IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 November 2023, accepted 2 December 2023, date of publication 7 December 2023,
date of current version 14 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340431

RESEARCH ARTICLE

Construction and Research of a Data-Driven Energy Consumption Evaluation Model for Urban Building Operation

GUOPING GAO [1] AND SHAOPAN YANG [2]
[1]School of Architectural Engineering, Huanghuai University, Zhumadian 463000, China
[2]School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China

Corresponding author: Shaopan Yang (shaopanyang2020@126.com)

This work was supported in part by the Science and Technology Project of Henan Province: Study on Migration Diffusion Law and Control Method in Impervious Cushion of Pollutants in Refuse Landfill under Project 212102310938.

ABSTRACT The bottom-up simulation method of building energy consumption, which mainly focuses on individual buildings, is difficult to apply to urban and regional level building energy consumption planning. Therefore, based on data-driven research, a simulation evaluation model for extreme gradient boosting at the mesoscale and K-means energy consumption at the macro scale is proposed and validated. The experimental results show that in the mesoscale model simulation prediction, in terms of training time comparison, the support vector method and sequential model have a time of 138.69 seconds and 90.00 seconds, respectively, which is much higher than other algorithms. The random forest algorithm has the highest accuracy comparison, at 83%; In the comparison of accuracy recall rate, the gradient improvement decision tree algorithm has the highest accuracy, at 83%. The extreme gradient boosting model proposed based on it has a mean square error value of 11.45 in residential sample prediction under refrigeration load and 9.16 under heating load, both of which are better than the comparative model. Applying it to practice, it is found that the correlation between building heating and cooling load and the 8 variables of the building is clearly demonstrated. In the macro scale model simulation prediction, the overall effect of K-means model I is the best under the comparison of contour coefficients, with the highest value maintained at around 0.1 when the number of clusters is 10. By applying it in practice, the spatial distribution of load can be clearly demonstrated, and the predicted value of total energy supply in the constructed regression equation is highly consistent with the actual value. Overall, the simulation and evaluation model for urban building energy consumption proposed in the study at two scales is practical and can effectively accelerate the development and improvement of urban energy systems. In addition, the research method considers the relevant operational energy consumption throughout the entire life cycle and the dynamic energy consumption of urban transportation systems, and provides reference value for existing urban renovation, thus possessing innovation.

INDEX TERMS Data-driven, urban architecture, energy consumption evaluation model, simulation methods, ML.

I. INTRODUCTION

Currently, building energy consumption accounts for about one-third of the total energy consumption, with heating, cooling, and hot water consumption accounting for about 50% of the total building energy consumption [1], [2], [3]. Therefore, studying the influencing factors and distribution patterns of building energy consumption is of great significance for achieving the goal of energy conservation and emission reduction. The energy consumption of buildings mainly consists of two parts. Firstly, the energy demand generated during the construction phase includes the production,

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

transportation, and other energy demands during the construction phase of building materials. Secondly, the energy consumption during the construction operation period mainly refers to the energy consumption of heating, ventilation, air conditioning, lighting, and other aspects of the building [4], [5], [6]. It is worth noting that the current energy consumption of urban heating, residential energy, and public buildings has exceeded 3/4 of the total energy consumption of social building operation [7], [8], [9]. Further urbanization will further increase this proportion. In addition, building an efficient urban building energy consumption model is the foundation for studying the total energy consumption and distribution patterns of cities. However, for individual buildings, the complexity and diversity of urban building composition make the research on building energy consumption immature. At the same time, the bottom-up simulation method, which mainly focuses on individual buildings in current building energy consumption, is difficult to apply to urban and regional level building energy consumption planning. The development of machine learning (ML) has opened up new avenues for large-scale urban building energy consumption simulation applications. Applying ML models to urban building models can save a lot of simulation time, and at the same time, it can be presented in data format to visualize relevant information [10]. Based on this, the study proposes a building energy consumption simulation and prediction model for mesoscale and large-scale using ML on the basis of data-driven mode. The aim is to solve the practical problems of current simulation methods and provide guidance for reducing urban building energy consumption and urban planning layout.

The research is divided into four sections: a review of building energy consumption research, the construction of building operation energy consumption simulation evaluation models, performance analysis of evaluation models, and article summary. The first section is a summary and discussion of the current research on urban building energy consumption models. The second section is to analyze the simulation and evaluation model of urban building energy consumption based on data-driven analysis. The third sectionis to validate the energy consumption simulation evaluation model at two scales, and the fourth section is a summary of the entire article.

II. RELATED WORKS

The energy consumption model of urban buildings is still in a relatively early stage compared to the mature physical performance simulation of individual buildings, and scholars have conducted in-depth research on it. Zhang et al. conducted a detailed analysis on the development of urban building energy consumption models by setting up three scenarios: low-speed development, stable development, and high-speed development. Based on the integration of current medium and long-term energy consumption models, they effectively improved the development challenges [11]. Chen et al. optimized the current urban building energy consumption model using differentiation methods to address issues related to carbon emissions reduction and carbon trading, thereby providing theoretical data support for the environmental utility of low-carbon city pilot policies while promoting urban carbon emissions reduction [12]. Aravindhan et al. conducted a detailed analysis of urban building energy consumption models under uncertain conditions to address the related issues of residential energy consumption under current energy demand growth. This provided theoretical support for addressing the development of urban energy consumption while considering the urban environment [13]. Keleş et al. conducted a comprehensive survey of individuals living in different cities to address the issues related to energy demand and consumption in urban buildings. Through data analysis, they constructed corresponding urban building energy consumption models, providing data support for reducing building energy consumption [14].

In addition, Hashmi et al. conducted empirical tests on the issues related to energy conservation and environmental protection in infrastructure buildings during the process of real development, and conducted empirical tests on the relevant content of environmental degradation in the top ten clustered cities from 1960 to 2014. Based on the construction of relevant urban building energy consumption models, they effectively improved urban planning and promoted the sustainable development of urban environment [15]. Peng et al. focused on the issue of energy consumption in urban residential buildings and constructed a related building energy consumption model by utilizing implicit energy consumption and carbon emissions of residential buildings, thereby reducing both building carbon dioxide emissions and actual implicit energy consumption [16]. Deng et al. focused on the issues related to urban building energy consumption and urban energy conservation and emission reduction, and constructed a building energy consumption model using clustering and random forest methods on different geographic information system datasets. This effectively simulated urban building energy consumption and enhanced the potential for urban energy conservation [17]. Jin et al. constructed a building energy consumption model based on the parameterization of the double canyon effect to address issues related to urban climate and heat emissions. This effectively reduced building energy consumption while considering the internal heat generation of buildings [18].

From the research of domestic and foreign scholars, the bottom-up simulation method of building energy consumption, which mainly focuses on individual buildings, is difficult to apply to urban and regional level building energy consumption planning. Furthermore, from a macro perspective, current urban building energy consumption models limit the scope of urban energy consumption to the actual operational energy consumption of buildings, neglecting the relevant operational energy consumption of urban infrastructure throughout its entire lifecycle and the dynamic energy consumption of urban transportation systems. From a micro perspective, the current urban building energy consumption models face challenges in collecting and organizing urban data information, with

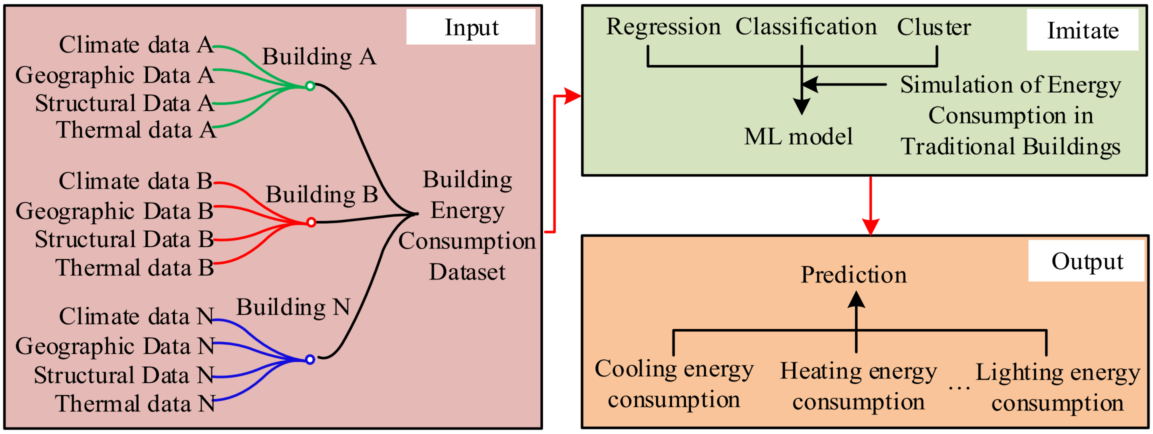


FIGURE 1. A data-driven simulation process for building energy consumption.

limited data-driven usage. Therefore, the study proposes innovative building energy consumption simulation and prediction models for mesoscale and large-scale using ML based on data-driven models.

In addition, the fact that the operating energy consumption of urban buildings accounts for an important proportion of the total energy consumption of social operations is beyond doubt. China's designated target plan for carbon emissions has also put forward urgent requirements for reducing building operating energy consumption. Therefore, the innovatively proposed model can not only promote the development of urban building energy consumption simulation field, but also promote horizontal research on urban energy consumption and socio-economic data, providing a basis for urban decision-making. And it can also make resource allocation more reasonable, and provide reference for mitigating urban heat island effects in combination with greening water bodies.

III. ANALYSIS OF A DATA-DRIVEN SIMULATION EVALUATION MODEL FOR BUILDING OPERATIONAL ENERGY CONSUMPTION

At present, traditional methods have increasingly high demands in terms of time, computational performance, and data acquisition. Therefore, it is necessary to develop a method that can adapt to the characteristics of large-scale building energy consumption simulation. This section mainly proposes corresponding building operation energy consumption simulation evaluation models from the meso and macro scales, and selects and modifies the corresponding datasets.

A. CONSTRUCTION OF ENERGY CONSUMPTION AND RESEARCH ON ML BASED ON DATA-DRIVEN MODELS

In response to the difficulty of applying bottom-up simulation methods, which mainly focus on individual buildings, to urban and regional level building energy consumption planning in current building energy consumption, a data-driven approach is proposed to evaluate urban building energy consumption, which is achieved through overall simulation. The existing building energy consumption simulation process is essentially a bottom-up energy consumption simulation system built around a limited number of individual buildings, with the basic process of "parameter input energy consumption simulation data output". The parameter input categories include: climate environmental parameters, building structural parameters, building thermal parameters, HVAC parameters, and building operating parameters; Energy consumption simulation mainly uses EnergyPlus as the main energy consumption simulation engine, with customized output time and output type; The data output terminal can output detailed energy consumption data based on energy usage, time interval, time accuracy, etc. The core advantage of a bottom-up simulation system is the ability to output highly analyzable energy consumption data including usage, time accuracy, etc. through simulation; Correspondingly, it involves complex modeling transformations, parameter settings, and simulation efficiency in the simulation process. Additionally, the current parameter setting in the process of building energy consumption simulation is mainly manual, lacking a relatively standard batch parameter setting method, and its timeliness is low, making it difficult to meet the timeliness requirements of urban scale building energy consumption simulation. The use of data-driven building energy consumption simulation models has standardized input formats, allowing for independent input of relevant energy consumption simulation parameters. The specific process is shown in Figure 1.

From Figure 1, the actual process of using data-driven building energy consumption models is the same as traditional models, following the basic principle of "input → simulation → output". However, in the comparison of the input end, the optimized model building energy consumption related database contains data attributes such as climate, geography, and structure, rather than a single building element. In the simulation comparison, the entire simulation is composed of existing building energy consumption modules and machine learning modules used to generate training data. In the output comparison, it mainly predicts building energy consumption through training related algorithm models. Overall, the optimized model only requires less data and existing learning algorithms to model existing building energy consumption, thereby improving simulation efficiency. At the same time, the training of machine learning models does not require high computational power, and their database attributes provide the possibility of migrating to the cloud. Compared to existing methods for building energy consumption simulation, data-driven building energy consumption simulation methods have inherent advantages in meeting the requirements of urban building energy consumption simulation. Essentially, they are urban building information databases with geographic coordinate attributes as their unique identity.

In addition, traditional single building energy consumption simulation methods belong to a bottom-up simulation method, which achieves energy consumption simulation based on physical facts by setting the thermal zone, air conditioning, walls, windows, and operating conditions of the building body. The use of this model requires detailed information on the structure, physics, and other aspects of the building. With the continuous increase in the scale of construction units, the demand for traditional methods in terms of time, computational performance, and data acquisition is also increasing. Therefore, it is necessary to develop a method that can adapt to the characteristics of large-scale building energy consumption simulation. Among them, the China Building Energy Consumption Special Climate Dataset (CWSD) currently provides the highest level of meteorological data for prefecture level cities, mainly in districts and counties [19]. Due to objective reasons, meteorological data for cities above higher levels cannot reflect the meteorological environment of the region as one meteorological data [20], [21], [22]. The limiting factor of building geographic information data is that at present, the geographic range of geographic information data that can provide complete contour and height information of buildings is mainly at the city level administrative region, with the highest being at the municipal level. Therefore, it is necessary to propose building energy consumption models at both meso and macro scales.

The construction of building energy models at the same scale is first defined by the scale, which is mainly constrained by two variables: climate environment data and building geographic information data. Climate data is mainly constrained by the research data source, namely CWSD. After summarizing two variables, the study defined the administrative regions of prefecture level cities as boundaries for scale division, defined building energy consumption simulation in prefecture level cities and below as mesoscale building energy consumption simulation, and defined building energy consumption prediction in administrative regions above prefecture level cities as macro scale building energy consumption simulation. Develop targeted building energy consumption simulation and prediction methods based on the needs and characteristics of urban building energy consumption simulation at different scales. Among them, in the analysis of mesoscale models, the key to the actual optimization goal of large-scale building energy consumption simulation is to standardize the input data and optimize the operational core, and the ML algorithm provides new ideas for achieving this goal. The main reason is that on the one hand, using computers to standardize data input is its natural advantage, and on the other hand, optimizing the computing core under specific conditions is the advantage of ML. The essence of applying ML to building energy consumption simulation is to train the computing core under certain conditions to meet accuracy requirements. In ML, it generally divides the relevant dataset into training, validation, and testing sets. The training set is used to construct the ML model. The validation set is used to assist in learning the model, and to evaluate the relevant performance of the model and adjust hyperparameters in the actual learning process. The test set is used to evaluate the trained model. Among them, feature engineering analysis is the most core part of the entire ML process, including data preprocessing, feature selection, and dimensionality reduction. The standardized expression in data preprocessing analysis is shown in equation (1) [23].

$$p_i = \frac{p_i - \delta}{\varphi} \tag{1}$$

In equation (1), p_i means the standardized data; δ denotes the mean of the data column; φ represents the standard deviation of the data column. The normalized expression in data pre-processing is shown in equation (2).

$$P^{\backprime} = \frac{P - \min}{\max - \min} * (mx - mi) + mi \tag{2}$$

In equation (2), $P^{\backprime}$ indicates the normalized data;min and max represent the minimum and maximum data column values; mx and mi represent the maximum and minimum values within the set interval. Therefore, the process of building energy consumption prediction using ML is shown in Figure 2.

From Figure 2, the initial settings of the energy consumption model (including climate, structural, and thermal parameter data) serve as the actual input ports, and the simulation core related to energy consumption is considered as a function of complexity, while the energy consumption data serves as the actual output result of this function. Overall, the model includes three processes: simulation, training, and prediction. In the simulation, random methods are used to select the relevant proportions of individual buildings, to conduct corresponding energy consumption simulations and obtain energy consumption data, thereby constructing the corresponding energy consumption data training set. In training, the relevant characteristics of building energy consumption are utilized to select a more suitable ML model and perform corresponding parameter adjustment and optimization evaluation, to screen out the optimal algorithm model. The remaining three body characteristic data of the building are

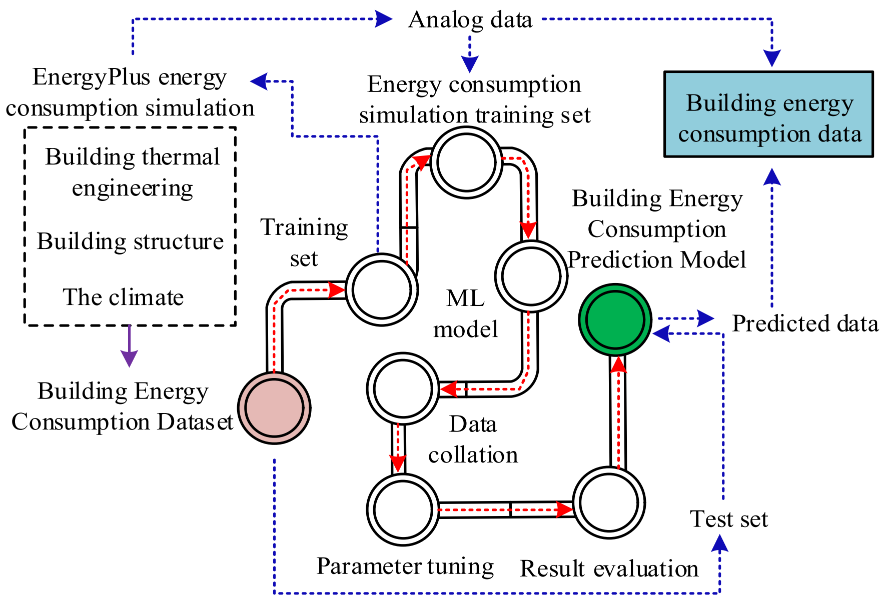


FIGURE 2. Flow chart of building energy consumption prediction based on machine learning.

input into the optimal algorithm model in the prediction to make corresponding predictions on the energy consumption data, to achieve the overall evaluation of urban building operation energy consumption.

At the macro scale, simulating the energy consumption of individual buildings in the target area one by one cannot meet the requirements of timeliness and computational performance. The study utilizes the concept of standard buildings, which calculates the energy consumption intensity of standard buildings in the target area, combined with socio-economic statistical data of the target area, including the number of permanent residents and unit living area, to achieve a preliminary estimation of regional building energy consumption. The core step of the prediction method is the calculation of building energy consumption intensity, and its influencing variables include climate factors, economic factors, and energy factors. Therefore in the analysis of regional building energy consumption models at the macro scale, the use of geographic information in building energy consumption simulation methods in mesoscale models is not applicable due to its corresponding geographic information dataset with relevant information on the actual contour and floor height of buildings. In response to this issue, a top-down prediction method for building energy consumption is proposed based on standard building energy consumption and statistical data. The process is shown in Figure 3.

From Figure 3, energy consumption simulation of individual building units in the target area at a macro scale cannot meet the requirements of time and computational performance. Therefore, the study utilizes the concept of standard buildings in conjunction with socio-economic statistical data of the target area to conduct preliminary estimates of building energy consumption. Reflected in the figure, climate factors

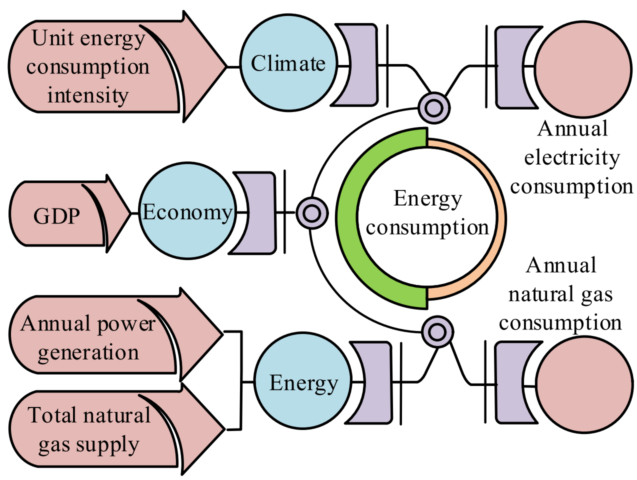


FIGURE 3. Flow chart of regional building energy consumption prediction at macro scale.

such as unit energy consumption intensity, economic factors such as regional GDP, and energy factors such as annual power generation and total urban natural gas supply are used to predict urban building energy consumption, that is, annual electricity consumption and natural gas consumption. The actual core of the prediction method is to calculate the energy consumption intensity of buildings. It mainly uses statistical concepts to select the most representative models among the main housing types in the target area for simulation, and calculates the weighted value of building energy consumption unit intensity based on the proportion of each type of building, and uses it as the standard building energy consumption intensity for the region.

In the final experiment, the ML algorithm is divided into "regression", "classification", and "clustering" algorithms based on whether the predicted results are continuous or discrete values. The ML algorithm is divided into ''supervised learning'' and ''unsupervised learning'' based on whether there is labeled data in the actual training data. The former represents classification and regression, while the latter represents clustering. Therefore, the performance evaluation of the classification model is judged from four dimensions: training time, accuracy, precision, and recall rate, and the corresponding expression is shown in equation (3) [24], [25].

$$\begin{cases} \psi = \dfrac{\varepsilon}{\varepsilon + \theta} \\ \xi = \dfrac{\varepsilon + \lambda}{\varepsilon + \lambda + \theta + \pi} \\ \vartheta = \dfrac{\varepsilon}{\varepsilon + \pi} \end{cases} \tag{3}$$

In equation (3), ψ denotes accuracy; ε indicates that this class is judged as a positive class; θ means that negative classes are judged as positive classes; ξ expresses accuracy; λ indicates that the negative class is judged as negative; π refers to that a positive class is judged as a negative class; ϑ serves as the recall rate. In the hyperparameter optimization of regression models, the Extreme Gradient Boosting (XGB) model is selected as the data analysis model, which is an improved model of the gradient boosting decision model. The biggest improvement is to apply second-order Taylor expansion to the loss function used in the actual fitting process, and introduce the regularization concept of trees to optimize the loss function and improve the fitting speed [26], [27], [28]. The expression of the relevant parameters for evaluating the model is shown in equation (4).

$$MSE = \frac{1}{\Re} \sum_{o=1}^{\Re} \left(\aleph - \bar{\aleph}\right)^2 \tag{4}$$

In equation (4), MSE means Mean Square Error (MSE); $\Re$ denotes the total number of samples; $\aleph$ expresses the true value; $\bar{\aleph}$ denotes the predicted value. Overall, at the mesoscale, the XGB model in the ML algorithm is chosen as the actual urban building energy consumption prediction model for research, while at the macro scale, the clustering algorithm is chosen as the core urban building energy consumption prediction model. At the same time, both models make further revisions to the dataset on a data-driven basis. Among them, the K-means algorithm in clustering algorithms performs best in dense clustering situations, and the actual clustering time complexity is relatively good. Therefore, this study chooses it as the preferred urban energy consumption prediction model at the macro scale.

B. SELECTION AND CORRECTION OF URBAN BUILDING ENERGY CONSUMPTION DATASETS

In data-driven building energy consumption modeling methods, the selection of relevant datasets is crucial. Therefore, the study continues to investigate the impact of different data-driven conditions on building energy consumption simulation from three aspects: climate and environmental data, building geographic information data, and building thermal data. Among them, the construction of building energy consumption datasets mainly utilizes Arc Geographic Information System Server (ArcGIS Server), Energy Plus as the core energy consumption simulation tool, and the conceptual characteristics of Geo Pandas geographic data. In the selection and correction of climate and environmental datasets, climate and environmental data is a prerequisite for simulating building energy consumption. Analyzing and studying the climate and environmental data in which buildings are located can provide people with a deeper understanding of the relationship between building energy consumption and the climate environment in which they are located. The climate and environmental data standards analyzed by the research institute are all in the format of Energy Plus Data (EPW), which covers the world and is used by Energy Plus energy consumption simulation software. The relevant climate data in a relatively standard EPW file includes three pieces of information: header, statistics, and hourly information. Hourly information is the main research object, and the relevant climate data is replaced according to needs.

The EPW format selected in the analysis of actual urban building energy consumption data is CSWD, and its main data includes outdoor design meteorological parameters and actual meteorological hourly data for simulation analysis. The basic idea for establishing a simulated meteorological year is to select the best 'average month' for each period, to establish a typical 'meteorological year'. The ''average month'' is selected based on the temperature, humidity, sunlight, wind speed and other indicators in the observation data. In selecting the weights of each parameter, it is necessary to comprehensively consider both its impact on building energy consumption and its role in practical applications, and to ensure consistency with the generation method and parameter weights used by CSWD to ensure the accuracy of data comparison. After determining the climate parameters and weights, it is necessary to standardize each parameter and obtain the weighted average of the corresponding parameters in the process of constructing a meteorological model for simulation. The corresponding expressions are shown in equations (5) and (6) [29], [30].

$$\upsilon_{j,k,l} = \left(\chi_{j,k,l} - \bar{\chi}_{j,k}\right) / \alpha_{j,k} \tag{5}$$

In equation (5), υ means the standardized data; χ denotes the average value of the parameter;χ represents the average value of the parameter over the years; α indicates the standard deviation; j is the actual number of the parameter; k expressess the month; l represents the year.

$$\kappa_k = \sum m_j \left|\upsilon_{j,k,l}\right| \tag{6}$$

In equation (6), κ means the weighted average of the standardized data; m_j denotes the parameter weight. Among the meteorological data provided by CSWD, there are five types of observed meteorological data: temperature, dew point temperature, relative humidity, solar radiation, and wind speed. Due to the high correlation between temperature and dew point temperature, four types of meteorological data, namely temperature, relative humidity, solar radiation, and wind speed,are selected as the research objects. The construction method is based on CSWD data and replaced by corresponding location data. After performing single and full permutations of four variables, a total of five typical meteorological year files were obtained for the study. In terms of actual energy consumption simulation, the control variable method is mainly used to explore the proportion of climate variables in building energy consumption. It replaces the four data of temperature, relative humidity, solar radiation intensity, and wind speed with the EPW file in CSWD to study the changes in building heating, cooling, and total energy consumption.

In the selection and correction of building geographic information data, starting from the perspective of building energy consumption simulation, the geographic data including basic building information is analyzed and screened, combined with the sources and merging methods of map mapping, to summarize the content that needs to be corrected. Specifically, geographic information data formed by merging raster data will result in buildings being segmented on the grid, which cannot reflect the actual contour of the building. In GIS data, the main facade contour of a building is not a straight line, but is composed of approximate straight lines connected by multiple lines, which increases the complexity of constructing and simulating energy models. Therefore, the study aims to correct the dataset problem of this data through two approaches, namely, to address the problem of building shapes being cropped due to raster data. Firstly, ArcMap software is used to partition the original geographic data into files based on layer attributes. Secondly, it uses the Python script compiler in ArcMap to summarize and process data at various levels to eliminate rasterization effects. In response to the impact of extreme values caused by drawing errors, the range of building types is determined based on attributes such as building plan area in the GeoPandas geographic data processing tool library, and reasonable building contours are selected as a standard.

In the selection and correction of building thermal data, energy model construction typically involves reading out building shapes through Shape files (Shp) with smaller ''concave'' folds, which poses two major challenges to Honeybee's energy model construction. Firstly, it causes excessive folding of building forms, increases computational complexity, and prolongs computational time; Secondly, due to the existence of a ''concave'' folded surface, Honeybee cannot accurately identify the Zone, resulting in simulation failure and reducing the effective sample size. Therefore, to improve the real-time performance and available sample size of the model, it is necessary to optimize the existing models. The principle of building contour optimization in Shp is to construct an abstract rectangle as a replacement model while keeping the three core parameters of the original building contour length, number of floors, and contour area unchanged (these three parameters can determine that the shape coefficient is unique). The corresponding conversion rules are expressed in equations (7) and (8) [31].

$$L\varpi_{i'} = \frac{B_{i'}}{4} + \left(\frac{B_{i'}}{16} - \varpi_{i'}\right) \wedge 0.5 \tag{7}$$

In equation (7), $L\varpi_{i'}$ indicates the actual length of the long side in the long direction; $B_{i'}$ denotes the length of the actual contour of the building; $\varpi_{i'}$ refers to the area of the actual contour of the building.

$$\varpi\varpi_{i'} = \frac{B_{i'}}{2} - L\varpi_{i'} \tag{8}$$

In equation (8), $\varpi\varpi_{i'}$ means the length of the actual short side of the rectangle. In the actual mesoscale prediction analysis of building energy consumption models, the content of the dataset mainly includes two types of data: climate characteristics and socio-economic data. Climate data includes data from 270 surface meteorological stations across the country, and socio-economic data mainly includes data from provincial statistics and statistical yearbooks. Among them, the characteristics of climate data have an important impact on building energy consumption. The study aims to explore the impact of meteorological factors on building energy consumption density with meteorological factorsand building energy consumption density as the entry points. Firstly, by analyzing the overall sensitivity of meteorological element data and energy consumption data, meteorological elements with weak correlation are eliminated. Secondly, by clustering the remaining meteorological parameters as a whole, the optimal clustering algorithm and number of classifications are selected. Finally, through local sensitivity analysis within the region, the trend of correlation between climate factors and building energy consumption in different regions is studied.

The actual clustering analysis process of the study first involves selecting a set of climate datasets and importing them into a clustering analysis comparison script. The contour coefficients and Calinski-Harabaz index coefficients of each algorithm in clusters ranging from 2 to 10 are calculated, and the clustering performance of each algorithm in each cluster is compared. The corresponding expressions are shown in equations (9) to (11) [32], [33], [34].

$$Z = \frac{c - d}{\max\{d, c\}} \tag{9}$$

In equation (9), Z means the contour coefficient;c denotes the average distance from the calculated object to all objects in the given cluster (reflecting the degree of separation); d expresses the average distance (reflecting cohesion) between the calculated object and all other objects in the cluster it belongs to.

$$M = \frac{\sum_{j'=1}^{D} \sqrt{\sum_{k'=1}^{N} \left(d_{k'} - c_{j'k'}\right)^2}}{D} \tag{10}$$

In equation (10), M refers to the contour coefficient of the cluster population; D denotes the total number of objects in the cluster to which it belongs; N indicates the total number of objects in a given cluster.

$$F(a) = \frac{tr(E_a)}{tr(W_a)} \cdot \frac{e-a}{a-1} \tag{11}$$

In equation (11), $F(a)$ denotes the numerical results of the Calinski-Harabaz index; E_a expresses the inter class covariance matrix; W_a indicates the covariance matrix of intra class data; e denotes the capacity of the data set; a expresses the number of categories in the cluster. Secondly, the elbow method and contour coefficient method are used to select the actual number of clusters, with the core being the sum of squares of errors, as expressed in equation (12).

$$G = \sum_{l'=1}^{a} \sum_{\Im \in C_{j'}} \left| \Im - \mu_{j'} \right|^2 \tag{12}$$

In equation (12), G refers to the sum of squares of errors; $\Im$ serves as the sample point;$C_{j'}$ expresses the set of sample points; $\mu_{j'}$ serves as the center point. Finally, after obtaining the optimal a value, necessary correlation analysis is conducted on the relevant climate feature data within the cluster to evaluate the trend of correlation changes between the relevant climate feature variables in different groups and energy consumption intensity. The expression of the correlation is shown in equations (13) to (15).

$$Cov(Q,I) = J\left[(Q - J[Q])(Q - J[Q])\right] \tag{13}$$

In equation (13), $Cov(Q,I)$ represents the correlation coefficient between variable Q and variable I. J represents the mean.

$$Var[Q] = \frac{\sum (Q - Q_{avg})^2}{T} \tag{14}$$

In equation (14), $Var[Q]$ denotes the total sample value of variable Q; Q_{avg} denotes the sample mean of the variable Q; T expresses the total sample size.

$$r(Q,I) = \frac{Cov(Q,I)}{\sqrt{Var[Q]\,Var[I]}} \tag{15}$$

In equation (15), r expresses the Pearson correlation coefficient.

IV. APPLICATION ANALYSIS OF SIMULATION EVALUATION MODEL FOR ENERGY CONSUMPTION IN URBAN BUILDING OPERATION

The effectiveness of the urban building operation energy consumption simulation evaluation model in practical application still needs to be verified. Therefore, this section mainly focuses on the simulation evaluation of building energy consumption models at the mesoscale and macro scale.

A. SIMULATION AND EVALUATION OF BUILDING ENERGY CONSUMPTION AT MESOSCALE

To verify the effectiveness of the proposed data-driven simulation and prediction model for urban building energy consumption, the study evaluated it through experiments and divided it into two levels, namely meso scale and macro scale, for analysis. At the mesoscale, the dataset is first initialized. 2000 sets of data were randomly selected as the test set and processed everywhere in the Shp. Then, a classification algorithm was selected to complete the classification of building types. Finally, a regression model was applied to predict building energy consumption (The reason for randomly selecting 2000 sets of data is to filter extreme values through area and body shape coefficient. Data with an area greater than 5 square meters and a body shape coefficient less than 1.5 were selected, resulting in a total of 20748 pieces of data. The training set was divided into a 9:1 format and a total of 2000 sets of data were tested.). The GeoPandas library, as a third-party library specialized in processing and analyzing geographic information data in Python, has both Pandas data processing and spatial data capabilities. Compared to ArcMap, GeoPandas has the advantages of faster data analysis, editability, and scalability. Filter extreme value data by filtering field data. In addition, building geographic information data mainly includes contour and layer information. The method of converting the dataset into image samples is to first print the building contour sample data through geopandas.plot, and convert the layer information into image grayscale. The image size is unified as (2cm * 2cm, dpi=50), and the training sample data is 2142. Secondly, the image is converted into matrix data through Opencv and normalized. Finally, through the train_ Test_ Split to split the sample data, test_ Size is 0.25, with 1605 training sets, 536 testing sets, random_ A state of 1 ensures that during repeated training, the classification state remains the same as the previous one.Among them, classification algorithms were analyzed from the classification of building contour data and image classification of building graphics. The former selection tested the predictive performance of eight classification models in ML, including naive Bayes, K-nearest neighbors, logistic regression classification, random forests, decision trees, gradient enhanced decision trees, support vectors, and sequential models (represented by A-H), on building geographic information data. The latter chose the LeNet5 network (The reason for choosing 8 ML algorithms is firstly that they include regression, clustering, and classification algorithms, which basically include all types of ML algorithms and are more commonly used and mainstream. Secondly, for the research and analysis of building energy consumption, machine learning algorithms can be more suitable for building geological information systems and can cope with different working conditions. In addition, classification algorithms constructed using different mathematical models will exhibit certain orientations in training samples. Therefore, the main content of the study is the analysis of building related datasets. After screening, these 8 most suitable machine learning algorithms were selected, which have higher suitability compared to HMM models, LDA models, and neural networks. Their application in building geographic information data has better predictive performance). In addition, in the experimental hardware environment, select the CPU

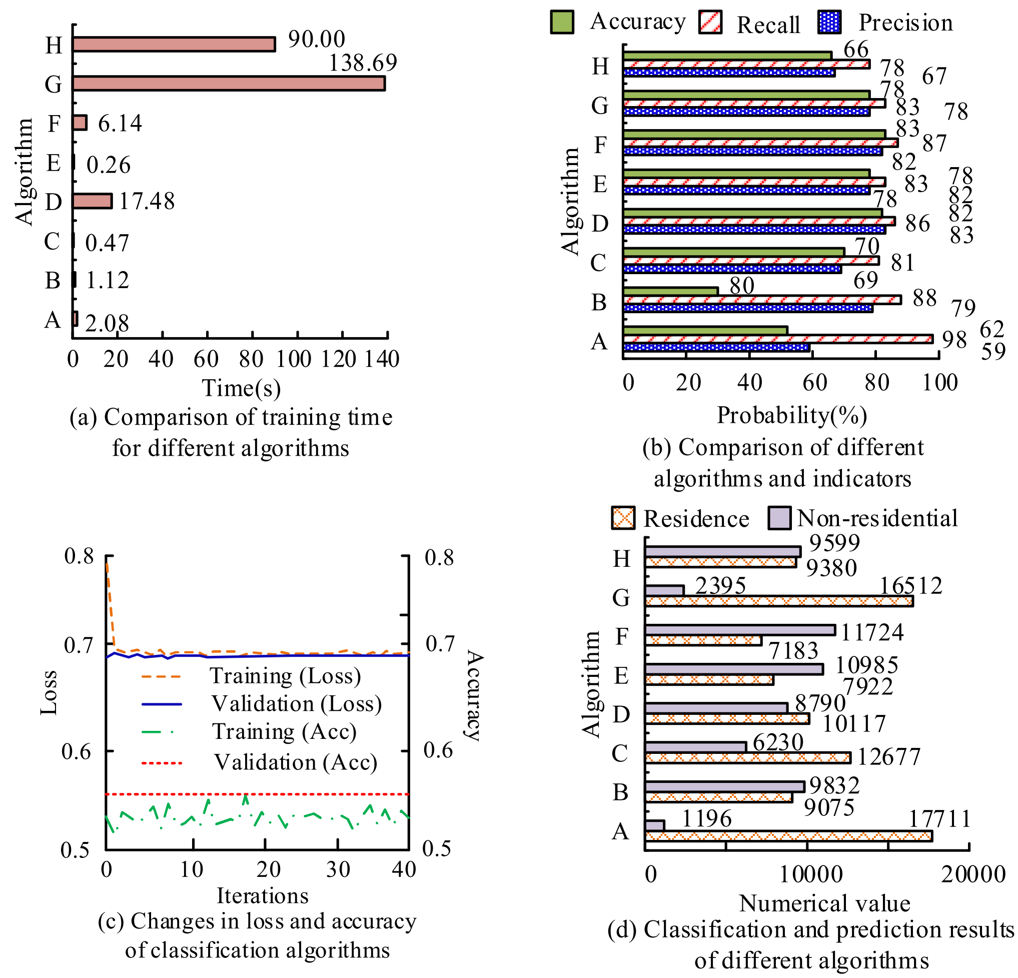


FIGURE 4. **Comparison results of evaluation data for different machine algorithms.**

model of Intel (R) Core (TM) i7 and the Tesla P100GPU. The operating system in the software environment is Windows 10 with the model Ubuntu 16.04, the programming language is Python 3.8 and Visual Studio 2019, and the algorithm programming environment is Python 1.4.0 and Open CV3.2. The results are shown in Figure 4.

From Figure 4 (a), in terms of training time comparison, the time of algorithms G and H were 138.69s and 90.00s, respectively, which were much higher than other algorithms. From Figure 4 (b), Algorithm D had the highest accuracy comparison, at 83%. Algorithm F had the highest accuracy recall rate in comparison, at 83%. From Figure 4 (c) and (d), the training loss value of the LeNet5 network tended to be flat at 69%, exhibiting cross prediction performance and overall significant errors. Overall comparison showed that algorithms B, D, E, and F all exhibited high prediction performance, while in actual prediction, algorithm D had better prediction performance, was more uniform and accurate, and had higher effective classification for residential clusters. Therefore, using algorithm D, also known as the random forest algorithm, as a classification model, could significantly improve the accuracy of subsequent energy consumption simulations. On this basis, the study began to evaluate the feasibility of building energy consumption prediction models at a moderate scale. 1000 pieces of data were selected as actual samples for the energy consumption model, and the relevant data was standardized. Subsequently, training and testing sets were selected to construct a regression algorithm model. Finally, the model was evaluated to select the optimal algorithm for predicting and analyzing the building energy consumption dataset.

The study analyzed and compared the test size of 0.25, with 543 training sets and 181 test sets. Linear regression (L) and fully connected layer model (F) were introduced to predict building energy consumption compared with the XGB model (X) proposed in the study. Among them, after hyper-parameter tuning, the XGB model had a search frequency of 1000 times, a runtime of 279s, an MSE value of 11.34, an estimation of 245, and a maximum depth of 2. Therefore, the MSE results of the three algorithms for cooling load prediction in residential and non residential samples are shown in Figure 5.

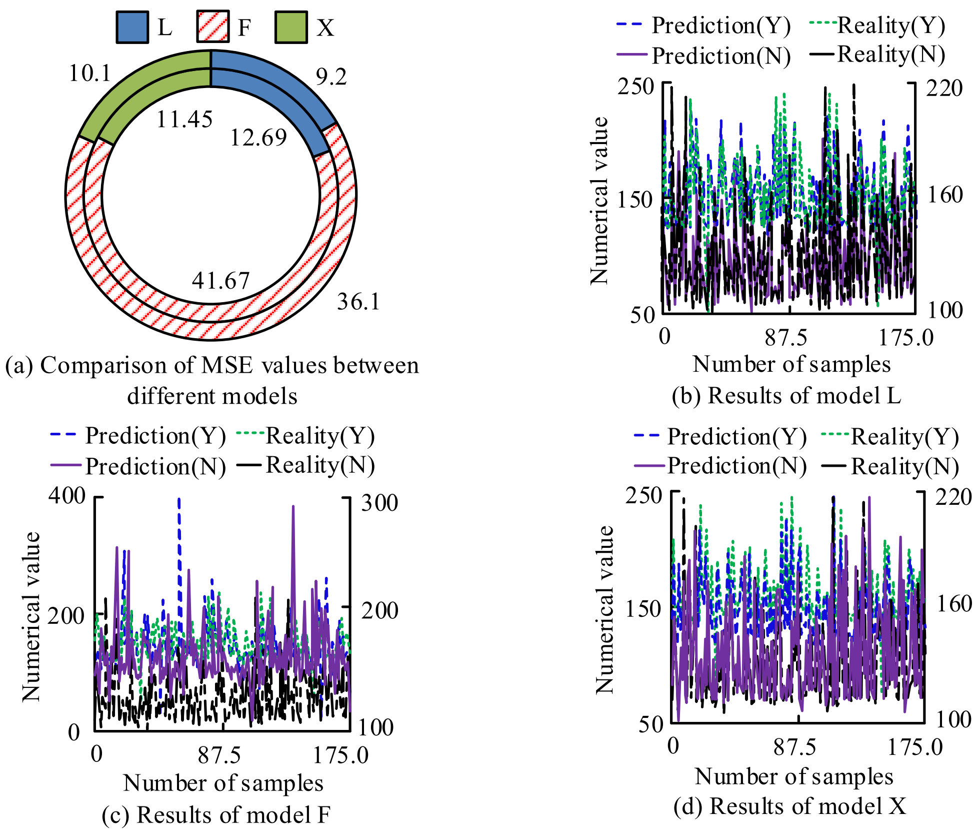


FIGURE 5. **Mean square error results of cooling load prediction using three algorithms in residential and non residential samples.**

In Figure 5, Y denotes residential and N means non residential. From Figure 5 (a), the MSE value of model X in residential sample prediction was 11.45, while the MSE value in non residential sample prediction was 10.10. The former was significantly lower than the comparison model, while the latter was slightly higher than model L, but significantly lower than model F. From Figure 5 (b), (c) and (d), model X and model L could predict the building load with high accuracy. The predicted results of Model X differed greatly from the actual situation and could not accurately reflect the actual building load. Overall, the XGB model was the most effective in predicting building loads. The results of heating load prediction are shown in Figure 6.

From Figure 6 (a), the MSE value of model c in residential sample prediction was 9.16, which was superior to the comparative model. The M color value in non residential sample prediction was 13.41, slightly higher than model L but better than model F.From Figure 6 (a), (b) and (c), among the predicted and actual curves, Model F had the best prediction effect. Based on Figures 5 and 6, the XGB model exhibited high performance in predicting building cooling and heating loads. Therefore, the XGB model was applied to integrate the heating and cooling load prediction data of different building types of datasets to form a building energy consumption dataset, and to analyze the correlation between the heating and cooling load of buildings and the eight variables of the building. The eight variables were body shape coefficient, building main direction angle, number of floors, contour area, contour length, surface area, volume, and building area (represented by 1-8). The correlation between building cooling load and 8 variables is shown in Figure 7.

From the four figures in Figure 7, it can be seen thatvariable 1 showed a significant positive correlation with the unit cooling load of the building. There was no obvious relationship between variable 2 and the unit cooling load. The remaining variables expressed a weak negative correlation with the unit cooling load of the building. Overall, the larger the building shape coefficient, the higher the unit cooling load demand of the building. The correlation between building heating load and 8 variables is shown in Figure 8.

From the four figures in Figure 8, it can be seen thatthe trend of building heating load and building cooling load was basically consistent, indicating that the building shape coefficient had a direct impact on the unit load of the building. The larger the shape coefficient, the higher the unit heating and cooling load, and the lower the energy-saving efficiency of the building. To further confirm the results and evaluate the effectiveness of using the data-driven XGB model on building energy consumption at the mesoscale, the study took the actual building energy consumption in a certain region as

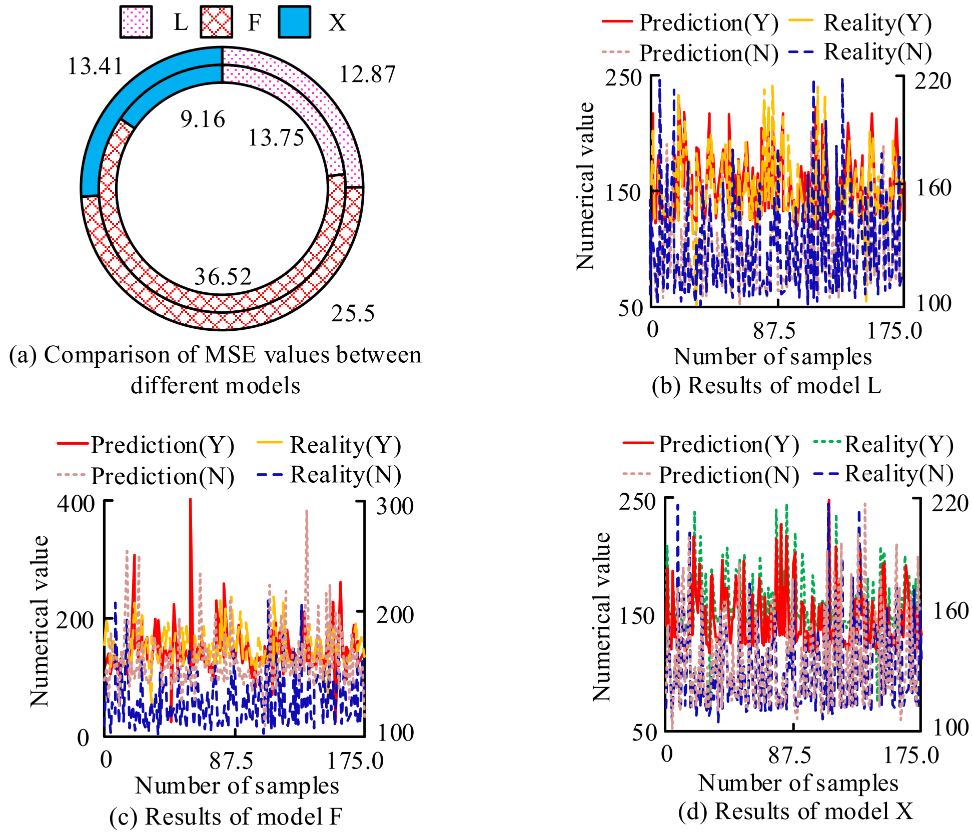


FIGURE 6. Mean square error results of heating load prediction using three algorithms in residential and non residential samples.

an example. The main variables in the dataset were set to Y for residential buildings and N for non residential buildings according to the distribution of building categories. The box distribution of energy consumption data is shown in Figure 9.

From Figure 9 (a), the coefficient of the residential system was between 0.28 and 0.72, while for non residential areas, it was concentrated between 0.19 and 0.40. From Figure 9 (b) and (c), the annual cooling load of residential units was concentrated between 127∼161kWh/m2, while non residential units were concentrated between 104∼120kWh/m2. The heating load of residential units was concentrated between 74∼135kWh/m2, while non residential units were concentrated between 46∼61kWh/m2. Overall, the distribution range of building shape coefficient in residential buildings was significantly higher than that in non residential buildings, consistent with the interval distribution of unit cooling and heating load. This result confirmed the impact of building shape coefficient on building energy consumption and indirectly verified the effectiveness of the XGB building energy consumption simulation and prediction model driven by data. To further validate the effectiveness of the mesoscale model constructed in the study for predicting building energy consumption, a prediction algorithm using time series regression, a model using reinforcement learning for building energy consumption prediction, and an improved whale algorithm were introduced for comparison (represented by a∼c). The comparison indicators were F1 value, accuracy, accuracy, and recall, and the results are shown in Table 1.

From Table 1, the F1 value, accuracy, accuracy, and recall values of the research algorithm are 89.5%, 91.2%, 90.1%, and 90.0%, respectively, which are higher than the comparison algorithm. Overall, the XGB building energy consumption simulation and prediction model driven by data is effective.

B. MACROSCALE SIMULATION AND EVALUATION OF BUILDING ENERGY CONSUMPTION

At a macro scale, the study utilized the regional energy consumption spatial distribution method simulated by building energy consumption engineering to analyze the annual heating and cooling of buildings in 270 regions of China(research on building energy consumption simulation models applicable to a wider range of building types in urban areas, mainly included in building types in 270 regions). Firstly, the performance of the K-means algorithm was verified, and small batch K-means, spectral clustering, birch, selection, and K-means algoriths were introduced for comparison (represented by X, P, I, Q, and K respectively). The results are shown in Figure 10.

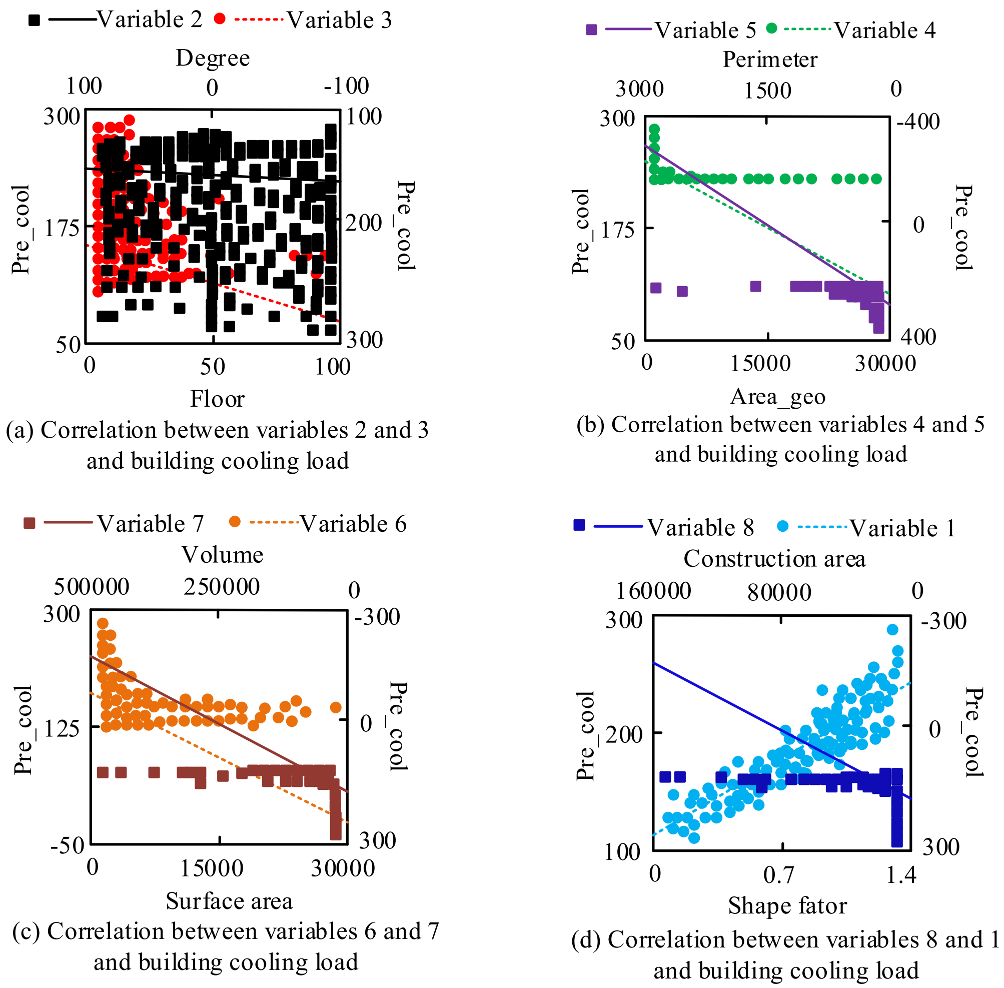


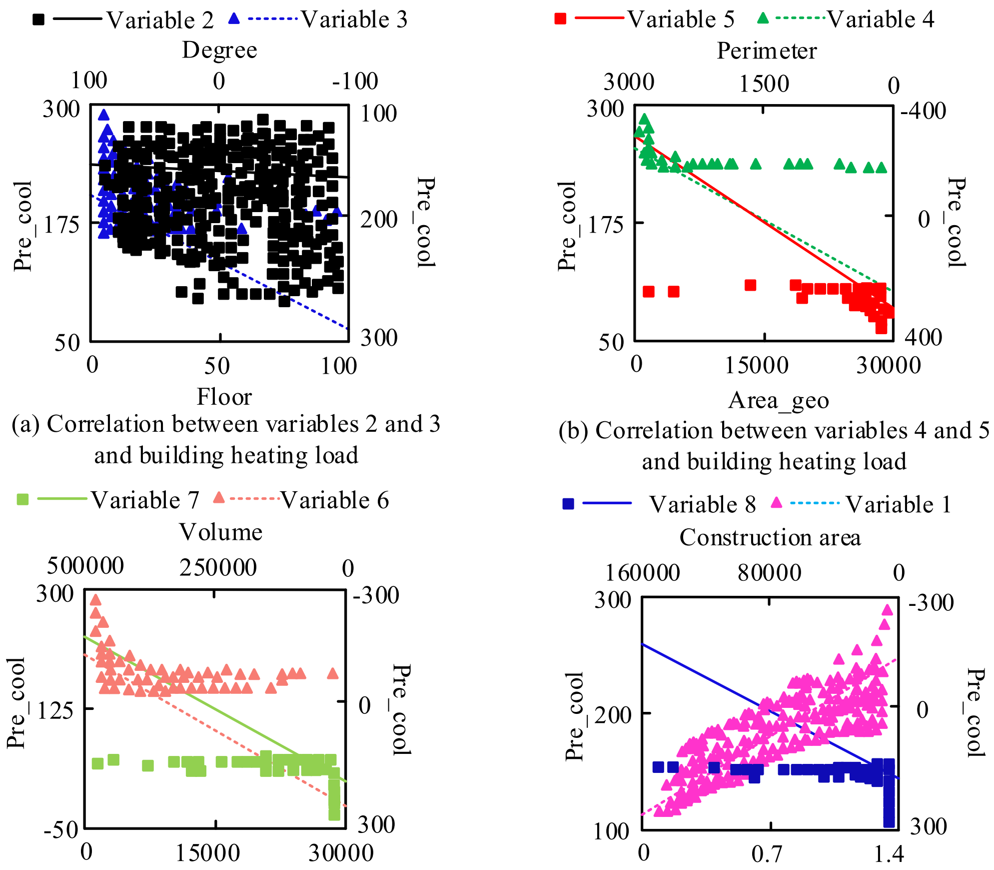
FIGURE 7. The correlation between building cooling load and 8 variables.

TABLE 1. Comparison results of different algorithms and indicators.

-	F1 value	Accuracy	Precision	Recall
a	82.2%	85.6%	81.5%	79.1%
b	81.5%	83.4%	80.5%	80.5%
c	83.4%	85.7%	84.6%	87.1%
Research algorithms	89.5%	91.2%	90.1%	90.0%

From Figure 10 (a) and (b), in the comparison of contour coefficients, algorithms K and I had the best overall effect, with algorithm K having the highest value when the number of clusters was 10, maintaining around 0.1. From Figure 10 (c), in the comparison of Calinski-harabaz scores, algorithm K also showed the best clustering performance. Overall, the K-means algorithm had the best actual clustering performance and was most suitable for clustering analysis of climate variable datasets. Therefore, it would be applied in subsequent analysis. In the elbow graph of the K-means algorithm in Figure 10 (c), the slope area was flat when the K-value was greater than 4, so the clustering effect was most suitable when the K-value was selected as 4. The K-means algorithm was applied to select a grouped dataset with a K-value of 4. The spatial distribution of clustering results obtained by taking the whole country as an example is shown in Figure 11.

From Figure 11, the study divided the country into four regions, with Zone 1 mainly distributed in higher latitude provinces, with Enable Heat Duty (EHD) as the main source, accounting for 94% of the Enable Duty (ED). EHD dominated in Zone 2, accounting for 82% of ED. The load in Zone 3 was relatively average, with Enable Cool Duty (ECD) and EHD accounting for 52% and 48% of ED. Zone 4 was mainly composed of ECD, accounting for 92% of ED. Based on the results in Figure 11, the proportion of Total Cool Heat Duty (TCHD) and Total Energy Consumption (TEC) of building heating and cooling



(a) Correlation between variables 2 and 3 and building heating load

(b) Correlation between variables 4 and 5 and building heating load

(c) Correlation between variables 6 and 7 and building heating load

(d) Correlation between variables 8 and 1 and building heating load

FIGURE 8. **The correlation between building heating load and 8 variables.**

TABLE 2. **The proportion results of different indicators in different regions.**

-	TCHD	TEC	UP	GDP
Zone 1	11.03%	8.18%	6.93%	5.84%
Zone 2	23.19%	24.63%	21.05%	19.41%
Zone 3	51.65%	49.30%	53.48%	55.68%
Zone 4	14.09%	17.84%	18.50%	19.02%

demand in China was analyzed. The results are shown in Table 2.

In Table 2, UP represents the average income per user. From Table 1, the TCHD values obtained from engineering simulation and the TEC values obtained from statistical data maintained a high degree of consistency in the proportion of total partition load. Among them, the two accounted for the largest proportion in Zone 3, with 51.65% and 49.30% respectively. The proportion in Zone 1 was the smallest, at 11.03% and 8.18%, respectively. Overall, the total energy consumption of Region 3 accounted for 50% of the total national load, which was closely related to the level of UP value. On this basis, to observe the trend of changes in the impact of climate characteristic variables on building unit load in more detail, the sensitivity of climate variables within the group was analyzed. The partition sensitivity of ECD is shown in Figure 12.

From the three figures in Figure 12, it can be seen that,the correlation between temperature and ECD showed a monotonic upward trend, with a maximum value of 0.8. The correlation between radiation variables and ECD reached a maximum value of 0.6 between Zones 2 and 3, and then decreased. The correlation between ECD in Zone 2 of the humidity class variables was opposite to that of other zones, with a negative correlation. Overall, the impact of temperature variables on building load was gradually increasing,

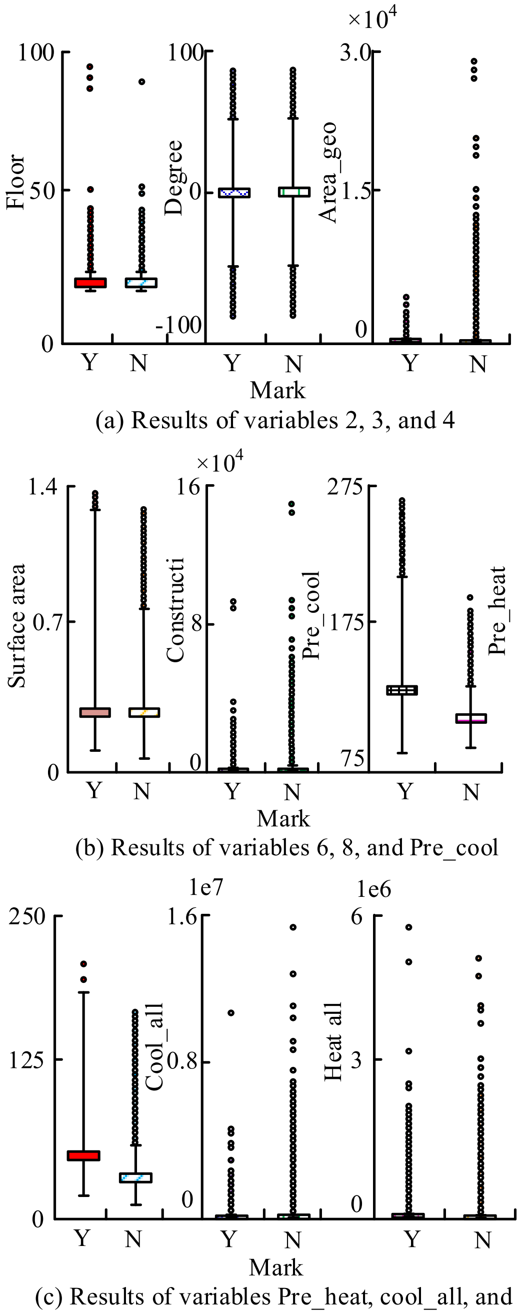


FIGURE 9. **Distribution of energy consumption data boxes for urban buildings.**

with radiation variables having the greatest impact in Zones 2 and 3. The partition sensitivity results of EHD are shown in Figure 13.

From the three figures in Figure 13, it can be seen that,the correlation of Zone 1 in the humidity category variables was opposite to that of other zones, showing a positive correlation. The overall performance of other variables was the same as that of ECD. Overall, different variables had different degrees and ways of influence in different regions, which indirectly verified the effectiveness of the four building energy

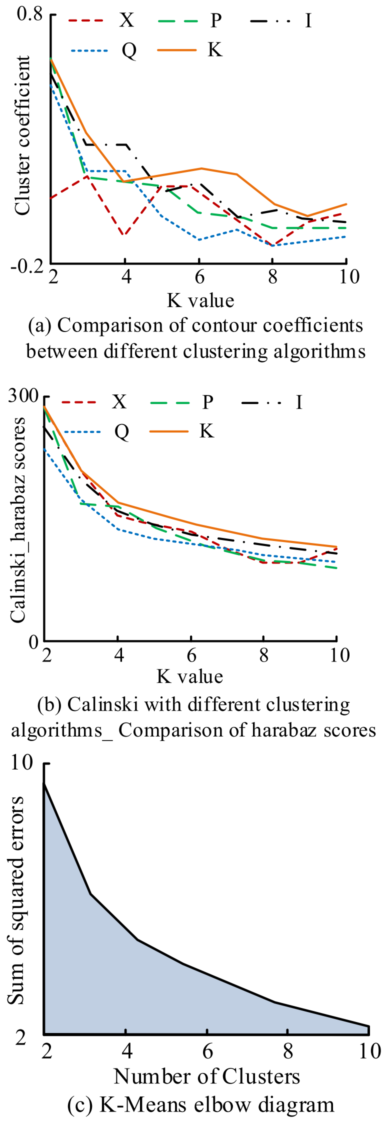


FIGURE 10. **Comparison results of different clustering algorithms and K-means elbow plots.**

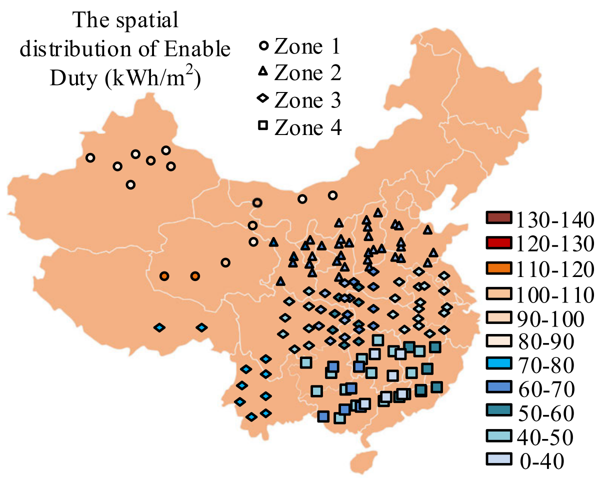


FIGURE 11. **Schematic diagram of spatial distribution of clustering results.**

consumption intensity regions with similar climate characteristics obtained using the K-means clustering algorithm

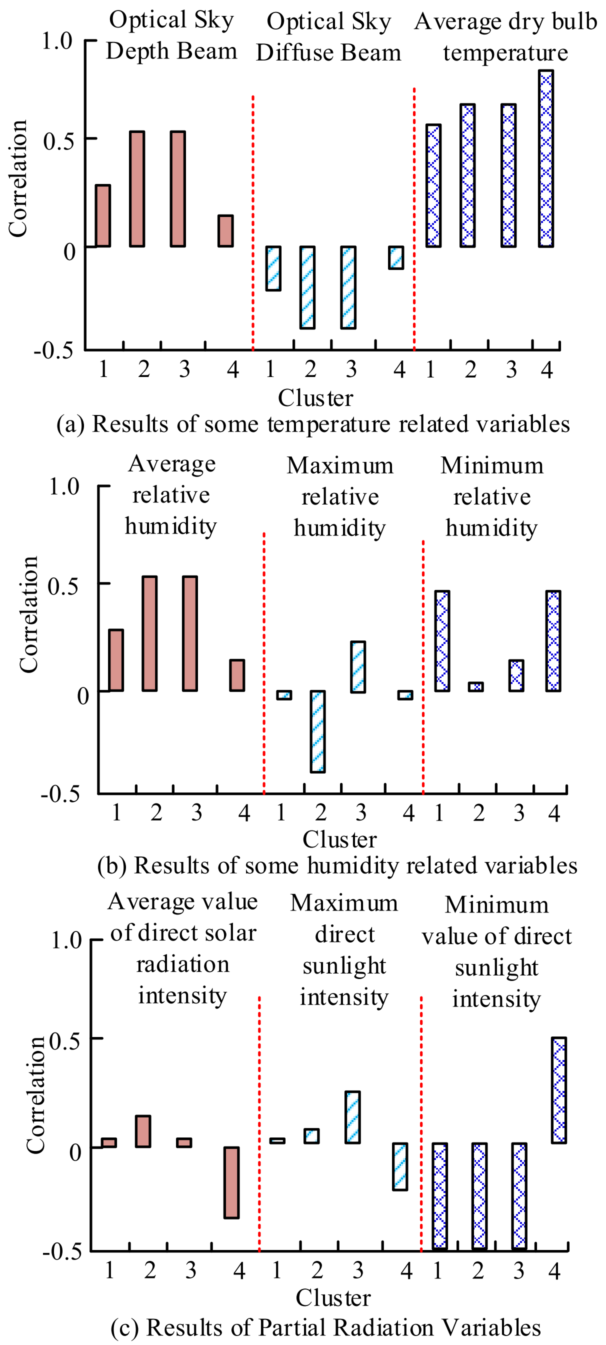


FIGURE 12. Schematic diagram of partition sensitivity of enable cool duty.

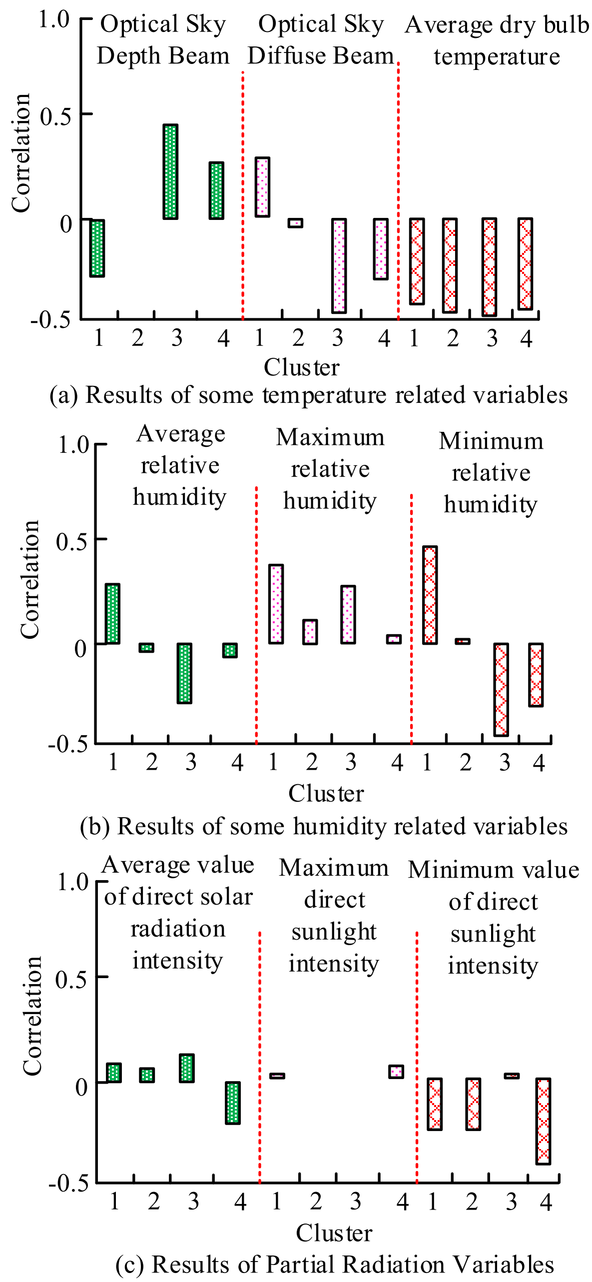


FIGURE 13. Schematic diagram of zoning sensitivity results for enable cool duty.

proposed in the study in actual energy consumption prediction. To further support this result, a quantitative regression analysis was conducted on the climate driven energy consumption of building heating and cooling, socio-economic level, energy supply capacity, and final energy consumption. The predicted TEC value was compared with the actual TEC value. The results are shown in Figure 14.

In Figure 14 (a) and (b), total Energy Supply (TES) represents total energy supply. Based on Figure 14, the analysis results of TEC, TCHD, and other indicators were obtained by using the K-means algorithm to obtain four areas of building energy consumption intensity with similar climate characteristics. The correlation coefficients between TEC and the other three indicators were 0.75, 0.93, and 0.94, respectively. In Figure 14 (c), therefore, a general regression equation was obtained, which could be used to predict and evaluate the actual energy consumption in the region. Moreover, the predicted TCE values were in good agreement with the actual values (the actual value data is sourced from the National Bureau of Statistics' total residential building load data for 31 provinces in 2022). Overall, using the K-means algorithm to simulate and predict energy consumption under data-driven conditions was practical and feasible.

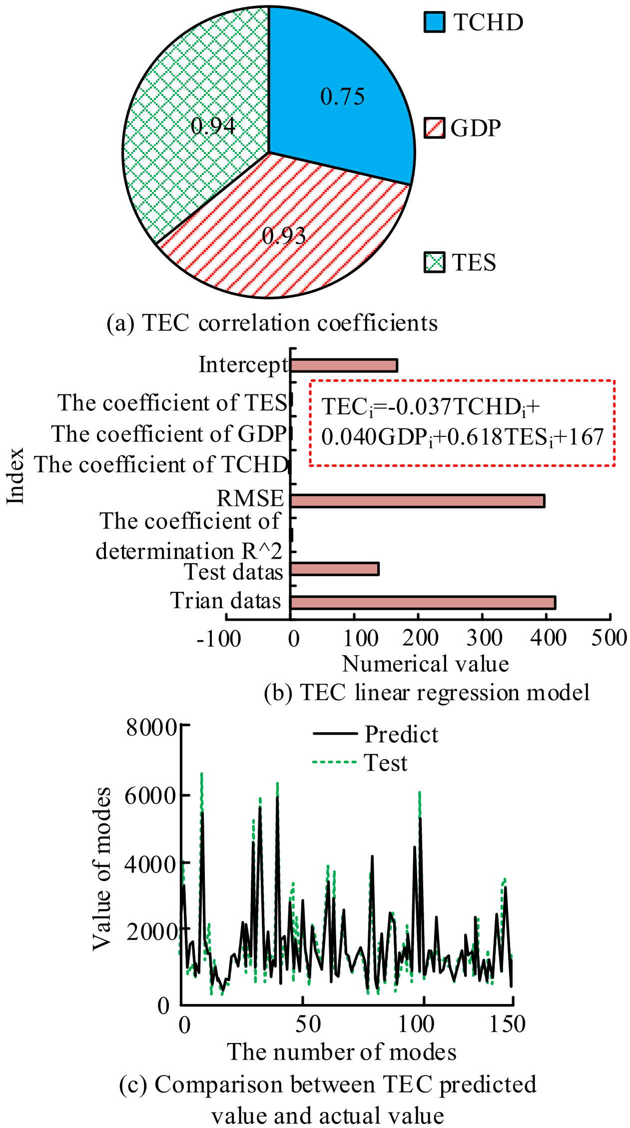


FIGURE 14. **Total energy consumption linear regression analysis results and correlation comparison results.**

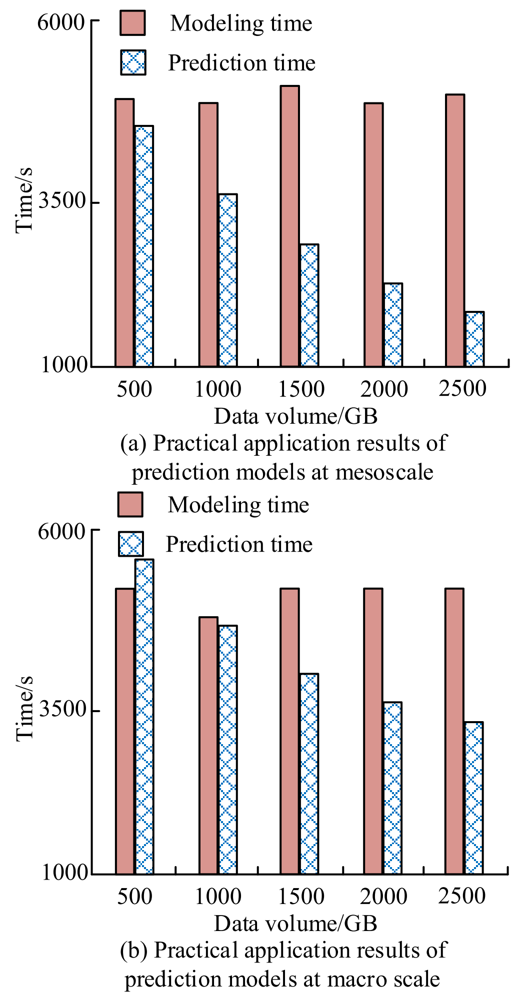


FIGURE 15. **Practical verification results of two models.**

To further verify the practicality of the prediction models at the meso scale and macro scale, the study applied the two to the energy consumption prediction of a building energy system in a central commercial district of a city in eastern China. The modeling time and prediction time were analyzed to test the prediction time cost and verify its practicality. Firstly, the prediction model is integrated into the building energy system of the central commercial district. Secondly, historical data and current relevant variable data are input, and finally, predictions are made. Finally, the actual modeling time and predicted time results are shown in Figure 15.

From Figure 15 (a), it can be seen that the modeling time fluctuation at the mesoscale is very small, maintaining around 5000 seconds. The modeling time decreases as the amount of data increases. From Figure 15 (b), At the macro scale, the modeling time remains around 5000 seconds, and the prediction time is relatively low. Overall, research on building energy consumption prediction models based on data-driven foundations is feasible. It can effectively reduce the actual calculation time in predicting the energy consumption of actual urban energy systems, thereby quickly solving current practical problems and saving a lot of decision-making time. In addition, on the basis of saving prediction time, it can be widely applied in the field of energy conservation, promoting the energy efficiency of the entire central commercial district building.

V. CONCLUSION

In response to the difficulty of applying bottom-up simulation methods, which mainly rely on individual buildings, to urban and regional level building energy consumption planning in current building energy consumption, a mesoscale XGB model and a macro scale K-means model were proposed in a data-driven mode to simulate and evaluate urban building energy consumption, and their effectiveness was verified through experiments. The experimental results showed that the classification algorithm D had the highest accuracy comparison at the mesoscale, at 83%. In the comparison of simulation prediction models built on the basis of classification algorithms, the MSE value of the XGB model for residential sample prediction in refrigeration load was 11.45, which was better than the comparison model. The application of XGB model in actual simulation prediction has found a clear analysis of the correlation between building energy consumption and related variables. In addition, in the comparison of contour coefficients at the macro scale, the K-means algorithm had the best overall effect, with the maximum value of algorithm K when the number of clusters was 10. Applying it to an example made the spatial distribution of load clear. In the sensitivity analysis of ECD, the correlation between temperature and ECD showed a monotonic upward trend, with a maximum value of 0.8. Heating analysis showed similar results, and in the final regression analysis, the predicted value of TCE was in good agreement with the actual value. Overall, the proposed urban building energy consumption simulation evaluation model based on data-driven research is feasible and practical. However, due to the incompleteness of building data information, the construction of high-precision building energy models is hindered, so it is necessary to improve the data information in the future. Due to the inability to quantitatively describe and apply the influencing factors of the surrounding environment of buildings to sample learning, the study only considers the impact of local climate and building factors on building energy consumption, avoiding the impact of microclimate on building energy consumption. Meanwhile, in the application of machine learning in the prediction of building energy consumption, the data-driven translation of building energy consumption models cannot record complete building information, and algorithm selection and parameter tuning still need to be optimized. Therefore, in the future, microclimate analysis and parameter optimization are needed.

REFERENCES

[1] K. Jain and A. Saxena, "Simulation on supplier side bidding strategy at day-ahead electricity market using ant lion optimizer," *J. Comput. Cogn. Eng.*, vol. 2, no. 1, pp. 17–27, Mar. 2022, doi: 10.47852/bonviewjcce2202160.

[2] A. Fnais, Y. Rezgui, I. Petri, T. Beach, J. Yeung, A. Ghoroghi, and S. Kubicki, "The application of life cycle assessment in buildings: Challenges, and directions for future research," *Int. J. Life Cycle Assessment*, vol. 27, no. 5, pp. 627–654, May 2022, doi: 10.1007/s11367-022-02058-5.

[3] M.-D. Ma, M.-X. Chen, W. Feng, and J.-W. Huo, "What decarbonized the residential building operation worldwide since the 2000s," *Petroleum Sci.*, vol. 19, no. 6, pp. 3194–3208, Dec. 2022, doi: 10.1016/j.petsci.2022.10.016.

[4] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12046–12063, Aug. 2021, doi: 10.1109/JIOT.2021.3078462.

[5] I. Skrzypczak, G. Oleniacz, A. Leśniak, K. Zima, M. Mrówczyńska, and J. K. Kazak, "Scan-to-BIM method in construction: Assessment of the 3D buildings model accuracy in terms inventory measurements," *Building Res. Inf.*, vol. 50, no. 8, pp. 859–880, Nov. 2022, doi: 10.1080/09613218.2021.2011703.

[6] C. Li, Z. Dong, L. Ding, H. Petersen, Z. Qiu, G. Chen, and D. Prasad, "Interpretable memristive LSTM network design for probabilistic residential load forecasting," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 6, pp. 2297–2310, Jun. 2022, doi: 10.1109/TCSI.2022.3155443.

[7] J. Chou and D. Truong, "Multistep energy consumption forecasting by metaheuristic optimization of time-series analysis and machine learning," *Int. J. Energy Res.*, vol. 45, no. 3, pp. 4581–4612, Mar. 2021, doi: 10.1002/er.6125.

[8] S. S. G. Perumal, R. M. Lusby, and J. Larsen, "Electric bus planning & scheduling: A review of related problems and methodologies," *Eur. J. Oper. Res.*, vol. 301, no. 2, pp. 395–413, Sep. 2022, doi: 10.1016/j.ejor.2021.10.058.

[9] Y. Yang, X. Wu, F. Liu, Y. Zhang, and C. Liu, "Promoting the efficiency of scientific and technological innovation in regional industrial enterprises: Data-driven DEA-Malmquist evaluation model," *J. Intell. Fuzzy Syst.*, vol. 43, no. 4, pp. 4911–4928, Aug. 2022, doi: 10.3233/jifs-220491.

[10] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and J. Henry, "UAV trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9540–9554, Sep. 2021, doi: 10.1109/TVT.2021.3102161.

[11] S.-C. Zhang, X.-Y. Yang, W. Xu, and Y.-J. Fu, "Contribution of nearly-zero energy buildings standards enforcement to achieve carbon neutral in urban area by 2060," *Adv. Climate Change Res.*, vol. 12, no. 5, pp. 734–743, Oct. 2021, doi: 10.1016/j.accre.2021.07.004.

[12] S. Chen, H. Mao, and J. Sun, "Low-carbon city construction and corporate carbon reduction performance: Evidence from a quasi-natural experiment in China," *J. Bus. Ethics*, vol. 180, no. 1, pp. 125–143, Sep. 2022, doi: 10.1007/s10551-021-04886-1.

[13] N. Aravindhan, M. P. Natarajan, S. Ponnuvel, and P. K. Devan, "Recent developments and issues of small-scale wind turbines in urban residential buildings—A review," *Energy Environ.*, vol. 34, no. 4, pp. 1142–1169, Jun. 2023, doi: 10.1177/0958305x221084038.

[14] A. E. Keleş, E. Önen, and J. Górecki, "Make saving crucial again: Building energy efficiency awareness of people living in urban areas," *Adv. Building Energy Res.*, vol. 16, no. 3, pp. 371–384, May 2022, doi: 10.1080/17512549.2021.1891134.

[15] S. H. Hashmi, H. Fan, Z. Fareed, and F. Shahzad, "Asymmetric Nexus between urban agglomerations and environmental pollution in top ten urban agglomerated countries using quantile methods," *Environ. Sci. Pollut. Res.*, vol. 28, no. 11, pp. 13404–13424, Mar. 2021, doi: 10.1007/s11356-020-10669-4.

[16] Z. Peng, S. Zhao, L. Shen, Y. Ma, Q. Zhang, and W. Deng, "Retrofit or rebuild? The future of old residential buildings in urban areas of China based on the analysis of environmental benefits," *Int. J. Low-Carbon Technol.*, vol. 16, no. 4, pp. 1422–1434, Oct. 2021, doi: 10.1093/ijlct/ctab070.

[17] Z. Deng, Y. Chen, J. Yang, and Z. Chen, "Archetype identification and urban building energy modeling for city-scale buildings based on GIS datasets," *Building Simul.*, vol. 15, no. 9, pp. 1547–1559, Sep. 2022, doi: 10.1007/s12273-021-0878-4.

[18] L. Jin, S. Schubert, D. Fenner, F. Meier, and C. Schneider, "Integration of a building energy model in an urban climate model and its application," *Boundary-Layer Meteorol.*, vol. 178, no. 2, pp. 249–281, Feb. 2021, doi: 10.1007/s10546-020-00569-y.

[19] X.-C. Tan, Y. Wang, B.-H. Gu, L.-S. Kong, and A. Zeng, "Research on the national climate governance system toward carbon neutrality—A critical literature review," *Fundam. Res.*, vol. 2, no. 3, pp. 384–391, May 2022, doi: 10.1016/j.fmre.2022.03.010.

[20] S. Hu, Y. Zhang, Z. Yang, D. Yan, and Y. Jiang, "Challenges and opportunities for carbon neutrality in China's building sector—Modelling and data," *Building Simul.*, vol. 15, no. 11, pp. 1899–1921, Nov. 2022, doi: 10.1007/s12273-022-0912-1.

[21] J. Lao, H. Song, C. Wang, Y. Zhou, and J. Wang, "Reducing atmospheric pollutant and greenhouse gas emissions of heavy duty trucks by substituting diesel with hydrogen in Beijing–Tianjin–Hebei–Shandong region, China," *Int. J. Hydrogen Energy*, vol. 46, no. 34, pp. 18137–18152, May 2021, doi: 10.1016/j.ijhydene.2020.09.132.

[22] D. Gao, Y. Li, and G. Li, "Boosting the green total factor energy efficiency in urban China: Does low-carbon city policy matter?" *Environ. Sci. Pollut. Res.*, vol. 29, no. 37, pp. 56341–56356, Aug. 2022, doi: 10.1007/s11356-022-19553-9.

[23] S. Park and Y. Choi, "Analysis of international standardization trends of smart mining technology: Focusing on GMG guidelines," *Tunnel Underground Space*, vol. 32, no. 3, pp. 173–193, Jun. 2022, doi: 10.7474/TUS.2022.32.3.173.

[24] C. Kumar A and D. M. N. Mubarak, "Classification of early stages of esophageal cancer using transfer learning," *IRBM*, vol. 43, no. 4, pp. 251–258, Aug. 2022, doi: 10.1016/j.irbm.2021.10.003.

[25] S. Soni, S. S. Chouhan, and S. S. Rathore, "TextConvoNet: A convolutional neural network based architecture for text classification," *Int. J. Speech Technol.*, vol. 53, no. 11, pp. 14249–14268, Jun. 2023, doi: 10.1007/s10489-022-04221-9.
[26] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, "A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability," *Energy Environ.*, vol. 33, no. 8, pp. 1583–1612, Dec. 2022, doi: 10.1177/0958305x211044998.
[27] H. Tao, S. M. Awadh, S. Q. Salih, S. S. Shafik, and Z. M. Yaseen, "Integration of extreme gradient boosting feature selection approach with machine learning models: Application of weather relative humidity prediction," *Neural Comput. Appl.*, vol. 34, no. 1, pp. 515–533, Jan. 2022, doi: 10.1007/s00521-021-06362-3.
[28] W. Cai, R. Wei, L. Xu, and X. Ding, "A method for modelling greenhouse temperature using gradient boost decision tree," *Inf. Process. Agricult.*, vol. 9, no. 3, pp. 343–354, Sep. 2022, doi: 10.1016/j.inpa.2021.08.004.
[29] T. Senapati, G. Chen, and R. R. Yager, "Aczel–Alsina aggregation operators and their application to intuitionistic fuzzy multiple attribute decision making," *Int. J. Intell. Syst.*, vol. 37, no. 2, pp. 1529–1551, Feb. 2022, doi: 10.1002/int.22684.
[30] A. Paul, A. Basu, M. Mahmud, M. S. Kaiser, and R. Sarkar, "Inverted bell-curve-based ensemble of deep learning models for detection of COVID-19 from chest X-rays," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16113–16127, Aug. 2023, doi: 10.1007/s00521-021-06737-6.
[31] A. B. Kahng, M. Kim, S. Kim, and M. Woo, "RosettaStone: Connecting the past, present, and future of physical design research," *IEEE Des. Test. IEEE Des. Test. Comput.*, vol. 39, no. 5, pp. 70–78, Oct. 2022, doi: 10.1109/MDAT.2022.3179247.
[32] A. Karna and K. Gibert, "Automatic identification of the number of clusters in hierarchical clustering," *Neural Comput. Appl.*, vol. 34, no. 1, pp. 119–134, Jan. 2022, doi: 10.1007/s00521-021-05873-3.
[33] M. T. García-Ordás, H. Alaiz-Moretón, J.-L. Casteleiro-Roca, E. Jove, J. A. Benítez-Andrades, I. García-Rodríguez, H. Quintián, and J. L. Calvo-Rolle, "Clustering techniques selection for a hybrid regression model: A case study based on a solar thermal system," *Cybern. Syst.*, vol. 54, no. 3, pp. 286–305, Jan. 2022, doi: 10.1080/01969722.2022.2030006.
[34] S. K. Mann and S. Chawla, "A proposed hybrid clustering algorithm using K-means and BIRCH for cluster based cab recommender system (CBCRS)," *Int. J. Inf. Technol.*, vol. 15, no. 1, pp. 219–227, Jan. 2023, doi: 10.1007/s41870-022-01113-6.

GUOPING GAO was born in Henan, China, in 1985. He received the bachelor's degree in management from the Henan University of Economics and Law, Henan, in 2007, the master's degree in economics from the Guizhou University of Finance and Economics, Guizhou, China, in 2012, and the Ph.D. degree from the Wuhan University of Technology, Hubei, China, in 2018.

Since 2018, he has been a Lecturer with the School of Architectural Engineering, Huanghuai University, Henan. He is the author of one book and more than five articles. His research interests include engineering informatization, smart construction, and civil engineering construction and management.

SHAOPAN YANG was born in Hubei, China, in 1988. He received the bachelor's degree in mathematics and applied mathematics from the East China University of Technology, Jiangxi, China, in 2012, and the master's degree in architecture and civil engineering from the Wuhan University of Technology, Hubei, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include precast concrete structures, earthquake engineering, and structural vibration control.