Build Simul
https://doi.org/10.1007/s12273-022-0885-0

Development of a key-variable-based parallel HVAC energy predictive model

Huajing Sha[1], Peng Xu[2] (✉), Chengchu Yan[3], Ying Ji[4], Kenan Zhou[5], Feiran Chen[5]

1. Terminus (Shanghai) Information Technology Co., LTD, Shanghai 200062, China
2. Department of Mechanical and Energy Engineering, Tongji University, Shanghai 201804, China
3. Department of Urban Construction, Nanjing Tech University, Nanjing 211816, China
4. Beijing Key Laboratory of Green Built Environment and Energy Efficient Technology, Beijing University of Technology, Beijing100124, China
5. China Southern Power Grid Guangdong Foshan Power Supply Bureau, Guangdong, China

Abstract

Building heating, ventilation, and air conditioning (HVAC) systems consume large amounts of energy, and precise energy prediction is necessary for developing various energy-efficiency strategies. Energy prediction using data-driven models has received increasing attention in recent years. Typically, two types of driven models are used for building energy prediction: sequential and parallel predictive models. The latter uses the historical energy of the target building as training data to predict future energy consumption. However, for newly built buildings or buildings without historical data records, the energy can be estimated using the parallel model, which employs the energy data of similar buildings as training data. The second predictive model is seldom studied because the model input feature is difficult to identify and collect. Herein, we propose a novel key-variable-based parallel HVAC energy predictive model. This model has informative input features (including meteorological data, occupancy activity, and key variables representing building and system characteristics) and a simple architecture. A general key-variable screening toolkit which was more versatile and flexible than present parametric analysis tools was developed to facilitate the selection of key variables for the parallel HVAC energy predictive model. A case study is conducted to screen the key variables of hotel buildings in eastern China, based on which a parallel chiller energy predictive model is trained and tested. The average cross-test error measured in terms of the coefficient of variation of the root mean square error (CV-RMSE) and normalized mean bias error (NMBE) of the parallel chiller energy predictive model is approximately 16% and 8.3%, which is acceptable for energy prediction without using historical energy data of the target building.

Keywords

HVAC energy prediction;
data-driven model;
sequential predictive model;
parallel predictive model;
key-variable screening;
sensitivity analysis

Article History

Received: 22 June 2021
Revised: 31 December 2021
Accepted: 03 January 2022

1 Introduction

Buildings account for more than 30% of the total energy consumption worldwide. The building heating, ventilation, and air conditioning (HVAC) system is one of the highest energy consumers in building service systems (Liu et al. 2019). Engineers try to reduce HVAC energy consumption using methods such as optimizing the design scheme and implementing efficient operating strategies. Most of these energy-efficient strategies rely on accurate energy consumption predictions. There are two main methods for building HVAC energy prediction (Sha et al. 2019). The first is the use of physical-based models, such as energy simulation tools. However, using simulation tools to precisely calculate HVAC energy is difficult. Current simulation tools have the following disadvantages:

(1) A large amount of information (i.e., model input parameters) is needed.
(2) Development of the building geometric model is time consuming.

E-mail: xupeng@tongji.edu.cn

(3) The computation time for buildings with complex forms or systems is too long (Hong et al. 2008).
(4) The simulation result has a large deviation from the actual value because of uncertainties caused by the input parameters and physical model simplification (Higdon et al. 2004).

Because of these disadvantages, researchers tend to perform energy prediction with data-driven models, which is much more efficient and precise (Zhao et al. 2020). Two types of data-driven models have been studied for predicting building energy (Sha et al. 2021). The first uses the historical energy data of the target building as training data to predict its future energy. This is referred to as the sequential predictive model. The parallel predictive model, on the other hand, involves the use of historical data from similar buildings. The parallel predictive model is suitable for newly built buildings or buildings without historical data records. However, this type of model has seldom been studied because its input features are difficult to identify. The data-driven model performance is significantly influenced by the model input features. The input features should contain the driving factors for target (the HVAC energy consumption) variation. For the sequential predictive model, variables such as meteorological parameters and occupancy schedule are the only driving factors with respect to a specific building. The building- and system-related variables were excluded because they remained the same. However, for the parallel predictive model, variables that distinguish between different buildings and systems should also be included as model input features. Previous studies have revealed that building HVAC energy is mainly determined by a few variables (Tian et al. 2018). These variables are referred to as *key variables* in this study. However, dozens of variables may influence HVAC energy but collecting all of them is difficult. Moreover, a large dimension of the input feature will cause a curse of dimensionality (Pardalos 2009) and degrade model performance when the training data size is small. In this paper, we propose a novel method for developing parallel HVAC energy predictive models. A key-variable screening toolkit was developed to facilitate the selection of input features of parallel HVAC energy predictive models. This key-variable screening toolkit was developed based on EnergyPlus and sensitivity analysis (SA). Finally, a parallel data-driven model for predicting the chiller energy of hotel buildings in eastern China was developed and tested to verify the feasibility of the proposed parallel model development method. The contributions and novelty of this study are summarized as follows:

(1) A novel framework for a parallel HVAC energy predictive model is proposed. This framework incorporates various input features regarding meteorological variation, occupancy activity, and key variables representing building and system characteristics. Moreover, this parallel model framework has a simple architecture and high flexibility for training data. Both the measured and simulated data can be used in this model to improve model generalizability.
(2) A parallel chiller energy predictive model that can be directly applied to the chiller energy prediction of hotel buildings in eastern China was developed and verified. The average coefficient of variation of the root mean squared error (CV-RMSE) is 16%, which is acceptable for engineering purposes.
(3) The key variables are different for different building types, climate zones, and prediction targets. There has been no comprehensive and systematic research on the identification of the key variables. A general key-variable screening toolkit was developed in this study. This toolkit has stronger versatility because it can analyze the impact on HVAC energy consumption of building geometry as well as building physics, construction quality, and system performance.

2 Literature review

In this section, a complete review of data-driven energy prediction model and building energy performance sensitivity analysis will be presented.

2.1 Parallel data-driven models for building energy prediction

In recent years, more smart meters have been installed in buildings to collect energy data, facilitate the development of relevant data-driven models, and perform energy analysis and prediction. As discussed in Section 1, building energy predictive models can be classified into sequential and parallel models with regard to model input features and training data. Sequential models have been widely studied and reviewed (Ahmad et al. 2018). As illustrated in Figure 1, both traditional machine learning algorithms and complicated deep learning models were explored to improve model prediction accuracy. In addition to advanced artificial intelligence (AI) algorithms, the influence of feature engineering on model performance has also been thoroughly analyzed (Sha et al. 2021). Apart from discussion on input features and AI algorithms for energy prediction model development, there are also papers focusing on acquisition of high-quality synthetic data which is beneficial to improve model quality. Lamagna et al. (2020) summarized the previous methods of acquiring high-resolution data from raw data and also presented an expeditious mathematical method to extract the building energy demand on an hourly basis from monthly energy bills. Fan et al. (2022) proposed

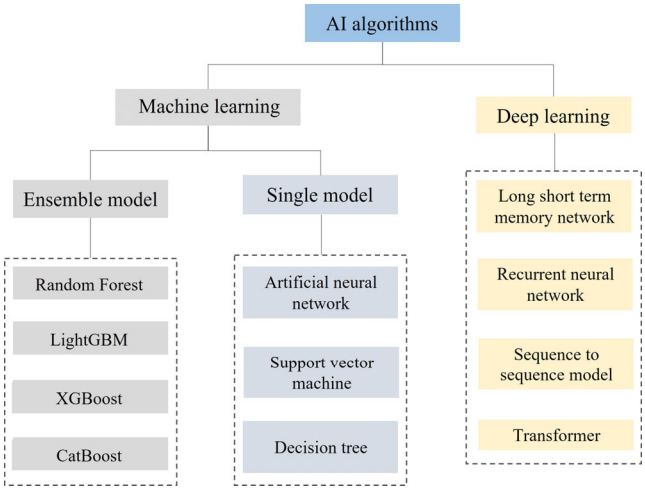


Fig. 1 AI algorithms used for building an energy predictive data-driven model

a deep generative modeling-based data augmentation method to solve the problem of data shortage in developing data-driven models.

The parallel energy predictive model, in contrast, has been less studied. The reasons are two-fold. First, the identification of input features for parallel models is difficult because the key variables accounting for building energy consumption differences should be included. The key variables were different for different types of buildings. Second, the collection of large-scale energy data is difficult. In previous studies, the selection of input features for a parallel predictive model is more experience-dependent and restricted to data that have been collected. Pan and Zhang (2020) used a new machine learning algorithm, CatBoost, to build an energy prediction model along with building type, energy star score, number of years, and number of floor areas of parking as input features. The prediction target was the energy use intensity rather than a time series of smaller granularity. Fan et al. (2020) employed transfer learning for 24 h-ahead energy prediction. The input features used included historical energy data, outdoor conditions, day type, month type, and primary building usage. The experimental results showed that the transfer learning-based model significantly improved the model performance. Li et al. (2021) also used transfer learning-based ANN model for one-hour ahead building energy prediction for buildings without sufficient historical data record. However, the transfer learning model architecture of the proposed model was too complex to popularize, and a large amount of data was required to build a pretrained model.

2.2 Sensitivity analysis for building energy performance analysis

SA is a commonly used and effective way to determine the prominent factors among input candidates. There are two main types of SA methods: global and local approaches (Saltelli et al. 2002). Global SA calculates the effects of all uncertain inputs simultaneously, whereas the local approach is focused on one or a few points. Therefore, the global SA is more reliable and widely used for building energy performance analysis. Table 1 lists several studies that have used global SA to perform building energy analyses in various applications. Demonstrably, the SA result is very sensitive to preconditions, i.e., candidate parameters, reference building, and output. The deductions from one case cannot be extended to other cases. Therefore, SA should be conducted in accordance with the specific analysis requirements. Moreover, SA of the building energy performance is largely conducted on a specific reference building. The effects on building geometry and layout, which can hardly be changed when using simulation parametric analysis tools, are seldom analyzed. A typical simulation-based SA process is illustrated in Figure 2 (Fan et al. 2020). The most time-consuming step is obtaining the corresponding output of the input variations. Tens of thousands of models must be developed and run to obtain an output dataset. For building energy sensitivity analysis, the fast parametric simulation tool jEPlus is often used to conduct batch simulations (Pang et al. 2020). However, jEPlus was designed for parametric analysis based on a specific reference building model. It cannot change the building model geometry or layout. In contrast, the key-variable screening method proposed in this paper can develop simplified building models in accordance with specific building dimensions (such as, building area, number of layers, and compactness ratio) and can be adapted to various building types and locations, which are more versatile.

3 Framework of parallel HVAC energy predictive model

The architecture of the proposed parallel HVAC energy predictive data-driven model is displayed in Figure 3. It contains three categories of input features: occupancy features, meteorological features, and key variables of building geometry, envelope, and energy systems. As mentioned above, weather conditions and occupant activity-related variables are recognized as input features for developing a sequential HVAC energy predictive data-driven model for a specific building. These two types of features are essential in allowing the model to capture the relationship between temporal variations in HVAC energy consumption during a period. Therefore, they are also employed in the parallel energy predictive model development. In this study, five extended features representing meteorological variation and occupancy activity were extracted from directly observable raw features

Table 1 Sensitivity analysis for building energy performance analysis

Reference	SA method	Objective performance	Sensitive parameters	Reference building/location
Tian et al. 2018	Sobol	• Heating energy • Cooling energy • Carbon emission	• Infiltration, equipment peak value, lighting power density • Occupancy density, equipment peak value, lighting power density • Equipment peak value, lighting power density	Office building/Tianjin, China
Li et al. 2018	Morris, FAST	Cooling energy	Building orientation, roof solar absorptance, window solar heat gain coefficient, overhang projection ratio	Zero carbon building/ Hong Kong, China
Petersen et al. 2019	Sobol	Total HVAC energy	Heating set point, window area, roof insulation, equipment power density, ventilation rate	Office building/Denmark
Delgarm et al. 2018	OFAT, variance-based method	• Annual cooling energy • Annual heating energy	• Window size, building orientation, glazing solar transmittance • Building orientation, window size, glazing visible transmittance	Thermal zone of office building/Iran
Tian et al. 2017	Regression (SRC), Sobol	• Annual cooling energy • Annual heating energy	• Window solar heat gain coefficient, chiller COP • Occupancy density, window *U*-value, heating set point	Office building/Tianjin, China
Heiselberg et al. 2009	Morris	Annual energy consumption	Lighting control, ventilation rate in winter	Office building/Denmark
Mechri et al. 2010	FAST	• Cooling energy • Heating energy	• Envelope transparent surface ratio, compactness ratio • Envelope transparent surface ratio, compactness ratio, building orientation, external shading reduction factor	Office building/Italy
Spitz et al. 2012	Sobol	Indoor air temperature	Infiltration, fiberglass thickness, heat exchanger efficiency, internal gains on the ground floor, fiberglass conductivity	Low-energy house/France

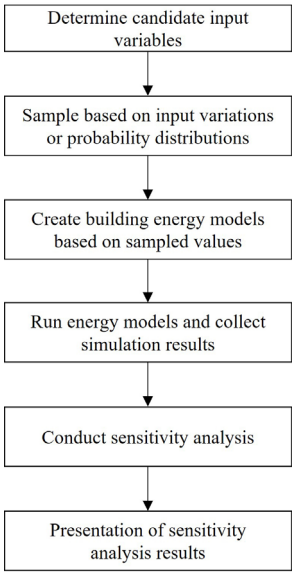


Fig. 2 Typical process of SA for building energy performance analysis

to improve model performance. A detailed explanation is provided in Section 3.1. In addition to the first two types of features, key variables summarizing the building envelope, geometrical, and energy system characteristics should also be included. These account for the difference in HVAC energy consumption between buildings. The key-variable identification process is explained in Section 4. Using all the above-mentioned variables as model input features, the parallel energy predictive model can be applied. Moreover, the available energy record data of similar buildings can be used as model training data for energy prediction in buildings without historical energy consumption data.

3.1 Extended features

(1) Time index and day type

Human activity is known to be a major factor influencing building energy consumption. However, real-time human activity is difficult to measure and quantify. As a substitute, time index features refer to categorical features such as the *i*th hour of the day, *i*th day of the month/week, and *i*th month of the year, which are commonly used as input features for building energy prediction to represent actual occupant number variation and activity characteristics. Moreover, energy consumption usually varies for different day types (for instance, during weekdays, weekends, and holidays). Therefore, day–type features, denoted as 0–1, are also used as input features in this study.

(2) Periodical factors and statistical factors

Historical energy consumption contains valuable information regarding building operating patterns. Periodical and statistical factors are dimensionless parameters extracted from historical energy data. Periodical factors are constructed under the assumption that human activity tends to display

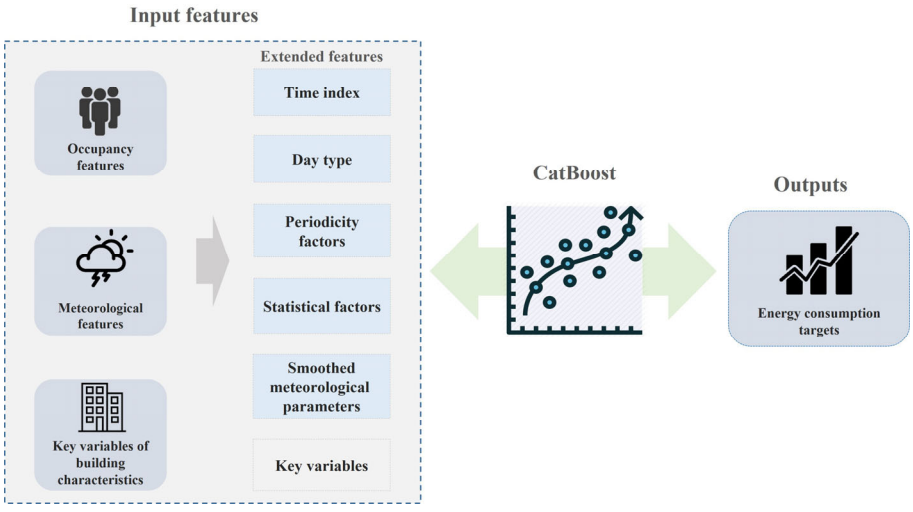


Fig. 3 Framework of parallel energy predictive model

similar patterns for the same type of day, which conforms to the regular rules. The daily periodic factor is calculated using the following equation:

$$dr_{i,j} = \frac{\overline{e}_{i,j}}{\overline{e}_j} \tag{1}$$

where $dr_{i,j}$ is the daily periodical factor in a month, with $i = 1, 2, \ldots, 7$ and $j = 1, 2, \ldots, 12$. $\overline{e}_{i,j}$ is the average energy consumption of the ith weekday in the jth month. $\overline{e}_j$ is the mean energy consumption of the jth month.

Statistical factors were extracted from historical energy data series using statistical methods. The calculation equation is given as follows:

$$t_{i,j} = T_i(Y_j) \tag{2}$$

where T_i is the statistical formula including the mean, median, maximum, minimum, skew, and standard deviation. Y_j represents the energy consumption data of the jth weekday, and $j = 1, 2, \ldots, 7$.

(3) Smoothed meteorological features

Meteorological parameters, including dry-bulb temperature and relative humidity, are driving factors behind building energy consumption and variations. Apart from those directly observable parameters, this study also adopts some smoothed meteorological parameters to handle lagging caused by building thermal mass and simultaneously avoid feature explosion. Savitzky–Golay filters (Savitzky and Golay 1964), which are commonly used to remove high-frequency oscillations in digital signals, were adopted to calculate the smoothed dry-bulb temperature and relative humidity in this study. The original and smoothed meteorological features are fed into the predictive model to extract more information.

3.2 Prediction algorithm

In this study, the machine learning algorithm CatBoost was used because of its high performance and efficiency. CatBoost is an algorithm for gradient boosting in decision trees. Compared with XGBoost and LightBoost, which are also gradient boosting algorithms, CatBoost has the following advantages:

- It is well-suited to building machine learning models with data involving categorical and heterogeneous data. Numerical features are created using the occurrence frequency of each categorical feature and some hyper-parameters.
- Composite categorical features that take advantage of the relationship between features are created to enrich the feature information.
- The ordered boosting method is adopted to cope with noise in the training data and avoid the deviation of gradient estimation, thereby improving the prediction accuracy.
- It can achieve high performance using default hyper-parameters.

3.3 Evaluation metrics

The coefficient of variation of the root mean squared error (CV-RMSE) and normalized mean bias error (NMBE) suggested by ASHRAE Guideline 14 (ASHRAE 2014) were used in this study to evaluate the model prediction performance. Compared with other popular metrics, such as root mean squared error (RMSE) and mean absolute error (MAE), CV-RMSE and NMBE is scale-independent, which is suitable for evaluating the performance of models built with different datasets. Lower CV-RMSE and NMBE indicate a higher accuracy. The CV-RMSE and NMBE

values can be calculated using the following equations:

$$\text{CV-RMSE} = \frac{\sqrt{\frac{\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2}{n}}}{\frac{\sum_{k=1}^{n} y_k}{n}} \tag{3}$$

$$\text{NMBE} = \frac{\frac{\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)}{n-1}}{\frac{\sum_{k=1}^{n} y_k}{n}} \tag{4}$$

where y_k is the test data, $\hat{y}_k$ is the predicted data, and n is the number of test data points.

4 Development of key-variable screening toolkit

The framework of this key-variable screening method is shown in Figure 4. It was developed based on EnergyPlus and sensitivity analyses. The objective of key-variable screening is to select a few variables that have a major impact on building HVAC energy consumption and variations. First, users should specify the following boundary parameters: (a) candidate variables, (b) building location (or weather file), (c) building type, and (d) prediction target. The key variables identified may vary according to the above parameters. The building HVAC energy consumption is influenced by not only the theoretical design parameters of building thermal characteristics and system configuration, but also by additional factors affecting construction quality and system operation level. Therefore, both types of variables are adopted as candidate variables from which the key variables are screened. One of the most significant challenges for sensitivity analysis is that many building models must be built and run in accordance with the input parameter samples. The *Sample Modelling* module was used to solve this problem. The simulation results were stored in a database for further sensitivity analyses.

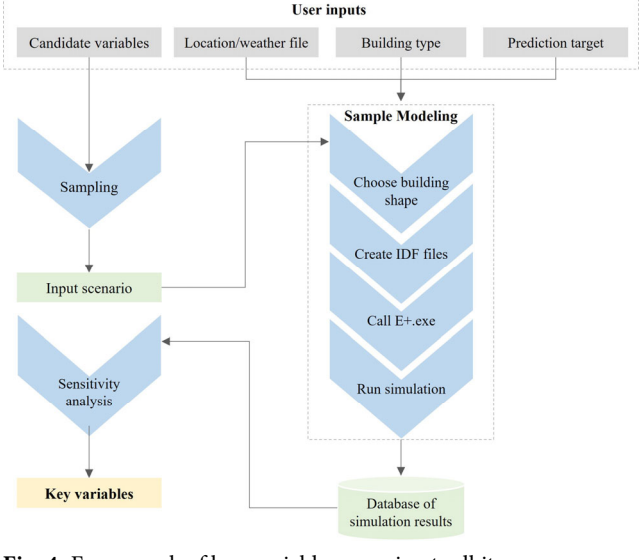


Fig. 4 Framework of key-variable screening toolkit

4.1 Sample model development

The *Sample Modelling* module was developed based on Python and Eppy (a scripting language for E+ IDF files). A detailed workflow is shown in Figure 5. It mainly comprises three parts: (1) base model files, (2) a geometric model generation module, and (3) a parameter alteration module. The base model files contain insensitive simulation information, such as basic simulation settings, typical schedules, and functional space allocation of the target building type. These parameters remained consistent throughout the analysis. The geometric model generation module builds a building model to match the sampled parameters. The parameter alteration module is designed to automatically change the model parameters (such as occupancy density, infiltration, and plant efficiency) in accordance with sample values and to create the corresponding IDF files. Following the aforementioned preparation steps, a modeling engine (EnergyPlus.exe) was used to conduct batch simulations and store the corresponding output results for further sensitivity analysis.

The building compactness ratio has a significant influence on building energy. This factor is a simplified mathematical representation of building shape. A higher compactness ratio indicates a building with less surface exposed to an outdoor environment. However, previous studies seldom focused on the influence of building shape on energy consumption because it is infeasible to manually build several models with different shapes. The geometric model generation module addresses the problem of automatic building shape alteration. A building shape database was created to represent the different building geometric characteristics. The database contains five types of basic building shapes, as shown in Figure 6. Building footprints

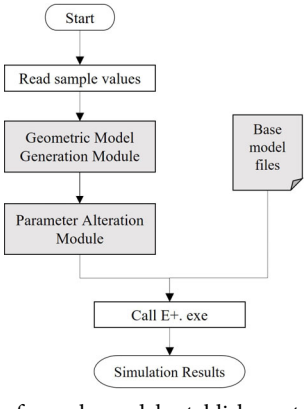


Fig. 5 Workflow of sample model establishment and simulation

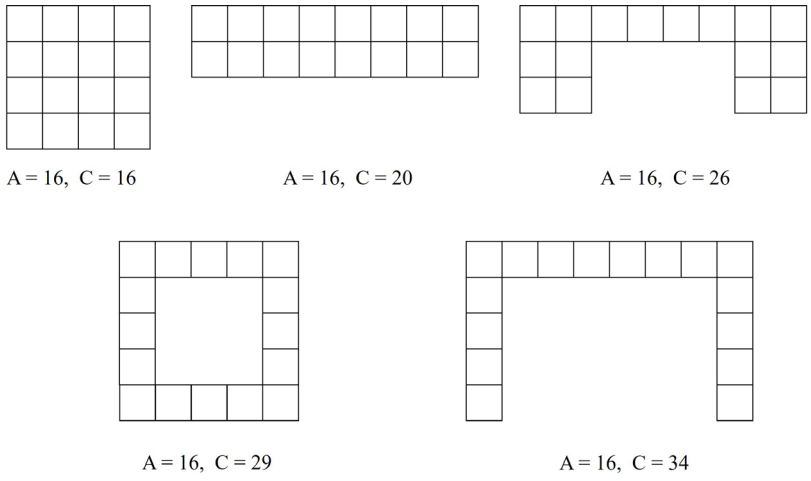


Fig. 6 Flat shape for different building compactness

a–e have equal areas but increasing perimeters; thus, they are arranged in order of decreasing compactness. We developed a factor *sigma* to represent the compactness of each shape.

$$sigma = C / \left(\frac{A}{16}\right) \tag{5}$$

The compactness ratio, which represents the compactness of a building, can be related to *sigma* using the following function:

$$\mathrm{CR} = f(sigma, A_{\mathrm{total}}, \mathrm{NL}) \tag{6}$$

where C is the perimeter of the building footprint, A is the area of the building footprint, A_{total} is the building area, and NL is the number of building layers. In this way, the geometric model generation module can find the most suitable building shape to match given parameters (i.e., building area, number of layers, and compactness ratio). The 3D model of different shapes developed by this module is shown in Figure 7.

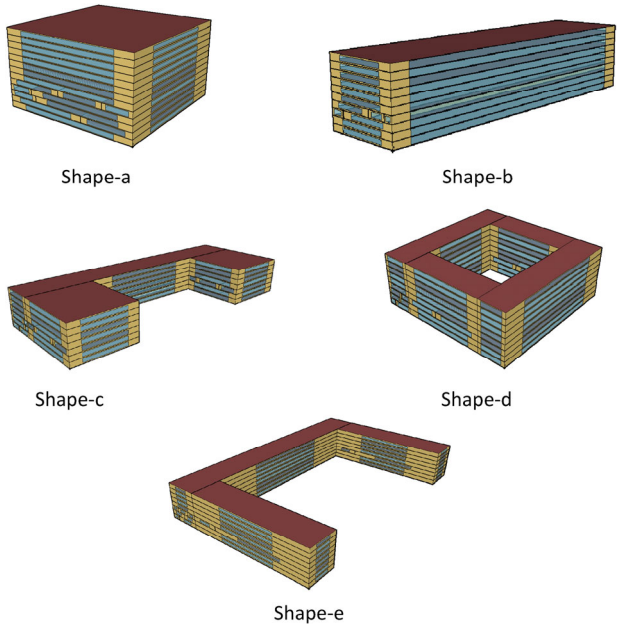


Fig. 7 3D model of different building shapes

4.2 Sensitivity analysis for key-variable screening

SA is the study of how uncertainty in a mathematical model or system output is assigned to different sources of uncertainty in its inputs (Douglas-Smith et al. 2020). The SA methods typically used for building energy analysis can be divided into global and local approaches. Global approaches focus on the impact of input parameters on whole input spaces, whereas local approaches are more interested in the influence of input parameters around a base point (Saltelli et al. 2002, 2008). Thus, global approaches are more reliable but time consuming and computationally intensive. The commonly used global SA methods for building energy performance analysis include the regression method, Morris method, Sobol, and FAST (Saltelli et al 2012). In this study, the regression method and Morris method were adopted because of their effectiveness and convenience.

The regression method is the most widely used because it is easy to interpret (Hopfe and Hensen 2011; Hygh et al. 2012). Using the regression method, the relation between the input and output was regressed using a linear equation. The coefficient of each input variable can be used to indicate the importance of the variables. Standardized regression coefficients (SRC) and partial correlation coefficients (PCC) are often used for this purpose. The calculation method for SRC and PCC can be found in Saltelli et al. (2002). However, these two indicators can only be used for linear

models (https://www.aatradingcourses.com/tag/simlab-2-2-tutorials-sensitivityanalysis/). The rank transformation of SRC and PCC (i.e., standardized rank regression coefficient (SRRC) and partial rank correlation coefficient (PRCC)) are often used for nonlinear models.

The Morris method is popular because it can handle computationally expensive models with a small number of model evaluations (Saltelli et al. 2002). The Morris method is also known as the elementary effect method. Assume that a model contains k independent variables, X_i, $i = 1, \ldots, k$ each of which is divided into p levels. Thus, the input space is divided into a p-level grid Ω. The elementary effect of the ith dimension of X is defined as

$$EE_i = \frac{[Y(X_1, X_2, \ldots, X_{i-1}, X_i + \Delta, \ldots, X_k) - Y(X_1, X_2, \ldots, X_k)]}{\Delta} \tag{7}$$

where Δ is selected from the collection of $\left\{\frac{1}{p-1}, \ldots, 1 - \frac{1}{p-1}\right\}$ to ensure $X + e_i\Delta$ is still inside Ω. e_i is a unit vector in the ith dimension. The distribution of the elementary effect is denoted as F_i—i.e., $EE_i \sim F_i$. The sensitivity indicator of the Morris method, μ, is the estimate of the mean of F_i. However, using μ as the indicator may be misleading when F_i contains both positive and negative values, i.e., the model is either nonmonotonic or has interaction effects. Thus, the estimate of the mean of the distribution of $|EE_i|$, which is denoted as μ^*, is more commonly used (Campolongo et al. 2007).

5 Case study

In this study, luxury hotel buildings in eastern China higher than four stars were analyzed. A reference hotel building model was also developed for key-variable identification. Then a parallel model for hotel chiller energy prediction which uses key variables selected previously as input features is trained and tested based on field-test data to validate the parallel energy predictive model developing methodology.

5.1 Reference building description

There are seven functional spaces in the hotel building model. The area ratio of each functional space was set to represent a typical hotel building: 0.1, 0.1, 0.075, 0.03, 0.025, 0.02, and 0.65 for the lobby, service room, dining room, kitchen, meeting room, gym, and guest rooms, respectively. The area ratio value for each functional space is set according to MOHURD (2014). The various schedules for each functional space were also set to be consistent with hotel characteristics (MOHURD 2015), as shown in Figure 8.

5.2 Selection of candidates for key-variable identification

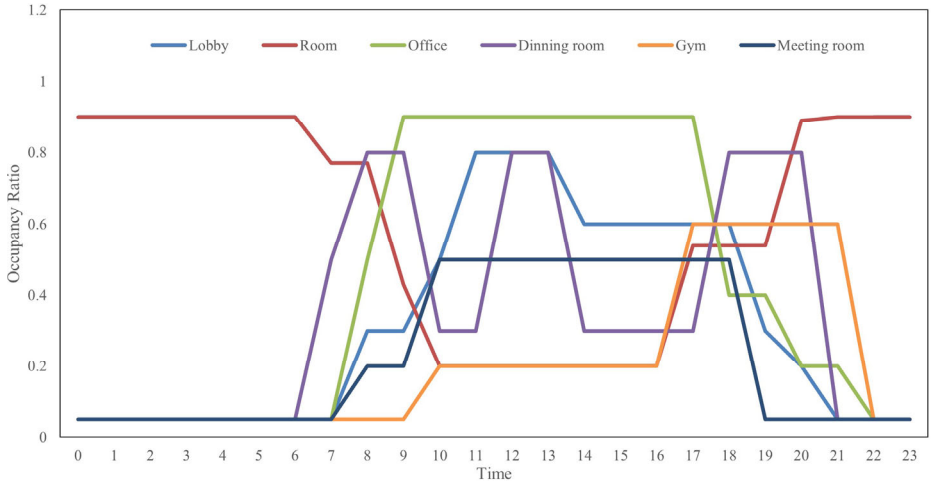
Building HVAC energy is influenced by both the building thermal load and system characteristics. Parameters related to the building thermal load, such as building window/wall ratio and wall U-value, are all numeric, whereas system-related parameters contain non-numeric parameters such as water pump type (variable speed or constant speed). If the two types of parameters are combined for sampling, the sample size becomes extremely large. Thus, in this study, sampling and sensitivity analyses were conducted twice to select the key influential variables separately from the building thermal load- and system-level parameters.

The potential influential variables, including theoretical variables and correction factors, are shown in Table 2. In total, 23 building thermal load-level variables and 11 system-level variables that can influence building HVAC energy were considered in this study. The probability distributions of the input variables depend on the research purpose. Uniform distribution is an appropriate assumption because SA is conducted for design purposes (Li et al. 2018). Variables, including building area, number of stories, and energy system type are determined according to energy audit reports, whereas the ranges of other numerical variables are mainly determined according to references (MOHURD 2015; Morrison Hershfield Limited 2016; Li et al. 2018).

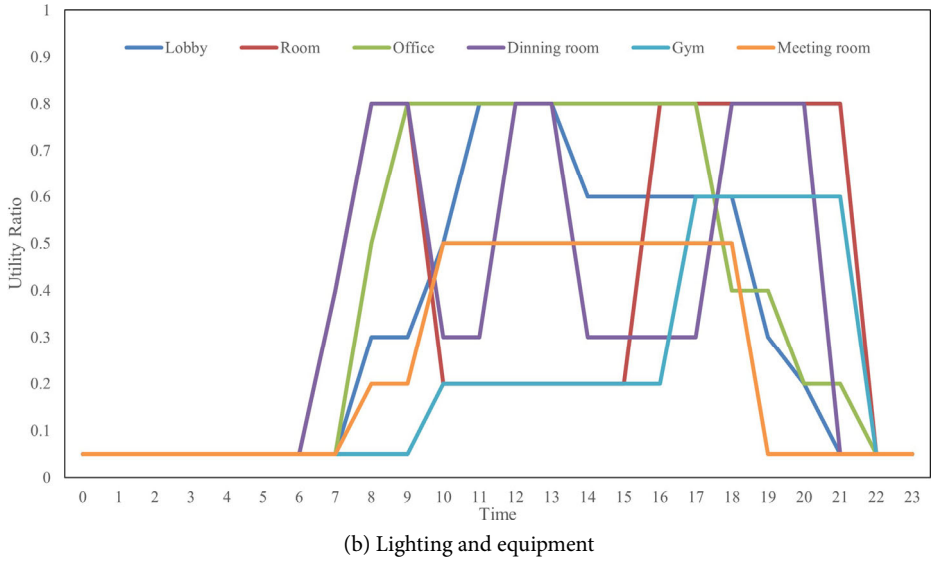
The 23 building thermal load-level variables were classified into four categories: building layout, envelope thermal characteristics, operation and occupancy, and construction quality. The building compactness ratio (defined as the ratio of the external surface area to the building area) is used to reflect the impact of different building geometric shapes on the building thermal load. The more compact a building is, the less heat is transferred through its envelope. This indicator is similar to the building shape coefficient but is easier to compute. As shown in Figure 5, different shapes a–e were adopted in this study to represent varying building compactness ratios. Construction quality (which mainly refers to the thermal bridge due to poor construction quality) has not been adequately considered in previous studies on building load simulation, and its impact on the building load is uncertain; therefore, it is considered as a potentially influential variable in this study. According to Morrison Hershfield Limited (2016), the impact of the thermal bridge on the building load can be transformed into increments of the building wall U-value (Eq. (8)):

$$U_T = \frac{\sum(\psi \cdot L)}{A_{tot}} + U_0 \tag{8}$$

where U_T is the wall U-value considering the impact of the thermal bridge (W/(m^2·K)). U_0 is the wall U-value



(a) People



(b) Lighting and equipment

Fig. 8 Schedules of hotel building model

(W/(m²·K)). A_{tot} is the total area of the opaque wall (m²). The impact of the thermal bridge was quantified using the thermal linear transmittance ψ (W/(m·K)). L is the length of the corresponding linear thermal transmittance (m).

For system-level parameters, four correction factors, which are listed as normal factors in Table 2-2, are included in addition to the theoretical design variables. A small temperature difference between the chilled supply and return temperatures is a common problem in commercial buildings in China. This phenomenon is referred to as low-delta T syndrome. The operational temperature difference is only 1–2 °C, which is much lower than the design temperature of 5 °C. The low-delta T syndrome considerably increases the water flow rate and energy use in a water supply system. The temperature difference between the chilled water supply and return water can be directly defined in EnergyPlus.

Heating and cooling coil fouling may reduce the overall heat transfer coefficient (UA), causing reduced coil capacity, resulting in unmet loads and/or increased water flowrate and reduced water-side temperature difference. In this study, we used the fouling factor for coil fouling. The overall UA factor of a fouled coil can be calculated using Eqs. (9)–(10) (EnergyPlus 2018).

$$R_{foul} = r_{air} / A_{air} + r_{water} / A_{water} \tag{9}$$

$$UA_{fouled} = 1/(1/UA_{air} + R_{foul} + 1/UA_{water}) \tag{10}$$

where R_{foul} is the overall fouling factor (K/W), r_{air} is the air-side fouling factor ((m²·K)/W), r_{water} is the water-side fouling factor ((m²·K)/W), $A_{air/water}$ is the air/water-side coil surface area (m²), and $UA_{air/water}$ is the heat transfer coefficient of the coil on the air/water side.

Cooling tower fouling is common in cooling tower

Table 2-1 Potential influential variables (building thermal load level)

Type	Parameter name	Abbreviation	Distribution	Range	Unit
Building layout	Window wall ratio (north)	NWWR	Uniform	0.2–0.8	
	Window wall ratio (south)	SWWR	Uniform	0.2–0.8	
	Window wall ratio (east)	EWWR	Uniform	0.2–0.8	
	Window wall ratio (west)	WWWR	Uniform	0.2–0.8	
	Area	AREA	Uniform	20000–200000	m^2
	Number of stories	NL	Uniform	5–40	
	Compactness ratio	CR	Uniform	0.1–0.9	
Envelope thermal characteristics	Wall *U*-value	WALLU	Uniform	0.09–5	$W/(m^2·K)$
	Wall specific heat	WSP	Uniform	800–2000	J/(kg·K)
	Roof *U*-value	RU	Uniform	0.09–4.8	$W/(m^2·K)$
	Window *U*-value	WINU	Uniform	0.2–7	$W/(m^2·K)$
	Window solar heat gain	SHGC	Uniform	0.1–0.9	
	Wall solar absorption coefficient	WSA	Uniform	0.1–0.9	
	Roof solar absorption coefficient	RSA	Uniform	0.1–0.9	$W/(m^2·K)$
Operation & occupancy	Setpoint temperature for cooling	SPC	Uniform	22–28	°C
	Setpoint temperature for heating	SPH	Uniform	18–24	°C
	Lighting power density	LPD	Uniform	3–15	W/m^2
	Occupancy density	OPD	Uniform	0.02–0.05	$people/m^2$
	Infiltration rate	INFIL	Uniform	0.05–0.5	ACH
	Interior shading rate	ST	Uniform	0.1–0.9	
Construction quality	Floor linear transmittance	FLT	Uniform	0.007–1.842	W/(m·K)
	Glazing linear transmittance	GLT	Uniform	0.03–1.058	W/(m·K)
	Corner linear transmittance	CLT	Uniform	0.036–0.684	W/(m·K)

Table 2-2 Potential influential variables (system level)

Type	Parameter name	Abbreviation	Distribution	Range	Unit
Normal factors	Air-side system type	AST	—	Constant volume system (CAV), variable volume system (VAV), fan coil system (FCU)	
	Water-side system type	WST	—	Constant primary flow system (CP), variable primary flow system (VP), constant primary variable secondary flow system (CPVS)	
	Supply air temperature	SAT	Uniform	8–18	°C
	Chilled water supply temperature	CWST	Uniform	5–10	°C
	Fan efficiency	FEffi	Uniform	0.3–0.8	
	Pump efficiency	PEffi	Uniform	0.3–0.8	
	Chiller COP	COP	Uniform	3–7	
Correction factors	Temperature difference of chiller supply and return water	TDW	Uniform	1–6	°C
	Coil fouling factor	CFF	Uniform	0–200	$(m^2·K)/W$
	Cooling tower fouling ratio	CTFR	Uniform	0.5–1	
	Air filter fouling ratio	AFFR	Uniform	1–2	

operations. It occurs when deposits become clogged, which is usually caused by poor water quality and treatment. Reportedly, the removal of scale deposits is one of the largest expenses in cooling tower maintenance. Scale deposits can reduce the overall UA, affecting both tower effectiveness and energy efficiency. In this study, we used a UA reduction factor called the cooling tower fouling ratio (CTFR) to describe the fouling severity. It is defined as the ratio between the UA value in the fouling case and that in the fault-free case.

Air filter fouling may increase the air-loop system resistance, resulting in a different system curve. This directly

affects the operation of the corresponding fans. Specifically, it may lead to variations in the fan pressure rise, fan energy consumption, and enthalpy of the fan outlet air. It may also lead to a reduction in the airflow rate and, thus, affect the performance of other system components. In this study, we used the air filter fouling ratio to describe the pressure rise variations of the fan associated with the fouling air filter. It is used as a multiplier for the fan design pressure increase. A curve that describes the relationship between the fan pressure rise and air flow rate should be defined.

5.3 Data description of parallel chiller energy predictive model development

The energy data used for the parallel energy predictive model training contained both measured and simulated data. The measured energy data came from a building submetering platform, which was established to monitor the energy consumption of commercial buildings in Shanghai, China. This energy monitoring platform also provides meteorological parameters, including dry-bulb temperature and relative humidity, which are gathered in real-time from a local weather station at Hongqiao Airport. In this study, the daily chiller energy consumption data of six luxury hotel buildings were adopted. The values of key variables come from the energy audit reports of each hotel building as are shown in Table 3. A total of 146 field-test energy data points ranging from May 1, 2019 to September 24, 2019, were selected for each building, contributing a total of 876 measured data points for model training. However, the amount of available measured data is insufficient to build a high-performance data-driven model. Therefore, this study integrates simulated energy data with field-test data to expand the training data. A total of 100 hotel building energy models were built to generate 18400 simulated chiller energy consumption data points. The values of the model input parameters are sampled from the value ranges in Table 2 using the Latin hypercube sampling method. The energy data profile is shown in Figure 9. It can be found that patterns of measured data are similar to those of simulated data. Moreover, the amounts and ranges of the simulated data were much larger than the measured data. In this regard, we believe that the integration of simulated data may improve parallel model performance by reducing under-fitting.

6 Results and analysis

In this part, separate analyses are made on key-variable identification and parallel model performance. The key

Table 3 Building and HVAC system information of hotels

No.	Area (m^2)	Stories underground	Stories above ground	Lodging ratio	SPC (°C)	INFIL (ACH)	LPD (W/m^2)	WST	AST	COP	OPD ($people/m^2$)
1	136800	36	3	0.6	24	1.05	7.6	CP	FCU	6.1	0.1
2	57000	31	1	0.56	24.5	0.83	8	CPVS	FCU	5.5	0.12
3	10549	20	2	0.7	24	0.5	7.5	CP	FCU	5.7	0.11
4	47193	23	2	0.56	24.5	0.6	8.3	VP	FCU	4.7	0.16
5	58899	20	1	0.7	23.8	0.45	7	CPVS	FCU	5.5	0.18
6	61055	16	1	0.66	24.6	0.4	8	CPVS	FCU	5.6	0.09

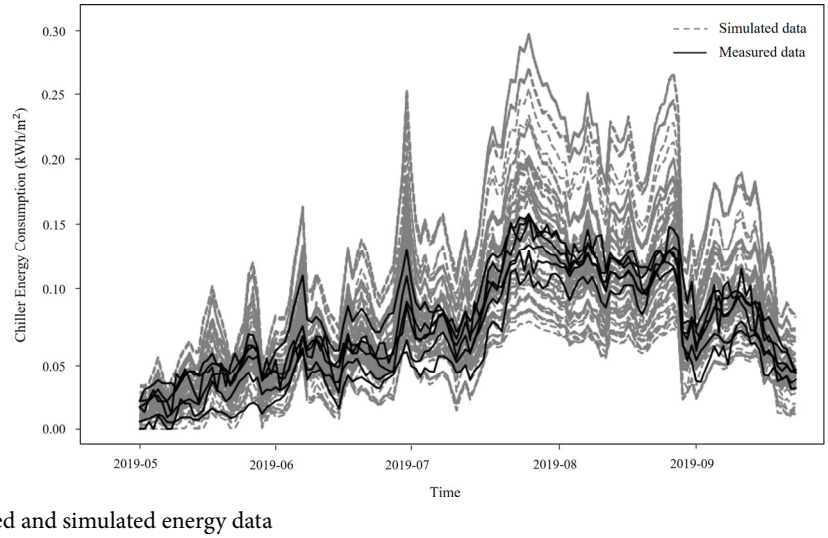
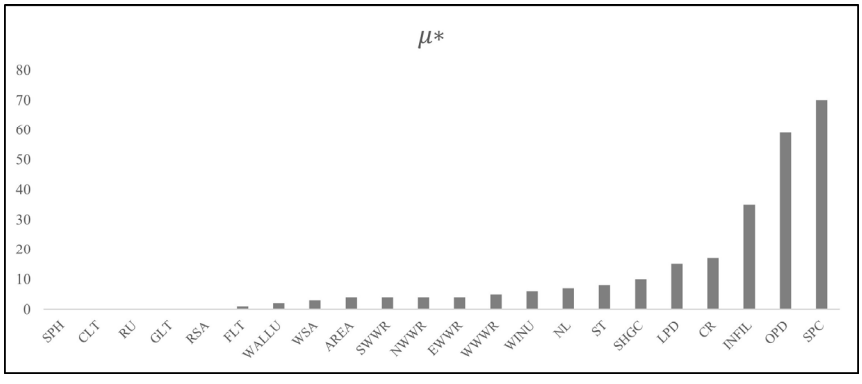


Fig. 9 Profile of measured and simulated energy data

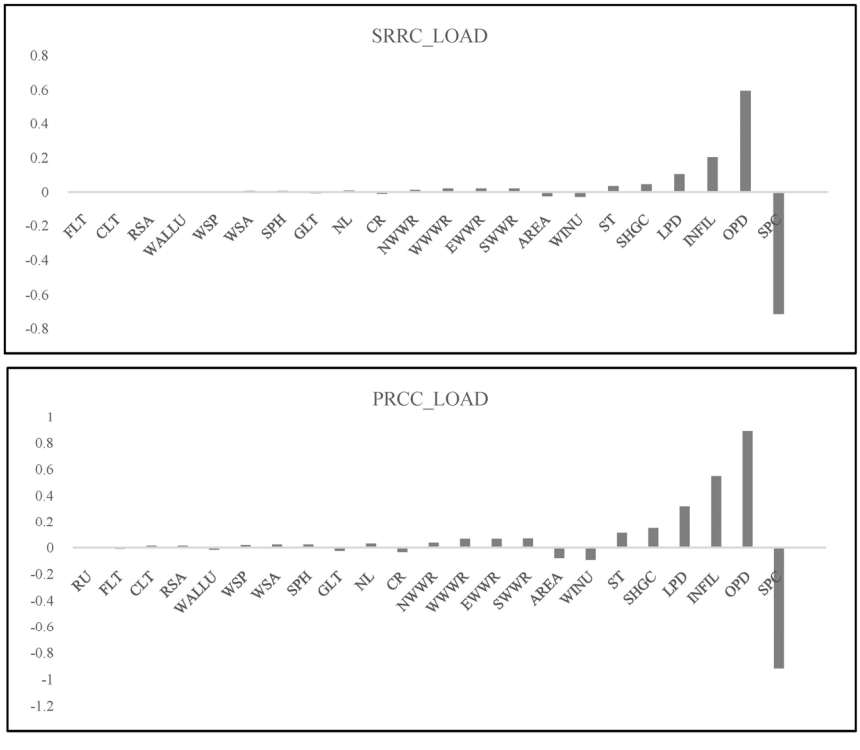
variables selected in the first procedure which represent differences regarding building geometry, envelop and system operation are used as input features for parallel energy predictive model. Besides, variables proposed in Section 3.1 are also incorporated as input features for parallel energy predictive model training.

6.1 Identification of key variables

The results of the sensitivity analysis are extremely sensitive to the preconditions. For a given system with given input parameters, the sensitivity analysis result may vary if the objective output parameter changes. Thus, we must specify the performance objective, which can be either cooling energy, heating energy, or total energy consumption, for building HVAC energy sensitivity analysis. In this section, we define the cooling energy (chiller electricity consumption) as the output parameter. In the first stage, both the Morris method and regression method were adopted to conduct the sensitivity analysis to select the key influential variables of the building thermal load level. For the Morris method, the number of effects per parameter is set to be 8, so 240 simulations obtained by sampling all 23 input parameters within their ranges were conducted for sensitivity analysis. The results are presented in Figure 10(a). As is explained in Section 4.2, μ^* is employed to represent parameter importance, the higher the μ^* value of a parameter is, the more sensitive it is. For the regression method, the Latin



(a) μ^* ranking of Morris method



(b) SRRC and PRCC ranking of regression method

Fig. 10 Results of sensitivity analysis (building thermal load level)

hypercube sampling method (Helton and Davis 2003) was used to sample the 23 input parameters, and 6000 samples were generated for sensitivity analysis. Because a building thermal system is highly nonlinear, SRRC and PRCC are calculated as sensitivity indicators, as shown in Figure 10(b). A positive value indicates the changes in this variable and performance objective go in the same direction and vice versa. The absolute value of each parameter indicator represents its importance no matter its influence is positive or negative. The input parameters in Figure 10(b) are sorted in increasing order of sensitivity. Both regression indicators (i.e., SRRC and PRCC) provide the same results. The results of the Morris method and regression also showed high consistency for the 10 most sensitive variables. Because we are interested only in high-sensitivity variables, the analysis results of both methods are considered valid. SPC, OPD, INFIL, CR, LPD, and SHGC were chosen as the high-sensitivity variables of the building thermal load level. The value of the OPD is replaced by the empirical monthly occupancy rate to represent the average occupancy density. CR and SHGC were obtained from the design documents. Other parameters were measured in the field, which can be found in the energy audit reports.

Because the Morris method requires the same change level for each variable dimension, the sampling size may be too small for problems with non-numeric input parameters. Thus, in the second stage, the Morris method is not applicable to system-level analysis because the system type has only three change levels. Only the regression method was used in this study. A total of 600 samples were generated using the Latin hypercube sampling method for numeric variables, and a total of 5400 samples were obtained together with nine combinations of two non-numeric variables (air-side system type and water-side system type). The ranking of the regression method indicators is shown in Figure 11. The chiller COP, AST, and WST are selected as the high-sensitivity variables of the system level, and their values can be found in energy audit reports. The average value of the chiller COP was adopted. The SAT was excluded because its value was not available.

6.2 Cross test of parallel chiller energy predictive model

Data-driven model performance should be evaluated in terms of both accuracy and stability. This study proposes a cross-test method, as shown in Figure 12. For a dataset composed of n real buildings, n times of model training and testing were run. For each duration, we chose the measured data of one building as test data and the others (including simulated data and measured data of buildings except the target building) as training data. Compared with the traditional train-test split method, which uses a randomly selected test dataset, the cross test method used in this study can display the model performance more comprehensively

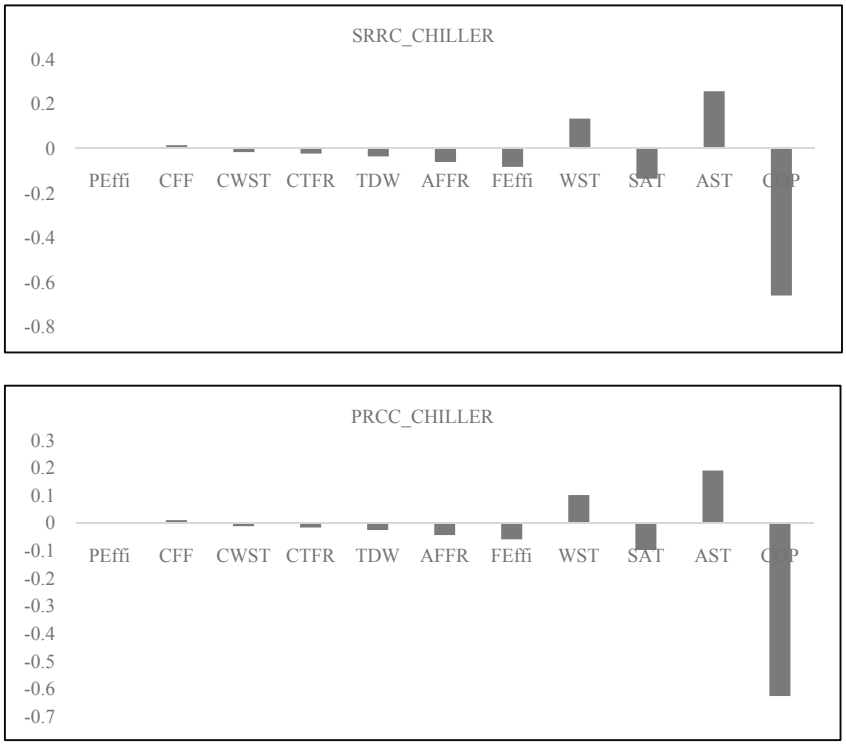


Fig. 11 SRRC and PRCC ranking of system level variables

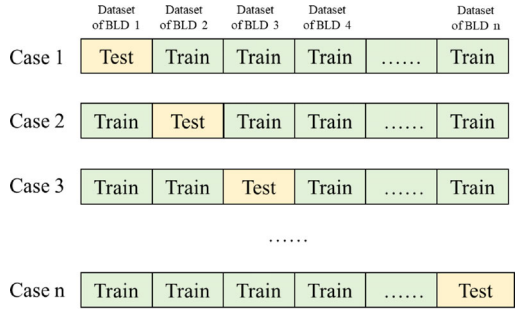
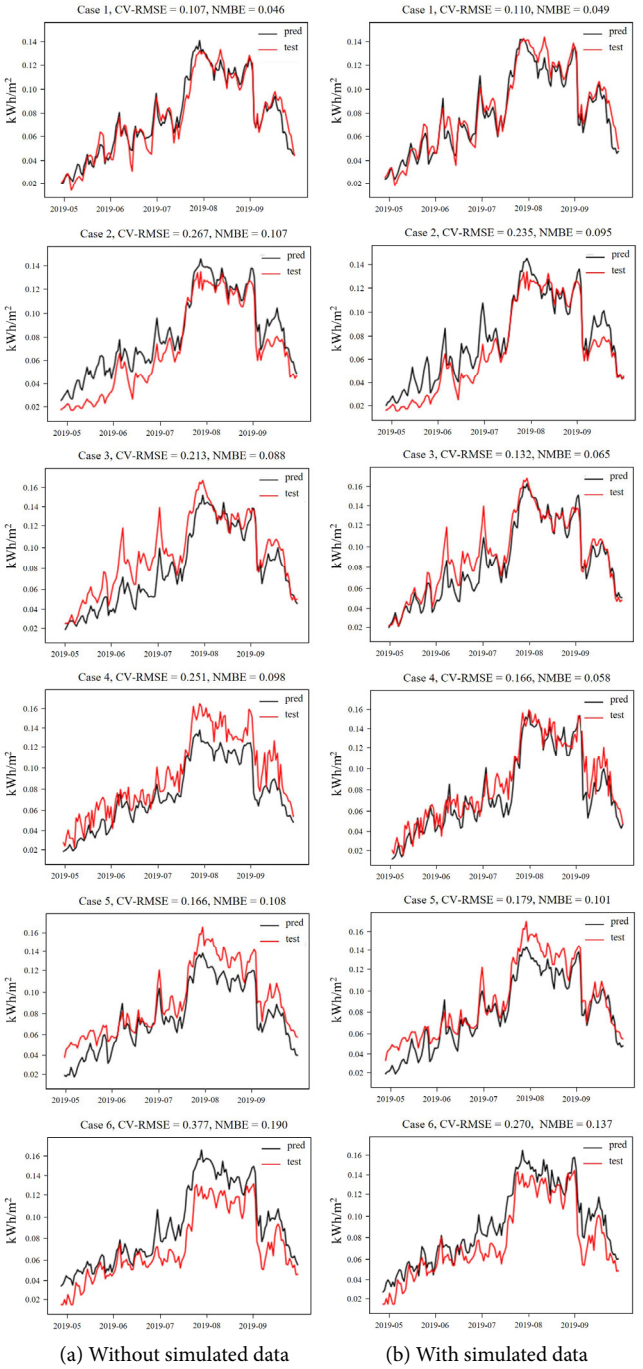


Fig. 12 Training and test data split for parallel energy predictive model cross test

from both accuracy and stability and avoid the deviation of model evaluation caused by random selection of the test dataset. The cross test results of the chiller energy predictive model are displayed in Figures 13–14. Scenario A in Figure 13 displays the performance of the models trained with merely measured energy data, whereas scenario B displays those trained with the integrated dataset. Noticeably, the parallel energy predictive model performs better when the simulated data is integrated into the training dataset. The average test CV-RMSE and NMBE of the model trained with the integrated training dataset are 0.16 and 0.083, which is significantly lower than that (i.e., 0.25 and 0.108) of the model trained with only measured data. The performance difference between the two scenarios is mainly caused by the training data scale, which has a significant impact on the data-driven model performance. Scenario B has approximately 20 times as much data as Scenario A. Moreover, the models can simulate the energy consumption of buildings under various conditions. In this regard, the ranges of simulated data are much larger, which enriches the training dataset and improves model generalizability. However, more simulated data cannot always provide better results because the distribution of the simulated data slightly differs from that of the measured data. A comparative experiment was conducted to determine the optimal simulated data size. As shown in Figure 15, with the accumulation of simulated training data, the model prediction error decreases gradually until the error reaches the minimum value when the number of simulated cases is 100. As the number of simulation cases continues to increase, the model prediction error begins to increase. In this study, we proved that if the amount of measured energy data is insufficient to build a well-behaved data-driven model, the integration of simulated data is helpful for improving model performance. However, it is quite hard to provide a concrete number of simulated data because it varies with the amount of measured data. So the optimal number of simulated data is different for different datasets, which should be carefully manipulated.



(a) Without simulated data (b) With simulated data

Fig. 13 Predicted and test chiller energy

7 Conclusions

The parallel HVAC energy predictive model proposed in this paper provides a convenient way of predicting the energy demand for buildings without historical energy data. The selection of input features is the primary step in the development of a data-driven energy model. The parallel HVAC energy predictive model incorporates variables pertaining to occupancy activity, weather conditions, and

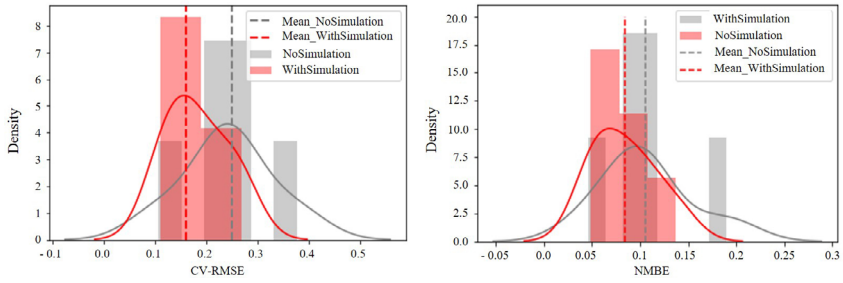


Fig. 14 Distribution of test CV-RMSE and NMBE

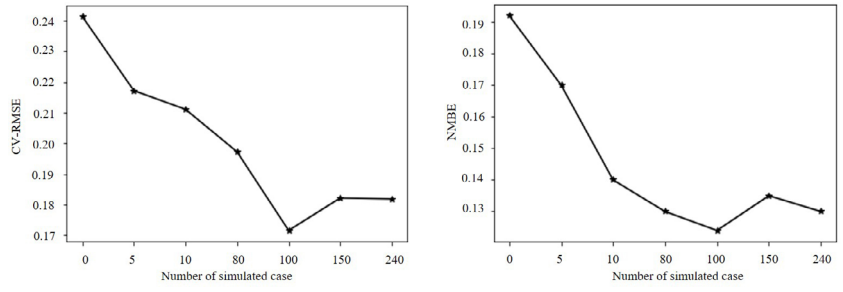


Fig. 15 Trend line of test CV-RMSE and NMBE with increasing number of simulated cases

HVAC key variables as input features. In addition, we propose a general key-variable screening toolkit to facilitate feature selection in parallel data-driven models. This toolkit is easy to use and suitable for various building types and prediction targets. In contrast to previous parametric analysis tools, we propose a factor *sigma* to link building geometric parameters (i.e., building area, number of building layers, and compactness ratio) to the corresponding building shapes, which makes this key-variable screening method flexible. To verify the feasibility of the proposed methods and models, a case study was conducted to develop a parallel chiller energy predictive model for hotel buildings in eastern China. Nine variables—SPC, OPD, INFIL, CR, LPD, SHGC, Chiller COP, AST, and WST—were identified as key variables for chiller energy consumption. Then, a parallel chiller energy predictive model was developed on the basis of key variables identified in the previous stage along with the other five extended features. Moreover, this study innovatively integrated simulated data and measured data to train a parallel chiller energy predictive model. The cross test result shows that the parallel energy predictive model obtains an average CV-RMSE of 16% when using integrated training data, which is acceptable for energy prediction without historical data. Finally, the following points should be considered when using the methods proposed in this paper.

a) The high-sensitivity variables found in this study are only sensitive to the chiller energy consumption. If users want to obtain the key variables of other objective performance, they should change the boundary parameters. Then, the simulation result of the *Sample Modelling* module can be changed to the target variables (pump energy, cooling tower energy, etc.). Moreover, this study used a hotel building in southeast China as a reference building model. Conclusions drawn under this condition cannot be directly extended to other cases. However, the methodology proposed in this study is universal. Users can easily switch to other cases by changing the boundary parameters.
b) The selection of the key variables was subjective. Users can choose any number of top-ranked variables according to specific purposes and limitations (financial, technical, etc.).
c) The possibledistribution form and variation range of the candidate variables influence the SA results. Thus, users should make decisions carefully. A uniform distribution is applicable for design purposes. However, for analyses involving retrofits of existing buildings, a normal distribution may be more suitable (Fan et al. 2020).
e) The amount of simulated data included in the integrated training dataset for parallel energy predictive model development should be carefully tuned for different applications.

Acknowledgements

This research is sponsored by China Southern Power Grid Technology Co. LTD (No. GDKJXM20200569).

References

Ahmad T, Chen H, Guo Y, et al. (2018). A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. *Energy and Buildings*, 165: 301–320.

ASHRAE (2014). ASHRAE Guideline 14—Measurement of Energy, Demand, and Water Savings. Atlanta, USA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Campolongo F, Cariboni J, Saltelli A (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, 22: 1509–1518.

Delgarm N, Sajadi B, Azarbad K, et al. (2018). Sensitivity analysis of building energy performance: A simulation-based approach using OFAT and variance-based sensitivity analysis methods. *Journal of Building Engineering*, 15: 181–193.

Douglas-Smith D, Iwanaga T, Croke BFW, et al. (2020). Certain trends in uncertainty and sensitivity analysis: An overview of software tools and techniques. *Environmental Modelling & Software*, 124: 104588.

EnergyPlus (2018). Engineering Reference. US Department of Energy.

Fan C, Sun Y, Xiao F, et al. (2020). Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Applied Energy*, 262: 114499.

Fan C, Chen M, Tang R, et al. (2022). A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions. *Building Simulation*, 15: 197–211.

Heiselberg P, Brohus H, Hesselholt A, et al. (2009). Application of sensitivity analysis in design of sustainable buildings. *Renewable Energy*, 34: 2030–2036.

Helton JC, Davis FJ (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81: 23–69.

Higdon D, Kennedy M, Cavendish JC, et al. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26: 448–466.

Hong T, Buhl F, Haves P, et al. (2008). Comparing computer run time of building simulation programs. *Building Simulation*, 1: 210–213.

Hong T, Kim CJ, Jeong J, et al. (2016). Framework for approaching the minimum CV(RMSE) using energy simulation and optimization tool. *Energy Procedia*, 88: 265–270.

Hopfe CJ, Hensen JLM (2011). Uncertainty analysis in building performance simulation for design support. *Energy and Buildings*, 43: 2798–2805.

Hygh JS, DeCarolis JF, Hill DB, et al. (2012). Multivariate regression as an energy assessment tool in early building design. *Building and Environment*, 57: 165–175.

Lamagna M, Nastasi B, Groppi D, et al. (2020). Hourly energy profile determination technique from monthly energy bills. *Building Simulation*, 13: 1235–1248.

Li H, Wang S, Cheung H (2018). Sensitivity analysis of design parameters and optimal design for zero/low energy buildings in subtropical regions. *Applied Energy*, 228: 1280–1291.

Li A, Xiao F, Fan C, et al. (2021). Development of an ANN-based building energy model for information-poor buildings using transfer learning. *Building Simulation*, 14: 89–101.

Liu J, Chen X, Cao S, et al. (2019). Overview on hybrid solar photovoltaic-electrical energy storage technologies for power supply to buildings. *Energy Conversion and Management*, 187: 103–121.

Mechri HE, Capozzoli A, Corrado V (2010). USE of the ANOVA approach for sensitive building energy design. *Applied Energy*, 87: 3073–3083.

MOHURD (2014). JGJ62-2014. Design Standard for Hotel Buildings. Ministry of Housing and Urban—Rural Development of China (MOHURD). Beijing: China Architecture & Building Press. (in Chinese)

MOHURD (2015). GB 50189-2015. Design Standard for Energy Efficiency of Public Buildings. Ministry of Housing and Urban—Rural Development of China (MOHURD). Beijing: China Architecture & Building Press. (in Chinese)

Morrison Hershfield Limited (2016). Building Envelope Thermal Bridging Guide. BC Hydro Power Smart.

Pan Y, Zhang L (2020). Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy*, 268: 114965.

Pang Z, O'Neill Z, Li Y, et al. (2020). The role of sensitivity analysis in the building energy performance analysis: A critical review. *Energy and Buildings*, 209: 109659.

Pardalos PM (2009). Approximate dynamic programming: Solving the curses of dimensionality. *Optimization Methods and Software*, 24: 155.

Petersen S, Kristensen MH, Knudsen MD (2019). Prerequisites for reliable sensitivity analysis of a high fidelity building energy model. *Energy and Buildings*, 183: 1–16.

Saltelli A, Tarantola S, Campolongo F, et al. (2002). Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. Chichester, UK: John Wiley & Sons, Ltd.

Saltelli A, Ratto M, Andres T, et al. (2008). Global Sensitivity Analysis: The Primer. Chichester, UK: John Wiley & Sons, Ltd.

Saltelli A, Ratto M, Tarantola S, et al. (2012). Update 1 of: Sensitivity analysis for chemical models. *Chemical Reviews*, 112: PR1–PR21.

Savitzky A, Golay MJE (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36: 1627–1639.

Sha H, Xu P, Yang Z, et al. (2019). Overview of computational intelligence for building energy system design. *Renewable and Sustainable Energy Reviews*, 108: 76–90.

Sha H, Xu P, Lin M, et al. (2021). Development of a multi-granularity energy forecasting toolkit for demand response baseline calculation. *Applied Energy*, 289: 116652.

Spitz C, Mora L, Wurtz E, et al. (2012). Practical application of uncertainty analysis and sensitivity analysis on an experimental house. *Energy and Buildings*, 55: 459–470.

Tian W, Liu Y, Zuo J, et al. (2017). Building energy assessment based on a sequential sensitivity analysis approach. *Procedia Engineering*, 205: 1042–1048.

Tian W, de Wilde P, Li Z, et al. (2018). Uncertainty and sensitivity analysis of energy assessment for office buildings based on Dempster-Shafer theory. *Energy Conversion and Management*, 174: 705–718.

Zhao Y, Zhang C, Zhang Y, et al. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, 1: 149–164.