APPLIED GEOPHYSICS, Vol.21, No.4 (December 2024), P. 1–14, 12 Figures.
DOI:10.1007/s11770-024-1120-9

Training data selection using information entropy: Application to heating load modeling of rural residence in northern China

Li-gai Kang[1,2], Hao Li[3,4], Zhi-chao Wang[3,4*], Dong-xiang Sun[1,2], Jin-zhu Wang[1,2], Yang Yang[1,2], and Xu Zhang[1,2]

Abstract: The selection of input variables and their amount has been an important issue in big data load forecasting. Taking heating load forecasting as an example, this paper proposed a method for data filtering based on information entropy. First, the heating data from an air source heat pump system adopted by a rural residence in northern China were employed. Moreover, the training data were classified based on linear or nonlinear variations of outdoor temperature and its changing ranges, while the validation data included three different types of weather conditions, namely, cold, cool, and mild. Then, the information entropy under 2-h, 4-h, 6-h and 8-h training window was quantified to be 1.811, 1.839, 1.877 and 1.856, respectively. For the employed rural residence, an equivalent three-resistance and two-capacity model was established to validate the effectiveness of the training window. Using the derived optimal thermal resistance and capacity, the various selection of outdoor temperature variation trend and range were compared and optimized. Results showed that 6 h of training data had the maximum information entropy and the most abundant information, the minimum errors between actual and forecasting data were observed under 6 h of training data, linear change, and lower outdoor temperature. The mean absolute percentage errors for the load forecasting of three typical days were 5.63%, 8.46%, and 12.10%, respectively.
Keywords: training data selection; information entropy; heating load model; rural residence

Introduction

The building sector represents a substantial energy consumer, and a substantial proportion of energy is allocated to heating, ventilation, and air conditioning. Statistical studies indicate that buildings account for 40% of global energy consumption (Kang et al., 2023; Lu et al., 2023). By accurately predicting short-term building load, energy utilization can be optimized to save energy and reduce emissions (Wang et al., 2023). The data involved in load forecasting have been expanding owing to the growing adoption of intelligent building technology and sensor equipment (Ledesma et al., 2023).

Manuscript received by the Editor May 14, 2024; revised manuscript received June 26, 2024.
This work was supported by the Opening Funds of the State Key Laboratory of Building Safety and Built Environment (grant number: BSBE-EET2021-01).
1. School of Civil Engineering, Hebei University of Science and Technology, Shijiazhuang, 050018, China
2. Engineering Technology Research Center for Intelligent & Low-carbon Assembled Building, Shijiazhuang 050018, China
3. State Key Laboratory of Building Safety and Built Environment, Beijing 100013, China
4. China Academy of Building Research, Beijing 100013, China
* Corresponding author: Zhi-chao Wang (wangzc@emcso.com).

Hence, the selection of input variables and their amount directly influences the accuracy and utility of the results of building load prediction.

The conventional approaches for input variable selection encompassed various algorithms and the use of principal component analysis (PCA) to decompose data. The cluster decomposition algorithm was generally employed in various algorithms for input variable selection. Li et al. (2022) combined unsupervised K-means clustering and supervised K-nearest neighbor classification methods to conduct data clustering and extract features from input data for load prediction. Fan et al. (2015) used cluster analysis to optimize the input data and enhance the predictive performance of the model. Panapakidis et al. (2014) proposed a method to study the electricity consumption of buildings through clustering techniques. Luo et al. (2020) employed clustering technology to extract features from input data, such as daily weather profiles. Chen et al. (2020) developed an enhanced pattern recognition prediction model based on fuzzy C-means clustering and nonlinear regression techniques. Ding et al. (2017) built a clustering method to optimize the accuracy of office building cooling load prediction for model input variables. Lin et al. (2022) used K-means algorithm to cluster the input data set to enhance the accuracy and stability of the model.

The random forest (RF) algorithm was also utilized for data processing and analysis. Liu et al. (2023) proposed RF to select high-influence parameters of building cooling load to lower the dimensionality of original input data and to strengthen the generalization ability of the model. Dong et al. (2021) examined RF and Pearson correlation analysis to identify features of the input data. Liu et al. (2023) used RF to select input data for cooling load prediction for large public buildings. In addition, some alternative algorithms were utilized to select input variables for load forecasting by scholars. Tan et al. (2023) developed a multistage input data feature filtering method for load prediction based on a synthesis correlation analysis algorithm. Amasyali et al. (2021) utilized neighborhood component analysis to conduct feature selection on input data. Xiao et al. (2023) proposed a time convolutional network to extract features and developed input feature sets for the prediction model. Zhou et al. (2021) adopted operating parameters determined by the ReliefF algorithm as input parameters to enhance model stability.

The technique of decomposing data to extract features for load forecasting was greatly used by scholars and researchers. Lu et al. (2020) employed a fully integrated empirical mode decomposition with adaptive noise to decompose raw data into multiple smooth datasets to forecast building load. Al-musaylh et al. (2018) and Wang et al. (2019) decomposed the input data into intrinsic mode functions and residual terms. Yang et al. (2022) used mutual information and PCA to conduct feature selection reduction on multidimensional weather influencing factors. Quanga et al. (2021) developed a hybrid online model for real-time data processing and decomposed the original load sequence into trend terms, seasonal terms, and residual terms.

The said methods mainly focused on the selection of input variables, while determining the optimal amount of data needed for load prediction is also crucial. Information theory has been employed for data processing and analysis for a considerably long time. Zhang et al. (2016) introduced a novel data feature selection method by merging fuzzy rough sets with information entropy theory. Dai et al. (2016) used feature selection algorithms for an interval-valued information system based on information entropy. Deng et al. (2022) performed the neighborhood fuzzy entropy method based on dual-similarity to select data features for label distribution learning. Zhu et al. (2023) presented an information screening method based on entropy to select samples with high information content. Zhang et al. (2023) explored the selection of data features for portion labeling of neighborhood rough set information theory. Zhao et al. (2023) suggested a partial label classification data outlier detection based on conditional information entropy, which improved the application scenarios and ranges of information entropy. He et al. (2024) presented an oscillating particle swarm optimization feature selection algorithm for mixed data based on mutual information entropy, which increased the classification accuracy by 5.8% compared with other algorithms. The advantages of data filtering based on information entropy are as follows:

1) Information entropy can be employed for feature selection to decrease data dimension and increase model efficiency.

2) Information entropy filtering facilitates lessening redundant information in the data and strengthening the generalization ability of the model.

3) Information entropy can be employed to optimize the attention model and enhance the information resolution.

Therefore, selecting the appropriate amount of training data is key to the successful construction of a heating load model. In this paper, a data filtering method based on information entropy is presented, and the information entropies of 2, 4, 6, and 8 h of training data are determined. The effectiveness of the data filtering using information entropy is proven through the prediction results of room temperature and heating load by the developed equivalent three-resistance, two-capacity (3R2C) model.

Modeling and Methodology

This section presents the experimental methodologies, the selection and classification of data, the method of information entropy for data filtering, and the development of an equivalent resistance–capacitance (RC) model.

Experimental methodologies

Fig. 1 shows that the detached rural residence is in Beijing, China, in a cold zone, with an average outdoor air temperature of −1.6°C in winter. The rural residence has a floor area of approximately 120 m^2. The building has four rooms, all with south-facing windows. The windows are double glass with aluminum frames. All the walls are made of 6 cm thick bricks and covered with plaster.

Fig. 1 View of a detached rural residence in Beijing.

The heating supply data are acquired from an actual heating project. The heating system is observed from November 15, 2019, to March 15, 2020. Heating load and outdoor air temperature are examined as inputs to the RC model. All data are sampled at intervals of 1 min, which can avoid overlooking thermal dynamics.

Methodology of information entropy

The term “information” is the interpretation or decoding of events, data, or signals that possess significance. In information theory, information content can commonly be expressed as the uncertainty of the contained data. To deepen the understanding of this uncertainty, employing the concept of information entropy for explanation is customary.

The idea of “information entropy” was introduced by Shannon in 1948 (Li et al. 2022). “Information entropy” is the level of uncertainty or amount of information contained in a random event (Rossini et al. 2019). It can be applied in numerous fields, such as geoscience, artificial intelligence, data filtering, and geoscience (Ghosh et al. 2017). The utilization of information entropy filtering mitigates superfluous data and improves the model’s generalization capacity.

A load forecasting case is used to analyze the application of information entropy filtering for data analysis in this paper. The process of information entropy filtering data is presented in Fig. 2. The general procedure for screening data by information entropy can include the following steps (using building load forecasting as an example):

1) Data collecting. Data collection involves gathering a diverse range of information relevant to load forecasting, covering historical load data, weather data, time data, and other relevant factors.

2) Data pre-processing. Data pre-processing entails cleaning, removing outliers, filling in missing values, and guaranteeing data integrity and accuracy to prime the gathered data for examination.

The method of data cleaning refers to the collected data to standardize the data format and correct the inconsistency. The removal of outliers refers to the detection and processing of extreme values that may cause bias in the analysis. The accumulated experimental data may possess unreasonable conditions, which are identified and eliminated by an Excel screening function. Filling missing values are data that have been supplemented owing to missing data. The average value

imputation approach is employed for data filling in this study.

3) Feature selection. Feature selection is a vital step in load forecasting, where remarkable features are identified from the collected data.

4) Construct feature vectors. Construction of feature

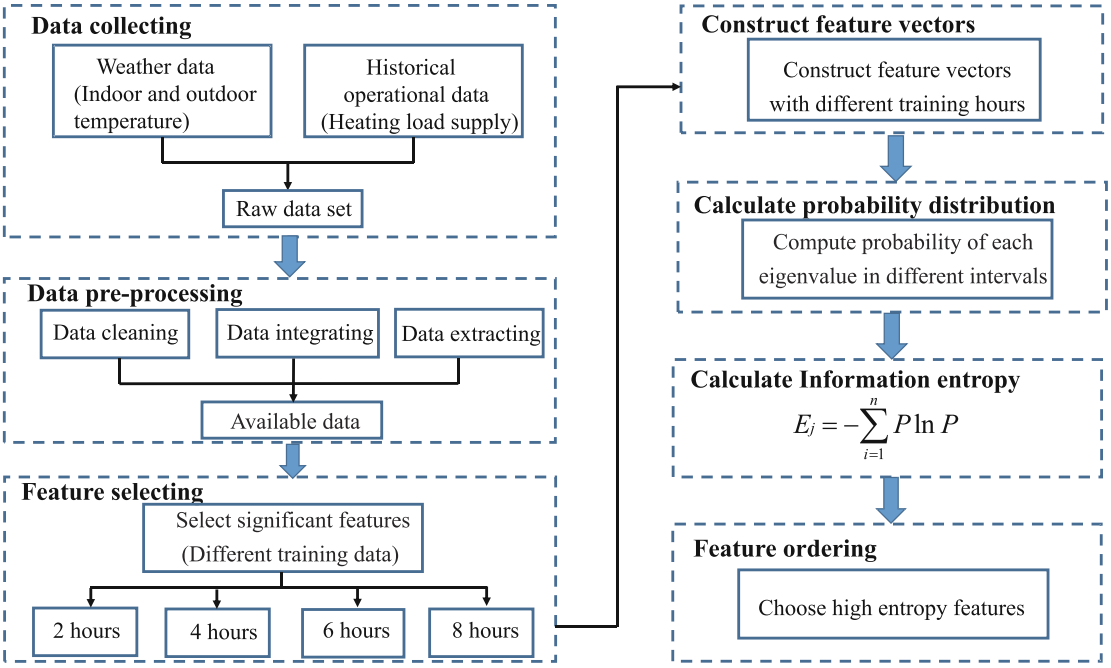


Fig. 2 Information entropy filtering data.

vectors is performed by arranging the eigenvalues of each data set in order of different interval probabilities to form distinct eigenvectors.

$$X = (x_{ij})_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ x_{31} & x_{32} & \cdots & x_{3n} \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \quad (1)$$

where m and n represent the number of rows and columns in the eigenvector, respectively, whereas i and j signify the position and index of element x within the eigenvector, respectively.

To acquire the eigenvalues of the dataset using PCA, the following steps are used: (1) Covariance Matrix Computation. Determining the covariance matrix of the dataset entails finding the covariance between each pair of variables in the dataset. (2) Eigenvalue Decomposition. Eigenvalue decomposition on the covariance matrix results in a set of eigenvalues and their corresponding eigenvectors. (3) Selection of Principal Components. The eigenvalues denote the amount of variance acquired by each corresponding eigenvector. The top eigenvalues and their associated eigenvectors are selected to keep the most substantial features. (4) Feature Vector Computation. The feature vectors are calculated by multiplying the original data matrix with the selected eigenvectors. Each row in the resulting matrix denotes a transformed data point in the reduced-dimensional space.

5) Calculate probability distribution. The occurrence probability of each eigenvalue in different intervals is determined for the eigenvectors of every dataset.

6) Calculate information entropy. The information entropy of the training sets is calculated to quantify the level of uncertainty and information across several durations of training. The entropy equation is used for computational purposes in this study.

$$E_j = -\sum_{i=1}^{n} P \ln P, \quad (2)$$

where P is the probability distribution, and E_j is the information entropy.

7) Feature ordering. The features are ordered according to the computed information entropy, and choosing features with higher information entropy is prioritized.

3R2C model

The advantages of the white box and black box models are merged in the development of the gray box model, which is based on the partial mechanism method. The refinement of such models also requires a limited number of data-driven conditions, while results can be interpreted with specific physical implications. An RC model is largely used in gray-box modeling for system characterization. It is principally employed for loading prediction and building thermal modeling. Hence, an RC model is used for model building in this paper.

Given the inadequate thermal insulation, simple structure, and small heat capacity of rural residences in northern areas, a 3R2C model is implemented for this building. A simplified RC model is demonstrated in Fig. 3. The 3R2C model is composed of three major components: outdoor air, opaque envelope, indoor air and internal thermal mass. The simplified RC model hypothesis is explained in reference (Li et al. 2022; Kang et al. 2023).

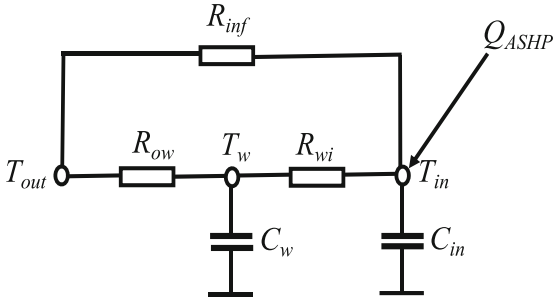


Fig. 3 Model structure of the 3R2C model.

Each node in a circuit, T_w and T_{in}, is distributed to a circuit node, and the flow to and from each node is balanced. The governing equation of the heat balance of each node is expressed as follows:

$$C_w \frac{dT_w}{dt} = \frac{T_{out} - T_w}{R_{ow}} + \frac{T_{in} - T_w}{R_{wi}}, \tag{3}$$

$$C_{in} \frac{dT_{in}}{dt} = \frac{T_{out} - T_{in}}{R_{\inf}} + \frac{T_w - T_{in}}{R_{wi}} + Q_{ASHP}, \tag{4}$$

$$Q_{ASHP} = c_p m \rho (T_s - T_r), \tag{5}$$

where ρ and c_p are the density and the specific heat capacity of the heat-carrying agent, respectively. This model has two states, namely, T_w and T_{in}, which correspond to the temperature of the building's opaque envelope and the mean indoor air temperature, respectively. Tout is the outdoor air temperature. R_{ow} is the thermal resistance between the exterior surface of an opaque envelope and the outdoor air. Accordingly, R_{wi} is the thermal resistance between the indoor air and the interior surface of an opaque envelope. C_w is the thermal capacitance of an opaque envelope. C_{in} is the thermal capacitance of the indoor air and internal thermal mass. R_{inf} is the thermal resistance between the indoor and outdoor air through windows. All thermal resistances and thermal capacitances are considered time-invariant. Q_{ASHP} is the heating load provided by air source heat pump system, which is computed by the measured water flow rate and the temperature difference between the supplied water and the returned water.

Model identification approach and the objective function

The constructed heating load model shows a nonlinear optimization to find the optimal values of R and C. Genetic algorithm (GA) is an optimization algorithm based on natural selection and gene recombination, which can automatically adjust parameters to find the optimal solution (Yi et al. 2022). Compared with other optimization algorithms, GA has the following advantages: 1) Parallel computing capability: GA can calculate multiple individuals in parallel, thus accelerating the optimization speed. 2) Robustness: GA has strong robustness to noise and discontinuity and can operate in complex, uncertain environments. 3) Adaptive: GA can automatically modify parameters to adapt to diverse problems and data sets to enhance optimization efficiency. Hence, GA is employed to search for the optimal values of R and C. The flow chart of parameter optimization by GA is illustrated in Fig. 4.

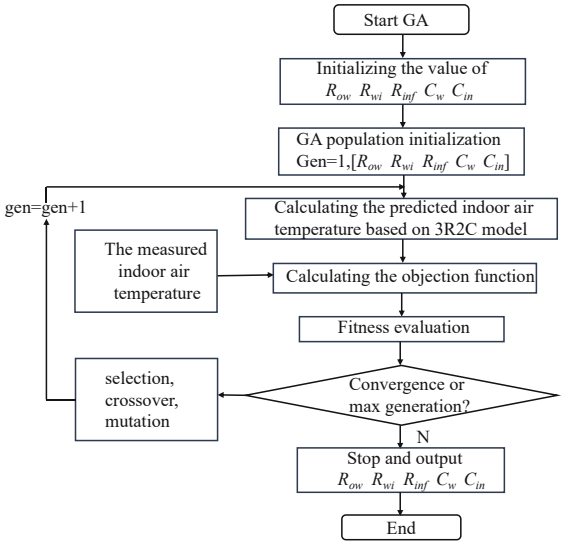


Fig. 4 Flow chart of parameter optimization by GA.

A fitness function of GA optimization is employed to evaluate the predicted results and the measurement results in this paper. This objective function J of the optimization aims to minimize the integral root mean square error ($RMSE$) described in equation (6):

$$J(R_{ow}, C_w, R_{wi}, R_{inf}, C_{in}) = minimize\sqrt{\frac{\sum_{i=1}^{n}(x_{act} - x_{pre})^2}{N}}, \quad (6)$$

where x_{act} is the measured indoor air temperature; x_{pre} is the predicted indoor air temperature; R_{ow}, C_w, R_{wi}, R_{inf}, and C_{in} are the parameters of the 3R2C model.

Evaluation metrics

The mean absolute error (MAE), mean absolute percentage error ($MAPE$), RMSE, and R-squared (R^2) are suggested to assess the accuracy of model prediction. The difference between the predicted results and actual values can be measured from several perspectives. The above performance criteria are expressed as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_{act} - x_{pre}\right|, \quad (7)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_{act} - x_{pre}}{x_{act}}\right|, \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{act} - x_{pre}\right)^2}, \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_{act} - \hat{x}_{pre})^2}{\sum_{i=1}^{n}(x_{act} - \bar{x}_{act})}, \quad (10)$$

Results and Discussion

To demonstrate the effectiveness of information entropy filtering data, this section introduces the prediction of indoor air temperature and heating load with different training windows based on the RC model.

Selection of data

Considering the small indoor thermal disturbance and the absence of solar radiation at night, a heating load model is developed using data collected after 22:00. The training data for 2, 4, 6, and 8 h are selected as the subjects. Table 1 describes the training time windows.

The training data of 8, 6, 4, and 2 h are the experimental data from 22:00 of the previous day to 6:00, 4:00, 2:00, and 0:00 of the next day, respectively, and 1 min denotes a set of experimental data. Hence, the training data of 2 to 8 h are expressed as 120 to 480 sets of experimental data, respectively.

Table 1 Description of training time windows

Training time window (h)	2	4	6	8
Time of duration (h)	22:00-0:00	22:00-2:00	22:00-4:00	22:00-6:00
Experimental data (sets)	120	240	360	480

Statistics reveal that during the 8 h from 22:00 of the previous day to 6:00 of the next day, the number of days exceeds half of the heating season when outdoor air temperature varies between 12°C and 1°C. Hence, this paper only considers the results of model training within this outdoor air temperature range. The results of outdoor temperature change under different training time windows (2, 4, 6, and 8 h) are presented in Fig. 5.

Fig. 5 shows that for 6 and 8 h of training data, outdoor air temperatures change between 3°C and 6°C on more than half of the days throughout the heating season. Thus, for 6 and 8 h of training data, model training is performed at an outdoor temperature interval of approximately 3°C–6°C. For 4 h of training data, the

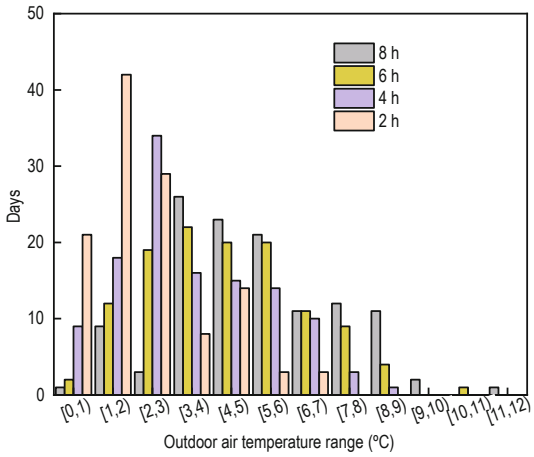


Fig. 5 Proportion of days with outdoor air temperature range.

outdoor temperature fluctuates between 2°C and 5°C on more than half of the days throughout the heating season. Hence, for 4 h of training data, model training is performed at a temperature range of 2°C–5°C. For 2 h of training data, the variation in outdoor temperature is mainly concentrated at 2°C. Thus, for 2 h of training data, model training is performed at an interval of 2°C.

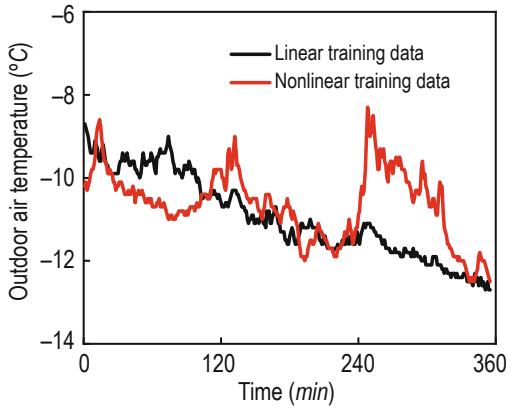


Fig. 6 Example of linear and nonlinear data.

Meanwhile, outdoor air temperature curve of training data can be classified into two types: similar to linear data and nonlinear. Two instances of linear data and nonlinear data are presented in Fig. 6. The linear data indicate minor fluctuations in outdoor temperature, whereas the nonlinear data reveal remarkable fluctuations. For convenience, L and N denote the linear and nonlinear training data, respectively, and their subscripts indicate diverse intervals of outdoor air temperature. A higher subscript number denotes a higher outdoor air temperature and vice versa.

Three representative days are chosen to validate the load prediction. The actual measurement data for these three validation days are shown in Table 2. These three categories of outdoor air temperatures, namely, cold, cool, and mild, are designated as test1, test2, and test3, respectively.

Table 2 Actual measurement data of three validation days

Weather type	Outdoor temperature range (°C)	Indoor temperature range (°C)	Supply water temperature (°C)	Return water temperature (°C)
cold	−11.9–−4.2	15.1–16.4	24.8–34.8	27.7–30.8
cool	−3.8–2.7	17.4–18.2	28.1–36.4	27.9–31.2
mild	0.7–5.2	20.9–22.9	28.0–36.5	28.1–31.6

Calculation result of information entropy

In practice, empirical guidelines can be obtained from historical data for the entire heating season (Datale et al., 2022). This paper employs four training windows (2, 4, 6, and 8 h) to compute information entropy. The calculation results are divided into seven diverse temperature error intervals. The probability distribution statistics are exhibited in Fig. 7.

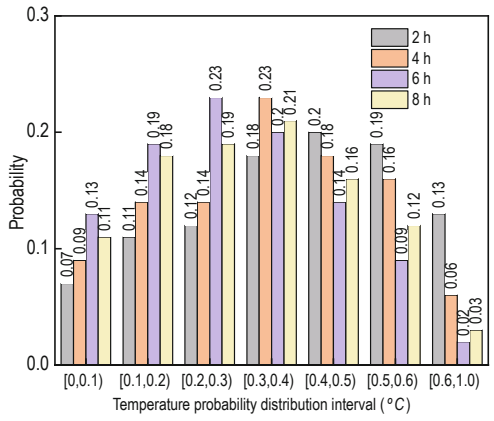


Fig. 7 Temperature probability distribution for different training windows.

Information entropy is determined according to equation (2), which is expanded to equation (11) in this paper.

$$E_j = p_1 \times \ln\frac{1}{p_1} + p_2 \times \ln\frac{1}{p_2} + p_3 \times \ln\frac{1}{p_3} + p_4 \times \ln\frac{1}{p_4} + p_5 \times \ln\frac{1}{p_5} + p_6 \times \ln\frac{1}{p_6} + p_7 \times \ln\frac{1}{p_7}, \quad (11)$$

where p_i is the probability distribution of temperature errors between the measured and predicted values, and subscript i denotes various error intervals, namely, [0,0.1), [0.1,0.2), [0.2,0.3), [0.3,0.4), [0.4, 0.5), [0.5, 0.6), and [0.6,1.0).

Fig. 7 presents that the probability distribution of 8-h training window within the temperature interval [0,0.1] is 0.18, and it increases to 0.19 within the intervals [0.1,0.2) and [0.2,0.3). The probability distribution in the interval [0.3, 0.4) is 0.21, it decreases to 0.16 in the interval [0.4, 0.5), further drops to 0.12 in the temperature range of [0.5, 0.6), and reaches a low value of 0.03 in the interval [0.6, 1.0).

The statistical findings reveal the following calculation results for the 8-h information entropy:

$$E_8 = p_1 \times \ln\frac{1}{p_1} + p_2 \times \ln\frac{1}{p_2} + p_3 \times \ln\frac{1}{p_3} + p_4 \times \ln\frac{1}{p_4} + p_5 \times \ln\frac{1}{p_5} + p_6 \times \ln\frac{1}{p_6} + p_7 \times \ln\frac{1}{p_7} \quad , (12)$$

The information entropy of 8 h of training data is 1.856. The information entropies of E_6, E_4, and E_2 are 1.877, 1.839, and 1.811, respectively.

The training window should ideally provide more informative content to enable the model to learn and predict with higher accuracy. If the training data demonstrate a higher level of information entropy, it indicates greater diversity and complexity within the dataset, enabling the model to gain insights into various scenarios and adapt better to unseen data. Such datasets enhance the model's generalization capabilities and its ability to control real-world complexities instead of solely overfitting existing data. Hence, the information entropy of 6 h of training data displays strong predictability.

Validation results of various training window

2 and 4-h training window validation

The outdoor air temperature varies during the 2-h training window from −12°C to 0°C, with intervals of 2°C. In Fig. 8, for 2-h training window, L1 (N1)–L6 (N6) are represented as −12°C–−10°C, −10°C–−8°C..., −2°C–0°C.

Figs. 8 and Fig. 9 present the actual measured temperature and 3R2C predicted temperature for the

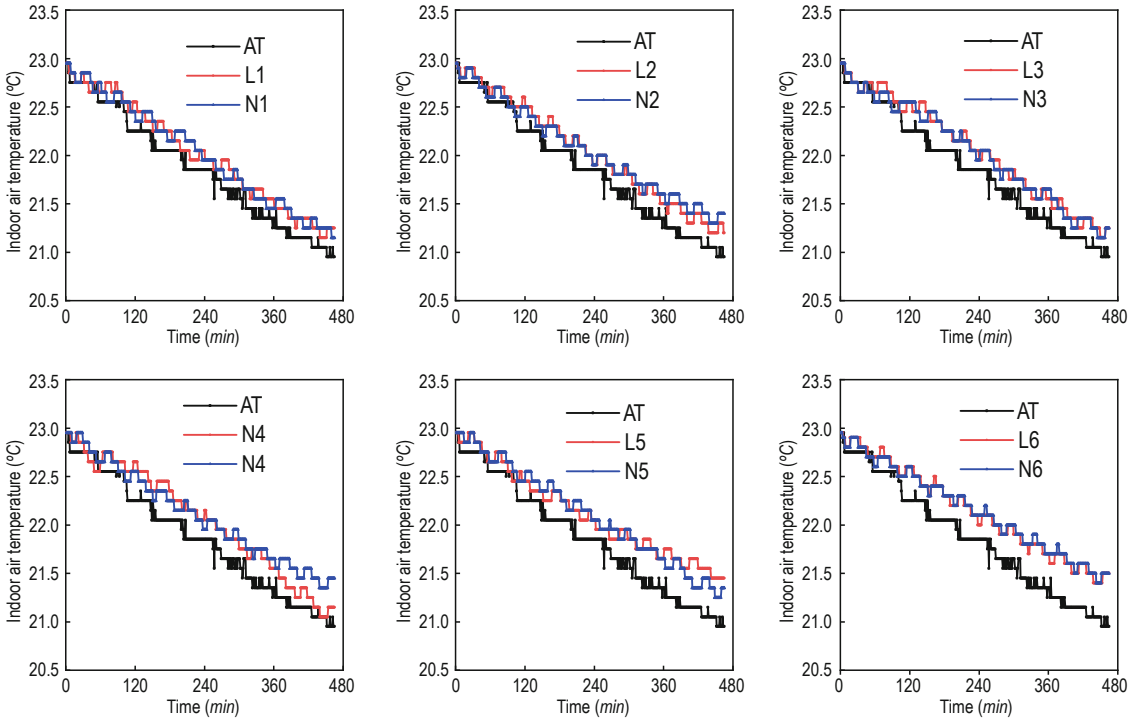
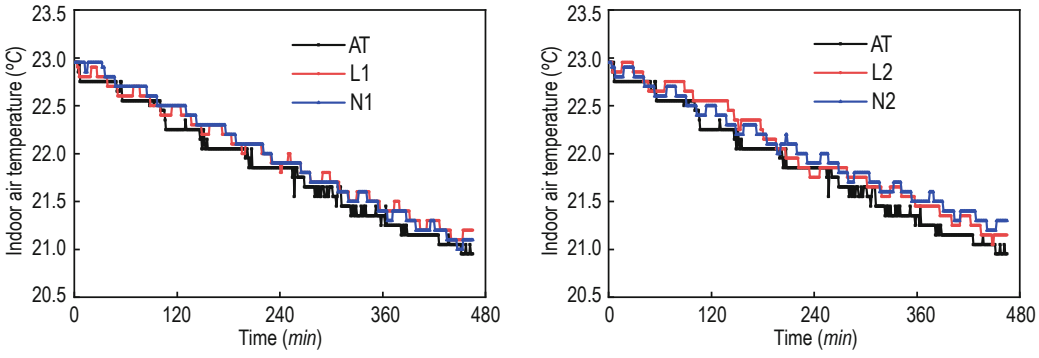


Fig. 8 Actual measured temperature and 3R2C predicted temperature (2 h).



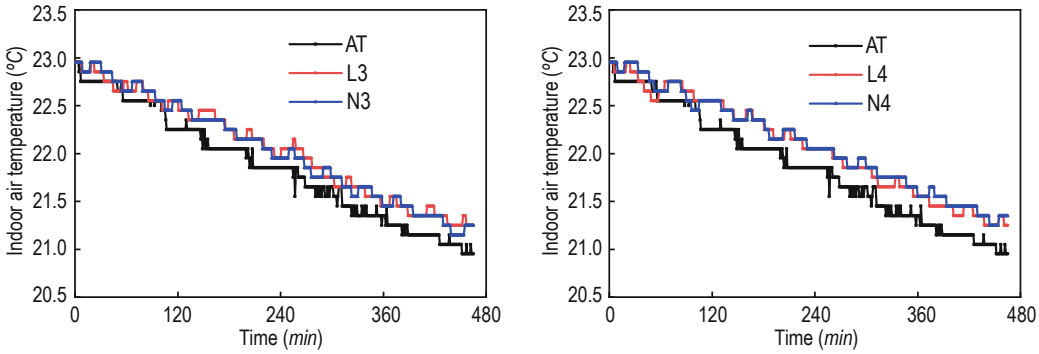


Fig. 9 Actual measured temperature and 3R2C predicted temperature (4 h).

training window of 2 and 4 h, respectively. The predicted indoor temperature is consistent with the actual indoor temperature decline trend, but errors are noted in several moments. The weak predictive performance may be owing to inadequate training data to capture the room's thermal behaviors accurately; hence, important building thermal dynamic properties are missed. Another possibility is that the temperature differences between indoors and outdoors are extremely small to drive the predictive performance of the RC model.

Table 3 Correlation performances and error results of test3 (2 h)

Outdoor temperature type	Outdoor temperature range (°C)	R^2 (%)	*MAE* (%)	*MAPE* (%)	*RMSE* (%)
L1	−12–−10	95.26	16.13	0.74	18.25
L2	−10–−8	95.00	18.30	0.84	20.40
L3	−8–−6	94.34	19.63	0.90	21.85
L4	−6–−4	93.15	22.36	1.03	26.52
L5	−4–−2	92.14	25.63	1.18	29.53
L6	−2–1	91.78	28.76	1.33	31.79
N1	−12–−10	95.24	16.15	0.75	18.45
N2	−10–−8	94.37	19.89	0.92	22.99
N3	−8–−6	94.68	18.81	0.87	21.41
N4	−6–−4	92.78	24.32	1.26	27.30
N5	−4–−2	92.12	24.77	1.14	27.11
N6	−2–0	91.25	29.83	1.38	33.42

Table 4 Correlation performances and error results of test3 (4 h)

Outdoor temperature type	Outdoor temperature range (°C)	R^2 (%)	*MAE* (%)	*MAPE* (%)	*RMSE* (%)
L1	−12–−9	96.86	12.39	0.57	14.46
L2	−9–−6	95.74	15.44	0.71	17.81
L3	−6–−3	94.63	19.27	0.89	21.79
L4	−3–1	93.18	22.22	1.02	23.96
N1	−12–−9	96.57	12.85	0.59	14.81
N2	−9–−6	95.34	16.30	0.76	18.78
N3	−7–−4	95.01	18.09	0.83	19.88
N4	−4–0	92.33	24.60	1.13	26.75

The error results acquired from validating 2- and 4-h training windows on test3 are presented in Tables 3 and 4, respectively. Table 3 shows the optimal result of forecasting indoor air temperature with a 2-h training window is L1; R^2 is 95.26%, *MAE* is 16.13%, *MAPE* is 0.74%, and *RMSE* is 18.25%. Table 4 reveals a high degree of accuracy for indoor air temperature prediction with a 4-h training window is L1; R^2 is 96.86%, *MAE* is 12.39%, *MAPE* is 0.57%, and *RMSE* is 14.46%.

6-h training window validation

Training windows of 6 h are split into three sets of linear and three sets of nonlinear data. Fig. 10 presents the actual measured temperature and 3R2C predicted temperature for test3. The measured indoor air temperature and predicted indoor air temperature fit well in most cases.

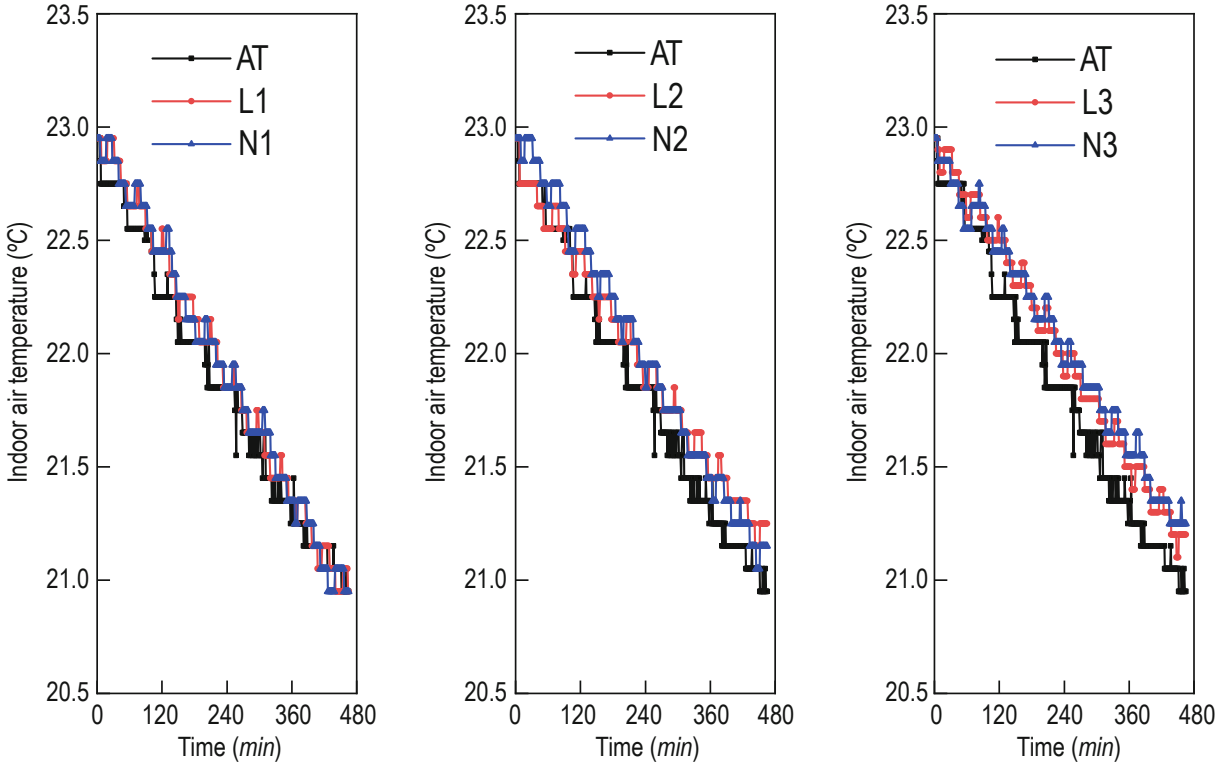


Fig. 10 Actual measured temperature and 3R2C predicted temperature (6 h).

The training windows of six groups are validated on test3, and the results are presented in Table 5. For 6-h linear training window, indoor and outdoor temperature differences range from 24.8°C to 28.1°C, 23.3°C to 25.1°C, and 20.2°C to 22.4°C. The RC model exhibits superior prediction performance owing to a greater temperature disparity between the indoor and outdoor temperatures. For example, the *RMSE* for L1, L2, and L3 are 11.73%, 16.45%, and 19.26%, respectively. Table 5 and Fig. 10 reveal that the predictive performance of the linear training window is better than that of the nonlinear window under the same outdoor temperature ranges. The

Table 5 Correlation performances and error results of test3 (6 h)

Outdoor temperature type	Outdoor temperature range (°C)	R^2 (%)	*MAE* (%)	*MAPE* (%)	*RMSE* (%)
L1	−13–−8	98.37	9.03	0.41	11.73
L2	−8–−3	97.42	13.44	0.62	16.45
L3	−4–2	95.10	17.27	0.80	19.26
N1	−13–−8	98.22	9.66	0.44	12.16
N2	−8–−4	96.29	15.25	0.70	17.24
N3	−4–1	94.61	19.74	0.91	22.34

RMSE for linear training window is 11.73%, whereas that for nonlinear training window is 12.16%.

8-h training window validation

Fig. 11 shows the actual measured indoor air temperature and 3R2C predicted indoor air temperature for test3. In most cases, the predicted indoor air temperature matches the actual measured indoor temperature well. Correlation performances and error results for test3 are presented in Table 6.

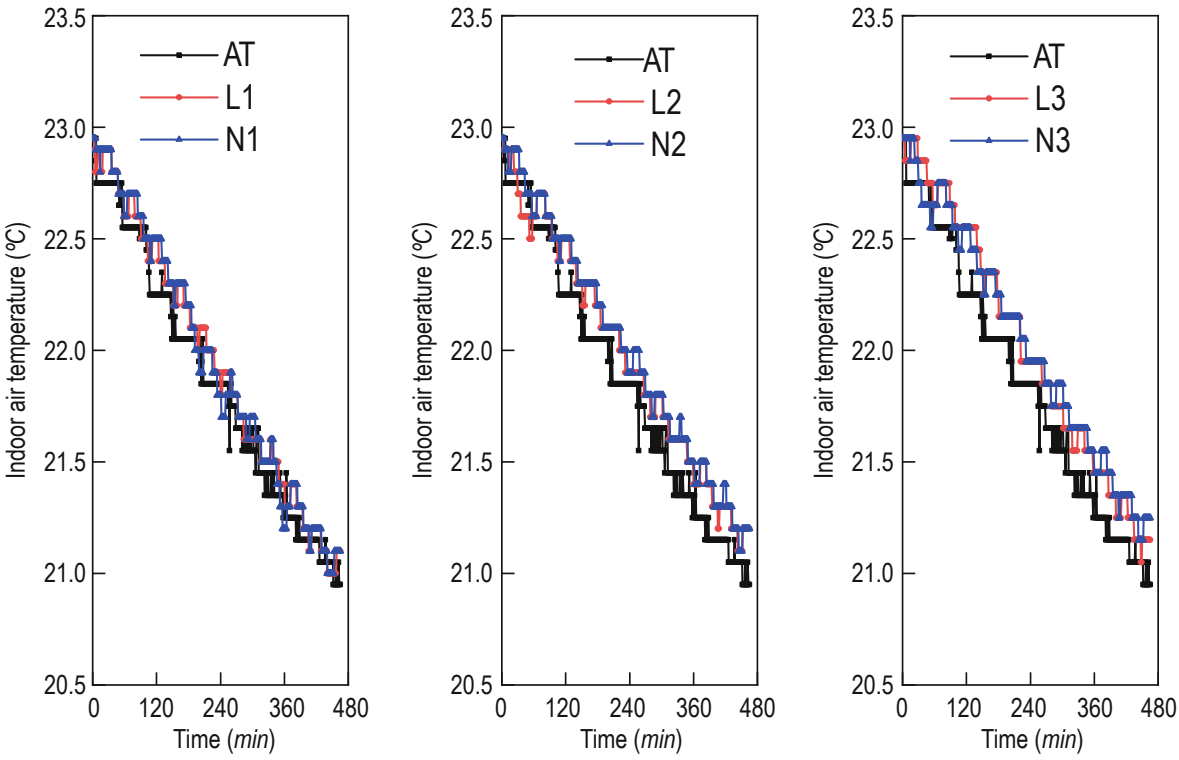


Fig. 11 Actual measured temperature and 3R2C predicted temperature (8 h).

Table 6 Correlation performances and error results of test3 (8 h)

Outdoor temperature type	Outdoor temperature range (°C)	R^2 (%)	*MAE* (%)	*MAPE* (%)	*RMSE* (%)
L1	−13–−8	98.16	9.81	0.45	11.80
L2	−8–−4	96.42	14.45	0.66	16.32
L3	−5–1	95.27	16.98	0.78	18.87
N1	−13–−8	97.28	11.07	0.51	12.96
N2	−9–−4	96.23	15.88	0.73	17.77
N3	−4–1	94.88	18.58	0.86	20.46

Combining Fig. 11 and Table 6, it can be seen that L1 has the highest performance when predicting indoor air temperature, and R^2, *MAE*, *MAPE*, and *RMSE* are 98.16%, 9.81%, 0.45%, and 11.80%, respectively. For the same outdoor air temperature interval, the prediction results of the 6-h training window are similar to or better than those of the 8-h training window. The comparison of the performance between 6- and 8-h training windows for room temperature prediction demonstrates results consistent with those obtained using the information entropy method in the calculation results of the information entropy section.

Combined Table 3 to table 6, the better results of prediction are obtained by using a 6-h training window, which is consist with the derived result of information entropy. Therefore, in future study, the method of information entropy can be used for the data volume selection of prediction.

Accumulated heating load validation

This paper employs a 6-h training window of linear large temperature difference type for accumulated heating load analysis. The three typical days mentioned above are used to validate the accumulated energy consumption. Fig. 12 presents the predicted and measured accumulated heating load of three typical days. The overall trend of heating load profiles of RC predicted and the measured are similar.

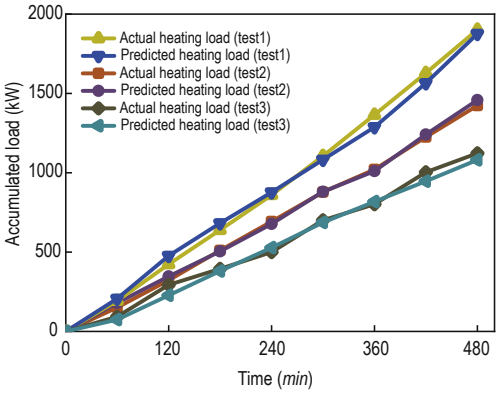


Fig. 12 Comparison of predicted and measured accumulated heating load.

Based on the measured heating load, the *MAPE* of 3R2C predicted heating load of three typical days are 5.63%, 8.46%, and 12.10%. Due to the greater variability in energy consumption compared with indoor air temperature and the inherent noise in energy consumption signals, the predicted indoor air temperature demonstrates less variability than the energy consumption prediction, making it more stable. In summary, the effectiveness of information entropy selecting data is proven by the prediction of indoor air temperature and heating load.

Conclusion

The effectiveness of load forecasting in big data is not necessarily improved by increasing the amount of training window employed. The key to successful construction of a model lies in selecting an appropriate amount of data. Thus, a method based on information entropy theory to filter the training data is proposed. The main conclusions are as follows:

(1) A training data filtering method based on information entropy is presented. The information entropies of 2, 4, 6, and 8 h are 1.811, 1.839, 1.877, and 1.856, respectively. The 6-h training window demonstrates the maximum information entropy, resulting in the largest amount of data.

(2) The load prediction improves when using a 6-h training window under the same outdoor air temperature conditions and ranges. Under cold, cool, and mild weather, MAPE for indoor air temperature predictions are 0.57%, 0.46%, and 0.41%, respectively.

(3) For the same training window and outdoor temperature range, linear data provide a more accurate prediction of room temperature than nonlinear data. For cold, cool, and mild weather conditions, the RMSE for 6 h of linear training data are 11.15%, 12.32%, and 11.73%, respectively, and the corresponding RMSE for 6 h of nonlinear training data are 12.04%, 12.68%, and 12.16%, respectively.

(4) The utilization of a training window with lower outdoor temperature can obtain more accurate predictions of room temperature under the same training window and outdoor temperature type. In cold, cool, and mild weather conditions, the MAE obtained using a 6-h training window are 8.94%, 7.68% and 9.03%, respectively.

(5) 6-h training window, linear change, and lower outdoor temperature type are used for load prediction. The MAPEs of three typical days are 5.63%, 8.46% and 12.10%, and the 6-h training window is validated to be abundant in information, thus possessing the maximum information entropy.

Acknowledgments

This work was supported by the Opening Funds of the State Key Laboratory of Building Safety and Built Environment (grant number: BSBE-EET2021-01).

References

Al-Musaylh, M. S., Deo, R. C., Li, Y., et al., 2018, Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting: Applied Energy, **217**, 422–439.

Amasyali, K., El-Gohary, N., 2021, Machine learning for occupant-behavior-sensitive cooling energy

consumption prediction in office buildings: Renewable and Sustainable Energy Reviews, **142**, 110714.

Chen, Y. B., Zhang, F. Y., Berardi. U., 2020, Day-ahead prediction of hourly subentry energy consumption in the building sector using pattern recognition algorithms: Energy, **211**, 118530.

Dai, J. H., Hu, H., Hu, Q. H., et al, 2016, Attribute reduction in interval-valued information systems based on information entropies: Frontiers of Information Technology and Electronic Engineering. 17, 919-928.

Datale, L. D., Svetozarevic, B., Heer, P., et al, 2022, Physically Consistent Neural Networks for building thermal modeling: Theory and analysis: Applied Energy, **325**, 119806.

Deng, Z. X., Li, T. R., Deng, D. Y., et al., 2022, Feature selection for label distribution learning using dual-similarity based neighborhood fuzzy entropy: Information Sciences, **615**, 385–404.

Ding, Y., Zhang, Q., Yuan, T. H., et al., 2018, Effect of input variables on cooling load prediction accuracy of an office building: Applied Thermal Engineering, **128**, 225–234.

Dong, Z. X., Liu, J. Y., Liu, B., et al., 2021, Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification: Energy and Buildings, **241**(2), 110929.

Chen, Y. B., Zhang, F. Y., Berardi, U., 2020, Day-ahead prediction of hourly subentry energy consumption in the building sector using pattern recognition algorithms: Energy, **211**, 118530.

Ghosh, S., Yadav, V. K., Mukherjee, V., 2018, Evaluation of cumulative impact of partial shading and aerosols on different PV array topologies through combined Shannon's entropy and DEA: Energy, **144**,765–775.

He, J. L., Qu, L. D., Li. Z. W., 2024, An oscillatory particle swarm optimization feature selection algorithm for hybrid data based on mutual information entropy: Applied Soft Computing, **152**, 111261.

Kang, L. G., Li, H., Wang, Z. C., et al., 2023, Investigation of Energy Consumption via an Equivalent Thermal Resistance-Capacitance Model for a Northern Rural Residence: Energies, **16** (23),7835.

Kang, L. G., Wang, J. Z., Yuan, X. X., et al., 2023, Research on energy management of integrated energy system coupled with organic Rankine cycle and power to gas: Energy Conversion and Management, **287**, 117117.

Ledesma S., Hernández-Pérez, I., Belma, J. M., et al., 2020, Using Artificial Intelligence to Analyze the Thermal Behavior of Building Roofs: Journal of Energy Engineering, **146** (4), 0000677.

Li, H., Li, Y. H., Wang, Z. C., et al., 2022, Integrated building envelope performance evaluation method towards nearly zero energy buildings based on operation data: Energy and Buildings, **268**, 112219.

Li, K. J., Zhang, J. X., Chen, X., et al., 2022, Building's hourly electrical load prediction based on data clustering and ensemble learning strategy: Energy and Buildings, **261**, 111943.

Li, Z. W., Wang, P., Mu, S., 2022, A strategy of improving indoor air temperature prediction in HVAC system based on multivariate transfer entropy: Building and Environment, **219**, 109164.

Lin, P. H., Zhang, L.M., Zuo, J., 2022, Data-driven prediction of building energy consumption using an adaptive multi-model fusion approach: Applied Soft Computing, **129**, 109616.

Liu, H. Y., Yu, J. Q., Dai, J. W., et al., 2023, Hybrid prediction model for cold load in large public buildings based on mean residual feedback and improved SVR: Energy and Buildings, **294**, 113229.

Liu, Z. Y., Yu, J. Q., Feng, C. Y., et al., 2023, A hybrid forecasting method for cooling load in large public buildings based on improved long short term memory: Journal of Building Engineering, **76**, 107238.

Lu, H. F., Cheng, F. F., Ma, X., et al., 2020, Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower: Energy, **203**, 117756.

Lu, S. L., Huo, Y. Q., Su, N., et al., 2023, Energy Consumption Forecasting of Urban Residential Buildings in Cold Regions of China: Journal of Energy Engineering, **149**(2), 4556.

Luo, X. J., Oyedele, L. O., Ajayi, A. O., et al., 2020, Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings: Renewable and Sustainable Energy Reviews, **131**, 109980.

Panapakidis, I. P., Papadopoulos, T. A., Christoforidis, G. C., et al., 2014, Pattern recognition algorithms for electricity load curve analysis of buildings: Energy and Buildings, **73**, 137–145.

Quanga, D. N., Anh, N., Thia, N., et al., 2021, Hybrid online model based multi seasonal decompose for short-term electricity load forecasting using ARIMA and online RNN: Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology, **41**(5), 5639–5652.

Rossini, R., Poccia, S., Candan, K. S., et al., 2019, CA-Smooth: Content Adaptive Smoothing of Time Series Leveraging Locally Salient Temporal Features,

MEDES19: 11th International Conference on Management of Digital EcoSystems.

Tan, M., Chen, J., Cao, Y. J., et al., 2023, A multi-task learning method for multi-energy load forecasting based on synthesis correlation analysis and load participation factor: Applied Energy, **343**, 121177.

Wang, D., Zheng, W. F., Wang, Z., et al., 2023, Comparison of reinforcement learning and model predictive control for building energy system optimization: Applied Thermal Engineering, **228**, 120430.

Wang, S., Sun, Y. H., Zhou, Y., et al., 2019, A New Hybrid Short-Term Interval Forecasting of PV Output Power Based on EEMD-SE-RVM: Energies, **13**, 1–17.

Xiao, J. W., Cao, M. H., Fang, H. L., et al., 2023, Joint load prediction of multiple buildings using multi-task learning with selected-shared-private mechanism: Energy and Buildings, **293**, 113178.

Yang, W. W., Shi, J., Li, S. J., et al., 2022, A combined deep learning load forecasting model of single household resident user considering multi-time scale electricity consumption behavior: Applied Energy, **307**, 118197.

Yi, Y. K., Anis, M., Jang, K., et al., 2023, Application of machine learning (ML) and genetic algorithm (GA) to optimize window wing wall design for natural ventilation: Journal of Building Engineering, **68**, 106218.

Zhang, H. Y., Sun, Q. Q., Dong, K. Z., 2023, Information-theoretic partially labeled heterogeneous feature selection based on neighborhood rough sets: International Journal of Approximate Reasoning, **154**, 200–217.

Zhang, X., Mei, C. L., Chen, D. G., et al., 2016, Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy: Pattern Recognition, **56**, 1–16.

Zhao, Z. W., Huang, D., Li, Z. W., 2023, Outlier detection for partially labeled categorical data based on conditional information entropy: International Journal of Approximate Reasoning, **164**, 109086.

Zhou, Y., Liu, Y. F., Wang, D. J., 2021, Comparison of machine-learning models for predicting short-term building heating load using operational parameters: Energy and buildings, **253**, 111505.

Zhu, Y. Q., Sun, S. X., Liu, C. Y., et al., 2023, PoQ-Consensus Based Private Electricity Consumption Forecasting via Federated Learning: Computer Modeling in Engineering and Sciences, **136**(3), 3285–3297.

Zhichao Wang, male, from Beijing, PhD, director of Environmental Measurement and Control Optimization Research Center, Environmental Energy Institute, China Academy of Building Science, mainly research on green building, regional energy planning and performance optimization of building electromechanical systems. E-mail: wangzc@emcso.com.

Ligai Kang, female, from Shijiazhuang, PhD, Associate Professor, Hebei University of Science and Technology, mainly research on integrated energy systems and building energy conservation. E-mail: ligaikang@hebust.edu.cn.

Nomenclature

Symbols	
C	thermal capacitance
E	information entropy
R	thermal resistance
T	Temperature
cp	specific heat
density	
m	mass flow rate of water
Subscripts and Superscripts	
in	indoor air
out	outdoor air
p	probability distribution
s	supply water
r	return water
w	building opaque envelope
inf	infiltration
m	internal thermal mass
Abbreviations	
AT	Actual temperature
GA	Genetic algorithm
PCA	Principle component analysis
RC	Resistance–capacitance
RF	Random forest
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RMSE	Root-mean square error
R^2	R-squared