Energy & Buildings 273 (2022) 112372

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

A rule-based data preprocessing framework for chiller rooms inspired by the analysis of engineering big data

Ruikai He [a,b], Tong Xiao [b], Shunian Qiu [c], Jiefan Gu [b], Minchen Wei [a], Peng Xu [b,*]

[a] The Hong Kong Polytechnic University, Kowloon, Hong Kong
[b] Department of Mechanical and Energy Engineering, Tongji University, Shanghai 201804, China
[c] Zhejiang University of Science and Technology, Zhejiang, China

ARTICLE INFO

Article history:
Received 17 June 2022
Revised 24 July 2022
Accepted 7 August 2022
Available online 12 August 2022

Keywords:
Data pre-processing
Big engineering data
Building energy management

ABSTRACT

The rapid development of building energy consumption monitoring platforms makes engineering data more diverse, which facilitates the goal of reducing emissions. It is increasingly acknowledged that data preprocessing deserves the same attention as intelligent algorithms. In this work, the data quality issue of the engineering big data from non-demonstration complexes in China are analyzed thoroughly, and the analysis is based on clustering-based algorithms. We can conclude that the data of the hourly power of equipment groups are quality and stable, which is suitable for the benchmark to check other data. The quality of the data about pipes is acceptable. The number of data types about cooling towers is less, and the quality is worse. Regarding other data, the quality is unstable, so researchers should deal with those case-by-case. According to the above analysis, we proposed a convenient, rule-based data preprocessing framework that utilizes the law of physics, ensuring the strong coupling of multi-variants. After the data preprocessing, these engineering data are more reliable and can be used to improve performance or train models. Additionally, the proposed framework is more suitable for preprocessing multi-variant engineering data.

1. Introduction

1.1. Background

Emission saving has been one of the top trending topics, closely related to issues posing a threat to creature survival, natural resource depletion, and climate change. They are mainly attributed to energy-related greenhouse gas (GHG) emissions. Clean energy innovation plays an integral part in the prospect of net-zero emission. To achieve this long-term goal as soon as possible, organizations worldwide spare no effort toward that goal. The technology portfolio in public energy Research and Development is more balanced today than in previous decades, with far more money going to energy efficiency and renewable energy[1]. Meanwhile, the plan to achieve carbon neutrality by 2060 launched by Chinese officials further motivates a wide variety of professions to promote low-carbon technologies.

In this energy landscape, the commercial real estate industry has accounted for over 30 % of the global final energy consumption: more than 35 % fossil fuel and over 30 % electricity consumption[1]. Technology innovation needs to undergo four phases: prototype, demonstration, early adoption, and maturity. In buildings, the number of clean energy technologies entering the early adoption phase is much more than in other fields (transport, industry, power generation, and fuels transformation)[1]. The technique for data acquisition is outstanding among them. Widespread building energy consumption monitoring platforms (BECMPs) are among the most prominent examples[2–4]. Smart energy meters are quickly meters have been adopted across the world[refs][5]. More than 100 million buildings have adopted smart energy meters by 2019 in U.S [refs][6]. Sweden, Italy and Finland have achieved more than 90 % smart meter marketing share [refs][7]. Apart from the *political motivation*, the reasons why data acquisition related to building energy consumption can get attention from all walks of life in China are enlisted below:

1) *More Disturbance*: Covid-19 pandemic is a challenge for improving energy efficiency in buildings. According to heating, ventilation, air-conditioning, and cooling(HVAC) operation guidelines during covid-19[8–10], ventilation strategies may increase the energy consumption on account of *the safety issue* and *uncertainty of occupant behavior*[11,12]. Data acquisition can give us more

* Corresponding author.
E-mail addresses: 15651920118@163.com (R. He), gu_lavender@outlook.com (J. Gu), minchen.wei@polyu.edu.hk (M. Wei), xupeng@tongji.edu.cn (P. Xu).

https://doi.org/10.1016/j.enbuild.2022.112372
0378-7788/

Nomenclature

Abbreviation

ANN	Artificial Neural Network
BECMPs	Energy Consumption Monitoring Platforms
BIM	Building Information Model
BO	Building Occupancy
CES	Contemporary Energy Systems
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree
FDD	Fault Detection and Diagnosis
GNNs	Graph Neural Networks
GAN	Generative Adversarial Network
HVAC	Heating, Ventilation, Air-conditioning and Cooling
IEA	International Energy Agency
KNN	K-nearest Neighbor
LSTM	Long Short Term Memory
MAR	Missing at Random
MCAR	Missing Completely at Random
ML	Machine Learning
MPC	Model Predictive Control
NMAR	Not Missing at Random
RNN	Recurrent Neural Network
CH	Chiller
CHs	All Chillers in one BECMP
CHWP	Chilled Water Pump
CHWPs	All Chilled Water Pumps in one BECMP
CWP	Cooling Water Pump
CWPs	All Cooling Water Pumps in one BECMP
CTF	Fan in Cooling Towers
CTFs	All Fans in Cooling Towers in one BECMP
T	Temperature
C	Current
Hourly_P	Hourly Power
Q	Water Flow
H	Relative Humidity

Subscript

chws	Chilled supply water
chwr	Chilled return water
cws	Cooling supply water
cwr	Cooling return water
on/off	On/off status
m	Header pipe
out	outdoor
in	indoor
s	setpoint

profound insight into BO against the background of the recent pandemic.

2) *Model Foundation*: Computational algorithms and peripheral hardware[13] make data-driven models shine in low-carbon energy technology: machine learning (ML) algorithms, like Decision Tree (DT), LightGBM, and XGBoost. Recently, researchers have tried to apply more advanced algorithms, including attention-based models[14–16], Graph Neural Networks (CNN) [17], reinforcement learning[18] for fault detection and diagnosis (FDD) [19],optimization, model predictive control (MPC)[2], building load prediction[20] and non-intrusive monitoring (NIM)[21–24]. *Immerse data is needed to feed these models to achieve satisfaction* [25], so accumulated engineering data establish the solid foundation to put these state-of-art approaches into practice[26].

3) *Building Information Model*: As the US National Building Information Modelling Standard defines[27], a Building information model (BIM)[28,29] is "A digital representation of ...a shared knowledge resource for information about a facility forming a reliable basis for decisions during its life-cycle". The concept of digital twins[13,30,31] brings a whole new dimension to BIM, especially in buildings. By far, the digital twin is facilitated because of the advent of the Internet of Things (IoT)[32–34]. IoT makes the digital twin emphasize maintenance and operation, so *the real-time engineering data is indispensable to fulfilling BIM*.

4) *Economic Demand*: As we all know, retrofitting existing public buildings is integral to emission savings. *Big engineering data gives a new dimension to building energy audits*[35,36]. The data-intensive approach can radically change the status quo of traditional audits: time-consuming and labor-intensive, which can help companies save costs.

The richness of engineering data facilitates the fact that countless researchers have been dedicated to making components more intelligent in CESs[refs][37]. Consequently, how to maximize the potential of big data from CESs is the top priority for some researchers interested in interdisciplinary research.

This section concludes with the motivations for collecting and building big engineering data in China. **Section 1.2** illustrates general data preprocessing practices and the significance of our work. **Section 2** introduces the proposed data preprocessing framework. **Section 3** shows the application of the proposed framework. **Section 4** is about how to improve the collection of engineering data. **Section 5** summarizes the paper.

1.2. Literature over data preprocessing to building energy-related data

1.2.1. Problems in data preprocessing

Data quality issue is prevalent. According to the analysis based on text-mining in buildings[38], data preprocessing has not gotten enough attention, no matter what kinds of data. FDD seems to be the sub-field that pays enough attention to preprocess data[38]. Generally, researchers tend to require data quality before experiments, which can guarantee that data preprocess would not dramatically reduce the quantity of valid data[39].

In our work, Data preprocessing is more like diagnosing data. That is the first procedure among them[26]. There are two common problems in data quality: **data missing** and **anomalies**. In civil engineering and buildings, **data missing** is complicated and can be divided into different situations[40,41]:

1) missing completely at random (MCAR); data missing happens entirely by accident and is nothing to do with other attributions, like data type and value range. For example, the lost data of energy consumption is almost nothing to do with other data.

2) missing at random (MAR); the observed variables affect the data missing, but the unobserved variables do not. For instance, when the hourly power of a chiller is lost, that can be fixed by the current of the same chiller, which is the characteristic of MAR.

3) not missing at random (NMAR); the data missing is only about itself. Under some circumstances, researchers cannot get some data due to historical or private issues.

Researchers are likely to encounter more than one kind of data missing simultaneously. Apart from data missing, another issue is **anomalies**, including *global constant* and *moving average*[26]. We can also divide abnormal data into *point anomalies*, *contextual anomalies*, and *collective anomalies*[20,42].

1.2.2. Methods of data preprocessing

Before training data-driven models, researchers have explored various approaches to preprocessing data: *engineering approach* [43], *statistic methods*[44], and *intelligent algorithms*. Of course, more than one method can be used in data preprocessing. In this section, much emphasis are put on the intelligent algorithm because many recent papers about.

1) Engineering approach means that researchers depend on their knowledge when preprocessing data. The effort aims at taking advantage of mechanism, such as the heat and mass transfer process between building envelop and surrounding, in mathematical equations and simulation tools[43,45]. Mathematical equations require proficiency with expertise and reality. Simulation tools (E.g. EnergyPlus[46], DeST[47]) could describe buildings at a detailed level, but some parameters is hard to get, which makes an significant gap between simulation and field measured data [48].

2) Statistic methods include single upper limit, pauta criterion (3 δ law), local outlier factor, and interquartile range[20]. Statistic methods may be better for detection or imputation when the data missing values is mild[49]. These methods are common in competitions, like ASHRAE Great Energy Predictor[20]. Firstly, competitors entirely use statistical methods to process data missing and anomalies. For invisible abnormal data in time series, expertise is used to filtering out these data.

3) Intelligent algorithms can be divided into unsupervised clustering, supervised classification, and semi-supervised recognition, especially when data missing is severe[50]. Compared with linear statistical methods, almost intelligent algorithms are nonlinear [49], which means these algorithms are capable of dealing with more complicated situations. Clustering-based methods can be used in two ways. The first is regarded as a preliminary step to identifying data clusters, and then statistic methods are applied for outlier detection[51]. The second one is to use these clustering-based techniques to fill or remove data directly; Cui et al. make use of the k-nearest neighbor (KNN)[51] for the imputation of building energy data[52]. Armini et al. studied and compared the performance of different fuzzy-rough nearest neighbors on missing imputations[53]. Liu et al. used DBSCAN to identify data when the system is in transient operations[54]. More complicated algorithms are also used to detect abnormal and missing data imputations. Liang proposed an ensemble method to hint at different kinds of data, and an artificial neural network (ANN) was among his ensemble methods[55]. Ma et al. utilized LSTM to impute missing data, proving that his proposed method can be suitable even if the percentage of missing data is relatively high [49]. Dongyeon et al. used a factor analysis matrix to impute the missing data, which utilized electricity characteristics to fill the data of load[56]. Cao et al. proposed a bidirectional approach to imputing missing data by Recurrent Neural Network (RNN) and used air quality and healthcare data to verify[57]. Luo et al. use Generative Adversarial Network (GAN) to preprocess multivariate time series[58], but not in buildings.

1.2.3. The gap the proposed framework wants to fill

Given the features of big engineering data, *multi-variants*, *various working conditions*, and *highly correlated*[40], the features are made the full use of to design the rule-based framework for data preprocess. The proposed framework is explainable, so it will be suitable in engineering.

Apart from missing data and anomalies considered, **the mismatch between labels and data** is another problem that few previous papers refer to, but it is ubiquitous in engineering. The proposed framework takes this problem into consideration.

To enhance the generalization ability of the proposed framework, *the analysis of engineering big data is conducted before*, which is rare in previous work. The data involved are from 100 commercial non-demonstration complexes in China. Compared with demonstration buildings, non-demonstration complexes is larger in number, and data quality issue is more complicated because of not highly professional operation and maintenance.Fig. 1.

2. Methodology

2.1. Overview of the workflow and proposed framework

The overall workflow (Fig. 2) can be divided into three parts: analysis of data quality, data preprocessing, and data application. The first two are the core of this paper, and the last part is based on the first two parts.

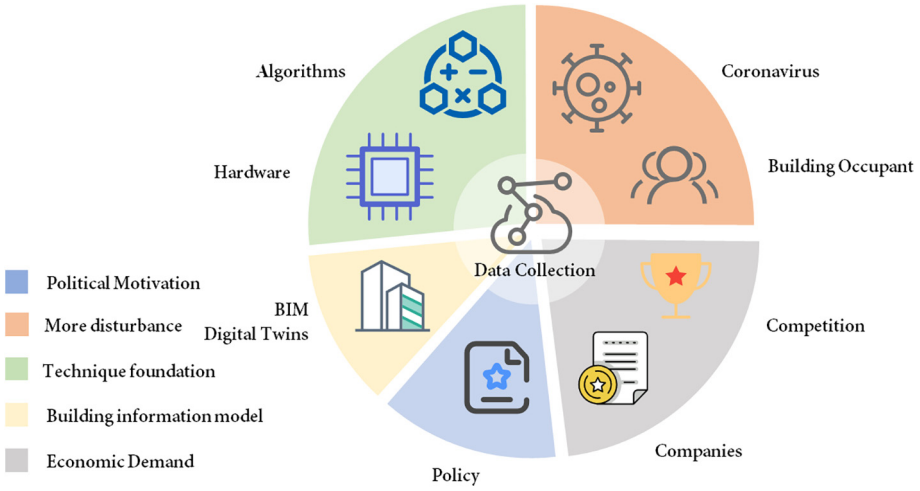


Fig. 1. The aspects facilitated engineering data from BECMPs.

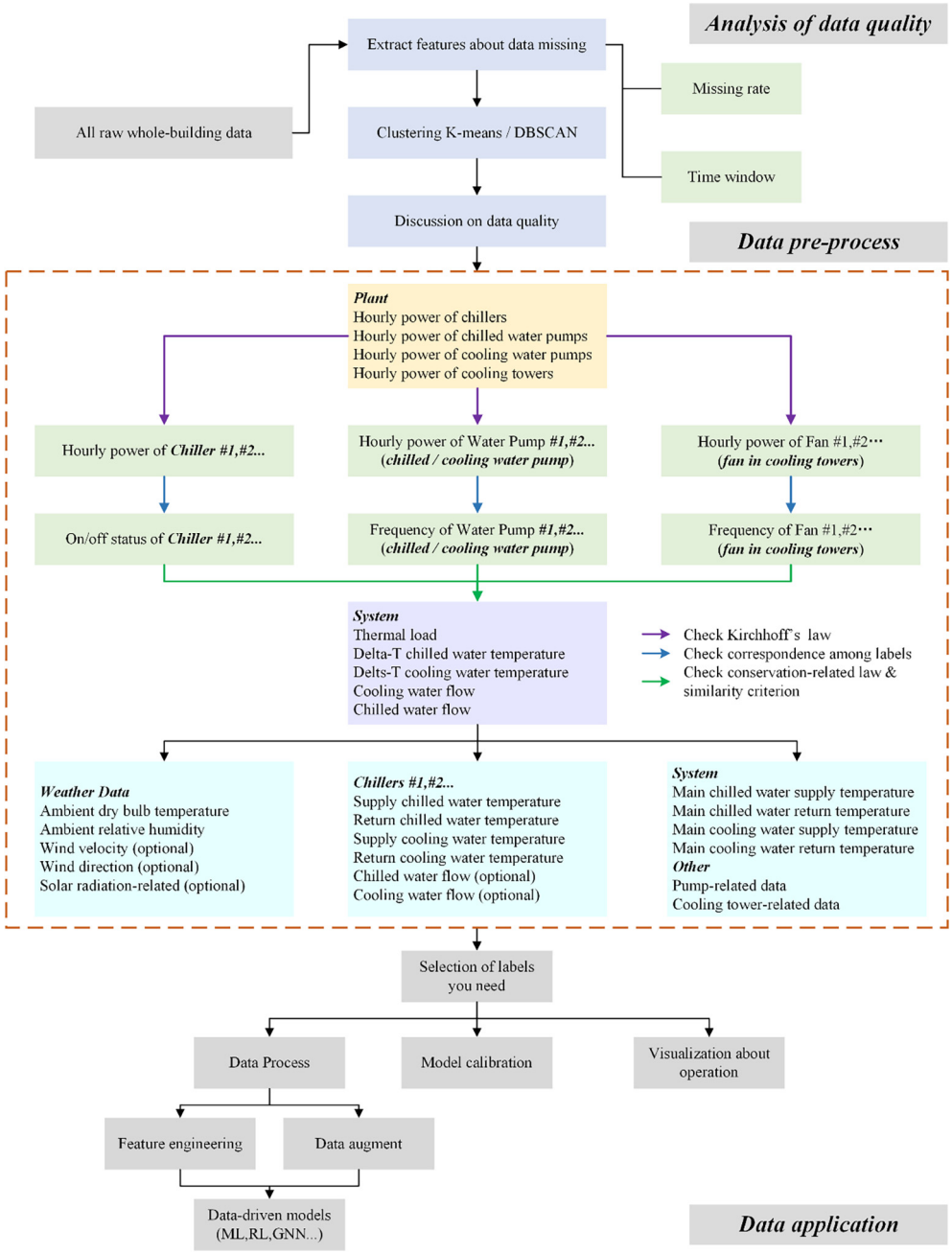


Fig. 2. The overall workflow.

Firstly, the data quality analysis aims to determine the benchmark of the data preprocessing. The clustering algorithms are used to analyze the features of data missing, including missing rates and time windows.

The second part is to process the engineering data considering the clustering results. The order of data processing is in the orange box in Fig. 2. Specifically, the "rule-based" is to take advantage of the laws of physics because CESs are physical systems, which can reduce dependence on expertise. In this paper, the laws of physics include Kirchhoff's law, the law of conservation of energy, and the similarity criterion in fluid mechanics.

After the data preprocessing, researchers can apply the checked data to fulfilling their goals. In this paper, energy efficiency analysis is used as the data application to verify the proposed frameowork.

2.2. Analysis of data quality

2.2.1. Method for analysis

The analysis includes feature extraction and clustering. First, 14 features are extracted (Table 1.) from raw data: Features #2 - #6 are statistical features of the vector of time window $X = (x_1, x_2, \cdots, x_k, \cdots, x_n)$.

where x_k is the number of consistent data missing, $k \in [0, N_{all}]$, $x_k \in [0, N_{all}]$.

Features #7 - #14 are used to find if data missing follows some rules related to time (month, the day of the week, hour in a day, minute in an hour).

Features #1 - #6 can tell if data missing happens intermittently or constantly and if massive data missing happens. Features

Table 1
The features about absence in data extracted from raw whole-building data.

No.	Features	Specification	Range
1	The overall missing rate	$r_{all} = N_{all\_missing}/N_{all}$	$[0, 1]$
2	The minimum time window	$N_{min} = \min\{x_1, x_2, \cdots, x_k, \cdots, x_n\}$	$[0, N_{max}] \bigcap N_{min} \leq N_{all\_missing}$
3	The maximum time window	$N_{max} = \max\{x_1, x_2, \cdots, x_k, \cdots, x_n\}$	$[N_{min}, N_{all}] \bigcap N_{max} \leq N_{all\_missing}$
4	The number of time windows	n	$[0, N_{all}]$
5	The average length of time windows	$x_{mean} = \sum_{k=1}^{n} x_k/n$	$[N_{min}, N_{max}]$
6	The variance of length of time windows	$s^2 = \sum_{k=1}^{n}(x_k - x_{mean})^2/n$	
7	The month with maximum data missing rate	the span ranges from May 2020 to December 2020	$[5, 12] \cap N^+$
8	The maximum missing rate by month	$N_{\max\_mon} = max\{x_{May}, x_{Jun}, \cdots, x_{Dec}\}$	$[0, N_{all\_mon}]$
9	The day of the week with maximum data missing rate	Monday: 0, Tuesday: 1, Wednesday: 2, Tuesday: 3, Friday: 4, Saturday:5, Sunday: 6.	$[0, 6] \cap N^+$
10	The maximum missing rate by the day of the week	$N_{\max\_wee} = max\{x_{Mon}, x_{Tue}, \cdots, x_{Sun}\}$	$[0, N_{all\_wee}]$
11	The hour with maximum data missing rate	0: 0.00 $\sim$ 0:59, 1: 1.00 $\sim$ 1.59, ..., 23: 23.00 $\sim$ 23.59	$[0, 23] \cap N^+$
12	The maximum missing rate by hour	$N_{\max\_hou} = max\{x_{0H}, x_{1H}, \cdots, x_{23H}\}$	$[0, N_{all\_hou}]$
13	The minute with maximum data missing rate	An arithmetic sequence spaced 15 min apart	$[0, 15, \cdots, 45]$
14	The maximum missing rate by minute	$N_{\max\_min} = max\{x_{0min}, x_{15min}, \cdots, x_{45min}\}$	$[0, N_{all\_min}]$

$N_{all\_missing}$ is the total number of absence in data.
N_{all} is the total number of data.
In feature #8, x_{May} means the total number of data missing in May 2020. And so on.
In feature #10, x_{Mon} means the total number of data missing in all Mondays in the second half of 2020. And so on.
In feature #12, x_{0H} means the total number of data missing in 0.00 $\sim$ 0.59 every day in the second half of 2020. And so on.
In feature #14, x_{15min} means the total number of data missing in hourly 15th minute in the second half of 2020. And so on.

Table 2
The details of 141 non-demonstration complexes themselves.

No.	Type of information		Know or not		Total
			known	unknown	
1	Name		0	141	141
2	Appearance		0	141	141
3	Drawings	Layout	0	141	141
		CES	3	139	141
4	Nameplates	Rough	140	0	141
		Detailed	1	0	
5	Overall Area		140	1	141
6	Location	Rough	141	0	141
		Detailed	0	0	

Rough information of nameplates means we only know the manufacturers of equipment.
Detailed information of nameplates means the nameplates of equipment are known.
Rough location means that we know which province where a complex is located.
Detailed location means the latitude and longitude of a complex.
Among 3 known CESs, 2 of them are primary pump systems, another is a secondary pump system. We have no idea about the type of terminal system for all CESs.

Table 3
The details of data from BECMPs in 141 non-demonstration complexes.

No.	Span	Time	Temporal	The number of complex
1	0.5 year	May 2020 - Dec 2020	Every 15 min	140
2	6 years	Jun 2015 - Oct 2020	Hourly	1

2 known primary pump systems are in #1.
Only 1 known secondary pump system is in #2.

#7 - #14 give us an insight into whether the data missing is highly correlated to time. Additionally, the value ranges of features #7 and #13 are confined to the data we have.

After choosing these features, normalization, the "min-max_scale" API in scikit-learn[59], makes all features range from 0 to 1. In order to make results convincible, we compare the results between K-Means and DBSCAN: K-Means is based on partition, while DBSCAN is based on density.

2.2.2. The engineering big data involved in analyzing

Static information: The engineering data involved are from 141 non-demonstration complexes, are provided by a technology company. Privacy issues are not allowed us to access all details. We only know the general area and location of almost every building. Static information is shown in Table 2.

Dynamic data: Data from different BECMPs have some differences in quantity, interval (Table 3), and labels (Table 4). The rough

Table 4
All labels in BECMPs in 141 non-demonstration complexes.

Level	Labels	
Weather	Ambient dry bulb temperature	Ambient relative humidity
Thermal comfort	Average indoor temperature*	
System	Main chilled water flow	Main cooling water flow
	Main supply chilled water temperature	Main supply cooling water temperature
	Main return chilled water temperature	Main return cooling water temperature
	Main supply chilled water pressure	Main supply cooling water pressure
	Secondary supply chilled water flow**	Secondary supply chilled water pressure**
	Secondary supply chilled water pressure setting**	Cooling load*
	The power of overall chillers*	The power of overall cooling towers*
	The power of overall chilled water pumps*	The power of overall cooling water pumps*
Chillers	Branch supply chilled water temperature	Branch supply cooling water temperature
	Branch return chilled water temperature	Branch return cooling water temperature
	Evaporating temperature**	Condensing temperature**
	Percent of current*	Supply chilled water temperature setting
	The power of individual chiller	On/off status
Primary chilled water pumps	The power of individual pump	On/off status**
	The frequency of individual pump*	
Secondary chilled water pumps	The power of single pump**	The frequency of individual pump**
	On/off status**	The frequency gear of individual pump**
Cooling water pumps	The power of individual pump	On/off status**
	The frequency of individual pump*	
Cooling towers	The frequency of individual fan*	On/off status of individual fan**
	The frequency gear of individual fan**	

**means ONLY #1 in Table 2 has these labels.
*means ONLY #2 in Table 2 has these labels.

locations of all sensors are shown in Fig. 3. These labels give us a deeper understanding of some HVAC systems:

1) If there is only one label "on/off status" for a water pump, it means the fixed frequency pump.

2) If there is a label "the frequency gear of individual pump", the only way to change the frequency of a pump is to adjust the gear.

3) If there is a label "the frequency of a water pump" and values are continuous, it means that the water pump is under the stepless control.

2.3. Data preprocessing

2.3.1. Check Kirchhoff's law

Given the clustering results, *the overall hourly power of equipment groups* becomes the benchmark of data preprocessing. After removing the missing data and anomalies in the overall hourly power of equipment groups, Kirchhoff's law is used to preprocess the hourly power of every piece of equipment. The strict subjection

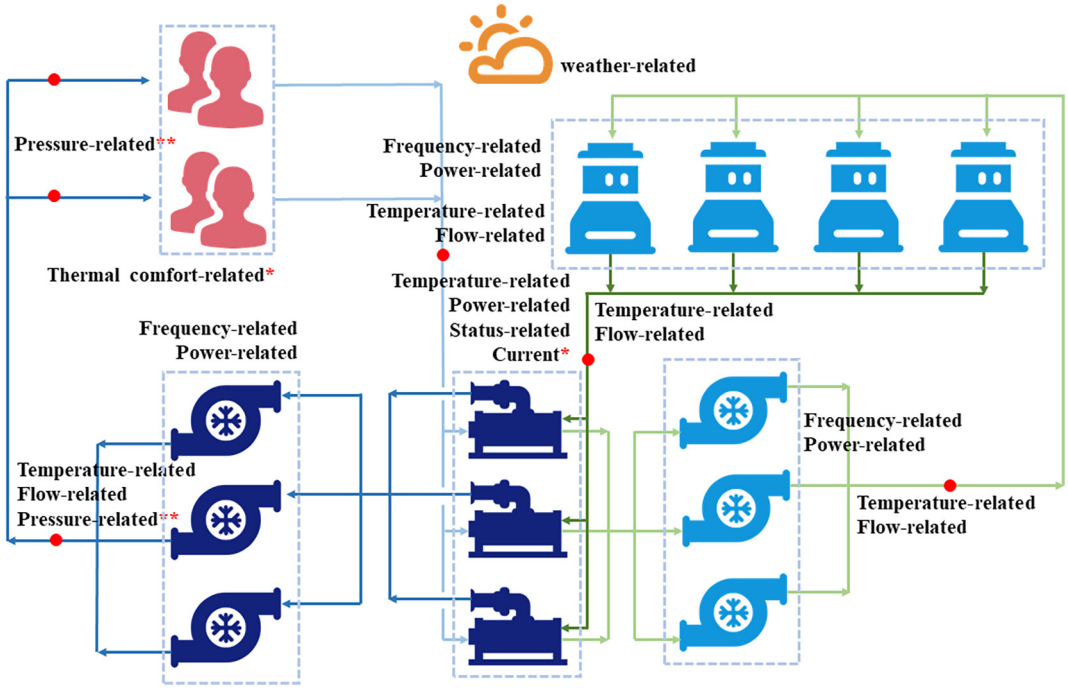


Fig. 3. The rough locations of sensors in a typical commercial HVAC system.

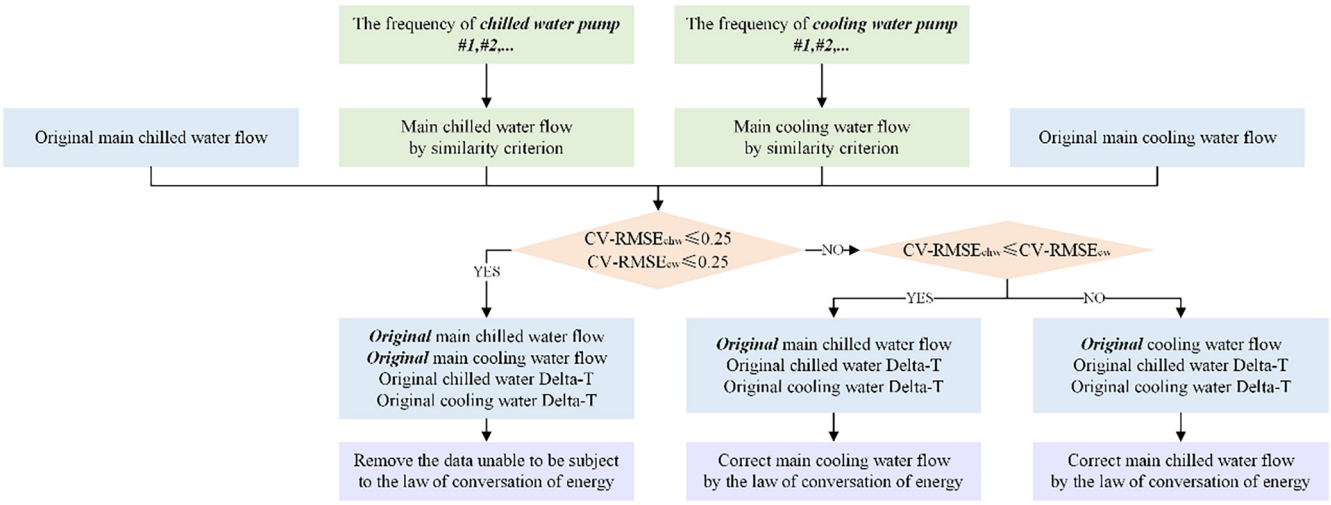


Fig. 4. The proposed framework for checking data in system level.

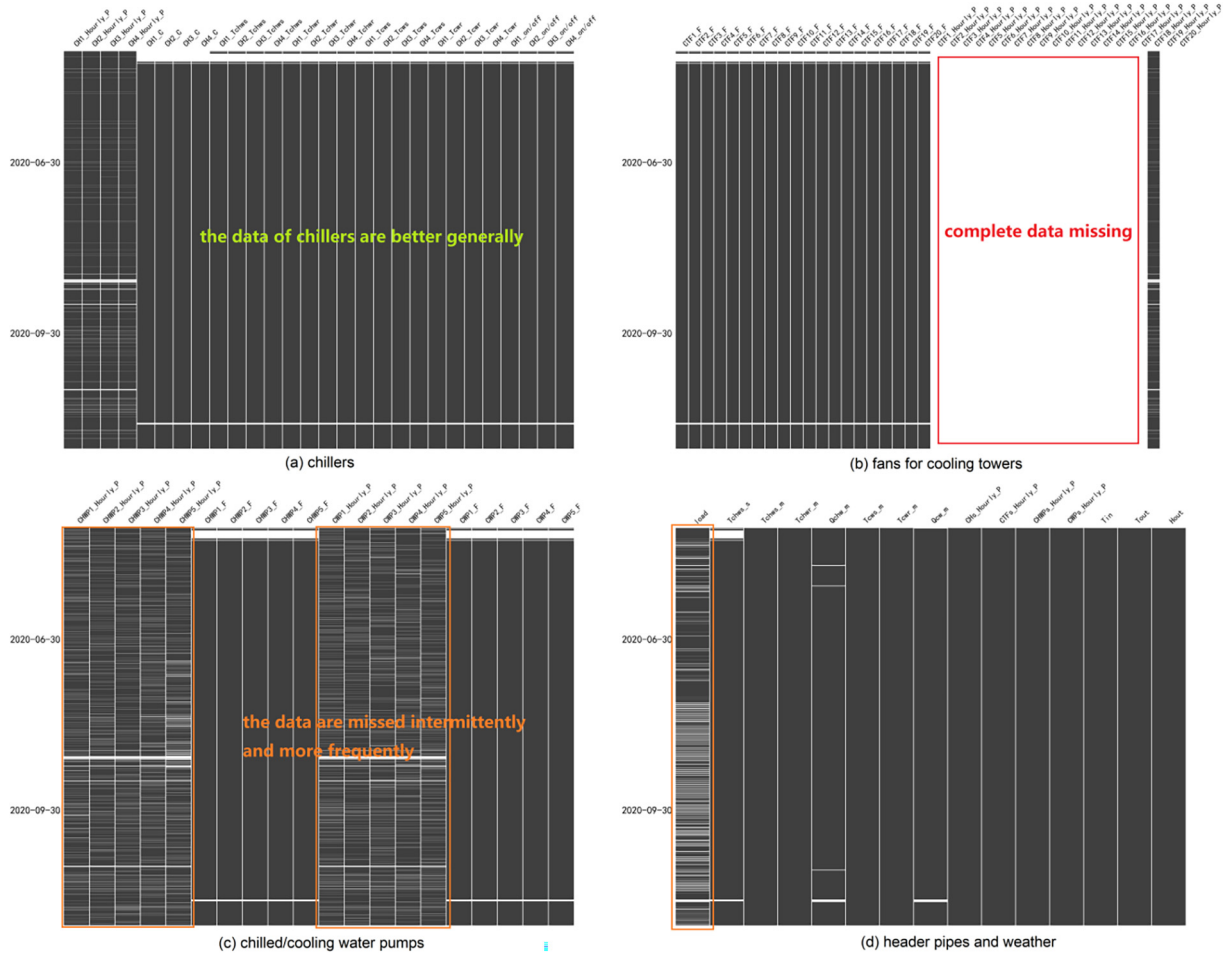


Fig. 5. The visualization of data missing of different data. (a) data of chillers, (b) data of fans for cooling towers, (c) data of water pumps and (d) data of header pipes. One bar is one label, the dark space means no data missing and the white space means data missing. The longer the length of white space, the more severe the data missing is.

to Kirchhoff's law is nearly impossible, so we set the threshold (Eq. (1)) to use data fully.

$$\begin{cases} \frac{|P_{all_{chs}} - \sum_{i=1}^{k} P_{chi}|}{P_{all_{chs}}} \le 0.25 i = 1, \cdots, k \\ \frac{|P_{all_{chwps}} - \sum_{i=1}^{m} P_{chwpi}|}{P_{all_{chwps}}} \le 0.25 i = 1, \cdots, m \\ \frac{|P_{all_{cwps}} - \sum_{i=1}^{n} P_{cwpi}|}{P_{all_{cwps}}} \le 0.25 i = 1, \cdots, n \\ \frac{|P_{all_{ctfs}} - \sum_{i=1}^{p} P_{ctfi}|}{P_{all_{ctfs}}} \le 0.25 i = 1, \cdots, p \end{cases} \quad (1)$$

where $P_{all_{chs}}$ is the hourly power of all chillers, $P_{all_{chwps}}$ is the hourly power of all chilled water pumps, $P_{all_{cwps}}$ is the hourly power of all cooling water pumps, $P_{all_{ctfs}}$ is the hourly power of all fans for all cooling towers, k is the total number of chillers, m is the total number of chilled water pumps, n is the total number of cooling water pumps and p is the total number of the fans for cooling towers.

2.3.2. Check correspondence among labels

After the above procedure, the data of every piece of equipment can be processed by the overall hourly power of equipment groups. Generally, there is an exclusive label to record *the on/off status of every chiller. The frequency* is necessary for every water pump and the fan for cooling tower. According to the expertise and similarity criterion in fluid mechanics, the data that should have been subject to Eq. (2) and (3) must preprocess.

In engineering, the frequency and the hourly power of some water pumps and fans cannot be in accordance with Eq. (2) because of the mismatch among labels. The example of the mismatch is that the label "the frequency of chilled water pump #1" actually monitors the frequency of chilled water pump #2. In terms of chillers, missing data of on/off status is more severe than that of the hourly power. We correct the on/off status of every chiller by Eq. (3). When the chiller is off, the meter fluctuates around 0, so the threshold is set to check the on/ff status of every chiller. The threshold need to be determined based on expertise and reality.

Table 5
Hyperparameters of K-Means and DBSCAN.

K-Means			
n_cluster	sample_weight	init	other
5	None	k-means++	default
DBSCAN			
eps	min_sample	metric	other
0.3	1000	minkowski	default

$$\begin{cases} \frac{f_1}{f_2} = \frac{n_1}{n_2} = \frac{Q_1}{Q_2} = V \\ \left(\frac{f_1}{f_2}\right)^3 = \frac{P_1}{P_2} \end{cases} \quad (2)$$

where f is frequency of a piece of equipment, n is rotate speed of a piece of equipment, Q is water flow for every water pump or air flow for every fan. V is speed ratio.

$$\begin{cases} S_{chi} = 1 if P_{chi} > 0 \\ S_{chi} = 0 if P_{chi} \le 0.025 P_{irated} \end{cases}, i = 1, \cdots, k \quad (3)$$

where k is the total number of chillers, P_{chi} is the hourly power of the chiller #i, P_{irated} is the rated power of the chiller #i, S_{chi} is the on/off status of chiller #i.

Generally, the data of the fans for cooling towers is worse than those of other equipment:

1) every cooling tower has more than one fan, and these fans can be controlled independently. In many BECMPs, there is no correspondence between fans and cooling towers in labels, which means we have no idea which fans are in the same cooling towers.

2) entirely data missing often happens to the data of the hourly power.

But Pearson correlation coefficient matrix, Eq. (4), can be used to tell whether all fans are controlled synchronously. If that's the

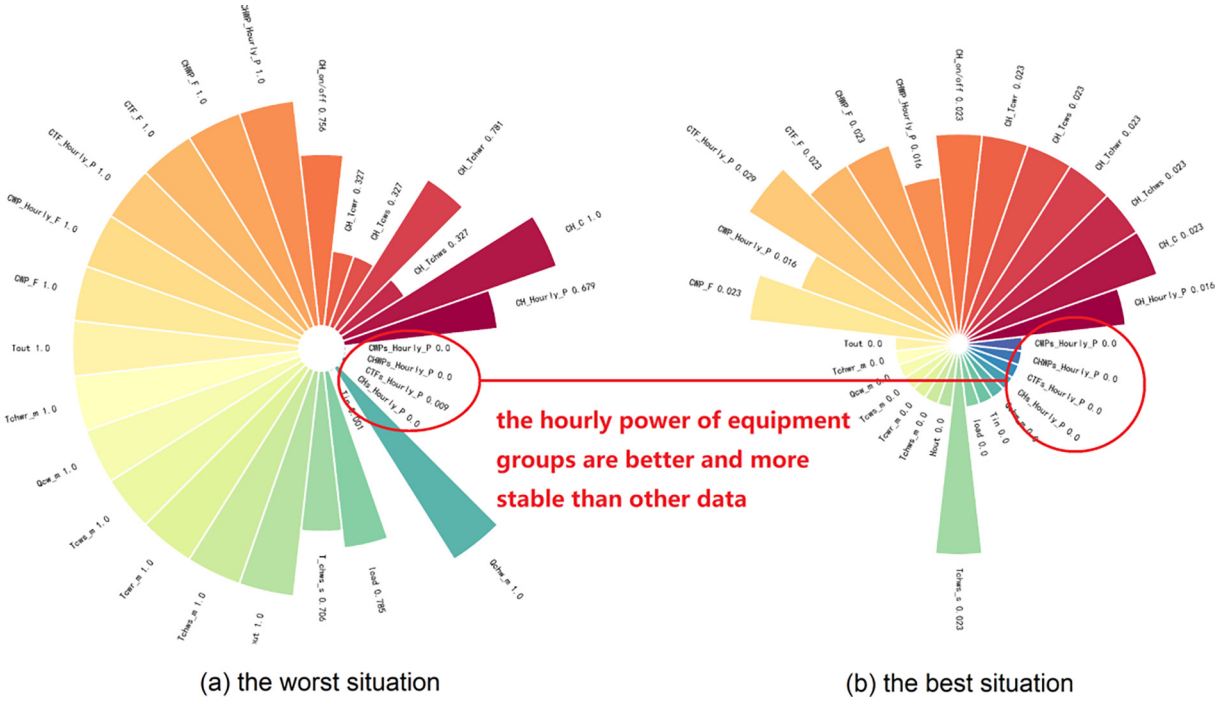


Fig. 6. The percent of data missing of different kinds of data. (a) the worst situation and (b) the best situation. The length of colored bar represents the missing rate, the longer the colored bar is, the higher the rate is.

Table 6
The centroids of different clusters (K-Means).

No. clustering	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
0	0.00458	0.000073	0.00121	15.616605	0.00014
1	0.99854	0.998538	0.99854	1	0.99854
2	0.46113	0.018894	0.22648	480.58832	0.06011
3	0.03078	0.000057	0.02172	31.968536	0.00229
4	0.10063	0.012422	0.07744	102.46261	0.02563
No. clustering	Feature #6	Feature #7	Feature #8	Feature #9	Feature #10
0	8.52e-07	5.435748	0.016392	0.130619	0.006681
1	2.14e-18	5	1	0.001969	0.99859
2	9.85e-03	7.351796	0.889354	2.597305	0.490453
3	3.96e-05	5.198959	0.17012	1.147859	0.051707
4	6.87e-04	5.252429	0.543724	3.258766	0.119578
No. clustering	Feature #11	Feature #12	Feature #13	Feature #14	
0	3.96e-01	0.007833	9.28e-01	0.00495	
1	1.09e-14	0.998547	−1.12e-12	0.99854	
2	6.49e + 00	0.526402	1.51e + 01	0.46367	
3	1.38e + 01	0.042414	1.45e + 01	0.0313	
4	2.14e + 00	0.116196	1.14e + 01	0.10283	

Features #1 - #14 are features #1 - #14 in Table 1.

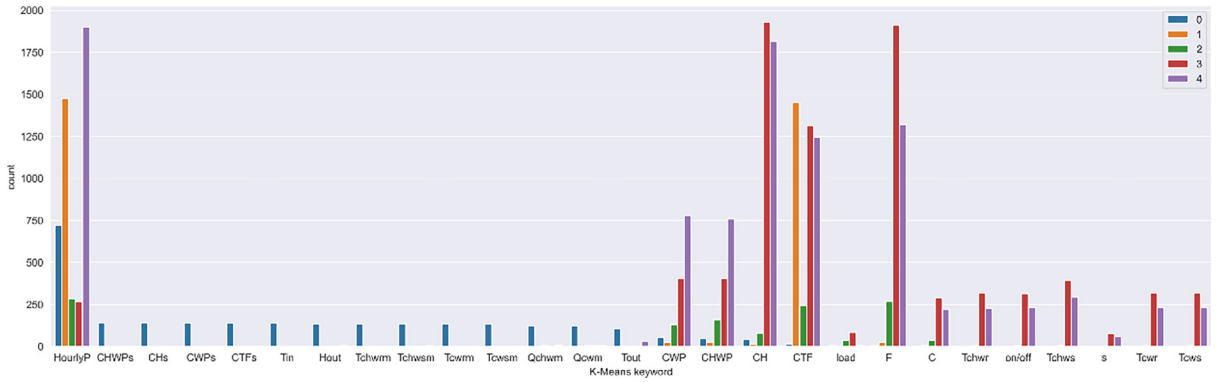


Fig. 7. The visualization of clustering result by K-Means. Different colored bar means different clusters. The length of colored bar means the count of keyword showing the corresponding cluster.

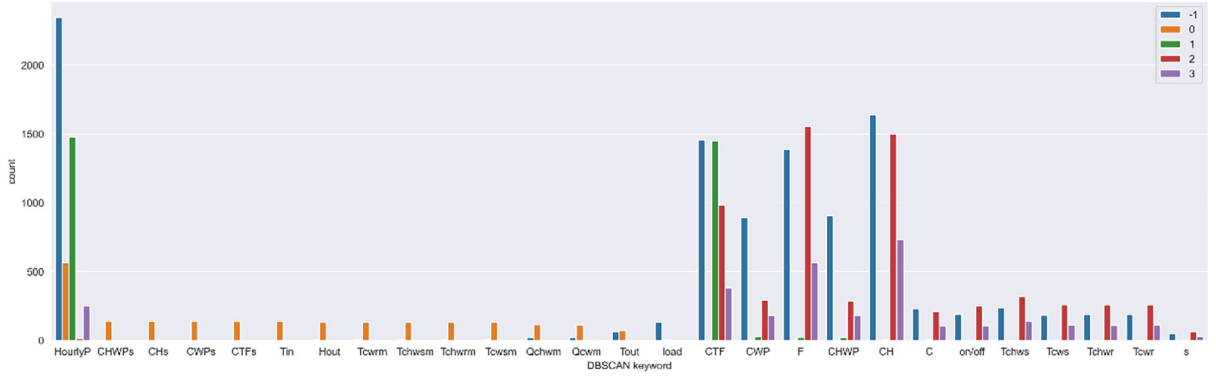


Fig. 8. The visualization of clustering result by DBSCAN. Different colored bar means different clusters. The length of colored bar means the count of keyword showing the corresponding cluster.

case, Eq. (2) and (3) can batch process these data of the fans for cooling towers.

$$r = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \tag{4}$$

Table 7
The key parameters of equipment for case #1.

Name	Quantity	Key parameters			
Chiller	3	Capacity	4462KW	Hourly_P	863KW
Chiller	1	Capacity	1371KW	Hourly_P	257.3KW
CHWP	2	Water flow	196 m^3/h	Hydraulic Head	32 m
		Hourly_P	–	Variable frequency	YES
CHWP	3	Water flow	639 m^3/h	Hydraulic Head	32 m
		Hourly_P	–	Variable frequency	YES
CWP	2	Water flow	281 m^3/h	Hydraulic Head	28 m
		Hourly_P	28KW	Variable frequency	YES
CWP	3	Water flow	911 m^3/h	Hydraulic Head	30 m
		Hourly_P	110KW	Variable frequency	YES
CT	–	Capacity	–	Water flow	–
		CTF_Hourly_P	5.5KW	The number of fans	–
CT	–	Capacity	–	Water flow	–
		CTF_Hourly_P	7.5KW	The number of fans	–

- "–" means that the parameters are unknown.

where $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n)$, X and Y are time series of frequency of every fan for cooling tower.

2.3.3. Check the law of the conversation of energy and similarity criterion

When the first two procedures are finished, the data about pipes can be checked sequentially.The detailed procedure is Fig. 4. The first step is to compare the water flow calculated by the similarity criterion and the original water flow in two loops. CV-RMSE Eq. (7) is used to tell if there are severe data quality issues happening to the original data of water flow. There are two scenarios.

1) When water flows in two loops meet the threshold of CV-RMSE, the law of conversation of energy, Eq. (8), is used to remove anomalies. The threshold of 0.25 aims to keep as much data as possible.

2) If one kind or both of water flow cannot meet the threshold, Eq. (6) is used to correct the original water flow. The threshold can be adjusted with the reality. The worse the data quality is, the bigger the threshold is. If the original data of water flow is far from the expertise, Fig. 4 should not be applied.

$$\begin{cases} Q_i = \frac{f_i}{f_{irated}} Q_{irated} \\ Q_{overall} = \sum_{i=1}^{n} Q_i \end{cases}, i = 1, \cdots, m \tag{5}$$

where Q_i is the water flow of chilled / cooling water pump #i, f_i is the frequency of chilled / cooling water pump #i, m means the total number of chilled / cooling water pumps. f_i is rated frequency (50 Hz in China) of chilled / cooling water pump #i, Q_{irated} is rated water flow of chilled / cooling water pump. $Q_{overall}$ is the main water flow in chilled / cooling water loop.

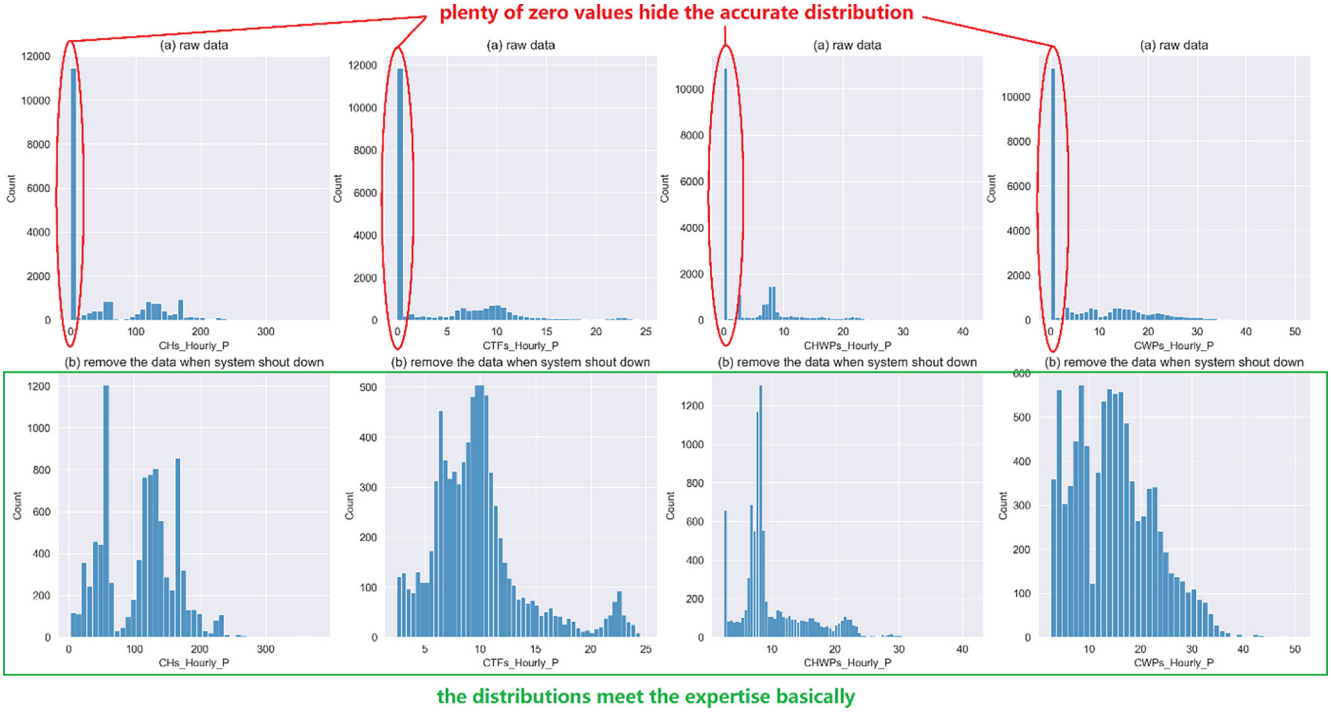


Fig. 9. The histogram about the overall hourly power of all chillers, pumps, and cooling towers: diagrams in the first row are from raw data and ones in the second row are data without zero values.

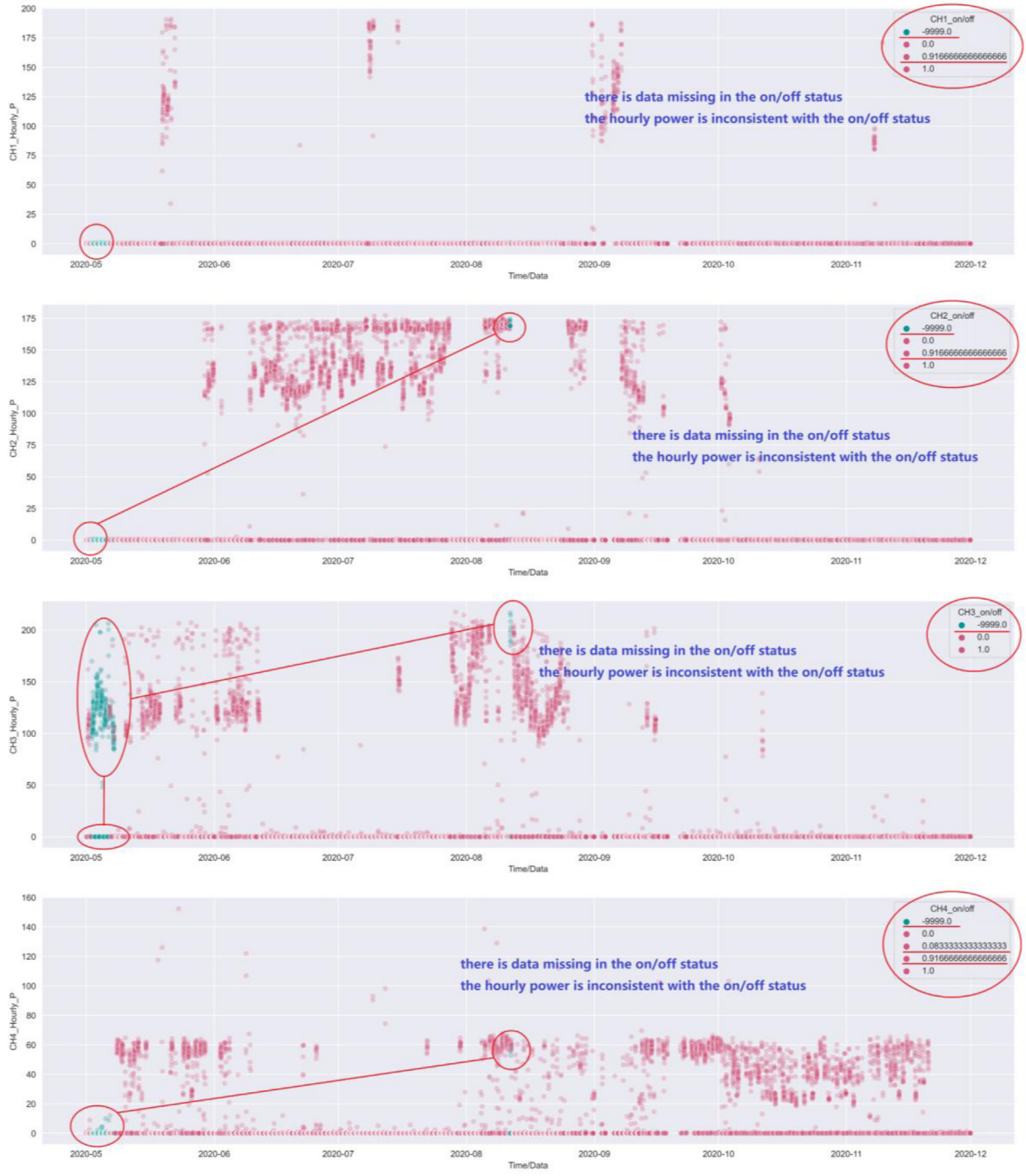
Fig. 10. The relationship between the summary of the hourly power of every chiller and the overall hourly power of all chillers. (a) raw data. (b) data without data missing. (c) after Kirchhoff's law.

Fig. 11. The relationship between the summary of the hourly power of every chilled water pumps and the overall hourly power of all chilled water pumps. (a) raw data. (b) data without data missing. (c) after Kirchhoff's law.

Fig. 12. The relationship between the summary of the hourly power of every cooling water pumps and the overall hourly power of all cooling water pumps. (a) raw data. (b) data without data missing. (c) after Kirchhoff's law.

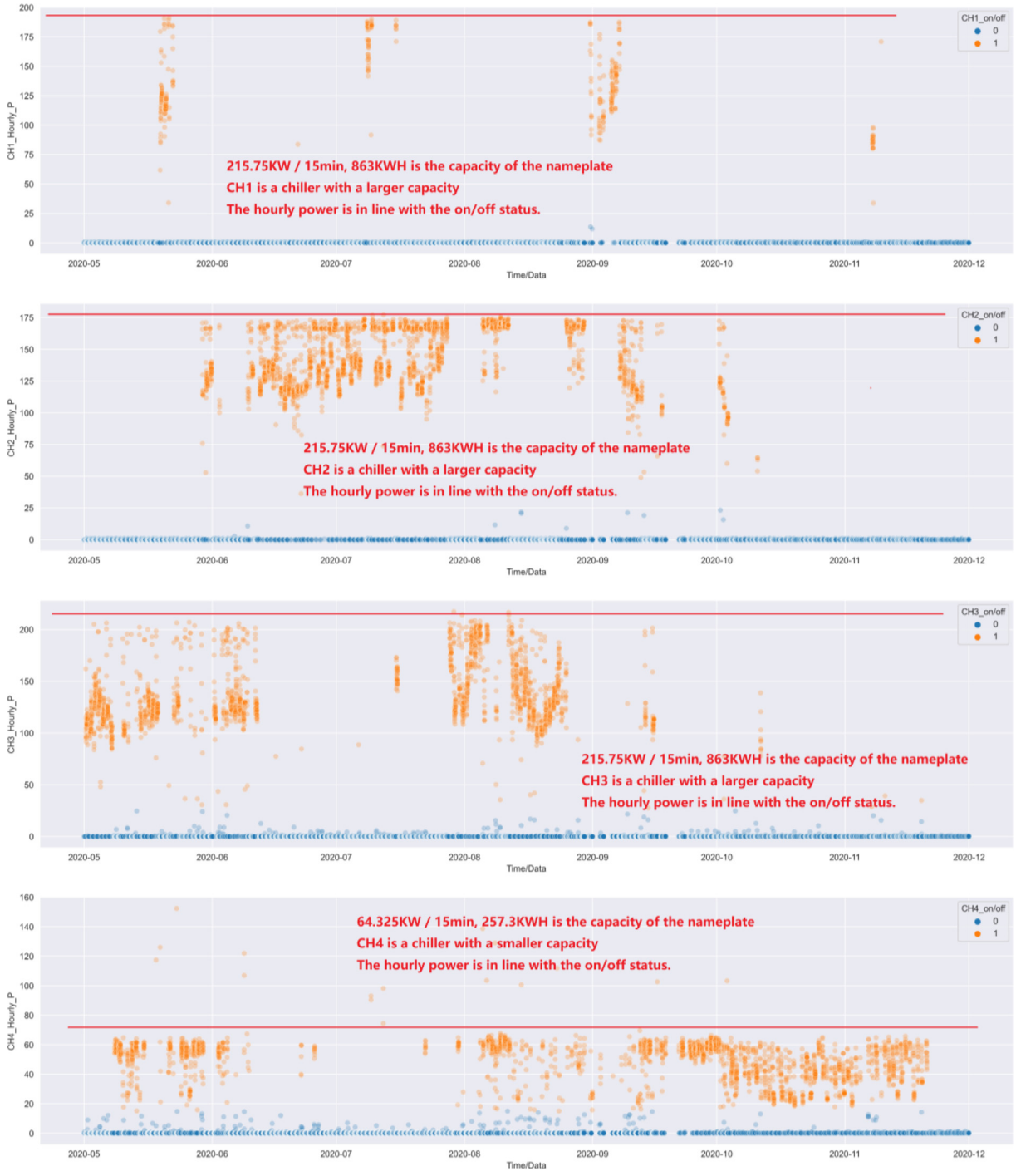
$$c_p \rho Q_{chw} \Delta t_{chw} + P_{all_{chs}} = c_p \rho Q_{cw} \Delta t_{cw} \tag{6}$$

Where c_p is specific heat capacity of water at constant pressure, $4.2kJ/(kg \cdot K)$ in engineering generally, ρ is density of water, $1000kg/m^3$ in engineering generally, Q_{chw} is main chilled water flow, Q_{cw} is main cooling water flow, P_{chs} is the overall hourly power of chillers. Δt_{chw} is chilled water temperature difference. Δt_{cw} is cooling water temperature difference.



(a)

Fig. 13. The relationship between power and on/off status (0: off 1: on). (a) raw data without data missing, and (b)data after processing.



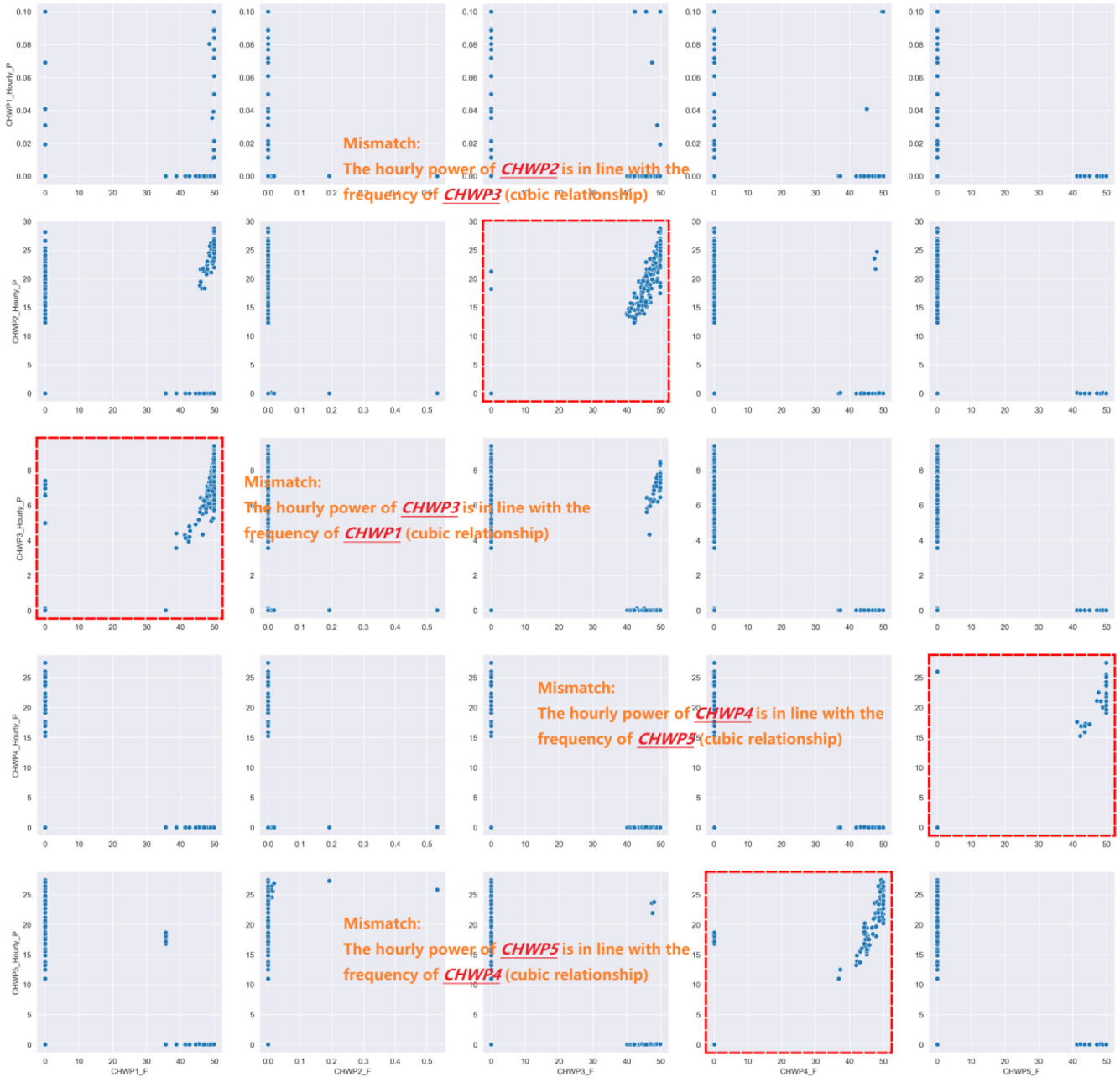
(b)

Fig. 13 (*continued*)

$$CV-RMSE=\frac{\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}{n}}}{\frac{\sum_{i=1}^{n}y_i}{n}} \tag{7}$$

$$-0.25\le\frac{\left(c_p\rho Q_{chwi}\Delta t_{chwi}+P_{chsi}-c_p\rho Q_{cwi}\Delta t_{cwi}\right)}{c_p\rho Q_{cwi}\Delta t_{cwi}}\le 0.25, i=1,\cdots,n \tag{8}$$

where y_i is the original hourly water flow, and $\widehat{y_i}$ is the calculated hourly water flow by frequency. n is the total number of data.



(a)

Fig. 14. The relationship between frequency and hourly power of individual chilled water pumps. (a) raw data (b) corrected data. In every subplot, x-coordinate is the frequency and y-coordinate is the hourly power. The subplots in the same line/column share the same y/x-coordinate.

2.3.4. Check other data

Other data, including the temperature of 4 ports of every chiller and weather data, are processed based on the above data and expertise.

3. Application

3.1. Analysis of data quality

3.1.1. Overall analysis

When we check the missing rate of different data *in one complex* (Fig. 5), we can see:

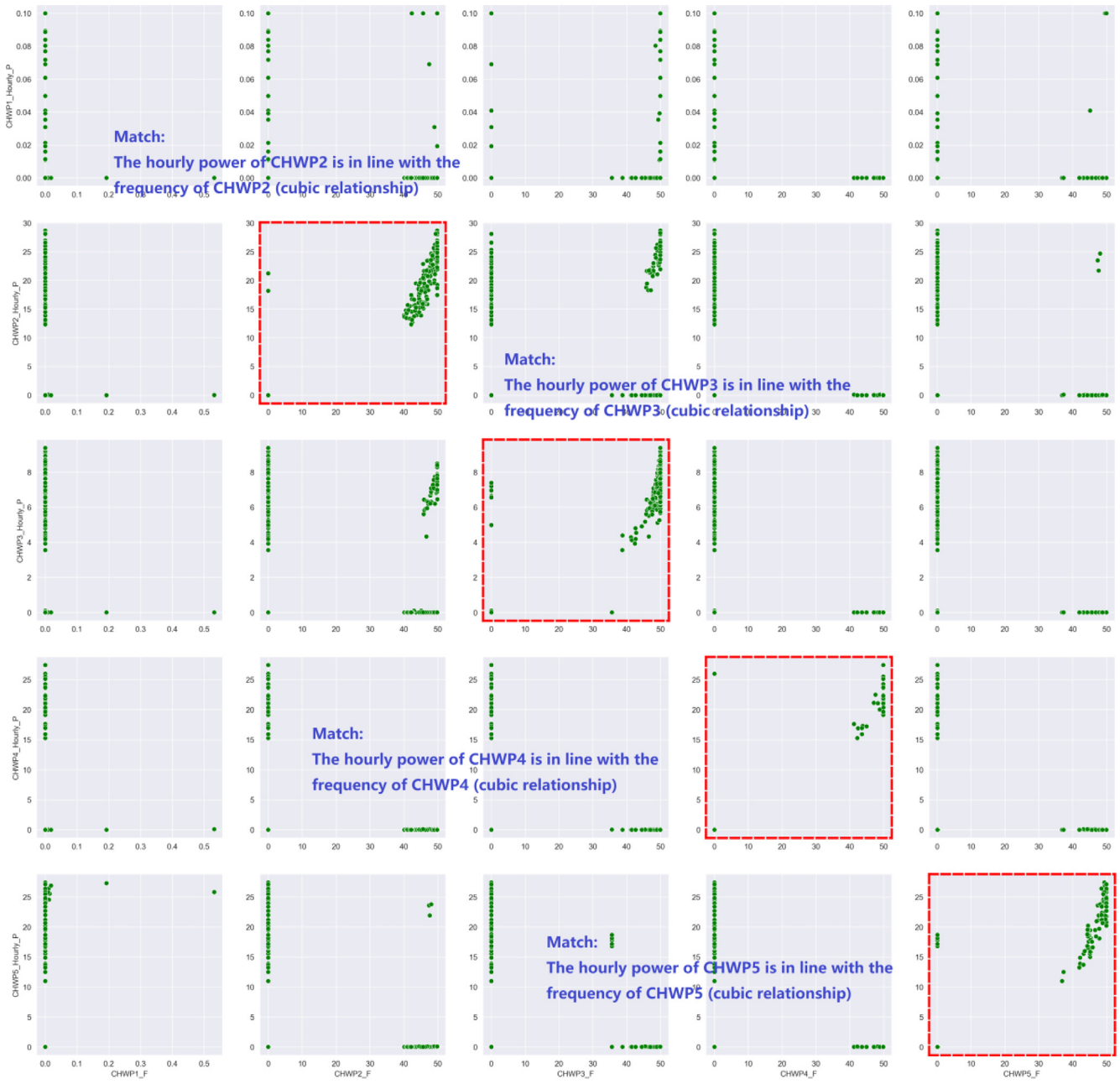
1) In terms of data missing, the data of chillers are better than those of other equipment.

2) For data of the hourly power of the fan for cooling tower, complete data missing happens in 19 out of 20 fans.

3) During the second half of 2020, the data of the hourly power of every water pump and load are missed intermittently and more frequently.

4) Data of temperature and water flow in header pipe, the overall hourly power of equipment group, indoor temperature, and weather are nearly intact.

When we check the missing rate of all data from *all complexes* (Fig. 6), we can see that:



(b)

Fig. 14 (*continued*)

1) In the best situation (Fig. 6 (b)), the missing rate of every kind of data can be low, even in the hourly power of the fans for cooling tower, the worst rate can go to 0.29 %.

2) In the worst situation (Fig. 6 (a).), the missing rate can rocket to 1.0 (complete data missing). But the missing rate is always low in the overall hourly power of equipment groups and indoor temperature.

According to macro- (complexes) and micro- (one complex) analysis, **the hourly power of equipment groups is suitable for the benchmark to process other data**. **The clustering results below prove it further**.

3.1.2. Cluster the features about missing rate

The 14 features are extracted from raw data; and then K-Means and DBSCAN are used to analyze if the data missing follows the patterns. Table 5 shows the hyper-parameters and Table 6 shows the characteristic of clusters:

1) the data quality of cluster #0 is excellent, and the overall missing rate (Feature #1) is low.

2) over the overall missing rate of cluster #1 is almost equal to 1, which means the data is missed entirely.

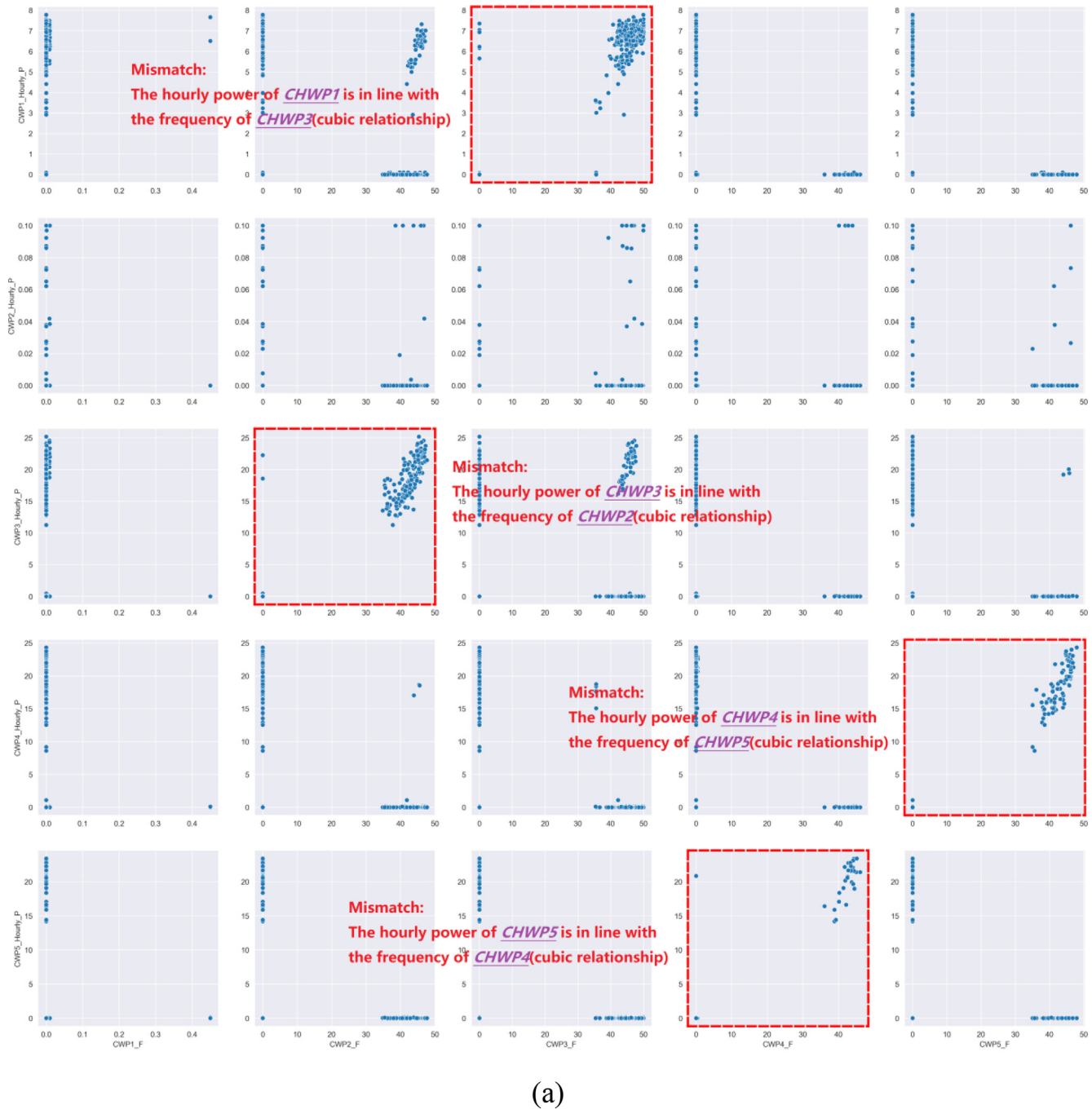


Fig. 15. The relationship between frequency and hourly power of individual cooling water pumps. (a) the raw data (b) the checked data. Other demonstrations are the same as those in Fig. 13.

3) In cluster #2, nearly half of the data is missing, and the data missing happens intermittently because the number of time windows (Feature #4) is much more than that of other clusters.

4) In clusters #3 and #4, the overall missing rates are more commonplace, but the time-related characteristics vary. Although both maximum missing rates (Feature #7) occurred in May 2020, the maximum missing rates (Feature #8) are 0.17 and 0.54, respectively.

The labels of data are processed before the number of occurrences of different labels is counted in each group (Fig. 7, Fig. 8). For instance, "CH1" becomes "CH" and "CHWP3" becomes "CHWP". Furthermore, to distinguish an equipment group from a piece of equipment, "CHs" means all chillers in a CES, and "CH" represent a chiller.

As shown in Fig. 7, all system-level data and indoor temperature belong to cluster #0. The hourly power of the fan for cooling tower is classified as cluster #1. Partial data of the hourly power and frequency of a piece of equipment are classified as cluster #2, but the majority of these data are classified as clusters #3 and #4 on almost average, which means there is a vast difference in the characteristic of data quality, including hourly power, and frequency of a piece of the equipment (chiller and water pump).

Fig. 8 is the result of DBSCAN, which is similar to that of K-Means. Cluster #0 corresponds to cluster #1 in Fig. 7. Cluster #1

(b)

Fig. 15 (*continued*)

corresponds to cluster #2 in Fig. 7. Cluster #3 corresponds to cluster #2 in Fig. 7. Cluster #-1, #2 correspond to cluster #3, #4.

According to the clustering results, the overall power of equipment groups is better than others, and that is in line with Section 3.1.2.

3.2. Case study

3.2.1. The reasons why the proposed framework is applied to two cases

This section will show the necessity of every step of the proposed general framework. Two cases are used to demonstrated how to apply the proposed framework to raw engineering data. There is marked differences between two cases. In case #1, the water flow is in accordance with expertise and reality, but the water flow is against expertise in case #2. Different methods are used to preprocess the water flows in two cases.

3.2.2. Case #1

(i) basic information.

Case #1 is located at Sichuan province, and that is hot summer and cold winter zone. The area of the complex is 106585 m^2. The BECMP belongs to #1 in Table 3, and Table 7 shows the key parameters of equipment.

(ii) check the benchmark for data preprocessing

The first step is to process data of the overall hourly power of equipment groups (Fig. 9). Without zero values, we can see that

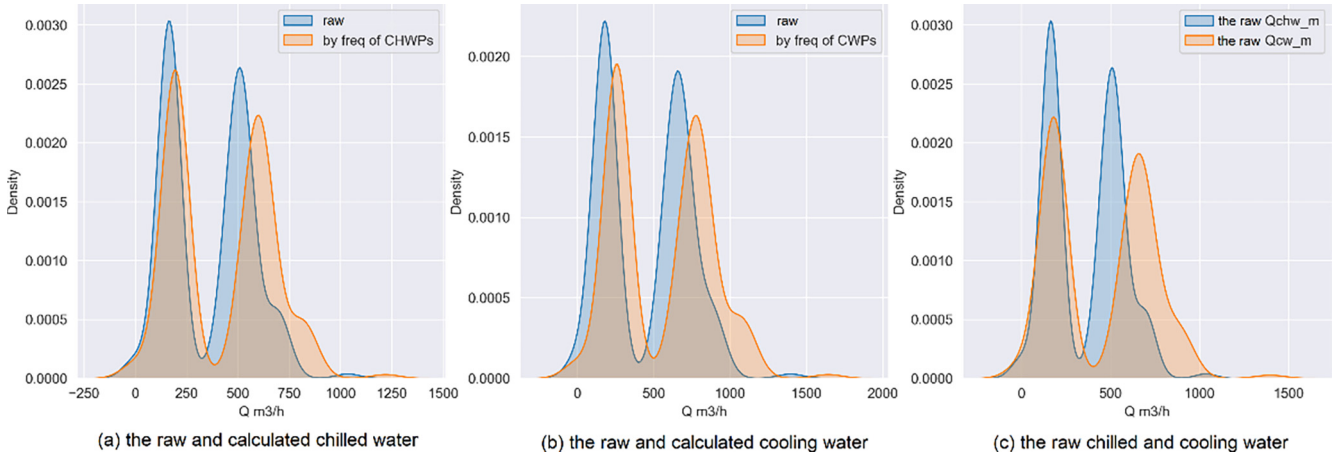


Fig. 16. Histogram of water flow in chilled/cooling water loop. The shapes of the distributions are similar.

Fig. 17. The scatter plots about energy conservation in two loops. (a) data with data missing. (b) before using the law of conservation of energy. (c) after using the law of conservation of energy.

the data of overall hourly power of equipment groups are quality: no data missing and sensible distribution.

(iii) check Kirchhoff's law.

After checking the data of the overall hourly power of equipment groups, the hourly power of every piece of equipment is processed by Kirchhoff's law. Theoretically, the summary of all hourly power of every chiller should equal the overall hourly power of all chillers (Fig. 10(c)). Furthermore, data of water pumps can be done similarly (Fig. 11, Fig. 12).

(vi) check correspondence among labels.

After processing the hourly power of every chiller, the on/off status of every chiller can be checked, ensuring the on/off status is in line with the hourly power. In Fig. 13(a), the raw data of every chiller's on/off status are weird, the checked data of the hourly power of every chiller are used to correct these data (Fig. 13(b)).

Next, the mismatch happening to water pumps should be solved. Fig. 14 and Fig. 15 show chilled and cooling water pump, respectively. For example, in the red box of the second row of Fig. 14(a), the raw data means that the combination between the frequency of pump #3 and the hourly power of pump #2 is in accordance with the similarity criterion, which does not make sense. And then, the proposed framework is used to solve these problems (Fig. 14(b)). In the red box of the second row in Fig. 14 (b), the checked data means that the combination between the frequency and the hourly power of pump #2 is in accordance with the similarity criterion, which does make sense.

(v) check the law of the conversation of energy and similarity criterion.

Firstly, the similarity criterion is used to check water flow in the header pipe (Fig. 16). There is not much difference between the original data and the calculated data (CV-RMSE$_{chw}$ = 0.21, CV-RMSE$_{cw}$ = 0.25), so **the original water flow is acceptable**. And then, Eq. (8) is used to process load and temperature differences in the header pipes (Fig. 17).

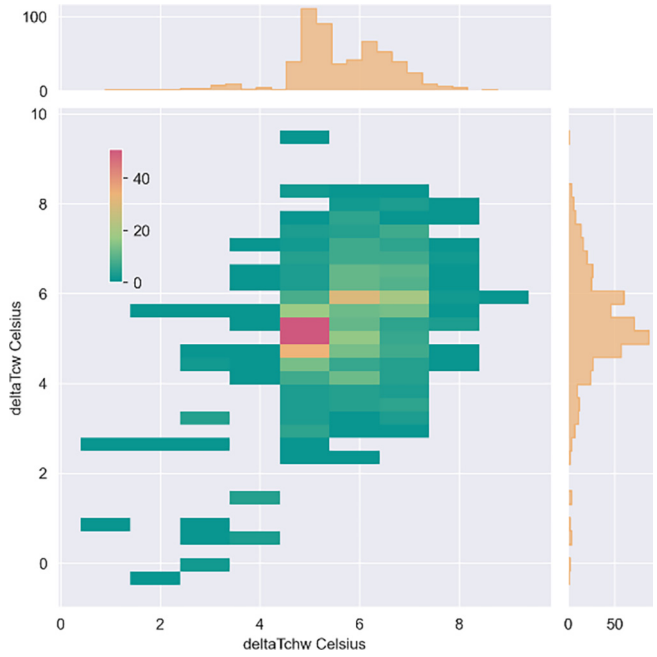


Fig. 18. The relationship about temperatures in header pipes in case #1.

(vi) check other data.

After the above procedure, the remaining data should be processed by expertise. In Fig. 18, the delta-Ts (difference in temperature between chilled and cooling water) are expected (5 degreesC approximately), which means temperature-related data can reflect the operation.

Fig. 19 shows that massive data missing happened to the hourly power of the fan for cooling tower, and that problem has been beyond the proposed framework.

(vii) energy efficiency analysis – verify the proposed framework.

Energy efficiency analysis is a good way to verify the checked data by the proposed framework. First, the checked data can show the relationships between chillers and water pumps but thr raw data cannot (Fig. 20). Fig. 20(a) and (b) are chilled and cooling water pumps, respectively.

These relationships are analyzed from two aspects: water pump frequency and hourly power. The right plots shows that the frequency is in line with the hourly power, but the left plots cannot show the thing. According to the checked data, it is clear that the relationships between chillers and water pumps are reasonable and in accordance with expertise.

Second, Fig. 21 shows that relationships between thermal load and the number of chillers under operation. In Fig. 21(b), the checked data shows that the thermal load becomes larger with the larger capacity of chillers under operation, but Fig. 21(a) shows the raw data is not in accordance with expertise: a chiller with larger capacity (the orange dots) cannot deal with 4000KW thermal load, which will lead to complaints about thermal comfort, but that did not happen in reality.

3.2.3. Case #2 - supplement

(i) basic information.

Case #2 is a supplement to Case #1. Case #2 is a complex in Guangxi province in hot summer and warm winter zones. The key parameters of equipment (Table 8) are more detailed than in Case #1.

(ii) check the law of the conversation of energy and similarity criterion.

All procedures are the same before we check the water flow data. **There are two severe problems in water flow** (Fig. 22), which has never been met in Case#1:

1) In (a), the raw chilled water (the blue points) flow surged in Nov 2020, which is unreasonable. The chilled water water should drop because the thermal load will drop when winter is approaching.

2) In (b), there is a massive gap between the raw chilled (the blue points) and cooling water (the orange points), which is unseasonable.3) In (c), the raw cooling water flow (the blue points) is beyond the upper limit of the rated water flow calculated by frequency (the orange points), CV-RMSE$_{cw}$ = 0.47. In reality, the cooling water flow cannot reach 3000 m^3/h approximately because of the on/off status of chillers, so the raw cooling water need to be corrected.

3) In (d), the water temperature differences between supply and return chilled/cooling water is 5 degrees Celsius approximately, which is in line with reality and expertise.

4) In (e), after removing the chilled water flow from Nov 2020 onward, the distribution of the raw chilled water flow is similar with that of the calculated water flow, so the raw chilled water is regarded as the benchmark to correct the cooling water flow and thermal load.

According to the above analysis, the cooling water flow is processed by the check chilled water flow, temperature (Fig. 22(f)).

(iii) energy efficiency analysis – verify the proposed framework.

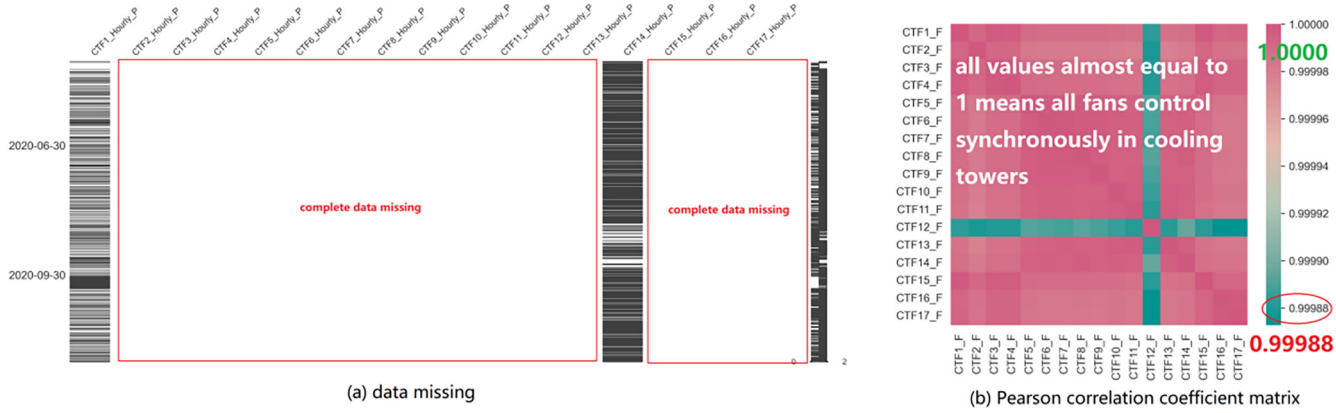
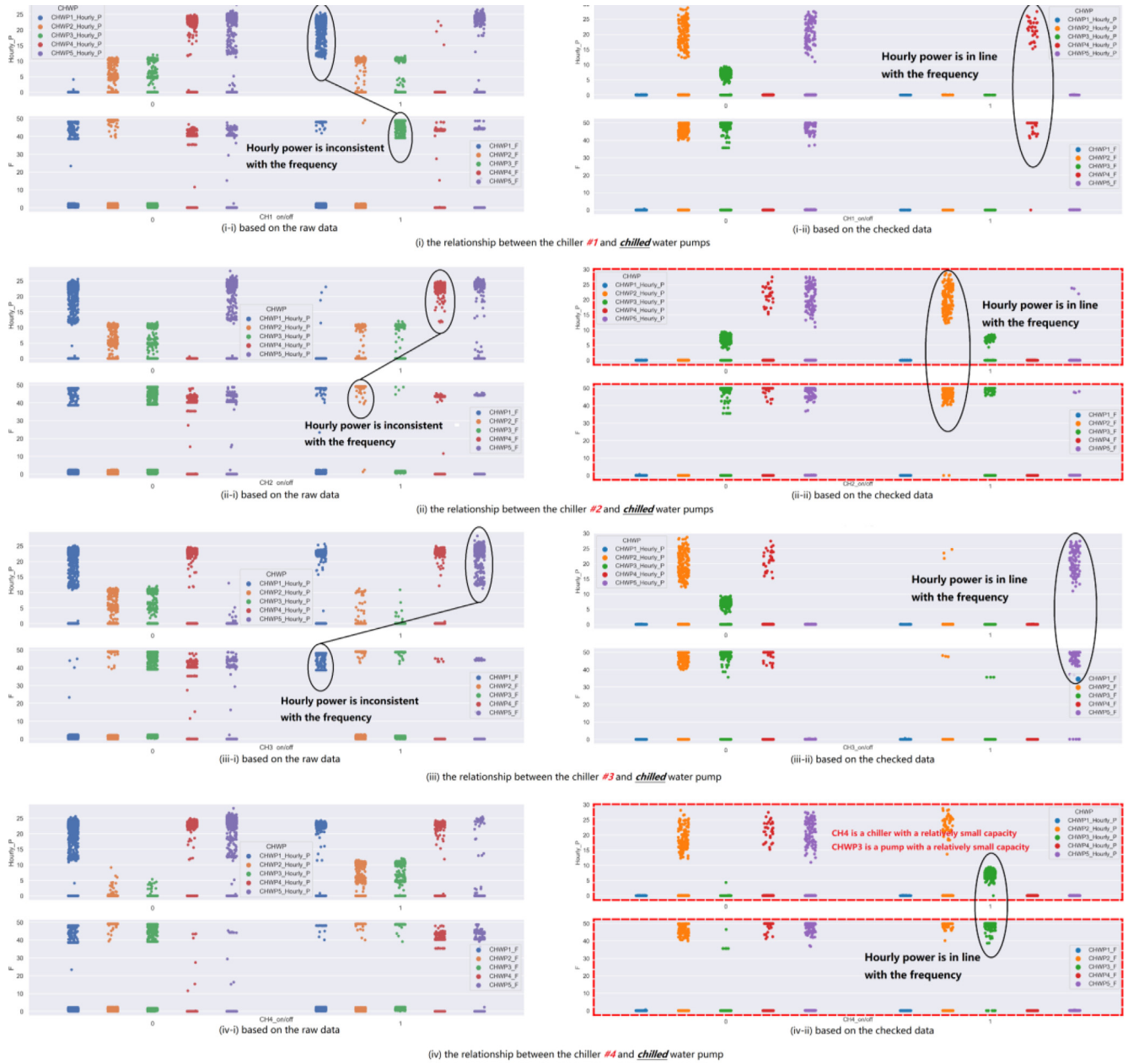


Fig. 19. Visualization for fans in cooling towers in case #1: (a) the hourly power of fans in cooling towers, (b) the frequency of fans in cooling towers.



(a) the relationships between chillers and chilled water pumps

Fig. 20. The relationships among chillers and water pumps. In every x-coordinate, 0 means the chiller is shut down and 1 means the chiller is running.

This part is the same as that in Case #1. Fig. 23(a) shows that the relationships between thermal load and the number of chillers unber operation is not accordance with expertise. After processing data by the proposed framework, the relationship is in line with expertise in Fig. 23(b).

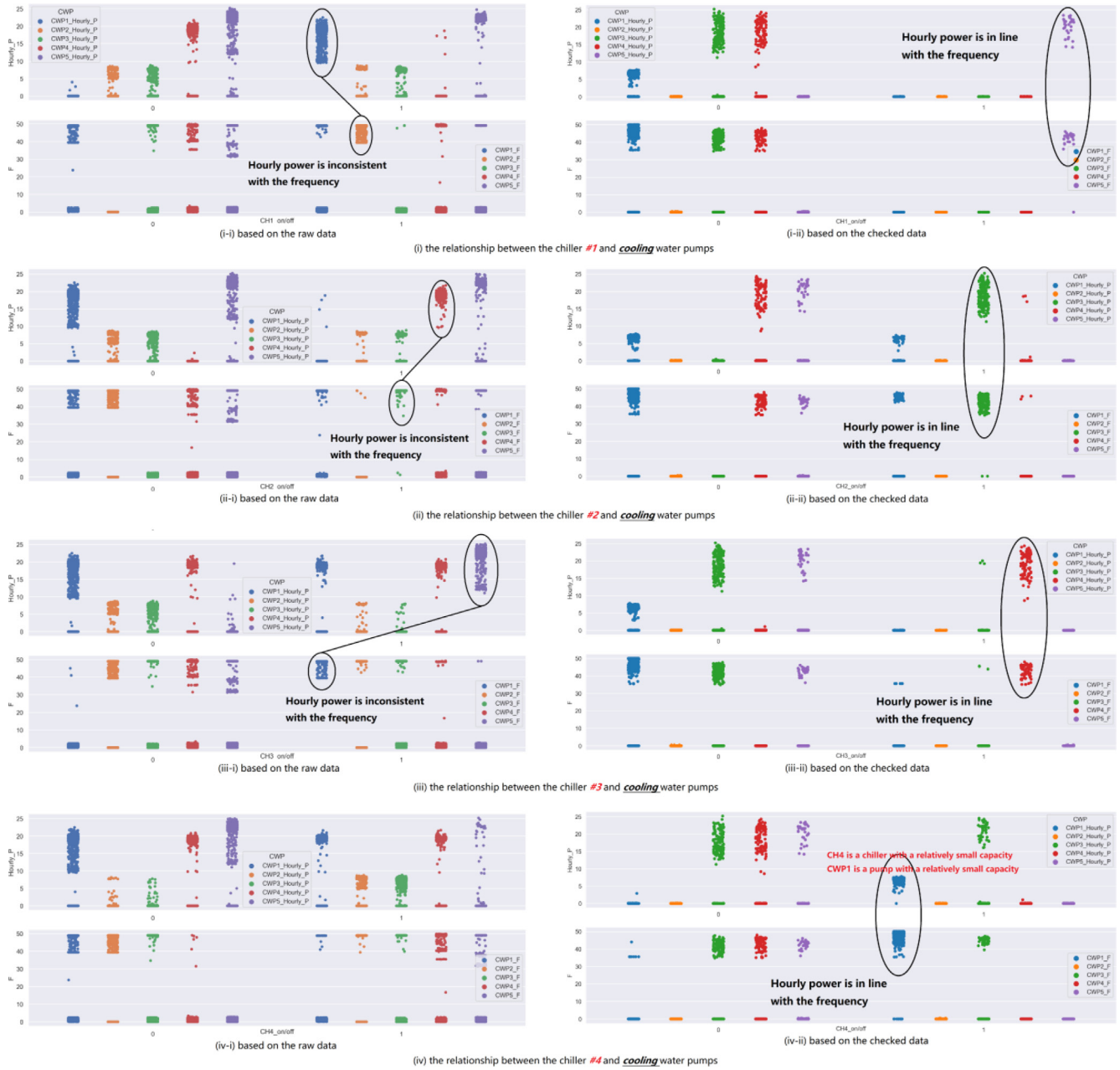
4. Suggestion of data collection

The distribution of sensors (Fig. 6) shows that much more emphasis is put on chillers and pipes than on other equipment: every chiller has six types of data, but other pieces of equipment only have two types of data. Worse, among 141 BECMPs, there is much difference in data quality.

Chillers: 1) The temperature data of four ports need engineers to check further. 2) It would be nice to be able to record the water flows of every chiller.

Water pumps: 1) Mismatch between data and labels is severe. 2) It would be better if the pressure data of the inlet and outlet could be recorded, which would help engineers maintain water pumps.

Cooling towers: 1) Massive data missing should get attention, which is a severe problem to solve. 2) Mismatch between data and labels confused the researchers. 3) it would be better to record the water flow and temperature data, which can help us tell if the performance degradation occurs in cooling towers.



(b) the relationships between chillers and cooling water pumps

Fig. 20 (*continued*)

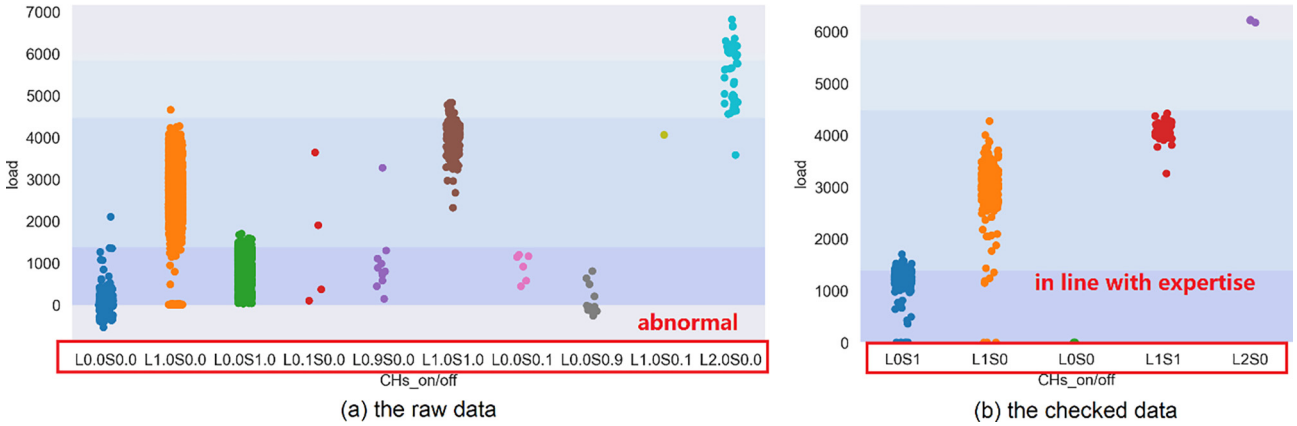


Fig. 21. Comparisons between the sum of nominal capacity of operating chillers and thermal load (case #1). L means the chiller with lager capacity and S means the chiller with smaller capacity. L1S1 means that 1 chiller with larger capacity and 1 chiller with smaller capacity are operating.

Table 8
The key parameters of equipment for case #2.

Name	Quantity	Key parameters			
Chiller	3	Capacity	3868KW	Hourly_P	721.9KW
		Tchws	6℃	Tchwr	12℃
		Tcws	37℃	Tcws	32℃
Chiller	1	Capacity	1363.9KW	Hourly_P	257.3KW
		Tchws	6℃	Tchwr	12℃
		Tcws	37℃	Tcws	32℃
CHWP	2	Water flow	196 m^3/h	Hydraulic Head	42 m
		Hourly_P	37KW	Variable frequency	YES
CHWP	3	Water flow	553 m^3/h	Hydraulic Head	42 m
		Hourly_P	110KW	Variable frequency	YES
CWP	2	Water flow	282 m^3/h	Hydraulic Head	32 m
		Hourly_P	45KW	Variable frequency	YES
CWP	3	Water flow	789 m^3/h	Hydraulic Head	32 m
		Hourly_P	45KW	Variable frequency	YES

Cooperation is necessary to improve the engineering data quality, although there are some interest conflicts. As a researcher, the top priority is to get as much data as possible. Still, the quality and quantity of data are not determined by themselves but by engineers who can check sensors to ensure big engineering data. Additionally, only a combination of engineering data and experimental data from manufacturers can help us know if the efficiency of building energy systems can be improved further Fig. 24.

5. Conclusion

Generally, our work can be divided into two parts: analyzing big engineering data quality and proposing a rule-based data preprocessing framework based on the analysis.

The analysis of data quality can be concluded:

1) The overall hourly power of equipment groups is excellent in quality because they are the basis of the electricity bills, so these data have the potential to be the benchmark for processing other data in the proposed framework.

2) The complete data missing happened to data of the fan for cooling tower, like the hourly power, which reminds engineers to check the data collection about cooling towers.

3) Mismatch between data labels puts an obstacle to data preprocessing, and the problem occurs in water pumps and cooling towers frequently, which exposes the fact that these equipment should have gotten enough attention. They play an integral role in improving building energy efficiency.

4) The data of weather are stable and excellent because they are collected by professional third party conpmaies.

And then, the rule-based data preprocessing framework is proposed. Lastly, two cases are used to verify the proposed framework and represent how to preprocess data when different problems happen to the engineering data.

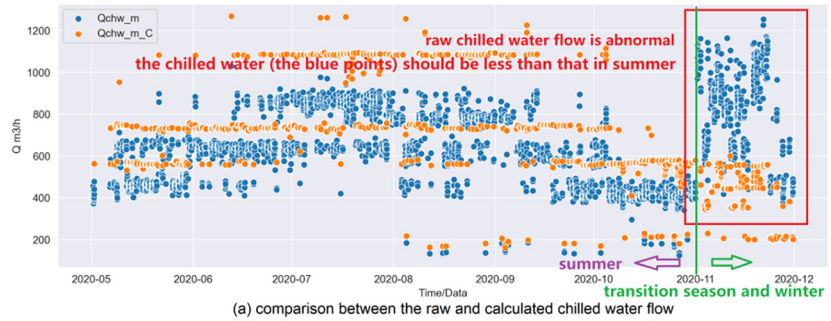
In the paper, the rule-based data preprocessing framework is proposed. The proposed framework makes full use of the laws of physics, which makes it explainable. The proposed framework does not need to train models in advances to process data, which makes it convenient. Last, the proposed framework is based on the analysis of engineering big data quality, so the proposed framework is suitable for engineering. But something needs to be improved in the future. The thresholds need to be set by expertise and reality. The framework can reduce but not be independence from the expertise.

CRediT authorship contribution statement

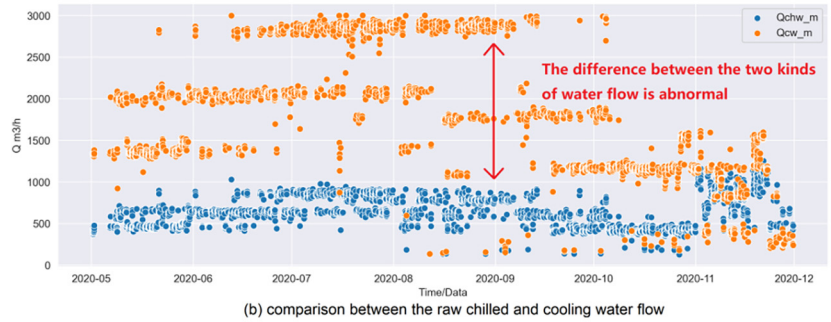
Ruikai He: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Tong Xiao:** Conceptualization, Writing - review & editing. **Shunian Qiu:** Data curation, Visualization. **Jiefan Gu:** Writing - review & editing. **Minchen Wei:** Writing - review & editing. **Peng Xu:** Supervision.

Data availability

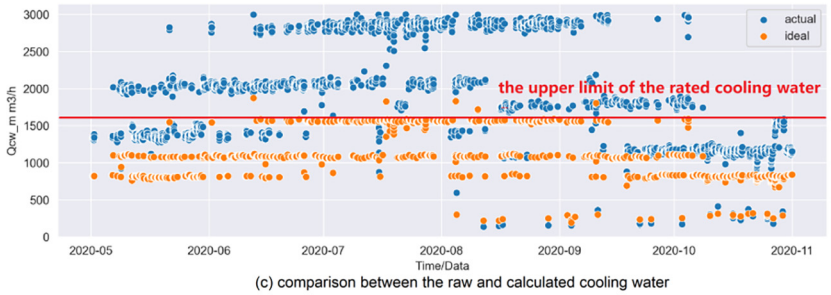
The authors do not have permission to share data.



(a) comparison between the raw and calculated chilled water flow

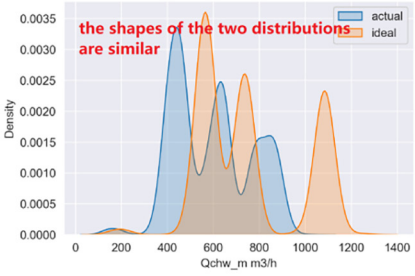


(b) comparison between the raw chilled and cooling water flow

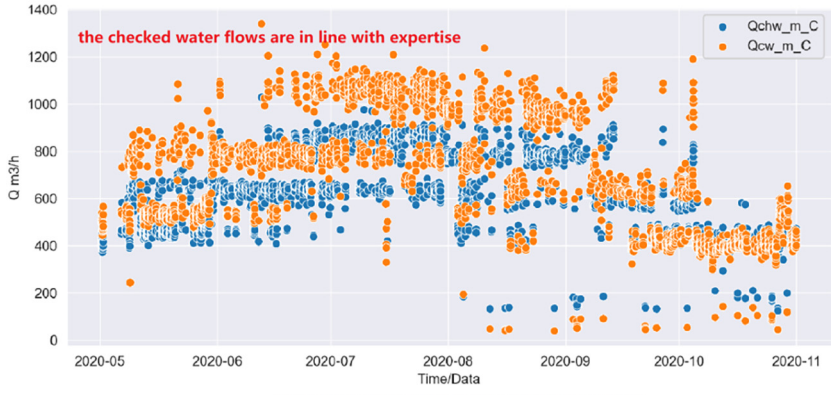


(c) comparison between the raw and calculated cooling water

(d) histogram of the deltaTs



(e) histogram of the raw and calculated chilled water after removing data in winter



(f) the checked water flows by the proposed framework

Fig. 22. Comparison among different kinds of water flow.

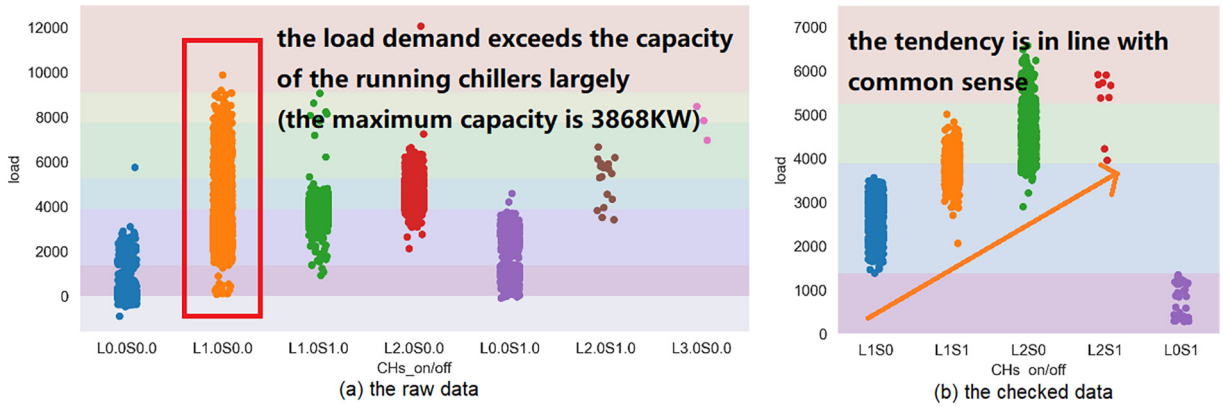


Fig. 23. Comparisons between the sum of nominal capacity of operating chillers and thermal load (case #2). Other demonstrations are the same as those in Fig. 20.

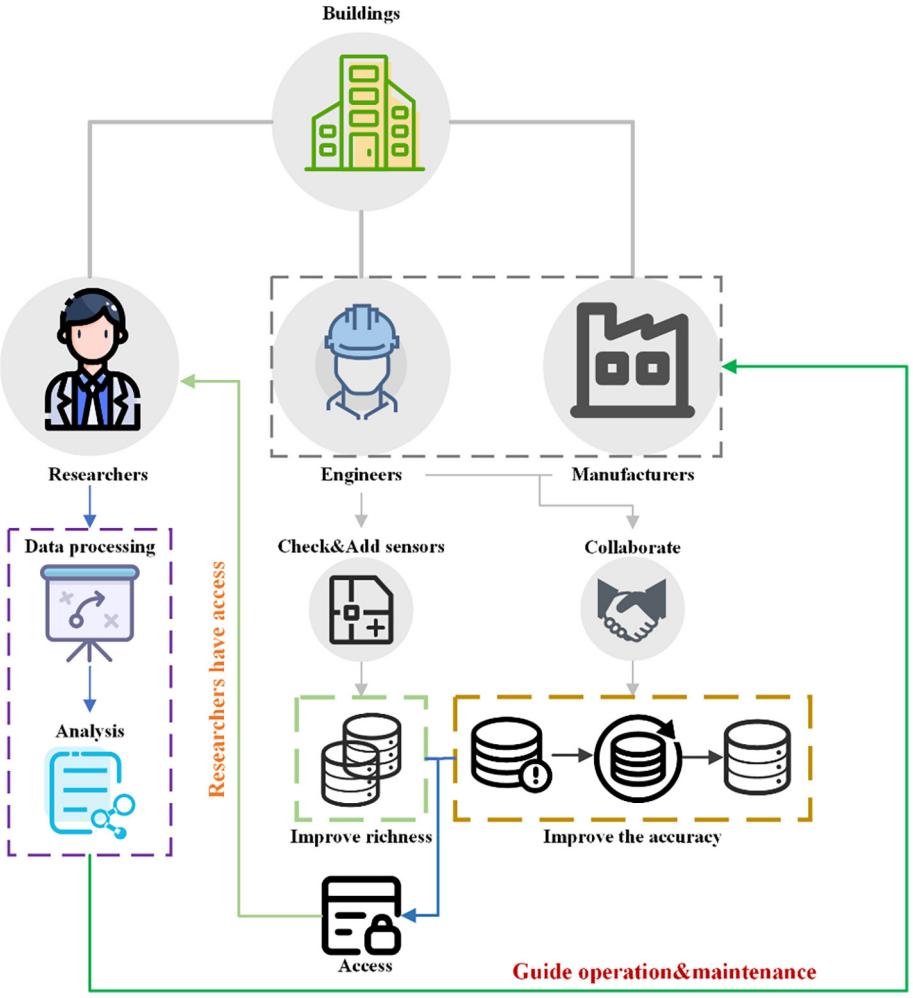


Fig. 24. The ideal situation of different parts involved building energy consumption. The collaboration among three parts is necessary to application of data-driven model in industry.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] IEA, Energy Technology Prospectives 2020, n.d. https://www.iea.org/reports/energy-technology-perspectives-2020.
[2] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E.P. Ollé, J. Oravec, M. Wetter, D.L. Vrabie, L. Helsen, All you need to know about model

predictive control for buildings, Annu. Rev. Control. 50 (2020) 190–232, https://doi.org/10/gj4mkt.
[3] J. Li, Y. Zhang, F. Kuang, Intelligent Building Automation and Control Based on IndasIBMS, in: 2013 Int. Conf. Serv. Sci. ICSS, 2013: pp. 266–270. https://doi.org/10/gn2dj3.
[4] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, F. Santos García, A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis, J. Build. Eng. 33 (2021), https://doi.org/10/ghqqsq 101692.
[5] J. Roth, R.K. Jain, Data-Driven, Multi-metric, and Time-Varying (DMT) Building Energy Benchmarking Using Smart Meter Data, in: I.F.C. Smith, B. Domer (Eds.), Adv. Comput. Strateg. Eng., Springer International Publishing, Cham, 2018: pp. 568–593. https://doi.org/10.1007/978-3-319-91635-4_30.
[6] U.S. Energy Information Administration, How many smart meters are installed in the United States, and who has them?, (2021). https://www.eia.gov/tools/faqs/faq.php?id=108&t=3.
[7] S. Zhou, M.A. Brown, Smart meter deployment in Europe: A comparative case study on the impacts of national policy schemes, J. Clean. Prod. 144 (2017) 22–32, https://doi.org/10.1016/j.jclepro.2016.12.031.
[8] A.M. Elsaid, H.A. Mohamed, G.B. Abdelaziz, M.S. Ahmed, A critical review of heating, ventilation, and air conditioning (HVAC) systems within the context of a global SARS-CoV-2 epidemic, Process Saf. Environ. Prot. 155 (2021) 230–261, https://doi.org/10/gmzz3r.
[9] M. Guo, P. Xu, T. Xiao, R. He, M. Dai, S.L. Miller, Review and comparison of HVAC operation guidelines in different countries during the COVID-19 pandemic, Build. Environ. 187 (2021), https://doi.org/10/gmdm6w 107368.
[10] Control of airborne infectious disease in buildings: Evidence and research priorities - Bueno de Mesquita - - Indoor Air - Wiley Online Library, (n.d.). https://onlinelibrary.wiley.com/doi/10.1111/ina.12965?__cf_chl_jschl_tk__=Mf9Z6yuVPiikpDr_DW5DDjTQeIfmpr_lnWW8U2EdB94-1641473434-0-gaNycGzNB-U (accessed January 6, 2022).
[11] T. Hong, D. Yan, S. D'Oca, C. Chen, Ten questions concerning occupant behavior in buildings: The big picture, Build. Environ. 114 (2017) 518–530, https://doi.org/10/f9qv4p.
[12] W. O'Brien, H.B. Gunay, The contextual factors contributing to occupants' adaptive comfort behaviors in offices – A review and proposed modeling framework, Build. Environ. 77 (2014) 77–87, https://doi.org/10/f56rx2.
[13] R. Parmar, A. Leiponen, L.D.W. Thomas, Building an organizational digital twin, Bus. Horiz. 63 (2020) 725–736, https://doi.org/10/gmg5zz.
[14] A. Li, F. Xiao, C. Zhang, C. Fan, Attention-based interpretable neural network for building cooling load prediction, Appl. Energy. 299 (2021), https://doi.org/10/gn2d7f 117238.
[15] W. Wang, T. Hong, X. Xu, J. Chen, Z. Liu, N. Xu, Forecasting district-scale energy dynamics through integrating building network and long short-term memory learning algorithm, Appl. Energy. 248 (2019) 217–230, https://doi.org/10.1016/j.apenergy.2019.04.085.
[16] Evolutionary double attention-based long short-term memory model for building energy prediction: Case study of a green building - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/S0306261921001902?via%3Dihub (accessed January 6, 2022).
[17] C. Fan, F. Xiao, M. Song, J. Wang, A graph mining-based methodology for discovering and visualizing high-level knowledge for building energy management, Appl. Energy. 251 (2019), https://doi.org/10/ggkb47 113395.
[18] K. Mason, S. Grijalva, A review of reinforcement learning for autonomous building energy management, Comput. Electr. Eng. 78 (2019) 300–312, https://doi.org/10/gmns65.
[19] R. He, P. Xu, Z. Chen, W. Luo, Z. Su, J. Mao, A non-intrusive approach for fault detection and diagnosis of water distribution systems based on image sensors, audio sensors and an inspection robot, Energy Build. 243 (2021), https://doi.org/10/gn2d7j 110967.
[20] T. Xiao, P. Xu, R. He, H. Sha, Status quo and opportunities for building energy prediction in limited data Context—Overview from a competition, Appl. Energy. 305 (2022), https://doi.org/10/gn2fz5 117829.
[21] E. Azizi, R. Ahmadiahangar, A. Rosin, J. Martins, R.A. Lopes, M.T.H. Beheshti, S. Bolouki, Residential energy flexibility characterization using non-intrusive load monitoring, Sustain. Cities Soc. 75 (2021), https://doi.org/10/gn2d6v 103321.
[22] Overview of non-intrusive load monitoring and identification techniques - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/S2405896315030566?via%3Dihub (accessed January 6, 2022).
[23] A. Meier, D. Cautley, Practical limits to the use of non-intrusive load monitoring in commercial buildings, Energy Build. 251 (2021), https://doi.org/10/gn2d63 111308.
[24] R. Enríquez, M.J. Jiménez, M.R. Heras, Towards non-intrusive thermal load Monitoring of buildings: BES calibration, Appl. Energy. 191 (2017) 44–54, https://doi.org/10/f933vs.
[25] The Role of Virtual Reality in Autonomous Vehicles' Safety | IEEE Conference Publication | IEEE Xplore, (n.d.). https://ieeexplore.ieee.org/document/8942308 (accessed January 6, 2022).
[26] Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches | SpringerLink, (n.d.). https://link.springer.com/article/10.1007%2Fs12273-020-0723-1 (accessed January 6, 2022).
[27] Ž. Turk, Ten questions concerning building information modelling, Build. Environ. 107 (2016) 274–284, https://doi.org/10/f84xf9.
[28] H. Gao, C. Koch, Y. Wu, Building information modelling based building energy modelling: A review, Appl. Energy. 238 (2019) 320–343, https://doi.org/10/gjsqn2.
[29] Building Information Modeling data interoperability for Cloud-based collaboration: Limitations and opportunities - Kereshmeh Afsari, Charles Eastman, Dennis Shelden, 2017, (n.d.). https://journals.sagepub.com/doi/10.1177/1478077117731174 (accessed January 6, 2022).
[30] Intelligent Building Construction Management Based on BIM Digital Twin, (n.d.). https://www.hindawi.com/journals/cin/2021/4979249/ (accessed January 6, 2022).
[31] J.J. Hunhevicz, M. Motie, D.M. Hall, Digital building twins and blockchain for performance-based (smart) contracts, Autom. Constr. 133 (2022), https://doi.org/10/gndb6n 103981.
[32] Design of a mass-customization-based cost-effective Internet of Things sensor system in smart building spaces - Sanguk Park, Sangmin Park, Jinsung Byun, Sehyun Park, 2016, (n.d.). https://journals.sagepub.com/doi/10.1177/1550147716660895 (accessed January 6, 2022).
[33] L. Chenyan, N. Jing, S. Hui-Wei, Research of Carbon Emission Reduction on the Green Building Based on the Internet of Things, in: 2019 Int. Conf. Smart Grid Electr. Autom. ICSGEA, 2019: pp. 83–87. https://doi.org/10/gn2f55.
[34] C. Miller, M. Abdelrahman, A. Chong, F. Biljecki, M. Quintana, M. Frei, M. Chew, D. Wong, The Internet-of-Buildings (IoB) — Digital twin convergence of wearable and IoT data with GIS/BIM, J. Phys. Conf. Ser. 2042 (2021), https://doi.org/10/gn2f56 012041.
[35] C.E. Kontokosta, D. Spiegel-Feld, S. Papadopoulos, Mandatory building energy audits alone are insufficient to meet climate goals, Nat. Energy. 5 (2020) 282–283, https://doi.org/10/gn2f6c.
[36] C.E. Kontokosta, D. Spiegel-Feld, S. Papadopoulos, The impact of mandatory energy audits on building energy use, Nat. Energy. 5 (2020) 309–316, https://doi.org/10/gmq2kr.
[37] N. Brown, A.J. Wright, A. Shukla, G. Stuart, Longitudinal analysis of energy metering data from non-domestic buildings, Build. Res. Inf. 38 (2010) 80–91, https://doi.org/10.1080/09613210903374788.
[38] Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/pii/S0378778821001699?via%3Dihub (accessed January 6, 2022).
[39] Z. Chen, Y. Chen, R. He, J. Liu, M. Gao, L. Zhang, Multi-objective residential load scheduling approach for demand response in smart grid, Sustain. Cities Soc. 76 (2022), https://doi.org/10/gn2f7q 103530.
[40] X. Meng, Research on Reconstruction and Repair of Missing Data in Building Energy Monitoring System, Master, Dalian University of Technology, 2021. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFDTEMP&filename=1021695457.nh&uniplatform=NZKPT&v=rTkhLdkobnlJMDEkAxlubejcApfNKZQ0dTquGfrSKW7NtTcQ9oymglKf7TmsCSki (accessed January 6, 2022).
[41] Review of Missing Data Processing Methods-All Databases, (n.d.). https://www.webofscience.com/wos/alldb/full-record/CSCD:7029646 (accessed January 6, 2022).
[42] The ASHRAE Great Energy Predictor III competition: Overview and results: Science and Technology for the Built Environment: Vol 26, No 10, (n.d.). https://www.tandfonline.com/doi/full/10.1080/23744731.2020.1795514 (accessed January 6, 2022).
[43] C. Zhang, L. Cao, A. Romagnoli, On the feature engineering of building energy data mining, Sustain. Cities Soc. 39 (2018) 508–518, https://doi.org/10.1016/j.scs.2018.02.016.
[44] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118, https://doi.org/10.1016/j.enbuild.2014.02.005.
[45] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renew. Sustain. Energy Rev. 23 (2013) 272–288, https://doi.org/10.1016/j.rser.2013.03.004.
[46] D.B. Crawley, L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, R.K. Strand, R.J. Liesen, D.E. Fisher, M.J. Witte, J. Glazer, EnergyPlus: creating a new-generation building energy simulation program, Energy Build. 33 (2001) 319–331, https://doi.org/10.1016/S0378-7788(00)00114-6.
[47] D. Yan, J. Xia, W. Tang, F. Song, X. Zhang, Y. Jiang, DeST — An integrated building simulation toolkit Part I: Fundamentals, Build. Simul. 1 (2008) 95–110, https://doi.org/10.1007/s12273-008-8118-8.
[48] P. de Wilde, The gap between predicted and measured energy performance of buildings: A framework for investigation, Autom. Constr. 41 (2014) 40–49, https://doi.org/10.1016/j.autcon.2014.02.009.
[49] J. Ma, J.C.P. Cheng, F. Jiang, W. Chen, M. Wang, C. Zhai, A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data, Energy Build. 216 (2020), https://doi.org/10/gn2f8q 109941.
[50] E.C. Malthouse, F.J. Mulhern, Book Review: handbook of data mining and knowledge discovery, J. Mark. Res. 40 (2003) 372–374, https://doi.org/10/b3ffdt.

[51] C. Fan, M. Chen, X. Wang, J. Wang, B. Huang, A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data, Front. Energy Res. 9 (2021) 77, https://doi.org/10/gn2f9j.
[52] Research on Preprocessing Technology of Building Energy Consumption Monitoring Data Based on Machine Learning Algorithm-All Databases, (n.d.). https://www.webofscience.com/wos/alldb/full-record/CSCD:6207036 (accessed January 6, 2022).
[53] M. Amiri, R. Jensen, Missing data imputation using fuzzy-rough methods, Neurocomputing. 205 (2016) 152–164, https://doi.org/10/gn2f9m.
[54] J. Liu, J. Liu, H. Chen, Y. Yuan, Z. Li, R. Huang, Energy diagnosis of variable refrigerant flow (VRF) systems: Data mining technique and statistical quality control approach, Energy Build. 175 (2018) 148–162, https://doi.org/10/gd9phc.
[55] L. Zhang, A pattern-recognition-based ensemble data imputation framework for sensors from building energy systems, Sensors 20 (2020) 5947, https://doi.org/10/gn2f9p.
[56] D. Jeong, C. Park, Y.M. Ko, Missing data imputation using mixture factor analysis for building electric load data, Appl. Energy. 304 (2021), https://doi.org/10/gn2f9s 117655.
[57] W. Cao, D. Wang, J. Li, H. Zhou, Y. Li, L. Li, BRITS: Bidirectional Recurrent Imputation for Time Series, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. CesaBianchi, R. Garnett (Eds.), Adv. Neural Inf. Process. Syst. 31 Nips 2018, Neural Information Processing Systems (nips), La Jolla, 2018. https://www.webofscience.com/wos/alldb/full-record/WOS:000461852001033 (accessed January 6, 2022).
[58] Y. Luo, X. Cai, Y. Zhang, J. Xu, X. Yuan, Multivariate Time Series Imputation with Generative Adversarial Networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. CesaBianchi, R. Garnett (Eds.), Adv. Neural Inf. Process. Syst. 31 Nips 2018, Neural Information Processing Systems (nips), La Jolla, 2018. https://www.webofscience.com/wos/alldb/full-record/WOS:000461823301057 (accessed January 6, 2022).
[59] sklearn.preprocessing.MinMaxScaler — scikit-learn 1.1.1 documentation, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html#sklearn.preprocessing.MinMaxScaler (accessed July 23, 2022).