = lb/hour * 0.305
gallons	1 liter = gallons * 0.264172

Table 5. Overview of measurement unit conversion process. All energy-related meters were converted to *kWh_{sum} or liters* from the various raw data units.

- `steam`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `water`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `irrigation`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `solar`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `gas`: Presence of this kind of meter in the building. `Yes` if affirmative, `NaN` if negative.
- `yearbuilt`: Year corresponding to when building was first constructed, in the format `YYYY`.
- `numberoffloors`: Number of floors corresponding to building.
- `date_opened`: Date building was opened for use, in the format `D/M/YYYY`.
- `sub_primaryspaceusage`: Energy Star scheme building description types subcategory.
- `energystarscore`: Rating of building corresponding to building Energy Star scheme (https://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager/understand-metrics/how-1–100).
- `eui`: Energy use intensity of the building collected from asset management data sources from the data donors. This metric is calculated from the utility bills of the building from data beyond the range of this data set, therefore there may be discrepancies from EUI's calculated in this data set. (kWh/year/m^2) (https://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager/understand-metrics/what-energy).
- `heatingtype`: Type of heating in corresponding building.
- `industry`: Industry type corresponding to building.
- `leed_level`: LEED rating of the building (https://www.usgbc.org/leed/).
- `occupants`: Design condition number of occupants in the building.
- `rating`: Other building energy ratings.
- `sqm`: Floor area of the building in square meters (m^2).
- `subindustry`: More detailed breakdown of Industry type corresponding to building.
- `timezone`: Site time zone.

Weather data. The building weather data file (`data/weather/weather.csv`) contains the time-series data for each building as it corresponds to the energy meters. These data have a time range from January 1, 2016, to December 31, 2017 - the same as the meter data files. A more detailed overview of these data can be found in the repository documentation (https://github.com/buds-lab/building-data-genome-project-2/wiki/Weather-Description). The following are the attributes or column headings and the description of the data found in the file:

- `timestamp`: Date and Time in the format `YYYY-MM-DD hh:mm:ss` in the local timezone.
- `site_id`: human name-animal-code-name unique identifier for the site.
- `airTemperature`: The temperature of the air in degrees Celsius (°C).
- `cloudCoverage`: Portion of the sky covered in clouds, in oktas (https://en.wikipedia.org/wiki/Okta).
- `dewTemperature`: The dew point (the temperature to which a given parcel of air must be cooled at constant pressure and water vapor content for saturation to occur) in degrees Celsius (°C).
- `precipDepth1HR`: The depth of liquid precipitation measured over a one hour accumulation period (mm).
- `precipDepth6HR`: The depth of liquid precipitation that is measured over a six-hour accumulation period (mm).
- `seaLvlPressure`: The air pressure relative to Mean Sea Level (MSL) (mbar or hPa).
- `windDirection`: The angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing (degrees).
- `windSpeed`: The rate of horizontal travel of air past a fixed point in (m/s).

Meter data. There are three sets of meter data found in the repository. The first is the *raw* data set that includes the most substantial data set that was formed after convergence of the data from each source and the initial cleaning, unit conversion, and other processing steps outlined in the Data Cleaning and Normalization

Section. The *cleaned* data set provides a data set with another phase of cleaning and processing described below. Finally, there is a data set that includes the 2017 data that matches with the *Kaggle* competition. This data set is included as several updates and conversions were performed on the BDG data sets after the competition. An overview of the differences between these data sets can be found in the repository documentation (https://github.com/buds-lab/building-data-genome-project-2/wiki/Meters-data-features).

Raw meter data. There are eight files containing the time-series data for each building meter type. These files contain a column for each building in the data set for that particular meter. These files are contained in the /data/meters/raw/ folder and includes the files electricity.csv, `hotwater.csv`, `chilledwater.csv`, `steam.csv`, `water.csv`, `irrigation.csv`, `solar.csv` and `gas.csv`. Each data file contains the data timestamp as the initial row in the format `YYYY-MM-DD hh:mm:ss` in the local timezone and one column per building in the data set in the units kWh_{sum} for the energy-related meters and *liters* for the non-energy meters. Each row represents one hour, and the reading is the energy or water sum across that hour. These data have a time range from January 1, 2016, to December 31, 2017 - the same as the weather data files. A more detailed overview of these data can be found in the repository documentation.

Cleaned meter data. This folder content and structure (`/data/meters/cleaned/`) is similar to the *raw* data folder, however, more outliers have been removed using the Twitter *AnomalyDetection* R library (https://github.com/twitter/AnomalyDetection), zero readings longer than 24 continuous hours are removed, and zero readings in electricity meters are removed.

Kaggle public test/validation data. This folder (`/data/meters/kaggle/`) includes a single file that contains the 2017 data of all the meters and sites from the GEPIII competition that was used as the public test/validation data set. This file can be used by those seeking to make a comparison to the training data found provided by the competition website. It can be used to train models and make submissions for the final score test data set (private leaderboard). This data set is provided as the other Building Data Genome 2 data sets have been transformed since the competition. This original form allows users not to have to reverse those transforms to use the data in the competition. More details of the connection from this repository and the competition can be found in the Usage Notes section.

Technical Validation

To illustrate to potential users the usefulness of the Building Data Genome 2 data set, several data quality screening techniques have been applied to the time-series meter data to show an overview of the normalized consumption patterns across the data set, the completeness and quality of the data, the relationship between the weather and meter data, and the volatility of the data in terms of shifts in steady-state. Each of these screening techniques was developed and applied to the previous BDG1 data set in earlier work[19]. These screening techniques are designed to validate the technical capacity for the data sets to meet the needs of various applications. A more detailed overview of the screening process can be found in the repository documentation and in the Usage Notes section (https://github.com/buds-lab/building-data-genome-project-2/wiki/Meters-data-screening).

Normalized consumption. The first screening technique applied is to visualize the meter data from a high level in a normalized way to see the general patterns and fluctuations across the data. The first step in this process is the summation of the hourly data across each day. The daily totals are then normalized once by dividing by the floor area (`sqm`) and then normalized again by scaling to the maximum and minimum for the time range for each meter data set. Figure 3 illustrates the panel of the eight-meter types with this screening process applied. This figure illustrates each meter type in its own heat map where the horizontal axis for each heatmap is the two year period, and the vertical axis represents all of the meters for each category sorted from top to bottom according to the metric. This visualization technique is used in Figs. 3–6. For the normalized energy consumption technical validation, the various meters have seasonal, cyclical patterns that are apparent for a certain range of each meter type.

Data quality. The next screening technique applied is a set of filters applied to the time-series data from the meters to categorize four different types of readings of the data: missing data, data with a reading of zero, outliers, and the remaining data that can be considered the most informational (labeled as *Good Data*). This process was applied to all the meter data sets, as shown in Fig. 4. The outliers for the heat map are calculated using the Twitter *AnomalyDetection* R library (https://github.com/twitter/AnomalyDetection). The resultant heat maps show a small percentage of the meters have a significant amount of missing data in certain time frames. These gaps are considered normal in meter data sets and can be the result of numerous technical or data collection issues. These data may also mean that the building was offline during certain periods. The meters still met the criteria to be included in the data set despite these gaps; therefore, the gaps are under a certain percentage of the overall data set as defined in earlier sections. The visualization also shows that there are a significant amount of zero readings for certain meter types, such as those related to heating, cooling, and irrigation. These zero measurement values make sense in those contexts, and these data are likely to be useful as they demonstrate periods when those systems are not in use. The screening shows few outliers as most of those data were filtered in previous cleaning steps outlined previously.

Weather data sensitivity. The next screening process illustrates and validates the relationship between the meter data and the associated weather data files that are included in the repository. This validation step shows the value of providing these data sets in tandem and the influence of weather on buildings' energy consumption. This metric is calculated by taking a cleaned version of the data set in which days with only zero readings are removed

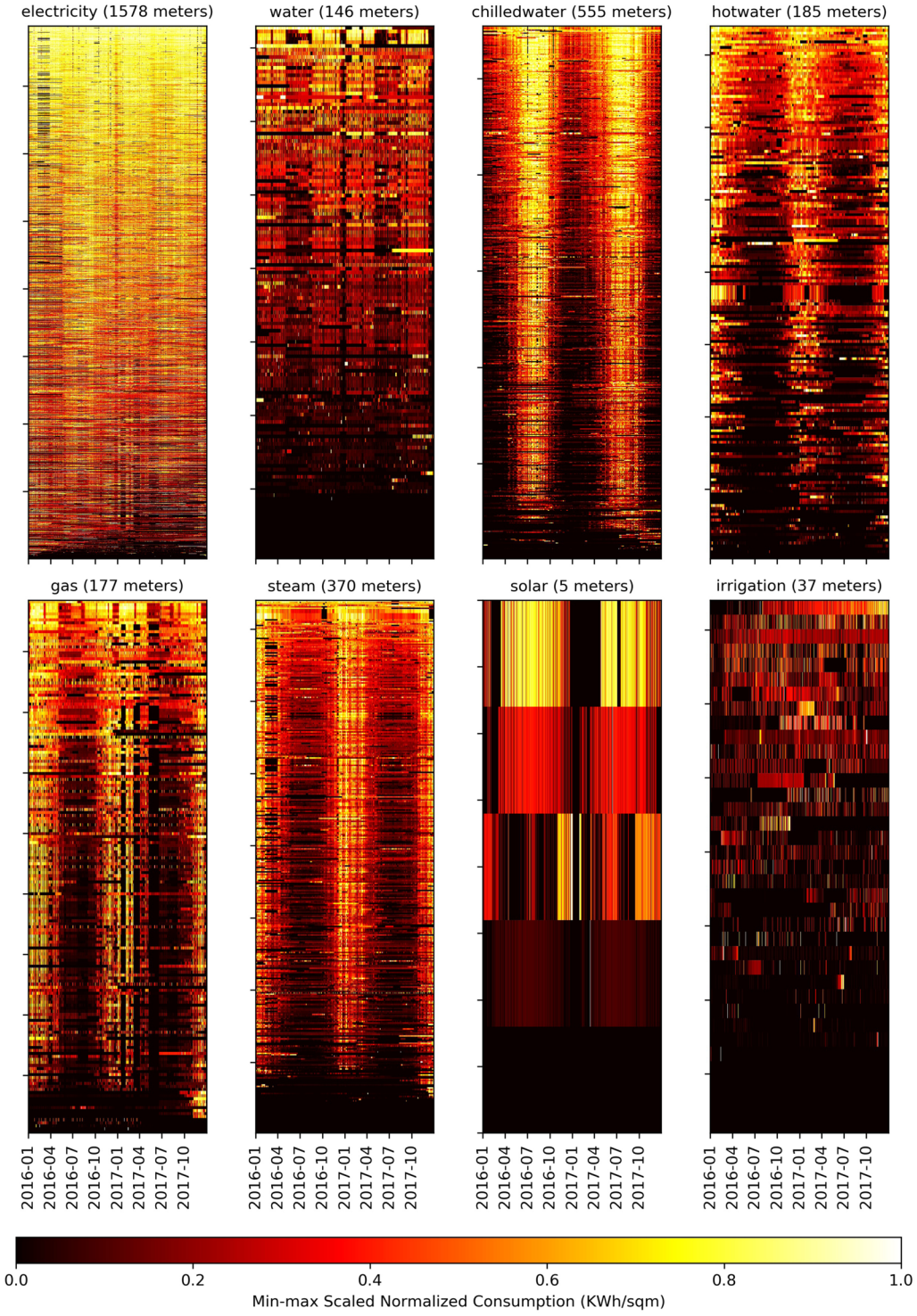

Fig. 3 Normalized meter consumption expressed as the daily energy consumption (kWh) per area unit (square feet) of the building that is then scaled to Min-max scaling (for a range of 0–1). Each heatmap corresponds to a meter type, the horizontal access for all graphics is the two year time range, and the vertical axis are the range of meters sorted anonymously from (bottom-to-top) from lowest to highest scaled daily normalized consumption.

and finding the Spearman rank-order correlation coefficient between the meter reading and the outside air temperature across each month. The resultant heat map visualization can be found in Fig. 5. The Spearman coefficient is a standard non-parametric measure of rank correlation. It shows which meter types are heavily positively correlated (related to cooling system energy influence) or negatively correlated (heating system energy influence). These heat maps illustrate the range of behavior for the various meters; the hot and chilled water and steam meters are heavily correlated, as expected, but also a significant number of electricity meters.

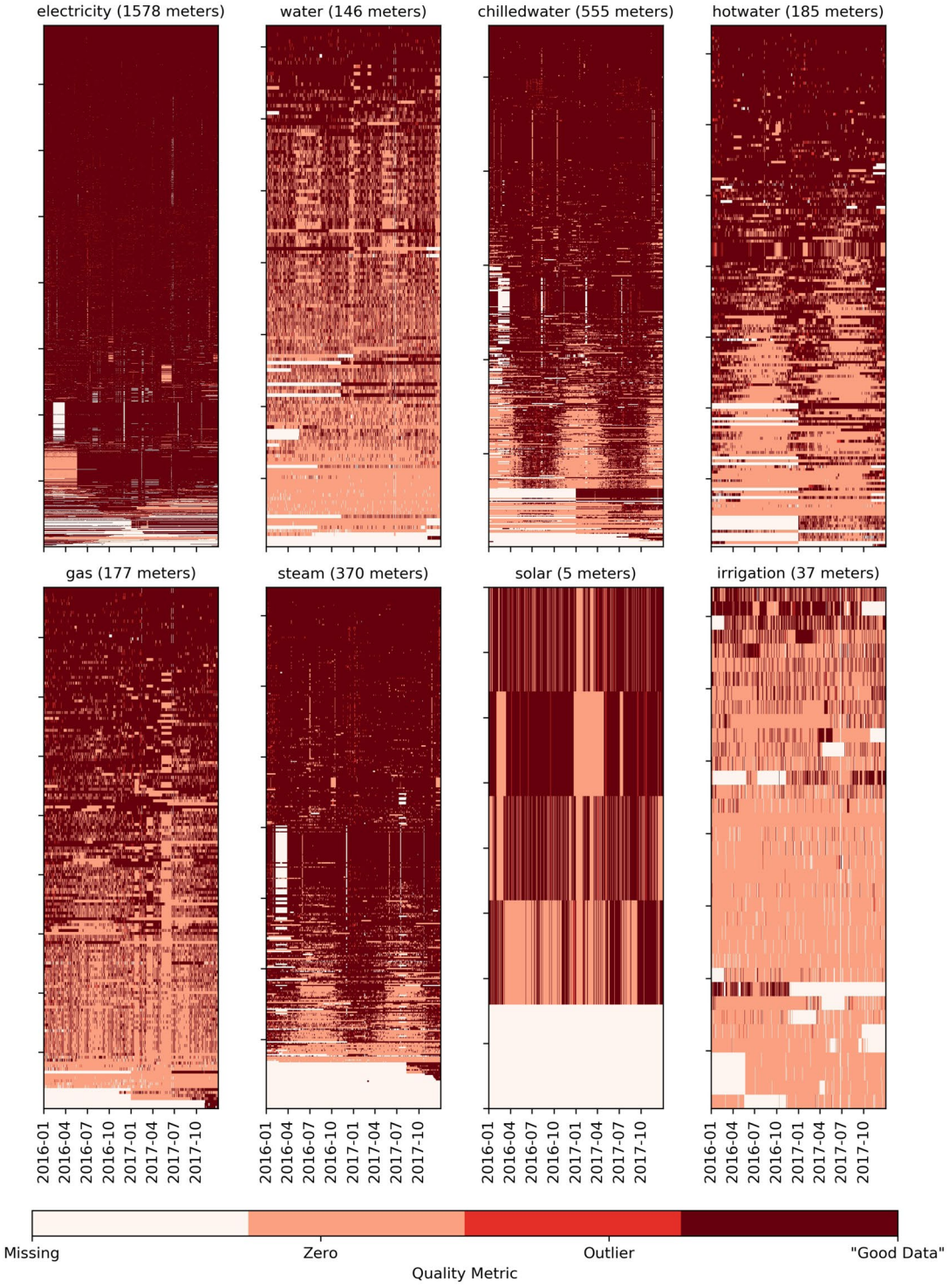

Fig. 4 Data quality plot of each meter type. Sorted (bottom-to-top) according to increasing number of *good data*.

Breakout detection. The final screening process shown in this publication is focused on quantifying the volatility of the time-series meter data through the use of breakout detection. A breakout is a time-series behavior that occurs when measurements have a shift from one steady-state behavior pattern to another. These shifts, or breakouts, are typically characterized by two steady states and an intermediate transition period. A breakout might be an example of a building operating in one type of schedule to another, such as commonly the case in educational buildings. For breakout detection, in this case, the Breakout Detection package developed by Twitter was used to detect the breakout shifts in an unsupervised way (https://github.com/twitter/BreakoutDetection). The critical parameter set for the model was that a steady state has to be at least 168 points long (a week) as a minimum. The resultant heat maps from this process can be seen in Fig. 6. These visualizations show the volatility

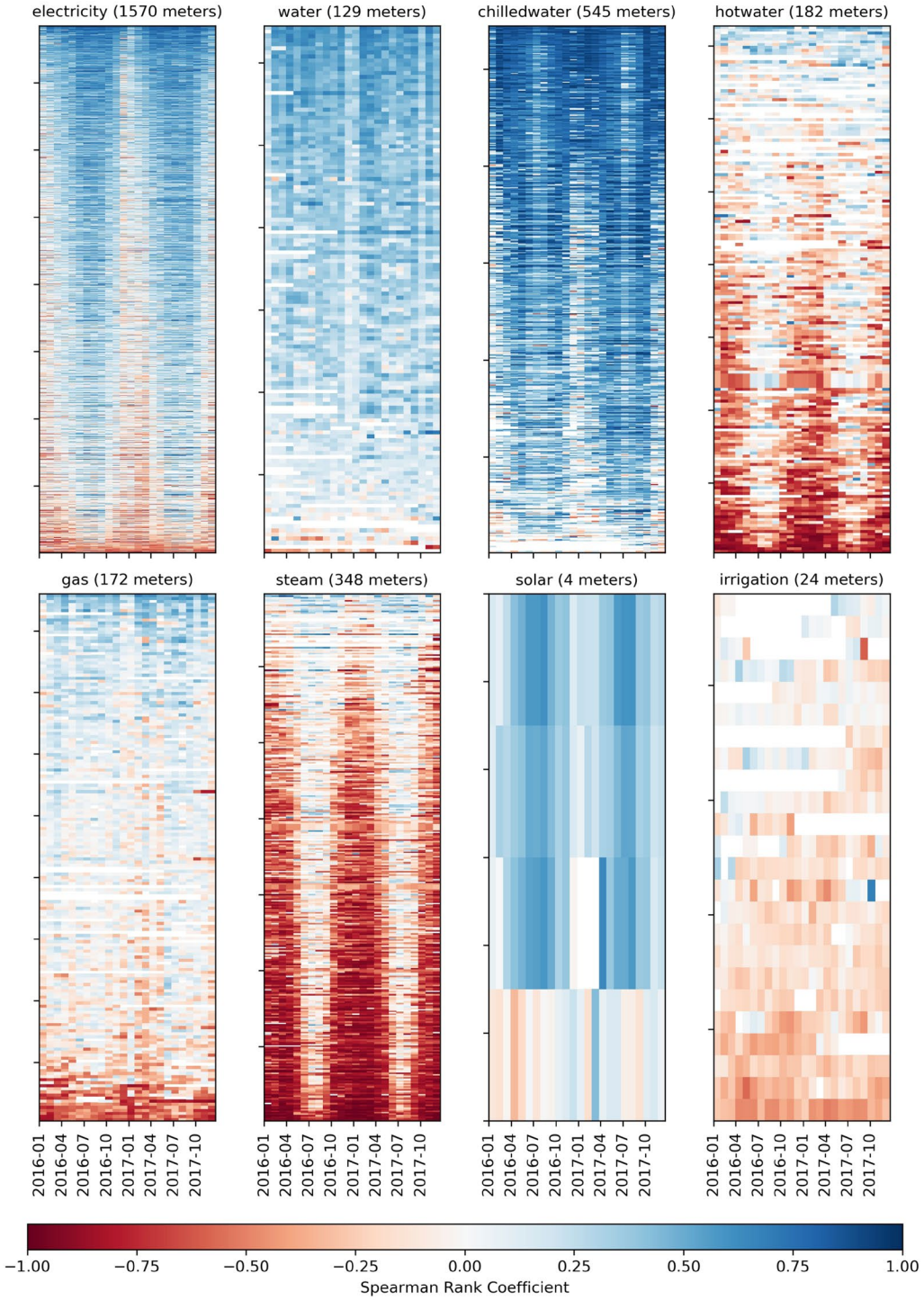

Fig. 5 Weather sensitivity plot of each meter type. Spearman rank coefficient was calculated between the meter reading (kWh or liters) and the outside air temperature (degrees Celsius) for each month. Sorted (bottom-to-top) according to increasing sum of coefficients.

of consumption based on the number of breakouts detected over the time range of two years. The steam and electricity meters show a broad range of volatility, while water and gas are more consistent comparatively.

Usage Notes

The usefulness of the Building Data Genome 2 data set can be understood in the context of several applications. The most obvious is in the context of the GEPIII competition and time-series meter data prediction in general. In this section, several examples of using the data set for various applications are discussed. It should be noted that gaps or removed outliers will make an impact on the summation at the daily, weekly, or annual basis. Therefore, care should be taken when calculating metrics such as energy use intensity (EUI) at those scales without filling gaps.

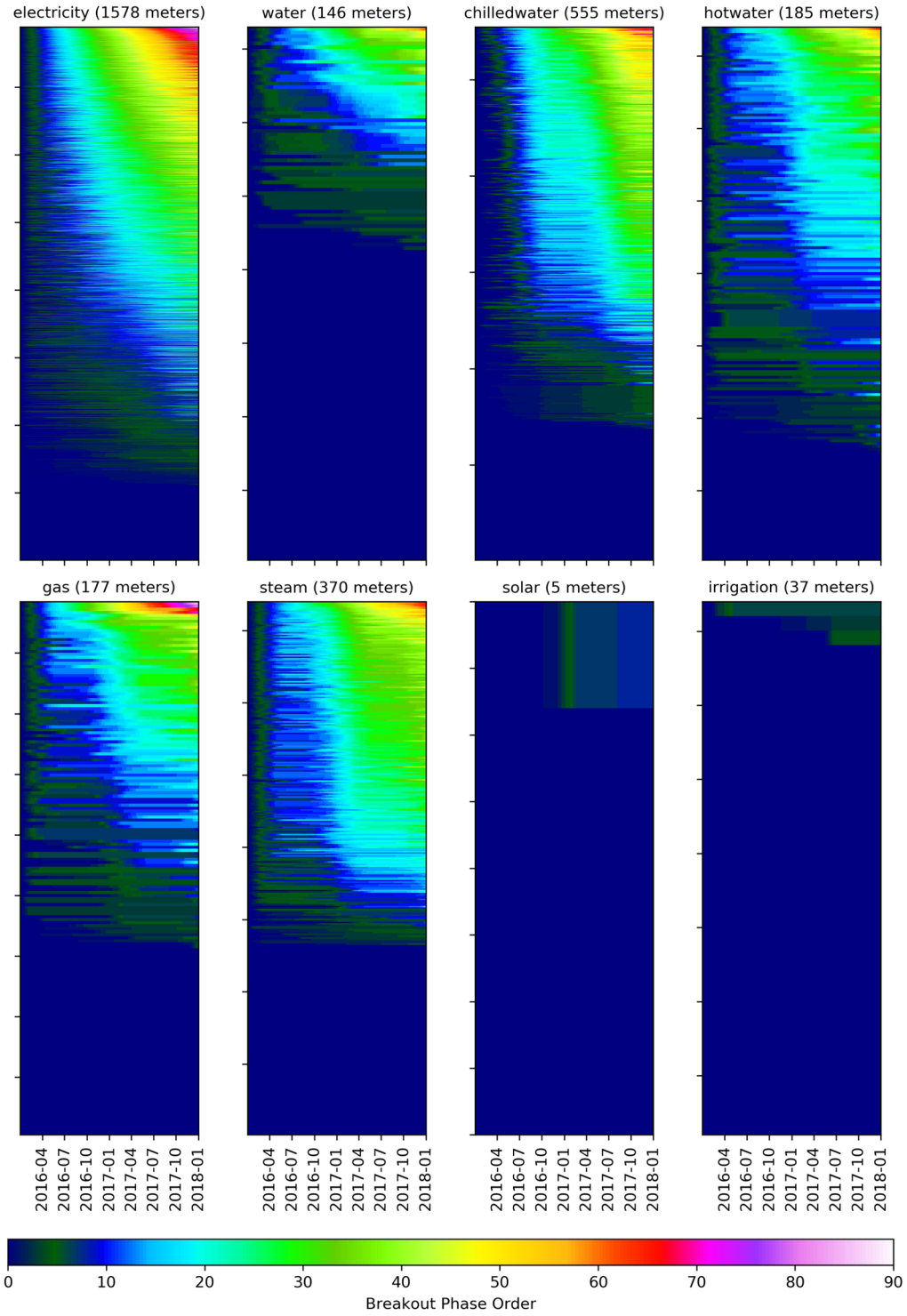

Fig. 6 Breakout detection heat map sorted (bottom-to-top) according to increasing number of breakouts detected. The more breakouts detected in a time-series data set, the more volatility is incurred in the data set.

Relationship with the GEPIII kaggle competition. The first application discussed is the use of the Building Data Genome 2 data set in the context of long-term data prediction. As mentioned, Building Data Genome 2 includes the data that were used in the GEPIII competition on the Kaggle machine learning platform. Users of this data set can map each of the unique building ID's on the Kaggle platform, represented as an integer, with the unique ID's created in this larger data set. The documentation for that mapping can be found on a Github documentation page for the repository. Table 1 includes a column that outlines which sites were used for the competition. The Building Data Genome 2 includes a folder (`/data/meters/kaggle/`) that includes the data for the validation data set (2017) that matches seamlessly with the training data found on the competition website.

The data contained in this folder have several differences as compared to the rest of the Building Data Genome 2 data sets. The first difference is that the Building Data Genome 2 data set only has timestamps in the local time zone, including the weather data. The weather data released in the Kaggle competition had a timestamp that was set to UTC, and the contestants had to come up with ways to find the right alignment for the weather data to use it properly. The other set of issues is related to several mistakes in unit conversion from the data sources and the Kaggle competition data set. Several meters that were assumed to be in kWh were in a different unit. Another issue is that several of the meters were converted from the wrong units. These mistakes have been fixed in the Building Data Genome 2 data sets (*raw* and *cleaned*), but were left as-is in the Kaggle data set.

A key consideration concerning the relationship between the Building Data Genome 2 and the GEPIII competition is that the third year (2018) of data from the competition is not released in this repository as some of those data are still used in the final test data (*private leaderboard*) component of the competition. The competition's structure was such that the first year was released as the *training* data, and the contestants were asked to produce predictions for the second and third years (2017 and 2018). In the competition, the second year was used to calculate the validation data set score (*public leaderboard*), and the third year was used for the final test score (*private leaderboard*). The final score test data, the third year (2018), is not released to enable users to use that year of data as the prediction objective to see how their methods match up to the contestants from the competition. Users now have two years of data from the Building Data Genome 2 project to predict the third year (2018) and, therefore, it should be noted that they have an advantage over the contestants who only had access to one year of training data at the time.

Long-term building hourly energy prediction model benchmarking. To create a curated example of meter data prediction similar to the Kaggle competition, the repository includes a well-documented instance of long-term energy prediction. These examples are described in detail in a documentation page on the repository (https://github.com/buds-lab/building-data-genome-project-2/wiki/Long-term-prediction). The example illustrated extracts various time-series feature from the meter and weather data and trains a model using one year of data to predict the following year. In this case, hourly data from 2016 is used to predict meter readings in 2017, and the accuracy as compared to ground truth is calculated using several metrics. This example is provided for users as a template for testing and incorporating their own machine learning process methods.

Short-term building hourly energy prediction model benchmarking. The next set of examples created in the repository are similar but focus on a shorter time-scale. A large body of research exists that is focused on short-term prediction with applications more aligned with grid-scale interactions, demand response, supervisory control systems, and anomaly detection[20]. The repository provides examples of short term prediction using the data set to use one month of hourly data to predict 72 hours ahead. The detailed documentation for these examples can be found on the documentation page in the repository (https://www.kaggle.com/claytonmiller/buildingdatagenomeproject2).

Building data genome project 2 kaggle data page. To create an environment where users of the data set can come up with new ideas for the use of the data set, a Kaggle Data Project has been created for a community to grow ideas focused on using this data set (https://www.kaggle.com/claytonmiller/buildingdatagenomeproject2). This project is independent of the Kaggle GEPIII competition and focused on the development of kernels (or notebooks) that process the data towards various objectives. This platform enables crowd-sourcing of analysis techniques, solutions, and processes. The page has a set of *Tasks* in a tab with that name that seed ideas of analysis beyond just short and long-term prediction. Some of the additional tasks outlined include time-series classification, anomaly detection, meta-data analysis, and data visualization techniques.

Code availability

The Building Data Genome 2 data set and the custom code used for its creation and analysis is hosted in a public Github repository (https://github.com/buds-lab/building-data-genome-project-2) and its v1.0 release has been deposited in Zenodo[18]. This codebase includes several Jupyter notebooks with Python and R data analysis workflows that can be easily reproduced.

Received: 12 June 2020; Accepted: 29 September 2020;
Published online: 27 October 2020

References

1. Granderson, J. *et al*. Accuracy of automated measurement and verification (M V) techniques for energy savings in commercial buildings. *Appl. Energy* **172**, 296–308 (2016).
2. Miller, C. More buildings make more generalizable models – benchmarking prediction methods on open electrical meter data. *Machine Learning and Knowledge Extraction* **1**(3), 974–993 (2019).
3. Dau, H. A. *et al*. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica* **6**, 1293–1305 (2019).
4. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. Preprint at https://arxiv.org/abs/1708.07747 (2017).
5. Mattson, P. *et al*. MLPerf: An industry standard benchmark suite for machine learning performance. *IEEE Micro.* **40**(2), 8–16 (2020).
6. Kriechbaumer, T. & Jacobsen, H. A. BLOND, a building-level office environment dataset of typical electrical appliances. *Sci. Data* **5**, 180048 (2018).
7. Kelly, J. & Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2**, 150007 (2015).
8. Mahdavi, A., Berger, C., Tahmasebi, F. & Schuss, M. Monitored data on occupants' presence and actions in an office building. *Sci. Data* **6**(1), 290 (2019).
9. Ruhnau, O., Hirth, L. & Praktiknjo, A. Time series of heat demand and heat pump efficiency for energy system modeling. *Sci. Data* **6**(1), 189 (2019).

10. Schweiker, M., Kleber, M. & Wagner, A. Long-term monitoring data from a naturally ventilated office building. *Sci. Data* **6**(1), 293 (2019).
11. Rashid, H., Singh, P. & Singh, A. I-BLEND, a campus-scale commercial and residential buildings electrical energy dataset. *Sci. Data* **6**, 190015 (2019).
12. Paige, F., Agee, P. & Jazizadeh, F. flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Sci. Data* **6**(1), 291 (2019).
13. Klemenjak, C., Kovatsch, C., Herold, M. & Elmenreich, W. A synthetic energy dataset for non-intrusive load monitoring in households. *Sci. Data* **7**(1), 108 (2020).
14. Roth, J., Lim, B., Jain, R. K. & Grueneich, D. Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. *Energy Policy* **139**, 111327 (2020).
15. Granderson, J. *et al*. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. *Appl. Energy* **173**, 296–308 (2016).
16. Miller, C. & Meggers, F. The building data genome project: An open, public data set from non-residential building electrical meters. *Energy Procedia* **122**, 439–444 (2017).
17. Miller C. *et al*. The ASHRAE Great Energy Predictor III competition: Overview and results, *Science and Technology for the Built Environment*, **26**(10), 1427-1447, https://doi.org/10.1080/23744731.2020.1795514 (2020).
18. Miller, C. *et al*. buds-lab/building-data-genome-project-2: v1.0. *Zenodo* https://doi.org/10.5281/zenodo.3887306 (2020).
19. Miller, C. & Meggers, F. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. *Energy Build.* **156**, 360–373 (2017).
20. Amasyali, K. & El-Gohary, N. M. A review of data-driven building energy consumption prediction studies. *Renewable Sustainable Energy Rev.* **81**, 1192–1205 (2018).

Acknowledgements

The authors acknowledge the individuals who assisted in the collection of data for inclusion in this data set. This list includes (alphabetical order) Adam Boltz, Adam Keeling, Ann Lundholm, Araz Ashouri, Catherine Patton, Doug Livingston, Gerry Hamilton, Ian Lahiff, James Ball, Jonathan Roth, Justin Owen, Kian Wee Chen, Maxime St-Jacques, Nate Boyd, Saptak Dutta, and Zach Wilson. The Kaggle platform technical and advisory team, including Addison Howard and Sohier Dane, were instrumental to the launch of the GEPIII competition. The ASHRAE competition planning team of (alphabetical order) Anthony Fontanini, Chris Balbach, Jeff Haberl, and Krishnan Gowri assisted in getting the competition launched and supported by the ASHRAE organization. Financial support for the development of the data set and travel support was provided by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) program. Additional research funding was provided by the Ministry of Education (MOE) of the Republic of Singapore (R296000181133). Financial support for the GEPIII competition monetary prizes was supported by ASHRAE. The Kaggle machine learning platform provided hosting as a non-profit competition.

Author contributions

C.M. coordinated the creation of the data set, led the data collection and preliminary analysis for 12 of the sites, and was the lead author of the publication. A.K., P.A. and J.Y.P. each led the data collection for one site and participated in the data cleaning and pre-processing before the GEPIII competition. B.P. transformed, cleaned, and prepared the data set for publication after the GEPIII competition. Z.N., P.R., B.H., Z.S. and F.M. each contributed data for one site and provided comments to improve the data set and its use. The majority of the contribution by Z.S was made at the National Research Council Canada, Ottawa ON, Canada. All authors reviewed the manuscript and take responsibility for its content.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00712-x.

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

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.