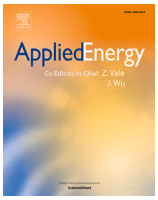
Applied Energy 395 (2025) 126144

Contents lists available at ScienceDirect

Applied Energy

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



Credible demand response capacity evaluation for building HVAC systems based on grey-box models

Siyu Jiang [a, b, c], Hongxun Hui [a, b, c,*], Yonghua Song [a, b, c]

[a] State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao, 999078, China
[b] Department of Electrical and Computer Engineering, University of Macau, Macao, 999078, China
[c] University of Macau Advanced Research Institute in Hengqin, Guangdong, 519031, China

HIGHLIGHTS

- Proposing a credible DR capacity evaluation framework.
- Developing a probabilistic model to estimate HVAC consumption baseline intervals.
- Adapting an equivalent thermal parameter model to quantify DR capacity.
- Capturing the non-linear relationship between uncertainties and DR capacity.

ARTICLE INFO

Keywords:
Demand response
Credible demand response capacity
HVAC systems
Probabilistic estimation

ABSTRACT

Demand response (DR) has been promising in recent years for maintaining the balance between power supply and demand in the power system. Evaluating the DR capacity is significant for the stable operation of the power system and for improving end-user participation. Heating, ventilation, and air conditioning (HVAC) systems account for about 40 % of the total power demand and have enormous potential. However, the power consumption of HVAC systems is subject to various uncertainties, making it difficult to evaluate the range of their DR capacity credibly. To solve this issue, this paper proposes a credible DR capacity evaluation framework based on grey-box models. This framework utilizes a probabilistic model to estimate HVAC consumption baseline intervals and leverages an adaptive equivalent thermal parameter model to derive credible DR capacity intervals. The intervals can reflect the non-linear relationship between multiple uncertainties and the DR capacity. A probabilistic model is proposed by combining a temporal convolutional network and ensemble conformalized quantile regression to estimate the baseline intervals. Additionally, an adaptive equivalent thermal parameter model is adapted to quantify the DR capacity under different regulation levels and different confidence levels. Finally, the effectiveness of the proposed framework in evaluating credible DR capacity is verified using realistic scenarios in Macao.

1. Introduction

The rapid growth of renewable energy sources, characterized by their inherent intermittency and stochasticity, has aggravated the imbalance between power supply and demand within power system [1]. Demand response (DR) guides end-users to change consumption patterns through electricity tariffs or incentive policies to relieve the imbalance [2]. DR offers faster ramp speeds and lower costs through widespread Internet of Things technologies, capturing significant attention in many countries [3]. In America, PJM claims that the 8451 MW DR capacity can be provided by approximately 2 million commercial and industrial customers [4]. Electric Reliability Council of Texas claims load resources with an aggregate capacity of approximately 8700 MW [5].

Building consumption approximately accounts for one-third of the global energy consumption [6]. Under net-zero emissions scenario, the building DR capacity can reach 259 GW [7]. Meanwhile, heating, ventilation, and air conditioning (HVAC) systems constitute approximately 40 % of the total energy usage in buildings by 2050 [8], which implies HVAC systems in buildings have substantial DR capacity. Although building users possess significant DR capacity, they prioritize meeting comfort requirements when participating in DR events. On the one hand,

* Corresponding author at: State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao, 999078, China
Email address: hongxunhui@um.edu.mo (H. Hui).

https://doi.org/10.1016/j.apenergy.2025.126144
Received 15 January 2025; Received in revised form 17 March 2025; Accepted 14 May 2025
Available online 3 June 2025
0306-2619/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

the DR organizers impose penalties on users whose actual DR capacity deviates significantly from the reported capacity. For example, PJM penalizes registrations whose deviations between real response capacity and day-ahead reported capacity are greater than 20 % [9]. In China, users whose deviation exceeds 20 % forfeit their incentive revenue [10]. On the other hand, the grid operators depend on accurate and credible quantification of building flexibility to develop demand response programs, such as peak load management [11]. Therefore, the accurate and credible evaluation of DR capacity before the event is crucial to ensure users can fully maximize their benefits.

DR capacity evaluation refers to quantifying a user's ability to adjust their electricity consumption during a DR event, typically through changes in device usage or shifting energy consumption to different time periods. This concept encompasses four primary capacity types: theoretical capacity, which evaluates the hourly capacity based on user availability, controllability, and storability; technical capacity, which accounts for appliance and operational constraints to refine the theoretical capacity; economic capacity, which considers investment and operational costs; and practical capacity, which evaluates the actual capacity based on user willingness and building characteristics in real DR events [12]. While these categories provide a useful framework for assessment, their evaluation is complicated by several inherent uncertainties, such as variable weather conditions, unpredictable occupant behavior, and fluctuating building operational parameters. These uncertainties often lead to inaccurate evaluation outcomes, which may result in severe penalties for users and undermine both their trust and the efficiency of the overall DR program. Therefore, developing a credible demand response capacity (CDRC) evaluation methodology that accounts for these uncertainties is crucial for enabling building users to participate effectively in DR events. Currently, the predominant evaluation approaches are generally classified into three categories: white-box models (physics-based models), black-box models (data-driven models), and grey-box models (physical data-driven fusion models).

White-box models: White-box models are built on established physical principles and implemented through simulation software or mathematical equations [13,14]. Tang et al. [15] design an optimized dispatch strategy for building DR capacity based on real-time TRNSYS-MATLAB simulation. Song et al. [16] leverage the mathematical high dimensional representative model by identifying different air conditioner parameters to evaluate the capacity. Jung et al. [17] propose an agent-based modeling framework, coupled with EnergyPlus simulations to quantify the capacity while considering the personal comfort. Ran et al. [18] develop a virtual flow-meter model, enabling a fast DR strategy for the HVAC system. Although the white-box models can display the concrete process with clear interpretability, they usually obtain every specific parameter with less transferability for each user.

Black-box models: With the smart sensors being applied widely [19], data can be obtained more easily to evaluate DR capacity for users. Black-box models aim to convey the uncertainty from observed historical data based on data-driven methods, such as neural networks or regression methods under different scenarios. In recent research, black-box models evaluate the DR capacity by predicting customer baseline based on data-driven methods. From the perspective of model generalization, Siddiquee et al. [20] and Yu et al. [21] utilize the unsupervised clustering technique K-means to estimate the different customer DR capacities. From the perspective of accuracy, Harikrishnan et al. [22] develop a heterogeneous ensemble learning method based on XGBoost-ANN to predict residential customer load consumption, which is beneficial for the upper utility to deploy the DR events. Liang et al. [23] design a data-driven method for the HVAC systems participating in incentive-based DR. Zhu et al. [24] construct a multi-layer perceptron layer to quantify DR capacity rapidly and accurately. From the perspective of user comfort, Amer et al. [25] develop a deep reinforcement learning method for home energy management systems to obtain DR capacity while considering user comfort. Zhang et al. [26] propose probability of comfort variation as a user comfort metric for residential buildings integrated with heat pump clusters, which is adopted to quantify DR capacity under different control strategies. Although the aforementioned black-box models have better predictive performance and can be used in different scenarios, when we need to obtain specific CDRC, we still need some interpretable process parameters to enhance physics-informed interpretability.

Grey-box models: Grey-box models are simplified physics-based models by integrating physical principles and observed data, such as equivalent thermal parameter (ETP) model [14]. Hui et al. [27] propose an evaluation method for the aggregated air conditioner capacity based on the ETP model. Considering the customer sensitivity to price, Kong et al. [28] develop a peak shaving response model based on customer psychology mechanism. Furthermore, Song et al. [29] utilize LSTM to predict the HVAC status as the input of ETP model, then introduce the social psychology to obtain theoretical DR capacity. Han et al. [30] evaluate the building DR capacity based on a physical-data fusion method. The physical parameters can be modified by the LSTM model to obtain the capacity. Although present grey-box models can inherit the model interpretability of white-box models and better predictive performance of black-box models, the DR capacity evaluation results are still deterministic and incredible.

Both meteorological and user behavior uncertainties contribute to fluctuations in user energy consumption, leading to highly uncertain DR capacity. This situation complicates credible deterministic evaluation. While previous studies have attempted to enhance the accuracy of deterministic DR capacity evaluations, this capacity is usually represented as a single-point expected value, which fails to capture the various uncertainties inherent in user energy consumption. Furthermore, since deterministic evaluation yields a single-point expected value, it does not account for potential errors. When grid operators make scheduling decisions based on these point evaluations, they can face two possible outcomes: if the evaluation is overly optimistic, it may compromise power system stability, whereas an overly conservative evaluation might result in suboptimal economic performance. In contrast, interval evaluation provides DR capacity estimates as an interval range, effectively quantifying the uncertainty of DR capacity. This approach offers critical, reliable, and comprehensive information to support power system planning, operational control, market transactions, and energy storage configuration and management.

Based on the aforementioned analysis, this paper acknowledges the challenge of performing a CDRC evaluation based on probability, which accounts for multiple uncertainties in meeting market requirements. To address this issue, this paper proposes a CDRC evaluation framework. This framework utilizes a probabilistic model to estimate HVAC consumption baseline intervals and leverages an adaptive equivalent thermal parameter model to derive CDRC probabilistic intervals. In comparison with other studies, this paper provides the following contributions:

1. We propose a CDRC framework to evaluate the day-ahead building credible demand response capacity. Notably, the definition of CDRC is represented as an interval under a specific confidence level, rather than as a determined point. The interval reflects the non-linear relationship between multi-uncertainties and response capacity.
2. An HVAC consumption baseline probabilistic estimation model is proposed. The model utilizes the temporal convolutional network to capture the multi-uncertainties. Then, it combines ensemble conformalized quantile regression to capture the HVAC energy consumption autocorrelation and heteroscedasticity. The proposed model can determine the HVAC system's variabilities in increasing consumption or decreasing consumption at different times and yield informative valid prediction intervals.
3. An adaptive equivalent thermal parameter model is adopted to quantify the DR capacity under different regulation levels at

different confidence levels. The adaptive model can generate the upper and lower bounds of DR capacity, taking into account multi-uncertainties transmitted by the baseline.

The remainder of this paper is organized as follows. Section 2 presents the credible DR capacity framework for building HVAC systems and the adaptive equivalent thermal model. Section 3 formulates the proposed probabilistic baseline estimation model. Section 4 conducts case studies. Section 5 concludes this paper. Section 6 discusses the future work directions.

2. Credible demand response capacity evaluation framework for building HVAC systems

2.1. Evaluation framework for building HVAC systems

The evaluation framework of CDRC is shown in Fig. 1. Buildings are provided with power by transmission lines to meet users' cooling demands. When renewable energy sources' intermittent output or sudden load disturbances induce power system imbalance, the distribution system operations detect the threat and announce the aggregator to prepare for the DR events. The aggregator transmits a DR signal to building users. If the building users choose to participate, they need to report their willingness signal, mainly the CDRC, to the aggregator. Rapid deployment of smart sensors allows buildings to gather abundant local building operation data, occupancy flow data, and real-time environmental data. These data enable local building users to evaluate the CDRC of HVAC systems. In detail, the local building user will evaluate the HVAC system's CDRC range, including the minimum capacity and the maximum capacity that can be provided to the power system during the DR events.

When the DR events start, the HVAC systems of buildings will change their operational power consumption. The difference between the adjusted operational power and baseline power is the realistic DR capacity provided to the power system. However, there are two main challenges:

(i) On the one hand, the baseline power is influenced by chaotic multi-uncertainties, such as occupancy flow, weather conditions, temporal variability, etc. So we utilize a black-box model to estimate the probabilistic credible baseline power considering these uncertainties, which is presented in Section 3.

(ii) On the other hand, the upregulation or downregulation power is the mapping from temperature settings, which is limited by the users' desired comfort level. The mapping relationship in the overall heat transfer process is influenced by the operational parameters of HVAC systems and the characteristic parameters of buildings, which cannot be measured by smart sensors. According to research conducted by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers [16,27], users' comfort levels differ due to occupancy flow, weather conditions, and temporal variability. Hence, we set different temperature settings according to occupancy flow and ambient temperature. Then an adaptive ETP model based on probabilistic credible baseline power is designed in Section 2.2 so that the downregulation or upregulation power can be obtained.

2.2. Adaptive equivalent thermal parameter model for HVAC systems

The classical equivalent thermal parameter (ETP) model for HVAC systems installed in buildings can be described as follows [16]:

$$C_{\text{in}} \frac{\partial \theta_{\text{in}}(t)}{\partial t} = \frac{\theta_{\text{out}}(t) - \theta_{\text{in}}(t)}{R_{\text{in}}} - Q_{\text{HVAC}}(t), t \in \mathcal{T}, \tag{1}$$

where C_{in} denotes the indoor thermal capacity; R_{in} denotes the indoor thermal resistance; $\theta_{\text{in}}(t)$ and $\theta_{\text{out}}(t)$ denote the indoor and outdoor temperature at time t, respectively; $Q_{\text{HVAC}}(t)$ denotes the HVAC systems cooling capacity.

The transfer process from the cooling capacity of HVAC systems to power consumption can be presented as:

$$P(t) = \frac{Q_{\text{HVAC}}(t)}{\eta}, t \in \mathcal{T}, \tag{2}$$

where $P(t)$ is the HVAC systems consumption at time t; η is the transfer efficiency of HVAC systems. When η is larger, the power consumption is lower for the same $Q_{\text{HVAC}}(t)$.

When the HVAC systems are at stable status, the θ_{in} will reach the initial setting temperature $\theta_{\text{set}}(t_0)$. Accordingly, the indoor thermal resistance can be derived from Eqs. (1) and (2) as:

$$R_{\text{in},\alpha} = \frac{\theta_{\text{out}}(t_0) - \theta_{\text{set}}(t_0)}{\eta P(t_0)}, t_0 \in \mathcal{T}, \tag{3}$$

where $P(t_0)$ can be replaced by baseline lower bound (BLB) or baseline upper bound (BUB) under the given confidence level α. BLB and BUB can be obtained from the credible probabilistic baseline estimation model in Section 3.

2.3. CDRC of building HVAC systems

Definition: CDRC can be defined as the guaranteed response capacity at arbitrary time t under a specific confidence level. The guaranteed capacity is not less than the minimum $CDRC_{\min,\alpha}$, and not larger than the maximum $CDRC_{\max,\alpha}$ under the given confidence level α.

$CDRC_{\max,\alpha}$, and $CDRC_{\min,\alpha}$ can be calculated by BLB and BUB under the given confidence level α based on the adaptive ETP model. When estimating BLB and BUB, we take weather conditions, occupancy flow, and temporal variability into consideration, so the corresponding CDRC result contains uncertainties. Additionally, we can obtain CDRC under different confidence intervals.

When building users agree to participate in DR events, they should change $\theta_{\text{set}}(t_0)$ to $\theta_{\text{set}}(t_1)$ during the response period, at the same time, the comfortable level should be maintained. We set $\Delta\theta$ as the temperature deviation range according to [16,27]. If the temperature is in the range of $[\theta_{\text{set}}(t_0)-\Delta\theta, \theta_{\text{set}}(t_0)+\Delta\theta]$, we can assume the temperature meets user comfort requirements. And the power consumption with the reset temperature $\theta_{\text{set}}(t_1)$ is:

$$P(t_{1,\bar{\alpha}}) = \frac{\theta_{\text{out}}(t_1) - \theta_{\text{set}}(t_1)}{\eta R_{\text{in},\bar{\alpha}}}, t_1 \in \mathcal{T}, \tag{4}$$

$$P(t_{1,\underline{\alpha}}) = \frac{\theta_{\text{out}}(t_1) - \theta_{\text{set}}(t_1)}{\eta R_{\text{in},\underline{\alpha}}}, t_1 \in \mathcal{T}, \tag{5}$$

where $P(t_{1,\bar{\alpha}})$ represents adjusted operation power upper bound under confidence level α; $P(t_{1,\underline{\alpha}})$ represents adjusted operation power lower bound under confidence level α.

Then, the CDRC can be calculated as follows:

$$\begin{aligned} CDRC_{\max,\alpha} &= \left|P(t_{0,\bar{\alpha}}) - P(t_{1,\bar{\alpha}})\right| \\ &= P(t_{0,\bar{\alpha}}) \left|\left(1 - \frac{\theta_{\text{out}}(t_1) - \theta_{\text{set}}(t_1)}{\theta_{\text{out}}(t_0) - \theta_{\text{set}}(t_0)}\right)\right|, t_0 \in \mathcal{T}, t_1 \in \mathcal{T}, t_0 \neq t_1, \end{aligned} \tag{6}$$

$$\begin{aligned} CDRC_{\min,\alpha} &= \left|P(t_{0,\underline{\alpha}}) - P(t_{1,\underline{\alpha}})\right| \\ &= P(t_{0,\underline{\alpha}}) \left|\left(1 - \frac{\theta_{\text{out}}(t_1) - \theta_{\text{set}}(t_1)}{\theta_{\text{out}}(t_0) - \theta_{\text{set}}(t_0)}\right)\right|, t_0 \in \mathcal{T}, t_1 \in \mathcal{T}, t_0 \neq t_1, \end{aligned} \tag{7}$$

where $P(t_{0,\bar{\alpha}})$ and $P(t_{1,\bar{\alpha}})$ represent BUB power before and after adjusting the setting temperature, respectively; $P(t_{0,\underline{\alpha}})$ and $P(t_{1,\underline{\alpha}})$ represent BLB power before and after adjusting the setting temperature, respectively.

In summary, when a confidence level α is specified, the baseline interval is constructed based on a probabilistic model (i.e., the proposed

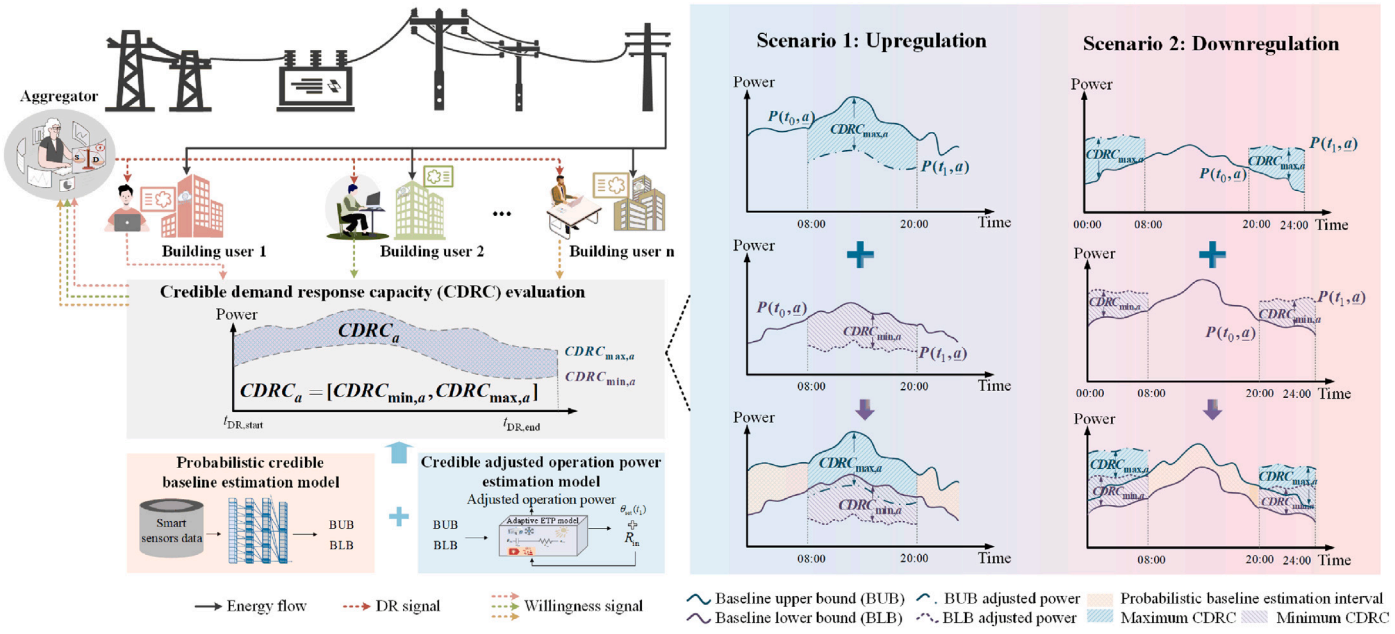


Fig. 1. The framework of CDRC evaluation.

method in this paper), where the interval is defined by the BUB and BLB. Then, the BUB power $P(t_{0,\bar{\alpha}})$ and BLB power $P(t_{0,\underline{\alpha}})$ are used as inputs for the adaptive ETP model to evaluate user response effectiveness. Specifically, taking the upregulation scenario as an example, when the HVAC's setting temperature is increased from $\theta_{\text{set}}(t_0)$ to $\theta_{\text{set}}(t_1)$, its operating power will decrease correspondingly from $P(t_{0,\bar{\alpha}})$ to $P(t_{1,\bar{\alpha}})$ for BUB, and from $P(t_{0,\underline{\alpha}})$ to $P(t_{1,\underline{\alpha}})$ for BLB, respectively. As shown in Fig. 1, the difference of BUB between $P(t_{0,\bar{\alpha}})$ and $P(t_{1,\bar{\alpha}})$ defines the maximum value of the CDRC at confidence level α, denoted as $CDRC_{\max,\alpha}$. Similarly, the difference of BLB between $P(t_{0,\underline{\alpha}})$ and $P(t_{1,\underline{\alpha}})$ defines the minimum value of CDRC at confidence level α, denoted as $CDRC_{\min,\alpha}$.

The thermal dynamics of HVAC systems are predominantly governed by the interplay between thermal resistance and thermal capacitance [31]. Thermal resistance dictates the rate of heat dissipation during transfer, while thermal capacitance determines the extent of indoor temperature variation as the system absorbs or releases heat. These properties are inherently reflected in the temperature dynamics driven by load fluctuations, where load variations effectively act as a dynamic representation of changes in thermal resistance and capacitance. As a result, their primary effects can be captured through load responses without the need for explicit modeling of their variations. This approach not only reduces model complexity but also ensures simulation accuracy, enabling more efficient applications in reliable capacity evaluation.

3. Credible probabilistic baseline estimation model

The power consumption of HVAC systems is influenced by both multi-time-scale historical data and external environmental factors, including weather variations, occupancy flow uncertainties, and temporal fluctuations. The complexity of these data leads to non-linear, non-stationary, and heteroscedastic relationships in the input-output mapping, while the analysis of multi-source, and heterogeneous data poses significant challenges. To address this issue, this section presents a model that combines a temporal convolutional network and a conformalized quantile regression model with the ensemble strategy to quantify uncertainties by obtaining confidence intervals. The model integrates temporal convolutional network for time series estimation and ensemble conformalized quantile regression to provide adaptive and valid intervals without requiring the data to be i.i.d. It assesses the confidence of predictions by calculating the conformity between a new predicted value and existing data, then constructs an interval that ensures the true value is contained within it at a specified confidence level.

3.1. Temporal convolutional network

Temporal Convolutional Network (TCN) is an effective model for time series estimation problems [32]. TCN consists of causal convolution, dilated convolution, and residual connections, as depicted in Fig. 2. Causal convolution addresses the data leakage issue inherent in traditional one-dimensional convolutions. Dilated convolutions expand the receptive field without altering the size of the output features, enabling TCN to capture dependencies at greater distances. The convolutional structure provides TCN with the advantage of parallel computation compared to long short-term memory (LSTM) architectures. Residual connections help TCN maintain long-term dependencies in time series data and enhance model learning capabilities by considering cross-layer information transfer. It should be noted that while the input sequence length of TCN can be arbitrary, the length of each convolutional layer must be strictly consistent with the length of the input layer.

Given an input series data to predict the corresponding outcome, only using observations made prior to that time t, a sequential model network usually refers to a function that produces the following mappings:

$$\hat{y}_0, \ldots, \hat{y}_t = f\left(x_0, \ldots, x_t\right), x \in \mathcal{X}, y \in \mathcal{Y}, \tag{8}$$

where f denotes the TCN structure; $\hat{y}$ is the output of TCN; x is the input series data.

In Fig. 2, the dilated convolutions at position t can be expressed as:

$$\mathcal{O}_{\text{con}}(x_t) = \sum_{k=0}^{K-1} f_k x_{t-dk}, x_t \in \mathcal{X}, \tag{9}$$

where f_k is the filter also named convolution kernel, which can represent the information decay rate; d represents the dilation rate, which widens the receptive field; x_{t-dk} represents dk series before x_t.

The receptive field depends on the depth of TCN and it can be depicted as:

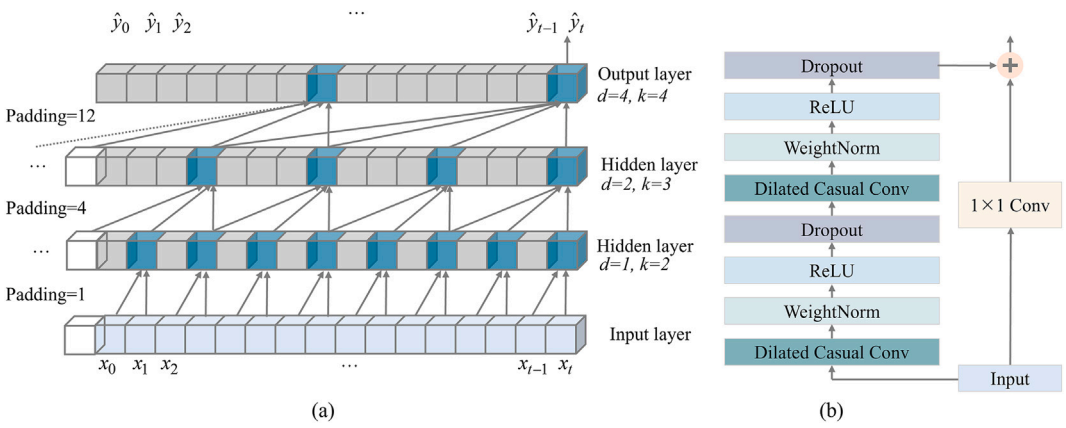


Fig. 2. The diagram of TCN structure: (a) The diagram of TCN with three layers; (b) The residual block structure of the first layer in TCN.

$$R_{\text{fie}} = 1 + 2 \cdot (K_i) \cdot i \cdot \sum_{i=0}^{2} d_i, \tag{10}$$

where d_i is equal to a^i; i is the layer number of the present layer; a is the dilation rate; K_i is the kernel size of the filter in i -th layer; especially, the number of the input layer is 0.

When i is larger than 1, the kernel size can capture the features of x/d, rather than the total input data. Kernel size keeps the time series data characteristics without increasing computational complexity.

In Eq. (10), we can see that the receptive field is determined by the depth of TCN, the dilation rate, and the kernel size. As the value of i increases, the disappearing gradients problem becomes more severe. To address this issue, a residual block is used, as shown in Fig. 2. The input data adds to the residual output and then becomes the input for the next layer. Notably, the output layer in Fig. 2 represents the output of the TCN structure rather than the estimation result. It will be the input to the fully connected layer behind it.

3.2. Conformalized quantile regression

Conformalized Quantile Regression (CQR) [33] combines probabilistic estimation methods: conformal prediction (CP) and quantile regression (QR) to obtain prediction intervals (PIs). PIs provide the future point x_i with an acceptable range of values consisting of lower bound and upper bound under an explicit confidence level α. under a specific confidence level, CP creates valid PIs which can be expressed as:

$$\begin{aligned}\hat{C}_{\alpha,\text{CP}}(x_i) = \big[&\hat{y}(x_i) - Q_{(1-\alpha)}(\mathcal{R}, \mathcal{I}_{\text{cal}}),\\ &\hat{y}(x_i) + Q_{(1-\alpha)}(\mathcal{R}, \mathcal{I}_{\text{cal}})\big], (x_i, y_i) \in \mathcal{I}_{\text{test}},\end{aligned} \tag{11}$$

$$\mathcal{R} = \left\{ r_i \mid r_i = |y_i - \hat{y}(x_i)| \right\}, (x_i, y_i) \in \mathcal{I}_{\text{cal}}, \tag{12}$$

where $\mathcal{I}_{\text{cal}}$ and $\mathcal{R}$ represent the calibration set and the residual set, respectively; $\hat{y}(x_i)$ is the estimation value; $Q_{(1-\alpha)}(\mathcal{R}, \mathcal{I}_{\text{cal}})$ is the $(1-\alpha)$ -th quantile value of residual set; x_i in test set $\mathcal{I}_{\text{test}}$ can be regarded as the future input data.

However, PIs generated by CP are almost fixed and CP assumes that the input data is independent and identically distributed (i.i.d.), which makes it unsuitable for autocorrelated and heteroscedastic HVAC systems' consumption.

In contrast, QR yields varying PIs due to the different input points as Eq. (13) shows. Therefore, QR can capture features of seasonal and heteroscedastic HVAC baseline.

$$\hat{C}_{\alpha,\text{QR}}(x_i) = \left[\hat{q}_{\underline{\alpha}}(x_i), \hat{q}_{\overline{\alpha}}(x_i)\right], \tag{13}$$

where $\underline{\alpha}$ is $\alpha/2$, $\overline{\alpha}$ is $1-\underline{\alpha}$; $\hat{q}_{\underline{\alpha}}$ and $\hat{q}_{\overline{\alpha}}$ are the empirical conditional distribution functions of $\underline{\alpha}$ and $\overline{\alpha}$, respectively.

However, the finite sample points cause the actual PIs to have a significant bias from the presumed confidence level, making the intervals invalid. CQR incorporates the advantages of CP and QR to construct valid and adaptive PIs. Specifically, CP yields valid PIs, which are constructed using the conditional mean estimates of the response variable. QR generates the adaptive PIs according to the input data. The expression of CQR PIs is as follows:

$$\begin{aligned}\hat{C}_{\alpha,\text{CQR}}(x_i) = \big[&\hat{q}_{\underline{\alpha}}(x_i) - Q_{(1-\alpha)}(\mathcal{S}, \mathcal{I}_{\text{cal}}),\\ &\hat{q}_{\overline{\alpha}}(x_i) + Q_{(1-\alpha)}(\mathcal{S}, \mathcal{I}_{\text{cal}})\big], (x_i, y_i) \in \mathcal{I}_{\text{test}},\end{aligned} \tag{14}$$

$$\mathcal{S} = \left\{ s_i | s_i = \max\left(\hat{q}_{\underline{\alpha}}(x_i) - y_i, y_i - \hat{q}_{\overline{\alpha}}(x_i)\right) \right\}, (x_i, y_i) \in \mathcal{I}_{\text{cal}}, \tag{15}$$

where $\mathcal{S}$ is the residual set improved from QR; $Q_{(1-\alpha)}$ is the $(1-\alpha)$ -th quantile value of set $\mathcal{S}$.

The empirical conditional distribution functions for $\hat{q}_\alpha(x)$ can be calculated by minimizing the pinball loss, which is the loss function for TCN in the Section 3.3:

$$\hat{q}_\alpha(x) = f(x, \hat{\theta}_\alpha), \tag{16}$$

$$\hat{\theta}_\alpha = \arg\min_{\theta} \sum_{x_i \in \mathcal{I}_{\text{train}}} Loss_{\alpha,\text{pin}}(x_i), \tag{17}$$

$$Loss_{\alpha,\text{pin}}(x_i) = \begin{cases} (1-\alpha) \cdot \left(\hat{q}_\alpha(x_i) - y_i\right), \hat{q}_\alpha(x_i) \geq y_i \\ (-\alpha) \cdot \left(\hat{q}_\alpha(x_i) - y_i\right), \hat{q}_\alpha(x_i) < y_i \end{cases}, (x_i, y_i) \in \mathcal{I}_{\text{train}}, \tag{18}$$

where f is the regression function based on quantile regression; $Loss_{\alpha,\text{pin}}$ is the total pinball loss under confidence level α for $(x_i, y_i) \in \mathcal{I}_{\text{train}}$.

3.3. Conformalized ensemble temporal convolutional quantile regression network

Although CQR can produce more valid PIs than QR, it still requires the input data to meet the i.i.d. assumption. Ensemble conformalized quantile regression (ECQR) [34] adapts the assumption, which assumes that the error of time series is stationary and mixing. Under this assumption, the HVAC energy consumption is i.i.d. ECQR heritages the advantage of homogeneous ensemble learning strategy, it yields E subsets to fit different subset learners. In detail, ECQR yields aggregated quantile results with upper bound and lower bound of PIs. The aggregated PIs yielded from ensemble learners are conformalized as follows:

Algorithm 1: Conformalized ensemble temporal convolutional quantile regression network (CETCQRN).

Input:
Dataset $\boldsymbol{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;TCN initial hyperparameters range; Confidence level $\alpha, \alpha \in (0, 1)$;Ensemble number E; Sequence length s
Output: The HVAC systems Consumption BLB $P(t_{0,\underline{\alpha}})$, BUB $P(t_{0,\overline{\alpha}})$.

Split data set $\boldsymbol{D}$ into $\mathcal{I}_{\text{train}}$, $\mathcal{I}_{\text{val}}$ and $\mathcal{I}_{\text{test}}$;
Calculate length of ensemble subset : $Len(\mathcal{I}_{\text{sub}}) = Len(\mathcal{I}_{\text{train}})/E$;
for $e = 1, \ldots, E$ **do**
  Split subset $\mathcal{I}_{\text{train},e} = \{\mathcal{I}_{\text{train}}\}_{Len(\mathcal{I}_{\text{train}})\cdot(e-1)}^{Len(\mathcal{I}_{\text{train}})\cdot e}$;
  Train subset TCN estimator with $\mathcal{I}_{\text{train},e}$ and obtain best regression model with $\mathcal{I}_{\text{val},e}$;
  if $Loss_{\alpha,pin}(x_i)$ *does not change in eval process within latest 10 epoch* **then**
    Stop the training process and update hyperparameters for TCN;
  else
    Reach maximum epoch then stop;
  end
  Compute PIs utilizing updated quantile TCN subset estimator as in Eq. (13);
end
Set $\underline{S} = \{\}, \overline{S} = \{\}$;
for $(x_i, y_i) \in \mathcal{I}_{train}$ **do**
  for $e = 1, \ldots, E$ **do**
    $\mathcal{I}_{\text{agg},e} = \complement_{\mathcal{I}_{\text{train}}} \mathcal{I}_{\text{train},e}$;
    Compute PIs for $\mathcal{I}_{agg,e}$ with subset estimator trained by subset $\mathcal{I}_{train,e}$
    Compute $\underline{S}, \overline{S}$ as in Eqs. (20) and (21);
  end
end
Calculate mean value of $[\hat{q}^e_{\underline{\alpha}}(x_i), \hat{q}^e_{\overline{\alpha}}(x_i)]$;
Calculate $Q_{(1-\alpha)}(\underline{S})$ and $Q_{(1-\alpha)}(\overline{S})$;
for $(x_i, y_i) \in \mathcal{I}_{test}$ **do**
  Compute $[\hat{q}^e_{\underline{\alpha}}(x_i), \hat{q}^e_{\overline{\alpha}}(x_i)]$ with each subset TCN estimator Update the latest s residuals;
  if *i-N=0 mod s* **then**
    for *j=i-s,…,i-1* **do**
      Compute $\underline{S}_j$ and $\overline{S}_j$:
      Intslashdollar$\underline{s}_j = \hat{q}_{\underline{\alpha}}(x_j) - y_j, \overline{s}_j = \hat{q}_{\overline{\alpha}}(x_j) - y_j$;
      Update residual set and reset index of $\underline{S}$ and $\overline{S}$ as follows:
      $\underline{S} = (\complement_{\underline{S}}\{\underline{s}_j\}_{j-1}^{j-N}) \bigcup \underline{S}_j$,
      $\overline{S} = (\complement_{\overline{S}}\{\overline{s}_j\}_{j-1}^{j-N}) \bigcup \overline{S}_j$;
    end
  end
  Calculate $\hat{C}_\alpha(x_i)$, as shown in Eq. (19);
  Return $BLB(x_i) = \min(\hat{C}_\alpha(x_i))$, $BUB(x_i) = \max(\hat{C}_\alpha(x_i))$.
end

$$\hat{C}_\alpha(x_i) = \left[\hat{q}_{\underline{\alpha}}(x_i) - Q_{(1-\alpha)}(\underline{S}),\ \ \hat{q}_{\overline{\alpha}}(x_i) + Q_{(1-\alpha)}(\overline{S})\right], (x_i, y_i) \in \mathcal{I}_{\text{test}}, \tag{19}$$

$$\underline{S} = \left\{\underline{s}_{\underline{i}} | \underline{s}_{\underline{i}} = \hat{q}_{\underline{\alpha}}(x_i) - y_i\right\}_{i=t-1}^{i=t-N}, (x_i, y_i) \in \mathcal{I}_{\text{cal}}, \tag{20}$$

$$\overline{S} = \left\{s_{\overline{i}} | s_{\overline{i}} = \hat{q}_{\overline{\alpha}}(x_i) - y_i\right\}_{i=t-1}^{i=t-N}, (x_i, y_i) \in \mathcal{I}_{\text{cal}}, \tag{21}$$

where $\underline{S}$ and $\overline{S}$ denote the latest N residuals before time t, respectively.

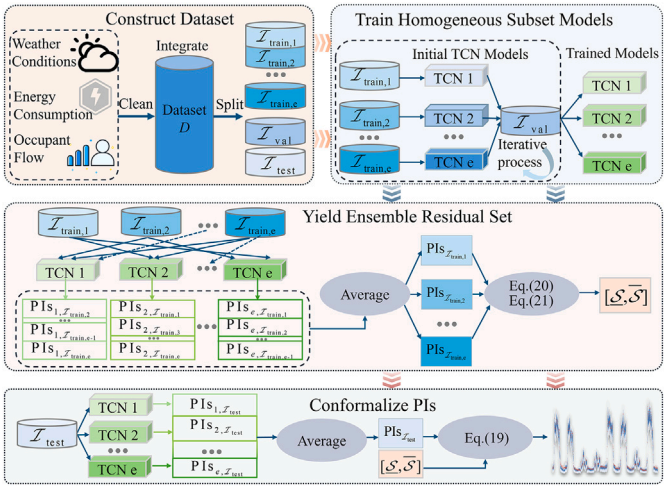


Fig. 3. The training process of CETQCRN.

Compared with single S in Eq. (15), $\underline{S}$ and $\overline{S}$ prevent the PI coverage error from being asymmetrically spread over the left and right tails, respectively. They independently control the coverage of the upper and lower quantile functions to obtain a more valid coverage guarantee.

As a result, this section combines TCN and ECQR to construct conformalized ensemble temporal convolutional quantile regression network (CETCQRN) to estimate the HVAC system's baseline, as shown in Algorithm 1.

The entire training process is presented in Fig. 3.

3.3.1. Construct dataset

The dataset is constructed by collecting historical energy consumption data from HVAC systems, along with additional energy usage information such as lighting and elevators. Furthermore, external factors, including weather conditions and occupant flow data, are incorporated. The collected data undergoes preprocessing, including outlier removal, missing value imputation, time scale alignment, and data integration to form the final dataset $\boldsymbol{D}$. Subsequently, the dataset is partitioned into training set $\mathcal{I}_{\text{train}}$, validation set $\mathcal{I}_{\text{val}}$, and test set $\mathcal{I}_{\text{test}}$. To facilitate ensemble training, the training set is further divided into E subsets, denoted as $\mathcal{I}_{\text{train},1}, \mathcal{I}_{\text{train},2}, \ldots, \mathcal{I}_{\text{train},e}$.

3.3.2. Train homogeneous subset models

Each subset, $\mathcal{I}_{\text{train},1}, \mathcal{I}_{\text{train},2}, \ldots, \mathcal{I}_{\text{train},e}$, is used to train a TCN model, as illustrated in Fig. 2, with all models initialized using identical hyperparameters. The subsets are fed into their respective TCN models (TCN 1, TCN 2, …, TCN e) for parallel training. Then, $\mathcal{I}_{\text{val}}$ is used to validate the model by evaluating the pinball loss, as defined in Eq. (18). If the loss falls within an acceptable range, the trained model is finalized. Otherwise, the process enters an iterative optimization stage, where further refinements are applied until the maximum number of iterations is reached, yielding the final trained model.

In the homogeneous subset models' training process, the stochastic search method is adopted to find the best hyperparameters by minimizing the pinball loss. This enables us to update the corresponding homogeneous subset model. In this stage, the E trained homogeneous subset models are generated.

3.3.3. Yield ensemble residual set

E trained homogeneous subset models are leveraged to generate the ensemble residuals set. Specifically, the input $\mathcal{I}_{\text{agg},e}$ for the e-th trained homogeneous subset model is the relative complement $\complement_{\mathcal{I}_{\text{train}}} \mathcal{I}_{\text{train},e}$ of $\mathcal{I}_{\text{train},e}$ in $\mathcal{I}_{\text{train}}$. For example, the input $\mathcal{I}_{\text{agg},1}$ for the trained TCN 1,

which consists of $\mathcal{I}_{\text{train},2}$, $\mathcal{I}_{\text{train},3}$, …, $\mathcal{I}_{\text{train,e}}$, excluding $\mathcal{I}_{\text{train},1}$. This leave-one-out approach enhances the model's generalization properties. Then, each subset generates its corresponding PIs from *e-1* models, based on Eq. (14). Furthermore, each subset averages the *e-1* PIs to generate its ensemble PIs. For example, the ensemble PIs for $\mathcal{I}_{\text{train},1}$ are denoted as $\text{PIs}_{\mathcal{I}_{\text{train},1}}$. Finally, the E ensemble PIs generate the ensemble residuals set $\underline{S}$ and $\overline{S}$ based on Eqs. (20) and (21).

3.3.4. Conformalize PIs

In this stage, the test set $\mathcal{I}_{\text{test}}$ is used as the input for the E trained homogeneous subset models. The corresponding results e PIs, denoted as $\text{PIs}_{\mathcal{I}_{\text{train},1}}$, $\text{PIs}_{\mathcal{I}_{\text{train},2}}$, … , $\text{PIs}_{\mathcal{I}_{\text{train,e}}}$ are obtained. The E PIs are then averaged to obtain $\text{PIs}_{\mathcal{I}_{\text{test}}}$. The latest N predicted points $(s_{j-1}, s_{j-2}, \ldots, s_{j-N})$ are used to update the residual sets $\underline{S}$ and $\overline{S}$, replacing the original initial values. Finally, the HVAC system's consumption BLB and BUB can be estimated according to Eq. (19).

4. Case studies

4.1. Simulation setup

4.1.1. Data description

This paper presents case studies using realistic HVAC energy consumption, occupancy flow, and numerical weather forecast data collected from a specific office building in Macao. The HVAC energy consumption baseline data covers the period from May to September 2021, with a time granularity of 15 min. We set the data from May to August into 4 subsets, with the validation set spanning from 1 st September to 15th September and the test set spanning from 16th September to 30th September. The HVAC consumption baseline fluctuations are caused by the data heteroscedasticity and autocorrelation, which means that the interference term at each observation point is generated from an independent but non-identical distribution. In Fig. 4, the values under the green dashed line represent the variances at different times. Variations in the variance of baseline power can indicate the presence of heteroscedasticity. The values indicated outside the green dashed line represent the covariance between any two distinct time points. The non-zero covariance between the baseline power at the same time on different days can indicate the presence of autocorrelation. It was found that from 08:00 to 20:00, the energy consumption exhibits significant autocorrelation and heteroscedasticity, which can be attributed to the flow of occupancy, weather conditions, and temporal variability. The significant heteroscedasticity and autocorrelation observed in HVAC energy consumption indicate that the data does not conform to the assumption of i.i.d.

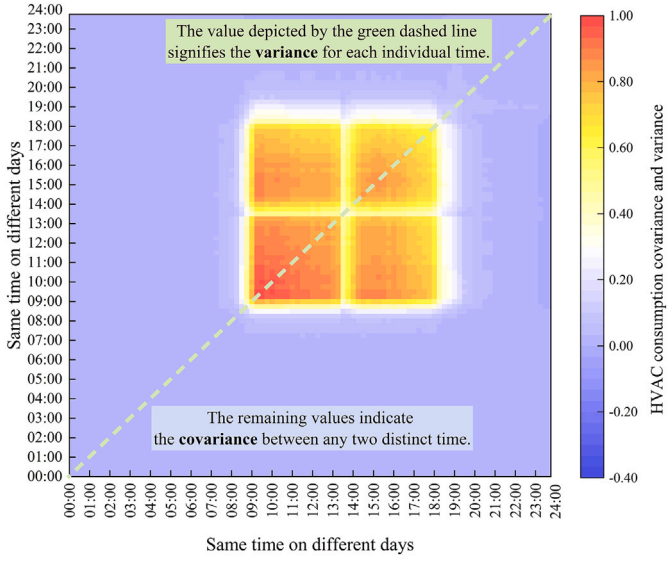


Fig. 4. Covariance and variance of building HVAC systems' baseline power.

The parameters of the HVAC systems and buildings are based on the Macao Standards (CSUS/GBC 07-2015) and the Chinese National Standards (GB/T 50378). The details are as follows: the set temperature θ_{set} is distributed $24 \sim 27\,^{\circ}\text{C}$ to meet various users' comfort requirements; the ambient temperature is based on real-time outdoor data collected by the meteorological station at the University of Macau; the maximum indoor temperature deviation is 1 °C; Taking into account that Macau is situated in the South Subtropical zone, where the HVAC systems coefficient of performance η is distributed among $4.4 \sim 5.9$. Moreover, the thermal capacity C_i and resistance R_i of the room are obtained as:

$$\begin{cases} C_{in} = c_{Air}\rho_{Air}V = c_{Air}\rho_{Air}Ah, & \forall i \in \mathcal{I}, \\ R_{in} = \frac{1}{A \cdot U} = \left[U\left(2A + 4h\sqrt{A}\right)\right]^{-1}, & \forall i \in \mathcal{I}, \end{cases} \tag{22}$$

where $c_{Air} = 1.005\,\text{kJ/(kg}\cdot^{\circ}\text{C)}$ and $\rho_{Air} = 1.205\,\text{kg/m}^3$ are the specific heat capacity and density of air, respectively; $U = 3.6\,\text{W/(m}^2\cdot^{\circ}\text{C)}$ denotes the heat transfer coefficient; V, A and h represent the volume, height, and surface area of the building, respectively. The area A of the building is assumed to be a square.

4.1.2. Evaluation metrics

Considering the effectiveness of proposed probabilistic methods, we use the following metrics to measure the results.

Prediction Interval Coverage Probability (PICP) aims to assess the effectiveness of PIs by calculating the percentage of predicted values $\hat{y}_i$ that fall within the prediction interval:

$$PICP = \frac{1}{n}\sum_{i=1}^{n} k_i, \tag{23}$$

$$k_i = \begin{cases} 1, \hat{y}_i \in \hat{C}_\alpha(x_i) \\ 0, \hat{y}_i \notin \hat{C}_\alpha(x_i) \end{cases}, \tag{24}$$

where n represents the number of estimation points and k_i represents whether the point real observation x_i falls into the estimation PIs.

Prediction Intervals Normalized Average Width (PINAW) offers insights into the precision of PIs by measuring average width relative to the range of observed values.

$$PINAW = \frac{1}{n(Y_{\max} - Y_{\min})}\sum_{i=1}^{n}(BUB(x_i) - BLB(x_i)), \tag{25}$$

where $Y = \{y_1, y_2, \ldots, y_i\}$.

To prevent the situation where we have to sacrifice the PINAW in order to obtain a higher PICP, we adapt the modified coverage width-based criterion (CWC) from [34] to provide a summary of the quality of the PIs. When the PICP is the same, the PINAW is lower, the CWC is bigger, and the yielded PIs are better.

$$CWC = (1 - PINAW)e^{-\lambda(PICP-(1-\alpha))^2}, \tag{26}$$

where λ is the penalty coefficient to penalize the points that fail to fall in the PIs; Based on the specific data characteristics and recent study [35,36], λ is 5 in this paper.

4.1.3. Hyperparameters setting

The TCN parameters are: learning rate is 0.001; batch size is 120; epochs are 200; the number of layers is 3; kernel size is [2,3,4]; number of channels is [64,128,64]; dropout is 0.2; optimizer is Adam. In Algorithm 1, ensemble number E is 4, N is 4, and s is 24, respectively.

Table 1
Abbreviations and full names of baseline comparison methods.

Abbreviation	Full name
ETCQRN	Ensemble temporal convolutional quantile regression network
CELSTMQR	Conformalized ensemble long short-term memory quantile regression
ELSTMQR	Ensemble long short-term memory quantile regression
CEXGBQR	Conformalized ensemble extreme gradient boosting quantile regression
EXGBQR	Ensemble extreme gradient boosting quantile regression
CELGBQR	Conformalized ensemble light gradient boosting quantile regression
ELGBQR	Ensemble light gradient boosting quantile regression
CEGBQRT	Conformalized ensemble gradient boosting quantile regression tree
EGBQRT	Ensemble gradient boosting quantile regression tree
CTCQRN	Conformalized temporal convolutional quantile regression network
TCQRN	Temporal convolutional quantile regression network

4.1.4. Environment setup

All the experiments are implemented using Ptyorch on a desktop with an Intel(R) Core(TM) i7-12700 CPU and NVIDIA GeForce RTX 3060TI GPU, with 64GB of RAM on a Windows 11 platform.

4.2. Estimation results of credible probabilistic baseline

In order to comprehensively demonstrate the proposed method, we select different benchmarks for both regression methods and PI yield methods.

4.2.1. Regression methods

We compare the TCN with four commonly used benchmarking regression methods to show its advantages: long short-term memory (LSTM), the light gradient boosting machine (LightGBM), the extreme gradient boosting (XGBoost), and the gradient boosting regression tree (GBRT).

4.2.2. PIs yield methods

We compare the ensemble conformalized quantile regression with three benchmarking PI yield methods to show its advantage: ensemble quantile regression, conformalized quantile regression, and quantile regression.

We utilize regression methods and PI yield methods to create eleven baseline comparisons. They are: ensemble temporal convolutional quantile regression network (ETCQRN), conformalized ensemble long short-term memory quantile regression (CELSTMQR), ensemble long short-term memory quantile regression (ELSTMQR), conformalized ensemble extreme gradient boosting quantile regression (CEXGBQR), ensemble extreme gradient boosting quantile regression (EXGBQR), conformalized ensemble light gradient boosting quantile regression (CELGBQR), ensemble light gradient boosting quantile regression (ELGBQR), conformalized ensemble gradient boosting quantile regression tree (CEGBQRT), ensemble gradient boosting quantile regression tree (EGBQRT), conformalized temporal convolutional quantile regression network (CTCQRN), and temporal convolutional quantile regression network (TCQRN). For ease of reading, the methods mentioned above are summarized in Table 1.

We can assess the efficacy of the TCN model by comparing the results from the following four methods, CELSTMQR, CEXGBQR, CELGBQR, and CEGBQRT. Furthermore, to showcase the PIs yield model, we compare the results from the following three methods, CETCQRN, CTCQRN, and TCQRN. Lastly, we comprehensively demonstrate the effectiveness of the PIs yield methods by comparing the results from different regression methods with ECQR and regression methods with CQR. We conduct case studies using different methods at five different prediction interval nominal confidence (PINC) levels, which are 95 % PINC, 90 % PINC, 80 % PINC, 70 % PINC, and 50 % PINC. PINC is the confidence level α.

Table 2 displays all the estimation results. It can be observed that the proposed method performs best in terms of CWC at different PINCs. Generally, when PINC is lower, the CWC is higher. In comparison to ETCQRN, the proposed method consistently exhibits lower PICP and PINAW, indicating that ETCQRN sacrifices PI width to ensure more points fall within the PIs. The results from different regression methods show similar patterns. In other words, the ensemble learning strategy facilitates the model in learning from data deeply.

We can find that in terms of CWC the proposed method is always over 0.8. Whatever at which PINC, the proposed method can keep the biggest CWC. And it increased by 10.485 % than CELSTMQR at different PINCs on average. The CELSTMQR performs more stable than CEXGBQR and CELGBQR at different PINCs. It is evident from Fig. 5 that as the PINC decreases, the proposed method CWC increases rapidly. From a 50 % PINC to a 95 % PINC, it is increased by 17.412 %, while the CTCQRN only increases by 4.530 %. This phenomenon demonstrates that the proposed method can produce reliable prediction intervals even when the PICP is low.

In order to display the different methods' results, we have chosen Sunday, September 19th, as the holiday situation depicted in Fig. 6; and Monday, September 20th, as the weekday situation depicted in Fig. 6. On holidays, the proposed method and CEXGBQR have median values that are closest to the real baseline, followed by CELGBQR and CELSTMQR. The worst performer is CEGBQRT, with median values always higher than the real baseline. Although CEXGBQR performs well in terms of median values, the PIs it generates are too wide from 08:00 to 12:00. This indicates a failure to capture the fast-decreasing baseline, or in other words, an inability to adjust to the varying baseline. On weekdays, CEGBQRT still has the worst performance in terms of median values, and its PIs are too wide during working hours (08:00 to 20:00). CEXGBQR also generates overly wide PIs during working hours, while only CELGBQR produces more valid PIs, although they are wider than those of the proposed method. We can conclude that the

Table 2
HVAC Consumption baseline probabilistic estimation results from different methods at different PICPs.

Method	95 %			90 %			80 %			70 %			50 %		
	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)
Proposed	95.3	0.176	**0.824**	88.8	0.122	**0.877**	80.6	0.087	**0.913**	72.0	0.058	**0.940**	48.2	0.031	**0.967**
ETCQRN	97.0	0.201	0.797	93.9	0.153	0.841	87.4	0.110	0.866	76.4	0.083	0.898	64.8	0.070	0.834
CELSTMQR	92.7	0.231	0.767	83.0	0.172	0.808	81.0	0.164	0.836	75.7	0.171	0.816	45.1	0.125	0.865
ELSTMQR	96.9	0.268	0.731	94.2	0.194	0.799	88.8	0.143	0.824	83.9	0.141	0.780	52.6	0.091	0.906
CEXGBQR	92.9	0.200	0.798	93.2	0.179	0.817	83.0	0.115	0.881	71.4	0.086	0.913	54.5	0.050	0.940
EXGBQR	97.6	0.231	0.766	95.1	0.194	0.796	88.2	0.127	0.844	81.8	0.097	0.842	61.5	0.059	0.881
CELGBQR	93.5	0.177	0.822	84.2	0.139	0.847	75.1	0.101	0.888	79.0	0.089	0.875	61.8	0.057	0.880
ELGBQR	93.4	0.196	0.803	87.3	0.145	0.852	77.3	0.103	0.894	74.1	0.084	0.908	53.9	0.052	0.941
CEGBQRT	94.1	0.284	0.716	89.4	0.157	0.843	78.4	0.100	0.899	74.7	0.075	0.915	48.6	0.038	0.961
EGBQRT	42.7	0.168	0.212	42.4	0.085	0.295	38.3	0.058	0.395	36.7	0.041	0.551	31.6	0.030	0.819
CTCQRN	96.0	0.203	0.797	92.3	0.139	0.859	82.3	0.092	0.906	73.4	0.068	0.927	55.9	0.040	0.943
TCQRN	96.2	0.253	0.746	88.4	0.186	0.813	79.1	0.143	0.857	65.0	0.114	0.875	44.9	0.082	0.906

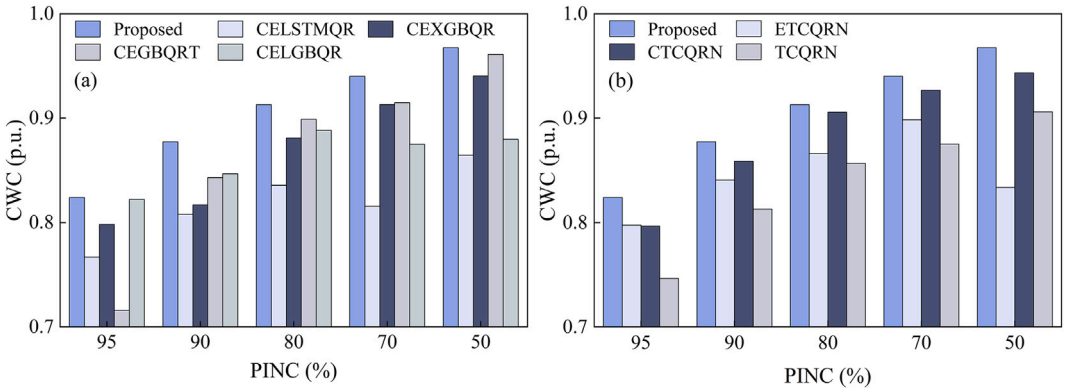


Fig. 5. HVAC consumption baseline probabilistic estimation results: (a) CWC from different regression methods at different PINCs; (b) CWC from different PI yield methods at different PINCs.

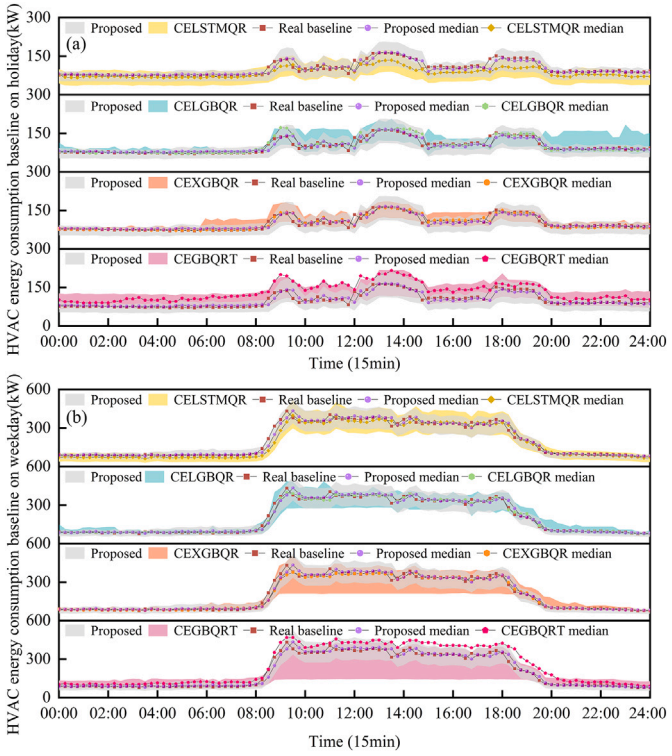


Fig. 6. HVAC consumption baseline probabilistic estimation results with different regression methods at 95 % PINC: (a) A specific holiday estimation results; (b) A specific weekday estimation results.

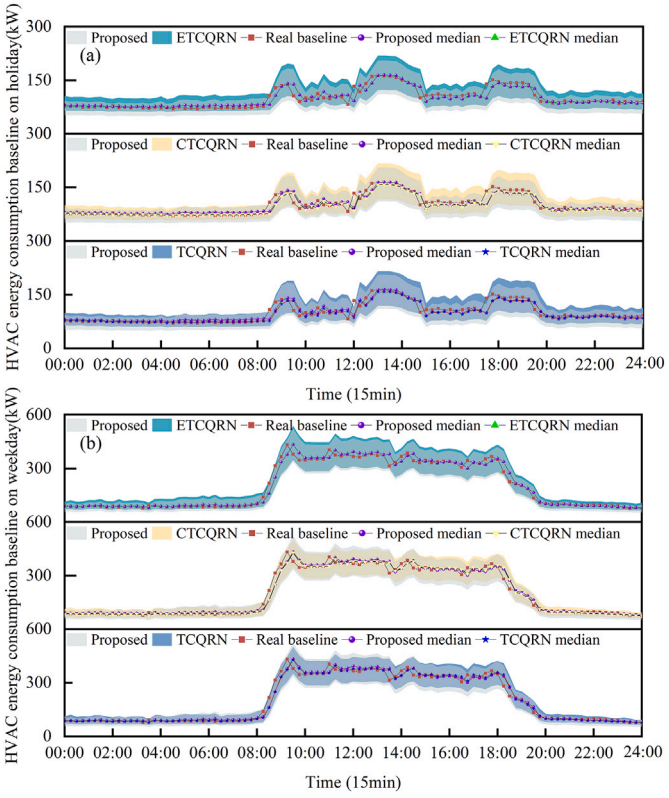


Fig. 7. HVAC consumption baseline probabilistic estimation results with different PI yield methods at 95 % PINC: (a) A specific holiday estimation results; (b) A specific weekday estimation results.

proposed method using TCN as a regression method can capture baseline variations from the historical baseline better than other regression methods.

We present the results compared to different yield PI methods in Fig. 7. Since the bottom regression method is TCN, the median values from ETCQRN and the proposed method are the same, causing their lines to overlap in Fig. 7. The main difference between the methods lies in the PIs. Similarly, the median values from TCQRN and CTCQRN are also the same. Overall, the median values from the proposed method and CTCQRN show similar performance. Specifically, when the baseline ramps up at 10:00 on holidays, the proposed method can track the change, while CTCQRN cannot. However, on weekdays from 12:00 to 14:00, the median values of the proposed method exceed the real baseline, whereas CTCQRN and TCQRN are closer to the real baseline. And the proposed method, ETCQRN, CTCQRN, and TCQRN have similar performances on yield PIs. Their PIs can contain most of the real baseline at 95 % PINC. To make it clearer, we also display the results at 50 % PINC in Fig. 8.

When employees arrive at the office, they start up all the HVAC equipment with a fixed temperature setting. This causes higher energy consumption during regular working hours. However, when they leave the meeting room or when the outside temperature is low, they may regulate the temperature setting higher to decrease the consumption. In Fig. 8, we can observe that, according to the proposed method, the length of the up error bar is shorter than the length of the down error bar during working hours. In contrast, fewer employees are working at night, leading to lower basic energy consumption. When employees need to work overtime, they turn on the HVAC equipment, which increases energy consumption. So during off-working hours, the length of the up error bar exceeds that of the down error bar. While the other methods

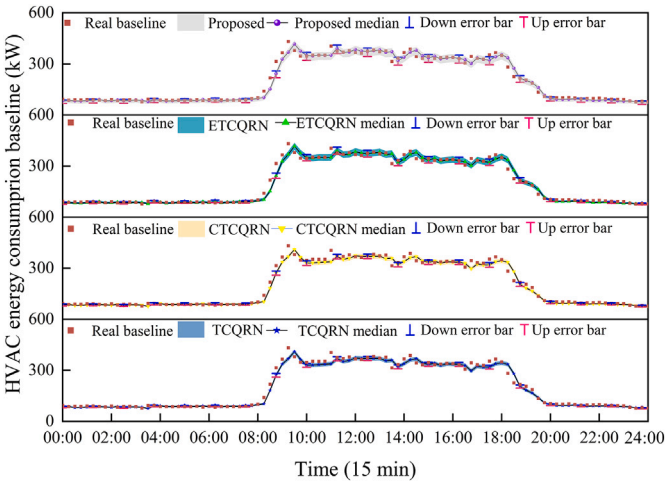


Fig. 8. HVAC energy baseline estimation results at 50 % PINC with different PI yield methods.

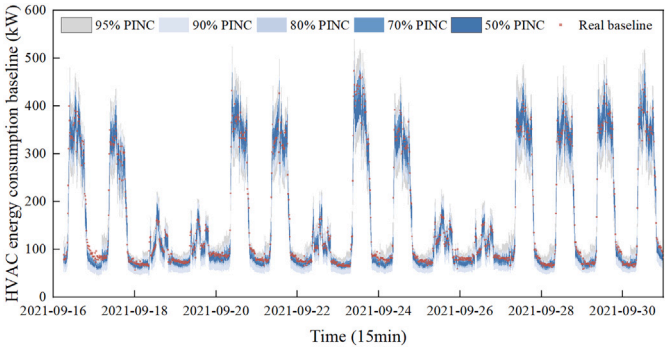


Fig. 9. HVAC consumption baseline probabilistic estimation results based on the proposed CETCQRN at different PINCs.

have nearly the same length for the up error bar and down error bar, they fail to accurately quantify the uncertainties within a day. As a result, although the PIs are narrower than the proposed method, few baseline points fall into them. In conclusion, the proposed method can yield more valid PIs by flexibly reflecting the direction that has greater variability, taking into account the level of uncertainty in different situations throughout the day.

From Fig. 9, we can observe that the PIs during off-working hours are lower than the PIs during working hours, indicating significantly less uncertainty during off-working hours. When the PINC is lower, the PIs are wider to ensure effective coverage ranges.

4.2.3. Discussion on more general scenarios

To further assess the adaptability of our method in broader contexts, we have included a case study on a gymnasium in addition to the original office building dataset from Macau. The choice of a gymnasium is motivated by its distinct operational characteristics compared to typical commercial buildings, and the results are shown in Table 3.

In this study, we select September 20 (weekday) and September 24 (holiday) as representative days to analyze how CETCQRN performs under different occupancy and operational conditions. A comparative analysis of baseline PIs generated by CETCQRN across these two scenarios is provided in Fig. 10. On weekdays, HVAC demand follows a more regular pattern, with stable occupancy-driven fluctuations. CETCQRN accurately captures the baseline trend and provides tight, valid prediction intervals. On holidays, CETCQRN successfully adapts to the changing conditions, maintaining a reliable prediction interval that reflects the uncertainty in demand.

To strengthen this discussion, we provide a comparative analysis in Fig. 10, illustrating the performance of CETCQRN across multiple scenarios. These results highlight the robustness of the proposed method and its potential to serve as a generalized approach for HVAC baseline estimation in diverse building environments.

4.3. Evaluation results of credible demand response capacity

In this section, we calculate the CDRC according to the PIs from the previous section. The PIs lower bound is BLB, and the upper bound is BUB. During working hours, which are from 08:00 to 20:00, the set temperature is set at 24 °C. During off-working hours, which are from 00:00 to 07:00 and from 20:00 to 24:00, the temperature is set at 27 °C. We apply different regulation levels of 1 °C, 2 °C, and 3 °C, respectively.

The range of CDRC downregulation refers to the interval in which power increases as the setting temperature decreases. The downregulation service begins during off-working hours and lasts for 15 min. In Table 4, we provide the mean values of $CDRC_{\min}$ and $CDRC_{\max}$ for different PINCs under various regulation scenarios. At different PINC levels, the $CDRC_{\min}$, which is the lower bound, shows an average increase of 30.076 kW, which is 1.09 times higher than the holiday $CDRC_{\min}$ of 27.721. And the $CDRC_{\max}$, which is the upper bound, shows an average increase of 42.235 kW, which is 1.15 times higher than the holiday $CDRC_{\max}$ of 36.605 kW. The range of CDRC upregulation refers to the interval in which power decreases as the setting temperature increases. It is evident that as the setting temperature increases, the weekday CDRC rises rapidly, as shown in Table 5. The upregulation service begins during working hours and lasts for 15 min. The evaluation results at different PINC levels are shown in Fig. 11. It is clear that the holiday CDRC is lower than the weekday CDRC due to employee activity.

Table 3
Gymnasium HVAC consumption baseline probabilistic estimation results from different methods at different PICPs.

Method	95 %			90 %			80 %			70 %			50 %		
	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)	PICP (%)	PINAW (p.u.)	CWC (p.u.)
Proposed	97.3	0.329	**0.669**	90.9	0.237	**0.763**	82.3	0.163	**0.835**	71.9	0.109	**0.889**	51.1	0.055	**0.944**
ETCQRN	97.5	0.344	0.654	93.7	0.245	0.750	76.5	0.171	0.824	76.5	0.113	0.868	56.9	0.068	0.910
CELSTMQR	97.7	0.447	0.551	86.2	0.349	0.646	68.5	0.205	0.744	74.6	0.191	0.800	64.0	0.128	0.791
ELSTMQR	98.3	0.504	0.493	89.2	0.319	0.681	66.0	0.175	0.748	81.4	0.172	0.776	49.1	0.074	0.926
CEXGBQR	98.0	0.470	0.528	96.5	0.344	0.642	90.3	0.224	0.736	86.1	0.142	0.754	78.7	0.087	0.605
EXGBQR	94.4	0.433	0.567	88.8	0.319	0.681	87.9	0.222	0.754	66.7	0.124	0.871	53.7	0.077	0.917
CELGBQR	97.7	0.439	0.559	94.3	0.332	0.662	80.3	0.206	0.794	77.8	0.158	0.817	54.5	0.109	0.882
ELGBQR	78.4	0.340	0.575	66.6	0.261	0.562	65.5	0.182	0.736	54.0	0.133	0.763	28.7	0.068	0.743
CEGBQRT	97.5	0.470	0.528	95.8	0.317	0.672	88.6	0.194	0.777	79.1	0.140	0.825	59.5	0.070	0.889
EGBQRT	59.0	0.198	0.420	61.0	0.139	0.565	58.1	0.090	0.716	56.4	0.077	0.841	45.0	0.093	0.896
CTCQRN	98.8	0.448	0.548	91.3	0.373	0.626	87.2	0.291	0.691	73.7	0.224	0.771	69.2	0.133	0.721
TCQRN	98.1	0.363	0.634	92.8	0.278	0.719	89.2	0.185	0.781	66.4	0.132	0.862	63.3	0.155	0.773

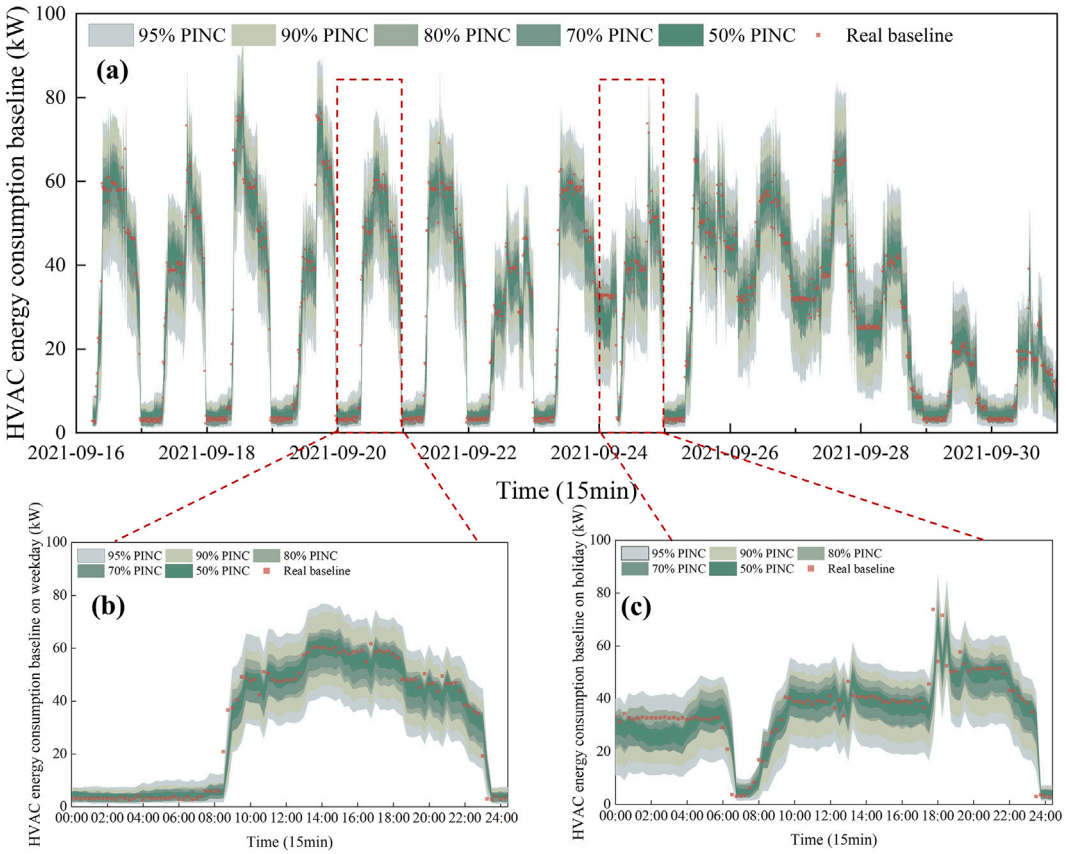


Fig. 10. Gymnasium HVAC consumption baseline probabilistic estimation: (a) results based on the proposed CETCQRN at different PINCs; (b) results on a weekday; (c) results on a holiday.

Table 4
Building HVAC systems downregulation CDRC at different PINCs.

Regulation level	Scenarios	95 %		90 %		80 %		70 %		50 %	
		$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)
Decrease 1 °C	Weekdays	15.526	26.593	12.510	21.228	15.004	20.033	15.849	19.619	16.302	18.115
	Holidays	14.294	23.839	11.571	19.515	13.738	18.304	14.561	17.959	15.163	16.931
Decrease 2 °C	Weekdays	31.053	53.186	25.019	42.457	30.008	40.065	31.697	39.239	32.604	36.231
	Holidays	28.581	47.664	23.136	39.021	27.469	36.598	29.115	35.909	30.320	33.856
Decrease 3 °C	Weekdays	46.579	79.780	37.529	63.685	45.012	60.098	47.546	58.858	48.906	54.346
	Holidays	42.871	61.496	34.704	58.531	41.203	54.897	43.673	53.863	45.480	50.785

Table 5
Building HVAC systems upregulation CDRC at different PINCs.

Regulation level	Scenarios	95 %		90 %		80 %		70 %		50 %	
		$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)	$CDRC_{min}$ (kW)	$CDRC_{max}$ (kW)
Increase 1 °C	Weekdays	26.800	42.406	27.277	38.945	29.329	37.727	30.318	37.348	31.436	36.150
	Holidays	10.751	19.645	9.412	17.120	11.045	15.722	11.731	15.447	12.535	14.870
Increase 2 °C	Weekdays	53.523	84.701	54.469	77.783	58.569	75.343	60.545	74.587	62.777	72.192
	Holidays	21.503	39.289	18.825	34.240	22.089	31.445	23.462	30.893	25.069	29.740
Increase 3 °C	Weekdays	80.285	127.052	81.703	116.675	87.854	113.015	90.818	111.881	94.166	108.287
	Holidays	32.254	58.934	28.237	51.360	33.134	47.167	35.193	46.340	37.604	44.610

On weekdays, the sudden increase in CDRC is caused by employees not resetting the HVAC setting temperature according to the decreased ambient temperature in time, resulting in a high potential for regulation. In contrast, although the holiday CDRC also grows, its absolute value remains insignificant. At different PINC levels, $CDRC_{min}$ shows an average increase of 57.933 kW, which is 2.61 times higher than the holiday CDRC of 22.190. And $CDRC_{max}$ shows an average increase of 76.867 kW, which is 2.32 times higher than the holiday CDRC of 33.121 kW.

In general, downregulation CDRC and upregulation CDRC have similar CDRC on holidays for office buildings. While on weekdays, the upregulation $CDRC_{min}$ is 1.93 times higher than the downregulation

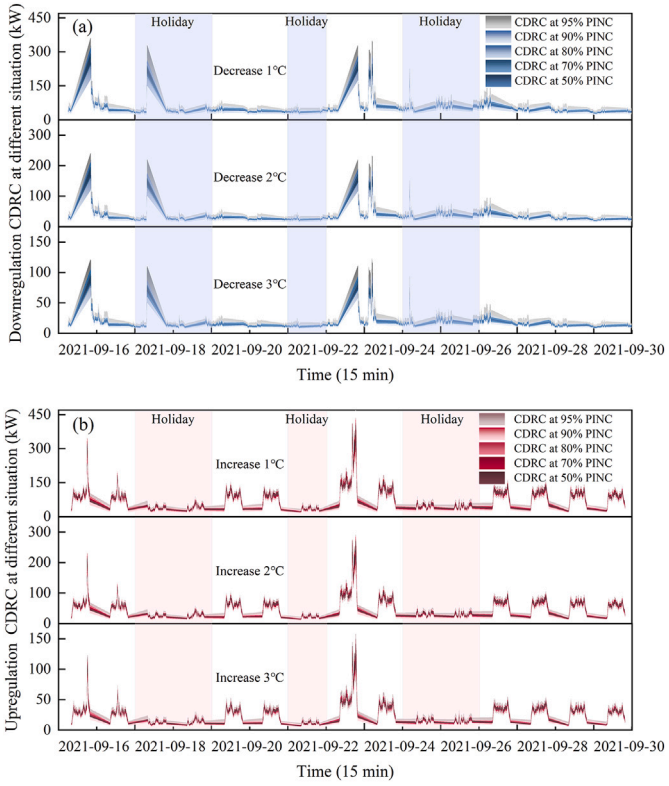


Fig. 11. HVAC credible demand response capacity evaluation results at different PINCs: (a) HVAC downregulation credible demand response capacity; (b) HVAC upregulation credible demand response capacity.

CDRC, and the upregulation $CDRC_{max}$ is 1.82 times higher than the downregulation CDRC.

We still select September 19th and September 20th to display typical CDRC results at 95 % PINC, as shown in Fig. 12. For clarity, we take the upregulation CDRC opposite number to figure the regulation over a single day effectively. On holidays, the single-day average downregulation CDRC ranges from [12.198, 20.074] with a 1 °C decrease, [30.200, 51.282] with a 2 °C decrease, and [36.594, 60.223] with a 3 °C decrease, respectively. And the average upregulation CDRC ranges from [10.770, 19.477] with a 1 °C increase, [21.541, 38.953] with a 2 °C increase, and [32.311, 58.430] with a 3 °C increase, respectively. On weekdays, the average downregulation CDRC ranges from [12.413, 21.869] with a 1 °C decrease, [24.827, 43.738] with a 2 °C decrease, and [37.240, 65.607] with a 3 °C decrease, respectively. And the average upregulation CDRC ranges from [23.844, 37.637] with a 1 °C increase, [47.688, 75.274] with a 2 °C increase, and [71.532, 112.911] with a 3 °C increase, respectively.

5. Conclusion

This paper proposes a CDRC evaluation framework based on the CETCQRN probabilistic baseline estimation model for building HVAC systems. The CETCQRN model and equivalent thermal parameter model are developed to provide regulation services for the power system. The proposed CETCQRN baseline evaluation model reflects the multi-uncertainties impact on the baseline and adapts to the historical baseline heteroscedasticity and autocorrelation. The results show that the proposed model can yield valid PIs by flexibly reflecting the direction that has greater variability, taking into account the level of uncertainty in different situations throughout the day. The evaluation framework shows that on weekdays, the upregulation CDRC range is approximately 2 times higher than the downregulation CDRC range. It can be concluded that there is almost no difference in the downregulation capacity range, whereas the upregulation CDRC range on weekdays is twice as much as it is on holidays.

6. Discussion and future work

6.1. Discussion

In this section, the paper compares and contrasts the proposed method with customer directrix load (CDL)-based DR, and explores two approaches for the rational evaluation of CDRC. CDL provides a top-down framework, which is designed to guide customers in adjusting

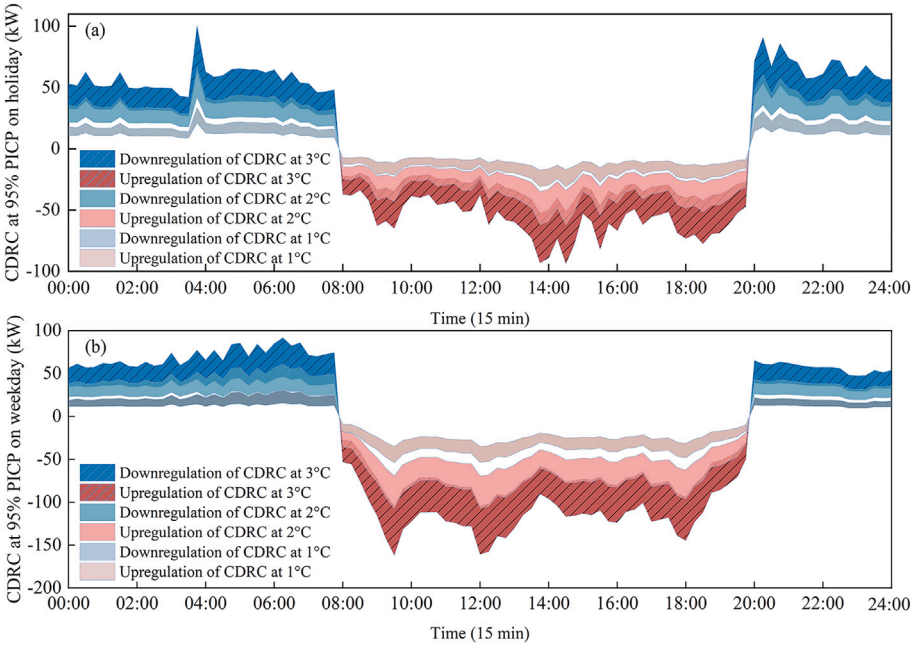


Fig. 12. HVAC credible demand response capacity evaluation results under different situations: (a) during the holidays; (b) during the weekdays.

their electricity consumption in response to system-level signals [37]. In this process, the Independent System Operator (ISO) or Regional Transmission Organization (RTO) computes the aggregated customer directrix load (ACDL) and sends it to the Load Serving Entity (LSE). Then, the LSE broadcasts the real-time activation signal to guide customers to adjust their consumption based on their own CDL, ensuring that the aggregate of all CDLs aligns with the ACDL. Building on this concept, Ref. [38] extends the CDL concept with multi-time scale CDL-based DR mechanism, incorporating both day-ahead planning and intraday optimization. This addresses the dual uncertainties stemming from renewable energy generation and customer behavior. Specifically, day-ahead CDL serves as the guiding target for DR in the day-ahead stage, factoring in the uncertainty of renewable energy. It is presented as a band-shaped curve to accommodate the variability in energy forecasts, allowing for more accurate and responsive consumption adjustments. In summary, CDL provides an up-bottom approach to DR by offering real-time, system-wide targets that effectively manage large-scale DR deployment.

In this paper, the proposed CDRC evaluation follows a bottom-up framework. This means that users can evaluate their own DR capacity and provide CDRC intervals to DSO. With the widespread deployment of edge sensors, the volume of data collected by edge users has significantly increased. Taking Shenzhen, China, as an example, since 2023, newly integrated buildings monitored by the power grid have predominantly consisted of public buildings, such as educational and commercial buildings [39]. This trend is expected to significantly reduce the problem of data scarcity. By leveraging local computations carried out by users, this approach helps to minimize computational delays. In this context, users are more inclined to participate in DR events without concerns regarding privacy leakage, and local computations performed by users will not result in extensive computational delays. Additionally, the interval-based results offer more reliable information to the grid, which helps the DSO create a credible and economically efficient dispatch plan. Although the proposed method and the CDL method come from different approaches, the proposed method can be adapted to the CDL framework. It can assist the RTO/ISO in customizing the ACDL baseline, taking into account uncertainties from renewable energy generation and customer behavior.

6.2. Future work

From the perspective of end-users, this study proposes a CDRC framework and provides an explanation of its definition and calculation methods. However, a significant area for future research is the effective allocation of the submitted intervals. Specifically, there is a need for further investigation into the methodologies for aggregating intervals across different users and establishing appropriate upper and lower boundaries of the intervals. Additionally, exploring the design of market mechanisms that facilitate high-return strategic behavior in the context of market participation and bidding, while utilizing the intervals, represents another important avenue for future research. Such advancements would contribute to enhancing both the economic efficiency and flexibility of the DR system.

CRediT authorship contribution statement

Siyu Jiang: Writing – original draft, Methodology, Data curation. **Hongxun Hui:** Writing – review & editing, Investigation, Funding acquisition, Conceptualization. **Yonghua Song:** Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hongxun Hui reports financial support was provided by National Natural Science Foundation of China. Hongxun Hui reports financial support was provided by Science and Technology Development Fund, Macao SAR. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper is funded in part by the Electric Power Research Institute of Guizhou Power Grid Co., Ltd for its support of project (GZKJXM20240010); in part by the National Natural Science Foundation of China (Grant No. 52407075); in part by the Science and Technology Development Fund, Macau SAR (File No. 001/2024/SKL); in part by the Chair Professor Research Grant of University of Macau (File No. CPG2025-00023-IOTSC).

Data availability

Data will be made available on request.

References

[1] Alobaidi AH, Fazlhashemi SS, Khodayar M, Wang J, Khodayar ME. Distribution service restoration with renewable energy sources: a review. IEEE Trans Sustain Energy 2023;14(2):1151–68.
[2] Liu N, Yu X, Wang C, Li C, Ma L, Lei J. Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers. IEEE Trans Power Syst 2017;32(5):3569–83.
[3] Conejo AJ, Morales JM, Baringo L. Real-time demand response model. IEEE Trans Smart Grid 2010;1(3):236–42.
[4] PJM. Demand response overview. 2023 Dec.
[5] ERCOT. Overview of demand response in ERCOT. 2023 Apr.
[6] Yang S, Lao KW, Hui H, Chen Y. Secure distributed control for demand response in power systems against deception cyber-attacks with arbitrary Patterns. IEEE Trans Power Syst 2024;39(6):7277–90.
[7] IEA. Demand response availability at times of greatest flexibility need and share of total flexibility under the net zero scenario, 2020 and 2030. 2023 Jul.
[8] IEA. More efficient and flexible buildings are key to clean energy transitions. 2024 Apr.
[9] PJM. Energy and ancillary services market operations (revision 130). 2024 Mar.
[10] Sichuan Provincial Development and Reform Commission. Implementation plan for market-oriented of electricity demand response in Sichuan province in 2023. 2023 Apr.
[11] Han B, Li H, Wang S. A probabilistic model for real-time quantification of building energy flexibility. Adv Appl Energy 2024;15:100186.
[12] Paula F GD. Review and assessment of the different categories of demand response potentials. Energy 2019;179:280–94.
[13] Liu Z, Chen Y, Yang X, Yan J. Power to heat: opportunity of flexibility services provided by building energy systems. Adv Appl Energy 2023;11:100149.
[14] Yin R, Kara EC, Li Y, DeForest N, Wang K, Yong T, et al. Quantifying flexibility of commercial and residential loads for demand response using setpoint changes. Appl Energy 2016;177:149–164.
[15] Tang H, Wang S. A model-based predictive dispatch strategy for unlocking and optimizing the building energy flexibilities of multiple resources in electricity markets of multiple services. Appl Energy 2022;305.
[16] Song M, Gao C, Shahidehpour M, Li Z, Yang J, Yan H. Impact of uncertain parameters on TCL power capacity calculation via HDMR for generating power pulses. IEEE Trans Smart Grid 2019;10(3):3112–24.
[17] Jung W, Jazizadeh F. Energy saving potentials of integrating personal thermal comfort models for control of building systems: comprehensive quantification through combinatorial consideration of influential parameters. Appl Energy 2020;268:114882.
[18] Ran F, Gao DC, Zhang X, Chen S. A virtual sensor based self-adjusting control for HVAC fast demand response in commercial buildings towards smart grid applications. Appl Energy 2020;269:115103.
[19] Wang Y, Chen Q, Hong T, Kang C. Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans Smart Grid 2019;10(3):3125–48.
[20] Siddiquee SMS, Agyeman KA, Bruton K, Howard B, O'Sullivan DT. A data-driven framework for quantifying demand response participation benefit of industrial consumers. IEEE Trans Ind Applicat 2024;60(2):2577–87.
[21] Yu X, Ergan S. Estimating power demand shaving capacity of buildings on an urban scale using extracted demand response profiles through machine learning models. Appl Energy 2022;310:118579.
[22] Harikrishnan GR, Sasidharan S, Binoy C N. Advanced short-term load forecasting for residential demand response: an XGBoost-ANN ensemble approach. Electr Power Syst Res 2025;242:111476.
[23] Liang H, Ma J, Lin J. Robust distribution system expansion planning incorporating thermostatically-controlled-load demand response resource. IEEE Trans Smart Grid 2022;13(1):302–13.
[24] Zhu J, Niu J, Tian Z, Zhou R, Ye C. Rapid quantification of demand response potential of building HAVC system via data-driven model. Appl Energy 2022;325:119796.
[25] Amer AA, Massoud AM. DRL-HEMS: deep reinforcement learning agent for demand response in home energy management systems considering customers and operators perspectives. IEEE Trans Smart Grid 2023;14(1).

[26] Zhang L, Good N, Mancarella P. Building-to-grid flexibility: modelling and assessment metrics for residential demand response from heat pump aggregations. Appl Energy 2019;233–234:709–723.
[27] Hui H, Ding Y, Liu W, Lin Y, Song Y. Operating reserve evaluation of aggregated air conditioners. Appl Energy 2017;196:218–28.
[28] Kong X, Wang Z, Liu C, Zhang D, Gao H. Refined peak shaving potential assessment and differentiated decision-making method for user load in virtual power plants. Appl Energy 2023;334:120609.
[29] Song Z, Shi J, Li S, Chen Z, Jiao F, Yang W, et al. Data-driven and physical model-based evaluation method for the achievable demand response potential of residential consumers' air conditioning loads. Appl Energy 2022;307.
[30] Han X, Zhang C, Tang Y, Ye Y. Physical-data fusion modeling method for energy consumption analysis of smart building. J Mod Power Syst Clean Energy 2022;10(2):482–91.
[31] Lu N. An evaluation of the HVAC load potential for providing load balancing service. IEEE Trans Smart Grid 2012;3(3):1263–70.
[32] Li Y, Song L, Zhang S, Kraus L, Adcox T, Willardson R, et al. A TCN-based hybrid forecasting framework for hours-ahead utility-scale PV forecasting. IEEE Trans Smart Grid 2023;14(5):4073–85.
[33] Romano Y, Patterson E, Candes E. Conformalized quantile regression. In: Advances in neural information processing systems; vol. 32. Curran Associates, Inc.; 2019.
[34] Jensen V, Bianchi FM, Anfinsen SN. Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Trans Neural Netw Learn Syst 2024:1–12.
[35] Liao W, Wang S, Bak-Jensen B, Radhakrishna pillai J, Yang Z, Liu K. Ultra-short-term interval prediction of wind power based on graph neural network and improved bootstrap technique. J Mod Power Syst Clean Energy 2023;11(4):1100–14.
[36] Stjelja D, Kuzmanovski V, Kosonen R, Jokisalo J. Building consumption anomaly detection: A comparative study of two probabilistic approaches. Energy Build 2024;313:114249.
[37] Fan S, Li Z, Yang L, He G. Customer directrix load-based large-scale demand response for integrating renewable energy sources. Electr Power Syst Res 2020;181:106175.
[38] Zhang Y, Meng Y, Fan S, Xiao J, Li L, He G. Multi-time scale customer directrix load-based demand response under renewable energy and customer uncertainties. Appl Energy 2025;383:125334.
[39] Shenzhen Municipal Bureau of Housing and Urban-Rural Development. Energy consumption monitoring report of public buildings in Shenzhen (2023). 2024 Jul.