Nonparametric probabilistic load forecasting based on quantile combination in electrical power systems

He, Yaoyao; Cao, Chaojin; Wang, Shuo; Fu, Hong

DOI:
10.1016/j.apenergy.2022.119507

Nonparametric probabilistic load forecasting based on quantile combination in Electrical power systems

Yaoyao He [a,b,*], Chaojin Cao [a,b], Shuo Wang [c], Hong Fu [a,b]

[a] *School of Management, Hefei University of Technology,Hefei 230009, China*

[b] *Key Laboratory of Process Optimization and Intelligent Decision-Making (Hefei University of Technology), Ministry of Education, Hefei 230009, China*

[c] *CERCIA, the School of Computer Science,The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

Abstract

Probabilistic load forecasting (PLF) aims to predict the future uncertainties of loads to reduce the potential risks in power system planning and operation. In the increasingly complex power market environment, exploring advanced approaches to obtain more accurate PLF is still a significant topic. Optimizing individual forecasting method is no longer the only direction to improve the accuracy of load forecasting in recent years. Researchers started to focus on combination methods because of their better accuracy in most cases than a single model. There are existing combination methods designed for parametric environment, where some results are based on the certain assumption (e.g., Gaussian distribution assumption of single prediction). Combining probabilistic forecasts in nonparametric environment is rarely investigated, because modeling the combination problem without assuming distributions of parameters is hard. This paper proposes a novel combined model for probabilistic forecasting tailored to nonparametric environments, which combines multiple quantile-based models by minimizing the overall loss function composed of continuous ranked probability score (CRPS) under kernel density estimation (KDE). We define a multilayer Gaussian mixture distribution, which is an extended form of Gaussian mixture distribution that can simulate any distribution type in nonparametric environment. Based on the multilayer Gaussian mixture distribution, the combined model is further formulated into a quadratic programming problem with linear restrictions that can be solved efficiently. Case studies are performed using benchmark and competition datasets from the United States and China. The results show that our proposed method outperforms the best individual model and other existing combination methods. In summary, this paper constructs a complete theoretical framework of nonparametric probabilistic combination forecasting and proves its effectiveness in practical application.

Keywords: Combination of quantile forecasts, Kernel density estimation, Continuous ranked probability score, Probabilistic load forecasting

*Corresponding author. School of Management, Hefei University of Technology, Hefei 230009, China
Email addresses: hy-342501y@163.com (Yaoyao He [a,b,*]), caochaojin_hfut@163.com (Chaojin Cao [a,b]),

1. Introduction

Due to the unique features of electricity production, electricity cannot be stored in large quantities. Utility companies need to carefully maintain the balance of power supply and demand in order to reduce power shortage and investment waste [1]. The economic development and social stability will be greatly affected once the power system oscillates and a blackout occurs [2, 3]. Load forecasting with high accuracy can help hybrid energy system to increase operational efficiency [4]. One major challenge in load forecasting is to quantify significant uncertainties in smart grids caused by the consolidation of distributed energy resources (DERs) and the deregulation of the electricity market [5]. Traditional deterministic forecasting only provides a single predicted value for load, which is unable to quantify the future uncertainty. Different from point forecasting, probabilistic load forecasting (PLF) can provide more detailed information about the variability of future power demand by depicting the load variation interval under certain confidence level and the probability of occurrence for each point within the interval. Therefore, PLF has become more dominating in this area [6].

Based on the form of prediction outputs, PLF can be classified into three types: prediction intervals (PIs), quantile and probability density function (PDF) forecasting [7]. Among them, traditional PIs methods assume the shape of the predictive distribution, such as Bayesian [8], Delta [9] and bootstrap [10]. However, the assumption based on prior knowledge of data or prediction errors is not always guaranteed to be correct because the actual load has strong seasonality and volatility. Unlike PIs, a quantile forecasting model is trained to minimize the pinball loss function, and the output of the trained model is utilized to build a set of quantiles. As a non-parametric estimation technique, the quantile type of methods does not assume any distributions of predictions [11]. Density forecasting can construct the probability density function to present more holistic and flexible information on future load than the above two forms. Therefore, the probability density prediction is also regarded as the most complete form of probabilistic forecasting [12]. Fortunately, quantile prediction can be converted to the form of probability density via nonparametric kernel density estimation (KDE) to obtain more comprehensive information [13].

The Global Energy Forecasting Competition 2014 (GEFCom2014) stimulated the development of quantile prediction [14]. Many machine learning algorithms used for point forecasting have been adapted for quantile forecasting. Nagy et al. proposed a method using gradient boosting machine (GBM) in quantile form, which has the superiority and robustness in performing regression tasks [15]. Other powerful tree-based models include LightGBM (LGB) and XGBoost (XGB), both of which have available and mature packages to facilitate quantile forecasting [16, 17]. However, a GBM cannot support quantile regression with multiple quantiles. Consequently, training GBM to obtain multiple quantile results becomes time-consuming.

s.wang.2@bham.ac.uk (Shuo Wang [c]), fuhong@hfut.edu.cn (Hong Fu [a,b])

As a classical deep learning technology, artificial neural network (ANN) has been widely used in prediction realm, such as weather forecasting [18], electricity price forecasting [19], wind power forecasting [20], and load forecasting [21]. For quantile forecasting, deep learning techniques establish another research stream. Pinball loss is used to guide the training of deep learning models to achieve quantile probabilistic forecasting [22]. Researchers have developed many advanced neural network models based on pinball loss, such as quantile regression neural network (QRNN) [23], quantile regression long short-term memory (QRLSTM) [24], quantile regression gated recurrent unit (QRGRU) [25], quantile regression minimal gated memory (QRMGM) [26], which have been successfully applied in various energy forecasts. It is worth noting that the prediction results under different quantiles from deep neural networks can be given simultaneously, due to the multi-output structure.

Although many of the individual forecasting methods have demonstrated their superior performance, no one can be the best for all datasets. The combination of different models is usually an effective approach to reduce the overall risk of selecting a poor model and obtain a smaller generalization error. Forecast combination is also known as a type of ensemble methods in machine learning, which can be divided into homogeneous and heterogeneous combinations, depending on whether the component models are of same or different types respectively [27]. Common homogeneous ensemble methods include random forest [28] and Adaboost algorithms [29]. Heterogeneous ensemble methods mainly include two forms: stacking and weight allocation. A stacking model consists of a base layer and a meta-layer. The results given by different models in base layer are combined as the input of model in meta-layer and the final prediction results are the output from the meta-layer model [30]. The other type of heterogeneous ensemble methods is based on weight allocation. Dudek developed a heterogeneous method by integrating 10 forecasting models with different weights and proved to be effective in improving the generalization ability [31]. Nowotarski et al. used different weight averaging methods to combine 8 sister load forecasts, including simple averaging, performance-based averaging, positive weights averaging, and so on [32]. Most of combination methods are designed to produce point load forecasting.

A few combination methods have been proposed to give probabilistic forecasting. In [33], Hall and Mitchell first brought together density forecasting and forecast combination by minimizing the Kullback-Leibler distance between the forecast and true. Bracale et al. [34] proposed a competitive ensemble method for the short-term probabilistic forecasting of photovoltaic power, which uses the continuous ranked probability score (CRPS) to guide the search for optimal weights. However, the literature on combining probabilistic load forecasting is still rarely discussed. The key challenge lies in the problem of how to deal with different distribution types in power data and combine individual models for the best results.

One related work was proposed in [35] to transform the probabilistic load forecasting results of all individual models into corresponding Gaussian distribution and combine them with the guidance of minimum CRPS. This is a parametric combination method, which assumes the forecasting results of the individual

models to be Gaussian distributed and needs to estimate the relevant parameters of the corresponding Gaussian distribution. However, no assumption is guaranteed to be correct under the strong fluctuation of actual load. Once there is too much difference between them, forced transformation to Gaussian distribution will cause information deviation. As the authors of this method mentioned at the end of the paper, the distribution assumption is a restriction of their method.

To overcome this restriction, this paper proposes a nonparametric combination method for probabilistic load forecasting, which introduces KDE into CRPS oriented optimization problems. KDE does not assume sample distributions [36]. Zhang et al. utilized KDE to transform the quantile forecast into the probability density curve for further forecast combination [12]. The combination of quantile regression neural network (QRNN) and KDE shows great effectiveness in improving probabilistic load forecast [13].

In this paper, we propose a new combining probabilistic load forecasting suitable for the nonparametric environment with a complete theoretical framework. First, the concept of multilayer Gaussian mixture distribution is defined and a derived proposition is proposed. Then, according the CRPS in the expectation form raised by Székely et al.[37] and the above proposition, we prove that the difference between two independent random variables in the CRPS under KDE obeys 4-layer Gaussian mixture distribution after quantile combination. On this basis, the CRPS integrated with KDE is used as the objective function of the combination problem, which can further be cast to a linearly constrained quadratic programming (QP) model. Compared with Ref. [35], our paper has two novelties: 1) The work in [35] converts the quantile forecasts by using a parameter estimation method—Gaussian approximation of quantiles (GAQ), while this paper finishes the process by utilizing KDE, which is a non-parametric approach. 2) The combination model constructed in [35] takes the CRPS with Gaussian distribution parameters as the objective function, while this paper selected the CRPS integrated with KDE as the objective function, which can complete the combining probabilistic prediction in the case of non-parameters. Finally, case studies are conducted on the real-world load data from ISO New England (ISO-NE) in the US and electrician mathematical contest in modeling (EMCM) in China.

The key contributions are outlined in the statement listed below.

(1) A novel method for combining probabilistic load forecasting is proposed, which is tailored to the nonparametric situation and does not assume the distribution types of quantile forecasts. All the information about the quantiles given by quantile-based models can be fully utilized in the combination process.

(2) In our combined model, the CRPS integrated with KDE is used as the objective function. To find the optimal solution, we construct a complete theoretical deduction and transform our model into a QP problem with linear constraint.

(3) An ensemble framework for combining quantile forecasts is developed for PLF. Four quantile-based techniques consisting of QRNN, QRLSTM, QRGRU and QRMGM are integrated for performance improvement.

(4) By comparing with base models and existing combination forecast methods (simple averaging, a combination method based on individual performance, and two combinations to find the optimal weight), our combined model indicate a significant improvement in probabilistic prediction as well as deterministic prediction.

The rest of the paper is organized as follows: Section 2 introduces the problems to be solved in the combination quantile forecast framework. In Section 3, a new combined model named KDE-CRPS-guided combined model (KCGC) is proposed and transformed into QP model. Section 4 briefly introduces four base models for generating quantile forecasts, and a feature selection technique. Section 5 summarizes the whole steps of producing probability density forecasts through combining quantile forecasts. Section 6 conducts case studies on load data to verify the proposed method. Conclusions and future works are given in Section 7.

2. Load density forecasting framework and problem formulation for model combination

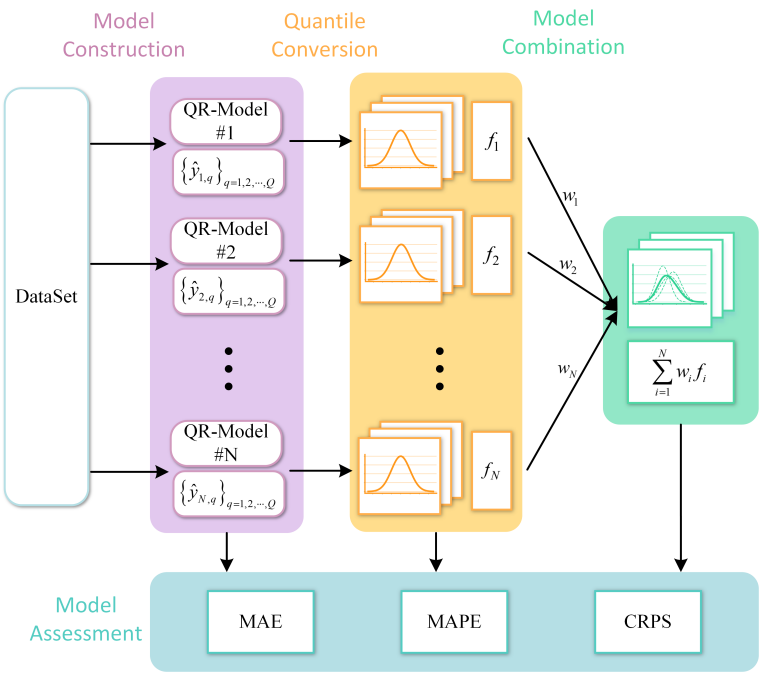

Fig. 1: Framework of probabilistic load forecast combination method. *Note*: $\hat{y}_{i,q}$ denotes the prediction result of the i-th QR-model at the q-th quantile, f_i is the PDF obtained by converting the quantile forecasts of the i-th model, and w_i is the weight of the i-th model.

2.1. Load density forecasting framework

As depicted in Fig. 1, we develop a framework of load density forecasts that combines multiple quantile forecasting models. It involves four stages: model construction, quantile conversion, model combination and model assessment. The four stages focus on the following issues:

(1) Construct a series of base models for the quantile forecasting results. In order to improve the prediction performance and the diversity of models, feature selection and the generation of different training sets are required. In addition, we need to make model construction more efficient.

(2) Convert different quantiles into probability density curves. The converting method should be non-parametric estimation for the best fits of different quantiles.

(3) Combine multiple converted forecast results to form the final probabilistic forecasts. The framework will find a set of optimal weights for the base models for the best combined performance.

(4) Make quantitative assessment of the combined model. The process involves performance evaluation in terms of effectiveness and reliability on an independent test set.

The subsequent sections will further explain these four stages. Specifically, the main innovations of this paper focus on stage 2 and 3, so they are first described in Section 3. The base model types and a feature selection method in the first stage are introduced in Section 4. Section 5 describes the complete implementation process of the four stages.

2.2. Problem formulation of model combination

Among the above four stages, how to combine the outputs of the probabilistic forecast models in stage 3 is the key problem. One of the common approaches is to assign weights to each base model. In order to find the optimal weights, the weight selection procedure can be formulated as an optimization problem in the whole time period $t = 1, 2, \cdots, T$:

$$\begin{aligned} \min_{w_i} \quad & \sum_{t=1}^{T} L\left(\sum_{i=1}^{N} w_i F_{i,t}, y_t\right) \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1 \\ & w_i \geq 0, \ i = 1, 2, \ldots, N \end{aligned} \tag{1}$$

where y_t is real value at time t; w_i is the weight for the i-th model; $F_{i,t}$ denotes a cumulative distribution function of $f_{i,t}$ for the time period t; and L denotes the loss function to evaluate the performance of the combined model.

3. KDE-CRPS-guided combined model (KCGC)

In this section, how KDE-CRPS-guided combined model (KCGC) implements the combined model under in a nonparametric situation minimizing Eq. (1) is discussed in detail. First, kernel density estimation (KDE)

is applied to convert different quantiles into probability density curves, which is a classic nonparametric method. After combining the converted result with a set of weights, the CRPS integrated with KDE, denoted by KDE-CRPS, is used to evaluate the combined result. Thus, in KCGC, the minimum total KDE-CRPS loss is set to be the objective function in Eq. (1). Finally, multilayer Gaussian mixture distribution and its derived proposition are leveraged to simplify the calculation of the objective function in KCGC. Then, the combination problem in KCGC is reformulated into a linearly constrained quadratic programming (QP) model.

3.1. Kernel density estimation (KDE)

KDE is a nonparametric method estimating unknown density functions without distributional assumptions. We use KDE to obtain the prediction distribution of base models. The PDF f_i^K of the i-th basis model using KDE is formulated as:

$$f_i^K(x) = \frac{1}{QB_i} \sum_{q=1}^{Q} K\left(\frac{\hat{y}_{i,q} - x}{B_i}\right) \tag{2}$$

where B_i is a bandwidth for the quantile forecasting result $\{\hat{y}_{i,q}\}_{q=1,2,\ldots,Q}$ of the i-th basis model, and $K(\cdot)$ is the kernel function.

In this paper, the Gaussian kernel is selected as the kernel function for its advantageous mathematical properties. It is widely used for probability density forecasting, showing good performance [38]. The Gaussian kernel function is defined as:

$$K(\eta) = \frac{1}{\sqrt{2}} e^{-\frac{1}{2}\eta^2} \tag{3}$$

By replacing $K(\eta)$ in Eq. (2) with Eq. (3):

$$\begin{aligned} f_i^K(x) &= \frac{1}{QB_i} \sum_{q=1}^{Q} \frac{1}{\sqrt{2}} \exp\left[\left(-\frac{1}{2}\right)\left(\frac{\hat{y}_{i,q} - x}{B_i}\right)^2\right] \\ &= \frac{1}{Q} \sum_{q=1}^{Q} \phi\left(x \,|\hat{y}_{i,q}, B_i\right) \end{aligned} \tag{4}$$

where $\phi(\cdot\,|\hat{y}_{i,q}, B_i)$ is the PDF of Gaussian distribution $N\left(\hat{y}_{i,q}, B_i^2\right)$. The combined PDF of N models with model weights $(w_1, \ldots, w_N)$ is:

$$\begin{aligned} f^K(x) &= \sum_{i=1}^{N} w_i f_i^K(x) \\ &= \frac{1}{Q} \sum_{i=1}^{N} \sum_{q=1}^{Q} w_i \phi\left(x \,|\hat{y}_{i,q}, B_i\right) \end{aligned} \tag{5}$$

The CDF of the combination can be obtained:

$$F^K(x) = \int_{-\infty}^{x} f^K(z)\, dz \tag{6}$$

3.2. Continuous ranked probability score (CRPS) integrated with KDE

CRPS is a scoring metric used for model evaluation in probabilistic forecasts [39].Given a cumulative distribution function (CDF) F of a random variable X and the real value y, the formula of CRPS can be expressed as:

$$CRPS\left(F,y\right)=\int_{-\infty}^{+\infty}\left(F\left(x\right)-\mathrm{I}\left(x-y\right)\right)^{2}dx \tag{7}$$

$$F\left(x\right)=\mathrm{P}\left[X\leq x\right]=\int_{-\infty}^{x}p\left(z\right)dz \tag{8}$$

$$\mathrm{I}\left(x-y\right)=\begin{cases}0 & x<y\\ 1 & x\geq y\end{cases} \tag{9}$$

where $p\left(x\right)$ is the probability density function (PDF) of X; Heaviside step function $\mathrm{I}\left(x-y\right)$ is used to simulate the "CDF" of the real value. In essence, CRPS evaluates the performance of probabilistic prediction by comparing the difference between the predicted and observed CDF. The smaller the CRPS, the better the performance of the probabilistic prediction is.

Due to the difficulty of solving the integral in Eq. (7), a closed form to estimate the integral is provided in [37]. In KCGC, the closed form of CRPS is integrated with KDE, denoted by KDE-CRPS, defined as follows:

$$\mathrm{CRPS}\left(F^{K},y\right)=E\left|H-y\right|-\frac{1}{2}E\left|H-H'\right| \tag{10}$$

where H and H' are independent copies of a random variable with CDF being F^{K}.

3.3. Multilayer gaussian mixture distribution

According to the statistical analysis of load data, there is no common probability distribution that can completely describe the variation of power load [40]. The Gaussian mixture distribution can simulate the complex probability density function by taking a weighted average of the finite normal distributions. Meanwhile, the Gaussian mixture model (GMM) has been widely used in the forecast field [41]. The probability density function (PDF) $f^{M}\left(x\right)$ of a random variable following the Gaussian mixture distribution is defined in Eq. (11) with the finite PDFs of Gaussian distribution $\left\{\phi\left(x\left|\mu_{a},\sigma_{a}\right.\right)\right\}_{a=1,2,\cdots,A}$.

$$f^{M}\left(x\right)=\sum_{a=1}^{A}w_{a}\phi\left(x\left|\mu_{a},\sigma_{a}\right.\right) \tag{11}$$

where A denotes the number of components in the Gaussian mixture distribution and w_a is the weight of the a-th Gaussian component of X, subject to $w_a\geq 0$ and $\sum_{a=1}^{A}w_a=1$.

We extend the concept of Gaussian mixture distribution to a multi-layer Gaussian mixture distribution, defined for the later calculation of the KDE-CRPS.

Definition 1. Let $\{x_a\}_{a=1,2,\cdots A}$ denote a set of random variables with Gaussian mixture distribution that is defined as follows:

$$
\begin{aligned}
x_1 &= w_{2,1}x_{1,1} + w_{2,2}x_{1,2} + \cdots + w_{2,B}x_{1,B} \\
x_2 &= w_{2,1}x_{2,1} + w_{2,2}x_{2,2} + \cdots + w_{2,B}x_{2,B} \\
&\cdots \\
x_A &= w_{2,1}x_{A,1} + w_{2,2}x_{A,2} + \cdots + w_{2,B}x_{A,B}
\end{aligned}
$$

where $\{x_{a,b}\}_{a=1,2,\cdots A;b=1,2,\cdots,B}$ are independent random variables with normal distribution $N\left(\mu_{a,b}, \sigma_{a,b}^2\right)$; $(w_{2,1}, w_{2,2}, \cdots, w_{2,B})$ is a set of weights, satisfying $\sum_{b=1}^{B} w_{2,b} = 1$ and $w_{2,b} \geq 0$ $(b = 1, 2, \ldots, B)$. Then, the random variable $X = w_{1,1}x_1 + w_{1,2}x_2 + \cdots + w_{1,A}x_A$ is called the 2-layer Gaussian mixture distribution, in which $\sum_{a=1}^{A} w_{1,a} = 1$, $w_{1,a} \geq 0$ $(a = 1, 2, \ldots, A)$. It has the structure shown in Fig. 2.

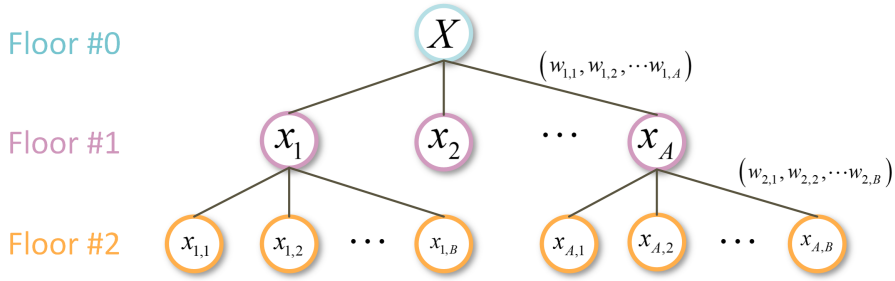

Fig. 2: The structure of 2-layer Gaussian mixture distribution.

Let the PDF of x_a be f_a:

$$f_a(x) = \sum_{b=1}^{B} w_{2,b}\phi\left(x \,|\mu_{a,b}, \sigma_{a,b}\right)$$

where $\phi\left(x \,|\mu_{a,b}, \sigma_{a,b}\right)$ is the PDF of $x_{a,b}$ with normal distribution $N\left(\mu_{a,b}, \sigma_{a,b}^2\right)$.

Therefore, the PDF of 2-layer Gaussian mixture distribution X can be calculated by:

$$f_X(x) = \sum_{a=1}^{A}\sum_{b=1}^{B} w_{1,a}w_{2,b}\phi\left(x \,|\mu_{a,b}, \sigma_{a,b}\right) \tag{12}$$

where $w_{1,a}$ and $w_{2,b}$ are subject to $w_{1,a} > 0$, $\sum_{a=1}^{A} w_{1,a} = 1$ and $w_{2,b} > 0$, $\sum_{b=1}^{B} w_{2,b} = 1$, respectively.

By analogy, let P be l-layer Gaussian mixture distribution whose PDF is defined:

$$f_P = \sum_{a=1}^{A}\sum_{b=1}^{B}\cdots\sum_{r=1}^{R} w_{1,a}w_{2,b}\cdots w_{l,r}\phi\left(x \,|\mu_{a,b,\ldots,r}, \sigma_{a,b,\ldots,r}\right) \tag{13}$$

where R denotes the number of weights in the l-layer; the sum of each layer of weights $(w_{i,1}, w_{i,2}, \ldots, w_{i,j})$ is 1, $i = 1, 2, \ldots, l$, $j = A, B, \ldots, R$; all weights are greater than or equal to 0.

Proposition 1. If X and Y are independent random variables with 2-layer Gaussian mixture distribution, their difference is a 4-layer Gaussian mixture distribution. In mathematical words, if

$$X = w_{1,1}x_1 + w_{1,2}x_2 + \cdots + w_{1,A}x_A \qquad Y = w'_{1,1}y_1 + w'_{1,2}y_2 + \cdots + w'_{1,C}y_C$$

$$f_X(x) = \sum_{a=1}^{A}\sum_{b=1}^{B} w_{1,i}w_{2,j}\phi\left(x\left|\mu_{a,b},\sigma_{a,b}\right.\right) \qquad f_Y(x) = \sum_{c=1}^{C}\sum_{d=1}^{D} w'_{1,c}w'_{2,d}\phi\left(x\left|\mu'_{c,d},\sigma'_{c,d}\right.\right)$$

$$\sum_{a=1}^{A} w_{1,a} = 1,\ \sum_{b=1}^{B} w_{2,b} = 1 \qquad \sum_{c=1}^{C} w'_{1,c} = 1,\ \sum_{d=1}^{D} w'_{2,d} = 1$$

$$w_{1,a} \geq 0,\ w_{2,b} \geq 0 \qquad w'_{1,c} \geq 0,\ w'_{2,d} \geq 0$$

then the PDF of $Z = X - Y$ is:

$$f_Z(z) = \sum_{a}^{A}\sum_{b}^{B}\sum_{c}^{C}\sum_{d}^{D} w_{1,a}w_{2,b}w'_{1,c}w'_{2,d}\,\phi\left(z\left|\mu_{a,b} - \mu'_{c,d}\,,\sqrt{\sigma_{a,b}^2 + \sigma'^2_{c,d}}\right.\right)$$

Proof: Since X and Y are independent, then

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{+\infty} f_X(y) f_Y(y-z)\,dy \\
&= \int_{-\infty}^{+\infty} \sum_{a=1}^{A}\sum_{b=1}^{B} w_{1,a}w_{2,b}\phi\left(y\left|\mu_{a,b},\sigma_{a,b}\right.\right) \sum_{c=1}^{C}\sum_{d=1}^{D} w'_{1,c}w'_{2,d}\phi\left(y-z\left|\mu'_{c,d},\sigma'_{c,d}\right.\right) dy \\
&= \sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C}\sum_{d=1}^{D} w_{1,a}w_{2,b}w'_{1,c}w'_{2,d} \int_{-\infty}^{+\infty} \phi\left(y\left|\mu_{a,b},\sigma_{a,b}\right.\right)\phi\left(y-z\left|\mu'_{c,d},\sigma'_{c,d}\right.\right) dy \\
&= \sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C}\sum_{d=1}^{D} w_{1,a}w_{2,b}w'_{1,c}w'_{2,d} f_{z_{a,b,c,d}}(z)
\end{aligned}$$

where $f_{z_{a,b,c,d}}(z)$ can be regarded as the PDF of $z_{a,b,c,d} = x_{a,b} - y_{c,d}$, in which $x_{a,b}$ and $y_{c,d}$ are the normal distribution $N\left(\mu_{a,b},\sigma_{a,b}^2\right)$ and $N\left(\mu'_{c,d},\sigma'^2_{c,d}\right)$ respectively. According to the additivity of normal distribution, $z_{a,b,c,d}$ is still a normal distribution and $z_{a,b,c,d} \sim N\left(\mu_{a,b} - \mu'_{c,d},\sigma_{a,b}^2 + \sigma'^2_{c,d}\right)$.

Therefore,

$$f_Z(z) = \sum_{a}^{A}\sum_{b}^{B}\sum_{c}^{C}\sum_{d}^{D} w_{1,a}w_{2,b}w'_{1,c}w'_{2,d}\,\phi\left(z\left|\mu_{a,b} - \mu'_{c,d}\,,\sqrt{\sigma_{a,b}^2 + \sigma'^2_{c,d}}\right.\right)$$

3.4. Problem reformulation

In order to calculate the KDE-CRPS shown in Eq. (10), $H - H'$ needs to be decided. According to Eq. (12), the combined PDF $f^K(x)$ in Eq. (5) can be treated as a 2-layer Gaussian mixture distribution because of $\sum_{q=1}^{Q}(1/Q) = 1$, $\sum_{i=1}^{N} w_i = 1$ and $w_i > 0$. Therefore $H - H'$ is the difference of two 2-layer Gaussian mixtures. Based on **Proposition 1** in Section 3.3, $H - H'$ is a 4-layer Gaussian mixture distribution. The PDF $f_{Z=H-H'}$ thus becomes:

$$f_{Z=H-H'}(z) = \frac{1}{Q^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sum_{q=1}^{Q}\sum_{q'=1}^{Q} \phi\left(z\left|\hat{y}_{i,q} - \hat{y}_{j,q'}\,,\sqrt{B_i^2 + B_j^2}\right.\right) \tag{14}$$

Moreover, the expectation of the absolute value of a Gaussian distribution $N\left(\mu,\sigma^2\right)$ is calculated as follows:

$$\begin{aligned} E\left|X\right| &= \int_{-\infty}^{+\infty} |x| f\left(x\right) dx \\ &= \int_{-\infty}^{0} -xf\left(x\right)dx + \int_{0}^{+\infty} xf\left(x\right)dx \\ &= 2\sigma\,\phi_s\left(\frac{\mu}{\sigma}\right) + \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right] \end{aligned} \tag{15}$$

where $\phi_s\left(\cdot\right)$ and $\Phi\left(\cdot\right)$ are the PDF and CDF of the standard Gaussian distribution.

Therefore,

$$\begin{aligned} E\left|H-H'\right| &= E\left|Z\right| \\ &= \frac{1}{Q^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sum_{q=1}^{Q}\sum_{q'=1}^{Q} \int_{-\infty}^{+\infty} |z| \cdot \phi\left(z \left| \hat{y}_{i,q} - \hat{y}_{j,q'}, \sqrt{B_i^2 + B_j^2}\right.\right) dz \\ &= \frac{1}{Q^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sum_{q=1}^{Q}\sum_{q'=1}^{Q} \left[2\sqrt{B_i^2+B_j^2}\cdot\phi_s\left(\frac{\hat{y}_{i,q}-\hat{y}_{j,q'}}{\sqrt{B_i^2+B_j^2}}\right) + \left(\hat{y}_{i,q}-\hat{y}_{j,q'}\right)\left(2\Phi\left(\frac{\hat{y}_{i,q}-\hat{y}_{j,q'}}{\sqrt{B_i^2+B_j^2}}\right)-1\right)\right] \end{aligned} \tag{16}$$

The $E\left|H-y\right|$ is:

$$\begin{aligned} E\left|H-y\right| &= \int_{-\infty}^{+\infty} |h| f_H\left(h+y\right) dh \\ &= \frac{1}{Q}\sum_{i=1}^{N}\sum_{q=1}^{Q} w_i \int_{-\infty}^{+\infty} |h|\,\phi\left(h\left|\hat{y}_{i,q}+y, B_i\right.\right) dh \\ &= \sum_{i=1}^{N}\frac{w_i}{Q}\sum_{q=1}^{Q}\left[2B_i\,\phi_s\left(\frac{\hat{y}_{i,q}-y}{B_i}\right) + \left(\hat{y}_{i,q}-y\right)\left(2\Phi\left(\frac{\hat{y}_{i,q}-y}{B_i}\right)-1\right)\right] \end{aligned} \tag{17}$$

Finally, the KDE-CRPS can be described explicitly:

$$CRPS\left(F^K, y\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} \alpha_{i,j,q,q'} w_i w_j + \sum_{i=1}^{N}\sum_{q=1}^{Q} \beta_{i,q} w_i \tag{18}$$

where

$$\alpha_{i,j,q,q'} = -\frac{\sqrt{B_i^2+B_j^2}}{Q^2}\cdot\phi_s\left(\frac{\hat{y}_{i,q}-\hat{y}_{j,q'}}{\sqrt{B_i^2+B_j^2}}\right) - \frac{\hat{y}_{i,q}-\hat{y}_{j,q'}}{2Q^2}\left(2\Phi\left(\frac{\hat{y}_{i,q}-\hat{y}_{j,q'}}{\sqrt{B_i^2+B_j^2}}\right)-1\right) \tag{19}$$

$$\beta_{i,q} = \frac{2B_i}{Q}\cdot\phi_s\left(\frac{\hat{y}_{i,q}-y}{B_i}\right) + \frac{\hat{y}_{i,q}-y}{Q}\left(2\Phi\left(\frac{\hat{y}_{i,q}-y}{B_i}\right)-1\right) \tag{20}$$

This is how the KDE-CRPS of probabilistic forecasts F^K after combining quantiles is calculated given the observation y. The discussion above is generalized to the entire time range $t = 1, 2, \ldots, T$. Then, the

total KDE-CRPS loss TL is calculated by:

$$\begin{aligned} TL &= \sum_{t=1}^{T} CRPS\left(F_t^K, y_t\right) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} \alpha_{i,j,q,q',t} w_i w_j + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{q=1}^{Q} \beta_{i,q,t} w_i \end{aligned} \tag{21}$$

Based on Eq. (21), the matrix expression of combination problem in KCGC becomes:

$$\begin{aligned} \min_{W} \quad & W^T G W + c^T W \\ \text{s.t.} \quad & \mathbb{1}^T W = 1 \\ & W \geq 0 \end{aligned} \tag{22}$$

where $W = [w_1, w_2, \ldots, w_N]^T$ is the optimal weights vector; G is a matrix with elements $G_{i,j} = \sum_{t=1}^{T}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} \alpha_{i,j,q,q',t}$, and $c = \left[\sum_{t=1}^{T}\sum_{q=1}^{Q} \beta_{1,q,t}, \sum_{t=1}^{T}\sum_{q=1}^{Q} \beta_{2,q,t}, \ldots, \sum_{t=1}^{T}\sum_{q=1}^{Q} \beta_{N,q,t}\right]^T$; $\mathbb{1}$ is a column vector whose elements are all equal to 1.

4. Base model generation and feature selection

In this section, four quantile regression (QR) models (QRNN, QRLSTM, QRGRU, QRMGM) are introduced as our base models in the combination framework. To improve the performance of these 4 QR-models, a feature selection method is provided to find a better set of representative input variables.

4.1. Quantile regression

Unlike traditional regression analysis which can only find the central trend of the dependent variable, quantile regression (QR) can infer its conditional probability distribution under the guidance of pinball loss (PL), which is defined as

$$L^P\left(y, \hat{y}\right) = \begin{cases} (y - \hat{y}) \times (1 - \tau) & y < \hat{y} \\ (y - \hat{y}) \times \tau & y \geq \hat{y} \end{cases} \tag{23}$$

In Eq. (23), $\hat{y}$ is the conditional quantile of the dependent variable y at $\tau\,(0 < \tau < 1)$ quantile.

For each quantile q and regression model i, QR-model $g_{i,q}\,(i = 1, 2, \ldots, N; q = 1, 2, \ldots, Q)$ can be generated by optimizing the following programming problem that minimizes PL:

$$\mathrm{W}_{i,q} = \arg\min_{\mathrm{W}_{i,q}} \sum_{t=1}^{T} L^P\left(y_t, g_{i,q}\left(\mathrm{X}_{i,t}, \mathrm{W}_{i,q}\right)\right) \tag{24}$$

where the parameter vector to be optimized is denoted by $\mathrm{W}_{i,q}$; $\mathrm{X}_{i,t}$ is the input feature vector and y_t is the real value at time t.

4.2. Quantile regression models

4.2.1. QRNN

Quantile regression neural network (QRNN) was put forward by Taylor [42]. It aims to overcome the shortcoming of the traditional linear model that is incapable of simulating the nonlinear relationship between variables. The structure of QRNN consists of an input layer, multiple hidden layers and an output layer. The neurons in each layer are fully connected to its previous layer. The output of each neuron in the hidden layer passes through the nonlinear activation function and then enters the next layer as the input.

4.2.2. QRLSTM

Long short-term memory (LSTM) is one of the variants of recurrent neural network (RNN) [43]. It introduces memory units and gated memory units to preserve historical information and long-term state and control the flow of information, which can effectively overcome the problem of gradient disappearance in RNN [44]. Therefore, it provides more accurate results for data with short-term or long-term dependence. Specifically, for each quantile q, the principle of LSTM forward propagation is as follows:

$$\begin{aligned}
\mathrm{I}_{t,q}^{L} &= \theta\left(\mathrm{W}_{\mathrm{I}^{L},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}_{t-1,q}^{L}\right]\right) \\
\mathrm{F}_{t,q}^{L} &= \theta\left(\mathrm{W}_{\mathrm{F}^{L},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}_{t-1,q}^{L}\right]\right) \\
\mathrm{C}_{t,q}^{L} &= \mathrm{F}_{t,q}^{L} * \mathrm{C}_{t-1,q}^{L} + \mathrm{I}_{t,q}^{L} * \tanh\left(\mathrm{W}_{\mathrm{C}^{L},q}\left[\mathrm{X}_{i,t}, \mathrm{H}_{t-1,q}^{L}\right]\right) \\
\mathrm{O}_{t,q} &= \theta\left(\mathrm{W}_{\mathrm{O},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}_{t-1,q}^{L}\right]\right) \\
\mathrm{H}_{t,q}^{L} &= \mathrm{O}_{t,q} * \tanh\left(\mathrm{C}_{t,q}^{L}\right) \\
\hat{y}_{i,t,q}^{L} &= \theta\left(\mathrm{W}_{\mathrm{L},q} \cdot \mathrm{H}_{t,q}^{L}\right)
\end{aligned} \tag{25}$$

where square brackets indicate that the two vectors are connected; asterisk indicates matrix multiplication; The symbol $\cdot$ indicates the product of matrix elements; $\theta(\cdot)$ and $\tanh(\cdot)$ represent the activation function of *sigmoid* and *tanh*; $\mathrm{W}_{\bullet,q}$ denotes parameter vector of certain unit; $\mathrm{I}_{t,q}^{L}$ and $\mathrm{F}_{t,q}^{L}$ are the input gate and the forget gate, which determine how much information goes into memory unit $\mathrm{C}_{t,q}^{L}$ at time t; The output gate $O_{t,q}$ is multiplied by the memory unit $\mathrm{C}_{t,q}^{L}$ with the *tanh* activation function to obtain the final updated information $\mathrm{H}_{t,q}^{L}$; $\mathrm{H}_{t,q}^{L}$ passes the last dense layer with the parameter $\mathrm{W}_{\mathrm{L},q}$, and finally outputs the quantile forecasting value $\hat{y}_{i,t,q}^{L}$.

4.2.3. QRGRU

Gated recurrent unit (GRU) is a simplified structure of LSTM [45]. The newly introduced update gate, denoted by $\mathrm{U}_{t,q}$, is equivalent to merging the input gate and the forget gate of the LSTM. The forget gates and the memory unit are replaced by the reset gate $\mathrm{R}_{t,q}$. More specifically, the principle of GRU forward

propagation is calculated as follows:

$$
\begin{aligned}
\mathrm{U}_{t,q} &= \theta\left(\mathrm{W}_{\mathrm{U},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}^{G}_{t-1,q}\right]\right) \\
\mathrm{R}_{t,q} &= \theta\left(\mathrm{W}_{\mathrm{R},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}^{G}_{t-1,q}\right]\right) \\
\tilde{\mathrm{H}}^{G}_{t,q} &= \tanh\left(\mathrm{W}_{\tilde{\mathrm{H}},q}\left[\mathrm{X}_{i,t}, \mathrm{R}_{t,q} * \mathrm{H}^{G}_{t-1,q}\right]\right) \\
\mathrm{H}^{G}_{t,q} &= (1-\mathrm{U}_{t,q}) * \mathrm{H}_{t-1,q} + \mathrm{U}_{t,q} * \tilde{\mathrm{H}}^{G}_{t,q} \\
\hat{y}^{G}_{i,t,q} &= \theta\left(\mathrm{W}_{\mathrm{L},q} \cdot \mathrm{H}^{G}_{t,q}\right)
\end{aligned}
\tag{26}
$$

Here, the information $\mathrm{H}^{G}_{t-1,q}$ of the previous moment and the new information $\tilde{\mathrm{H}}^{G}_{t,q}$ at present are used to generate the final output information $\mathrm{H}^{G}_{t,q}$. Similarly, the quantile forecasting result $\hat{y}^{G}_{i,t,q}$ is obtained by $\mathrm{H}^{G}_{t,q}$ through the dense layer.

4.2.4. QRMGM

Minimal gated memory (MGM) is a variant of LSTM to reduce training time and increase prediction accuracy [26]. Compared with LSTM, the input gate $\mathrm{I}^{M}_{t,q}$ and the forget gate $\mathrm{F}^{M}_{t,q}$ in MGM are coupled. Further more, the output gate in MGM is removed. The calculation steps of MGM forward propagation are as follows:

$$
\begin{aligned}
\mathrm{F}^{M}_{t,q} &= \theta\left(\mathrm{W}_{F^{M},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}^{M}_{t-1,q}\right]\right) \\
\mathrm{I}^{M}_{t,q} &= 1 - \mathrm{F}^{M}_{t,q} \\
\mathrm{C}^{M}_{t,q} &= \tanh\left(\mathrm{W}_{\mathrm{F}^{M},q} \cdot \left[\mathrm{X}_{i,t}, \mathrm{H}^{M}_{t-1,q}\right]\right) \\
\mathrm{H}^{M}_{t,q} &= \mathrm{F}^{M}_{t,q} * \mathrm{H}^{M}_{t-1,q} + \mathrm{I}^{M}_{t,q} * \mathrm{C}^{M}_{t,q} \\
\hat{y}^{M}_{i,t,q} &= \theta\left(\mathrm{W}_{\mathrm{L},q} \cdot \mathrm{H}^{M}_{t,q}\right)
\end{aligned}
\tag{27}
$$

Here, the hidden layer in MGM has only one set of weight matrix $\mathrm{W}_{\mathrm{F}^{M},q}$, compared with four in the LSTM and three in the GRU. This simplified structure in MGM can reduce a lot of calculation in the training of the model without harming prediction accuracy.

4.3. Incremental association markov blanket

Many feature selection methods consider features in isolation rather than as a whole, such as selection using Maximal Information Coefficient (MIC) [46]. They are problematic because the selected features are not used separately in the model training, but as a whole. In other words, the best performing feature selected by this type of methods may not work well with other selected features, while combining the rest of the features could outperform the best feature.

Therefore, we adopt Incremental Association Markov Blanket (IAMB) [47]. This is a Markov Blanket (MB) based feature subset selection algorithm. From the MB of a target T, denoted by $MB(T)$, we can determine a minimal set of features conditioned on which all other features are independent of the target

T. Koller et al. demonstrated that $MB(T)$ is the theoretically optimal set of features to predict the value of target T [48]. Thus, we can only use features in the $MB(T)$ instead of all the features for prediction. Meanwhile, IAMB considers cooperative relationships between features through the grow phase and shrink phase. Specifically, it first implements the grow phase to include all features that are dependent of a target variable, and then the shrink phase to remove invalid candidates in a separate step.

5. Model combination workflow

A more detailed implementation process for the proposed combination model is presented in this section, including data splitting, feature selection, base model construction, combination problem solution and combined model assessment. For a clearer illustration, the block diagram of the proposed model is given in Fig. 3.

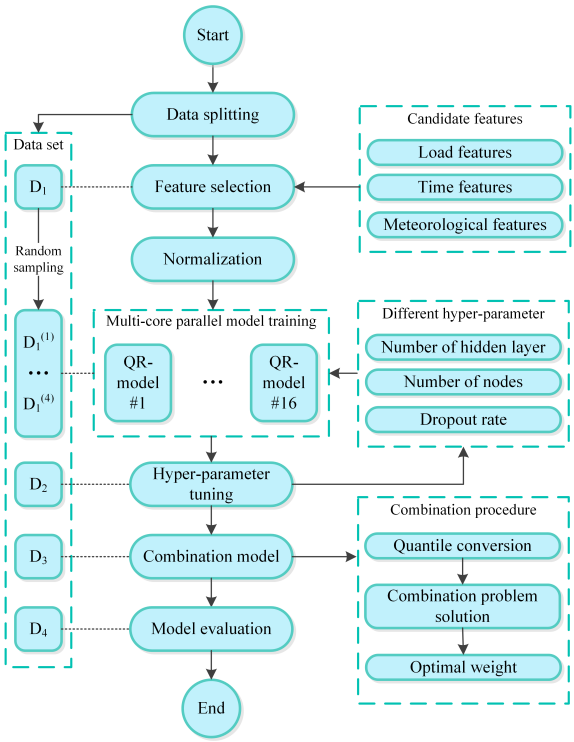

Fig. 3: The block diagram of the proposed combination model.

5.1. Data splitting

In order to evaluate the proposed combination method and avoid overfitting, we divide a data set into four parts, namely D_1, D_2, D_3 and D_4. Increasing the diversities of the base models can help to improve the forecasting performance of combined model [49]. One common way to achieve model diversity is to train

individual models with different sub-training sets through random sampling. This process of sampling is applied to D_1. In our case studies, the sampling proportion is 25% and four sub-training sets $\left(D_1^{(1)}, \ldots, D_1^{(4)}\right)$ are obtained by random sampling D_1 for 4 times. The D_2 set is used for hyper-parameter tuning. D_3 is used to determine component weights. The combined model is evaluated on D_4. Our splitting ratio is set to $D_1^{(m)} : D_2 : D_3 : D_4 = 10 : 1 : 1 : 1$, $m = 1, 2, 3, 4$.

5.2. Feature selection

The IAMB introduced in Section 4.3 is adopted to choose the input features. With the help of large sample size, the IAMB can produce more representative features. As a result, the selection is carried out on the D_1, rather than on the sub-training set. First, many alternative features are provided, including load features, time features and meteorological features, which are detailed in our case studies. Subsequently, IAMB selects a set of representative features from the alternative features for further QR-models training.

5.3. Parameter settings and parallel computing

There are several techniques that can improve the prediction of the QR-models and prevent overfitting, such as decayed learning rate [50], mini-batch mechanism [51], dropout [52] and multiple hidden layers. We include them into the models for our case studies. To diversify the base models, the hyperparameters including the dropout rate and the number of hidden layer and nodes are changed to feed to base models with different sub-training sets, while the remaining hyper-parameters are set to be the same in different models as shown in Table 1. The sub-training sets are normalized to eliminate the influence of dimension before model training. The models that perform well on D_2 are selected for further combination. In this paper, a total of 16 base models are picked, which is four of each type of QR models.

Table 1: Parameter settings.

Parameters	Details	Value	Value range
Decayed learning rate parameter	Initial learning rate	0.01	Common value [0.005,0.01,0.05,0.1,...]
	Decay rate	1.5	Common value [0.8,0.9,1.0,1.5,...]
	Decay steps	10	Common value [5,10,15,20,...]
	Minimum of learning rate	1.00E-04	A small value
Mini-batch parameter	Batch size	32	Common value [8,16,32,50,...]
	Epochs of training	200	Satisfying convergence

In the traditional ensemble framework, the training of the base models is carried out sequentially. Computational time increases significantly with the number of models. In order to improve the computational efficiency, parallel computing is used in the our framework to train the base models. This process takes place

on CPU with 16-thread in following cases, where the training tasks of each base model are independently run on different threads. Thus, 16 base models are built simultaneously.

5.4. Combination problem solution

After the construction of base models, the quantile prediction result $\{\hat{y}_{t,q}\}_{q=1,2,\ldots,Q}$ given by each of QR-models at each time t on D_3 is collected, and then the coefficient matrixes G and c in combination problem Eq. (22) are calculated. Thereinto, the choice of bandwidth B has an important effect on the final prediction results. The optimal bandwidth can be obtained by minimizing integrated mean squared error (IMSE). Silverman proposed a rule-of-thumb bandwidth estimator, which has been proved to be effective in many kernels [53]. Moreover, IMSE obtained by using Silverman's rule of thumb in different kernel functions has little difference. Thus, we adopt the rule to get the optimal choice for bandwidth, which can be calculated by:

$$B = \left(\frac{4\hat{\sigma}^5}{3Q}\right)^{\frac{1}{5}} \approx 1.05924\hat{\sigma}Q^{-\frac{1}{5}} \tag{28}$$

where $\hat{\sigma}$ is the standard deviation of a set of quantiles $\{\hat{y}_q\}_{q=1,2,\ldots,Q}$.

Although the problem (22) is a quadratic programming problem, it cannot be solved directly since all elements in G are negative according to Eq. (18). Therefore, we adopt the similar transformation proposed in [35] to make the problem (22) a convex quadratic programming problem. First, the matrixes in the objective function are divided into blocks:

$$G = \begin{bmatrix} G' & \tilde{G} \\ \tilde{G}^T & G_{N,N} \end{bmatrix}, W = \begin{bmatrix} W' \\ W_N \end{bmatrix}, c = \begin{bmatrix} c' \\ c_N \end{bmatrix} \tag{29}$$

where the shapes of G' and $\tilde{G}$ are $(N-1)\times(N-1)$ and $(N-1)\times 1$; W' and c' are column vectors of shape $(N-1)\times 1$; $G_{N,N}$, W_N and c_N are the last elements of the corresponding matrixes.

Therefore, equality constraint in Eq. (22) is transformed into:

$$W_N = 1 - \mathbb{1}^T W' \tag{30}$$

Then, two new coefficient matrixes are constructed:

$$\begin{aligned} G_{new} &= G' - \mathbb{1}\tilde{G}^T - \tilde{G}\mathbb{1}^T + G_{N,N}\mathbb{1}\mathbb{1}^T \\ c_{new} &= c' + 2\tilde{G} - 2G_{N,N}\mathbb{1} - c_N\mathbb{1} \end{aligned} \tag{31}$$

Substitute Eq. (30) and Eq. (31) into the problem Eq. (22), and the transformed problem is as follows.

$$\begin{aligned} \min_{W'} \quad & {W'}^T G_{new} W' + {c_{new}}^T W' \\ \text{s.t.} \quad & \mathbb{1}^T W' \leq 1 \\ & W' \geq 0 \end{aligned} \tag{32}$$

For convenience, some constants in the objective function are ignored in the transformation process.

Though it is not proven mathematically that G_{new} is positive definite, we observe that this characteristic exists in all following case studies. Thus, the global optimum can be obtained by solving the problem Eq. (32) in polynomial time. After determining the optimal weight, the combined PDF of N base models can be calculated by Eq. (5).

5.5. Evaluation of the combined model

5.5.1. Comparison design

To fully verify the performance of the proposed method, it is compared with other four model weighting methods in our experiment.

1. CRPS-guided combined model (CGC): To compare with KCGC, CRPS-guided combined model (CGC) proposed in [35] is considered in this work. Similarly, the CRPS is selected as the loss function in CGC. But it is a combination method in a parametric situation since they assume density distributions of the base models to be Gaussian distributed. Thus, unlike this paper, Gaussian approximation of quantiles (GAQ) is used in the their quantile conversion stage. the optimization problem in CGC can be formulated as:

$$\begin{aligned} \min \quad & \sum_{t=1}^{T} CRPS\left(\sum_{i=1}^{N} w_i F_{i,t}^{G}, y_t\right) \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1 \\ & w_i \geq 0, \ \ i = 1, 2, \ldots, N \end{aligned} \tag{33}$$

where $F_{i,t}^{G}$ is the converted output of i-th base model at time t by using GAQ.

2. MAPE-based model (MBM): The weights are determined by minimizing a similar problem as Eq. (1) with the objective function being mean absolute percentage error (MAPE) [54].

$$\begin{aligned} w_i &= \min_{w_i} \ \tfrac{1}{T} MAPE\left(\sum_{i=1}^{N} w_i \hat{y}_{i,t}^{P}, y_t\right) \\ &= \min_{w_i} \tfrac{1}{T} \sum_{t=1}^{T} \tfrac{\left|y_t - w_i \hat{y}_{i,t}^{P}\right|}{y_t} \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1 \\ & w_i \geq 0, \ \ i = 1, 2, \ldots, N \end{aligned} \tag{34}$$

where $\hat{y}_{i,t}^{P}$ is the point forecast result of i-th model at time t. In this paper, the median of quantile forecasts is selected as the point prediction result.

3. Simple average (SA): Each base model has the same weight as Eq. (35).

$$w_i = \frac{1}{N} \tag{35}$$

4. PL-weighted average (PLWA): The pinball loss (PL) is an indicator to represent the comprehensive performance of the QR-model. The smaller the indicator, the better the QR-model. Therefore, PLWA applies higher weights to models with smaller PL as Eq. (36).

$$w_i = \frac{1/\bar{L}_i^P}{\sum_{i=1}^{N} \left(1/\bar{L}_i^P\right)} \tag{36}$$

where $\bar{L}_i^P$ is the average PL and $\bar{L}_i^P = \frac{1}{T \times Q} \sum_{t=1}^{T} \sum_{q=1}^{Q} L^P\left(y_t, \hat{y}_{i,t,q}\right)$.

5.5.2. Method evaluation metric

After obtaining the optimal weight based on D_3, the performance of KCGC is verified on D_4 from three metrics: MAPE, mean absolute error (MAE) and CRPS. In particular, the results based on KCGC are compared to those of the combined models, including CGC, MBM, SA and PLWA, all of which are combined with the same QR-models but use different weights. As shown in Eqs. (33) and (21), the best combination of CGC and KCGC is based on GAQ and KDE, respectively. The median is not affected by the maximum and minimum extreme values, and as the point prediction result, it has well representative. In order to compare with the proposed model, the point forecasting results of CGC are the expectation of probability distribution obtained by GAQ, while those of other models are the median of probability distribution which are calculated using KDE. It should be noted that the probability distribution given by GAQ is the Gaussian distribution, so its expectation is also equal to the median.

6. Case studies

To validate the effectiveness of the proposed combination method, two case studies consider 24 h forecasting horizon to perform day-ahead load forecasting. Case 1 discusses the system load from ISO New England Inc. (ISO-NE) in the United States [55]. Case 2 uses data from electrician mathematical contest in modeling (EMCM) in China, which can be found in the attachment.

6.1. Data description

Case 1 contains 8 datasets collected from 8 regions ISO-NE operates: Maine (ME), New Hampshire (NH), Vermont (VT), Connecticut (CT), Rhode Island (RI), Southeast Massachusetts (SEMA), Western/Central Massachusetts (WCMA), and Northeast Massachusetts and Boston (NEMA), which includes hourly real-time load profiles, dry bulb temperature and dew point temperature from 2015 to 2019. Case 2 involves the load data sampled every 15 minutes and daily data of maximum temperature, minimum temperature, average temperature, humidity and rainfall. Their time span is from 2013 to 2014. There are only a few null values and outliers in the datasets for both cases, which are removed directly during data processing. Since

the resolution of load data and meteorological data in case 2 is inconsistent, the meteorological data are resampled to align the load data. Simultaneously, the whole processes of these case studies are performed in the computer with 16 threads and 32 GB RAM.

6.2. Case study 1

Using the data splitting method in Section 5.1, a total of 35030 data points from 2015 to 2018 are used as the training set D1. After random sampling for 4 times, each sub-training set $D_1^{(m)}$, $m = 1, \ldots, 4$, with 8400 data points is obtained. Each sub-training set is used to train the four QR-models described in Section 4.2, resulting in a total of $4 \times 4 = 16$ QR-models. The data from the first 15 weeks in 2019 are divided into 3 parts, each containing 840 data points and corresponding to D_2, D_3 and D_4 sequentially. The alternative features for load are lagged observations with 24, 25, 26, 27, 28, 48, 72, 96, 120, 144, and 168 hours lagged. The time features include hour of the day, day of the week, month of the year, and a binary feature which is assigned to 1 when the corresponding day is the weekend, otherwise equivalent to 0. The meteorological features include the dry bulb temperature and the dew point temperature. In the same case study, different datasets consider the same alternative features.

6.2.1. Results in the RI area

Table 2 shows numerical results in RI area, including the performances of sixteen QR-models in different metrics and the weights in the five combination methods. Concretely, the columns 2 to 4 provide the performances evaluated by the MAPE, MAE and CRPS as evaluated on set D_4. The last five columns show the weights given by different combination methods. The weight acquisition process of all combination methods is carried out on D_3.

The best model in D_4 is QRMGM1, with the lowest MAPE, MAE and CRPS, which are shown in bold. Although it is weighted in most combination methods, the value is relatively small. The reason is that the methods obtain the optimal weight on D_3, where QRMGM1 is not the best model. Moreover, Table 2 shows that the base models picked by CGC are also included in KCGC, because they have the same goals that find minimum CRPS. Since MBM is more concerned with minimizing MAPE after combination, the chosen models are different with other combination methods.

For SA and PLWA, in order to avoid being affected by the QR-model with poor performance, they chose the best five models in D_3, and then assigned weights according to Eq. (35) and (36) respectively.

Fig. 4 provides real loads through two weeks starting from Mar 21, 2019 and predictions of QR-models selected by KCGC. Obviously, no one can always perform best, especially at peak and trough load forecasting. Nevertheless, this makes it possible to reduce the overall prediction deviation when considering the prediction results of multiple models. In fact, that is exactly what KCGC does. Overall, the trend of the prediction

Table 2: QR-models' performances and their weights in different combination methods.

Model	MAPE (%)	MAE (MW)	CRPS (MW)	K-weights	C-weights	M-weights	S-weights	P-weights
QRGRU1	3.995	32.092	22.211	0	0	0.254	0	0
QRGRU2	4.807	37.074	26.425	0	0	0	0	0
QRGRU3	4.431	34.544	23.898	0	0	0	0	0
QRGRU4	5.777	42.087	37.045	0	0	0	0	0
QRLSTM1	4.058	32.344	22.652	0.226	0.241	0.314	0.200	0.190
QRLSTM2	4.115	32.604	22.240	0	0	0.075	0	0
QRLSTM3	4.422	34.485	23.936	0	0	0	0.200	0.196
QRLSTM4	5.484	41.327	27.534	0	0	0	0	0
QRMGM1	**3.621**	**29.095**	**19.877**	0.075	0.042	0	0.200	0.198
QRMGM2	4.128	31.908	23.231	0.273	0.269	0.335	0	0
QRMGM3	6.968	50.836	36.488	0	0	0	0	0
QRMGM4	7.283	55.723	40.654	0	0	0.022	0	0
QRNN1	3.876	30.359	21.007	0.424	0.448	0	0.200	0.211
QRNN2	4.288	33.053	22.983	0.001	0	0	0.200	0.204
QRNN3	6.167	46.087	33.384	0	0	0	0	0
QRNN4	6.849	51.039	36.402	0	0	0	0	0

Note: In the columns 2 to 4, smaller values are the better and the best value is in bold. The last five columns provide the weights of the QR-models in different combination methods, which sequentially correspond to KCGC, CGC, MBM, SA and PLWA. The weight greater than zero is highlighted with gray fill.

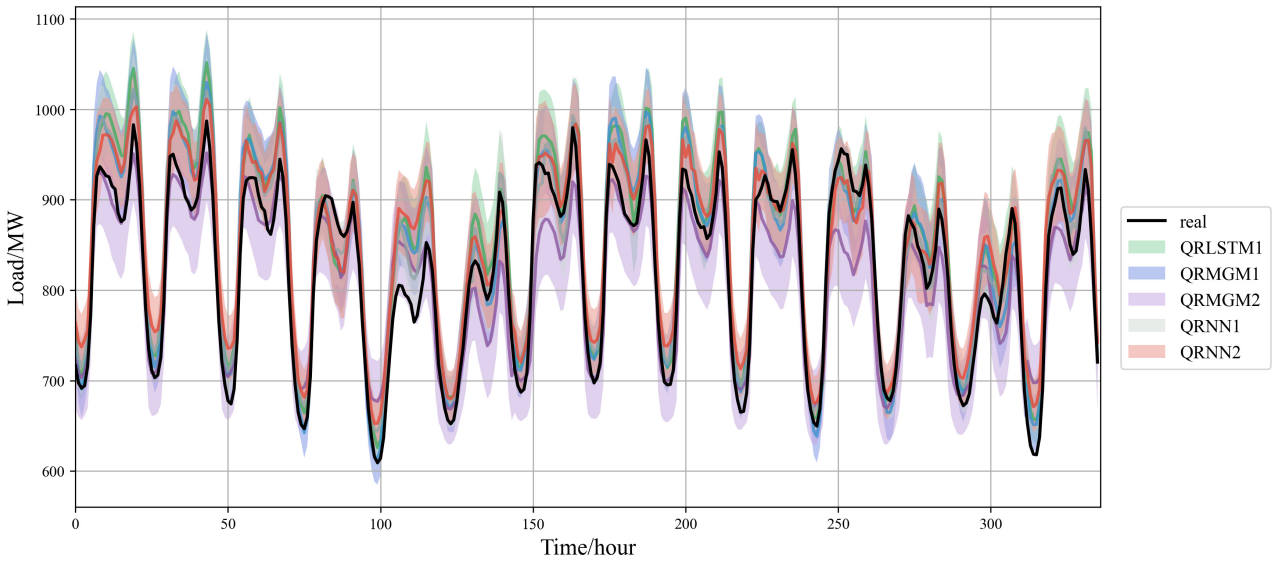

Fig. 4: Load forecasting for RI area by QR-models with weight greater than 0 in KCGC, from 21-Mar-2019 to 25-Mar-2019. *Note*: the solid color line is point forecasting results and the shaded area indicates 80% confidence interval forecasting results.

given by each QR-model is close to the real load and their fluctuation ranges cover the real loads most of the time, indicating that the predictions of five QR-models are reasonable.

Table 3: Performance metrics of combination methods.

Models	MAPE(%)	MAE(MW)	CRPS(MW)
KCGC	**3.342**	**25.845**	**18.509**
CGC	3.457	26.670	18.616
MBM	3.405	26.380	18.754
SA	3.956	31.007	21.020
PLWA	3.953	30.976	21.001

Note: Smaller values are the better. The best value in each column is in bold.

Table 3 shows the performance metrics of the different combination methods on D_4. It is shown that the KCGC, CGC and MBM perform better than the best QR-model (QRMGM1, with 3.621% MAPE, 29.095MW MAE, and 19.877MW CRPS), while SA and PLWA struggles with poor performance models in the base model set. KCGC, CGC and MBM are the combination methods based on overall optimization. When assigning weights, they do not just rely on the performance of a base model, but consider whether it can work better with other base models and produce the overall optimal result. In contrast, the combined

models, such as SA and PLWA, which only rely on the performance of the base model to determine the weights, may get much worse prediction results.

Using the best QR-model QRMGM1 as the benchmark, KCGC shows the largest improvement in all aspects compared to other combined methods, with 7.705% improvement in MAPE, 11.170% improvement in MAE and 6.882% improvement in CRPS. Notably, MBM aims to minimize MAPE of the overall ensemble and achieves indeed a higher MAPE improvement than CGC, but it is lower than that of KCGC. It shows that the combination strategy of KCGC can not only obtain better probability prediction results, but also better point prediction results. Although the optimization achieved by KCGC appears unremarkable on the surface compared to other combined models, it is still meaningful and worthy because the combined model needs to break through the performance limits of individual models. In particular, KCGC can get the best results in both point prediction and probabilistic prediction, which is a unique advantage.

6.2.2. Results in eight areas

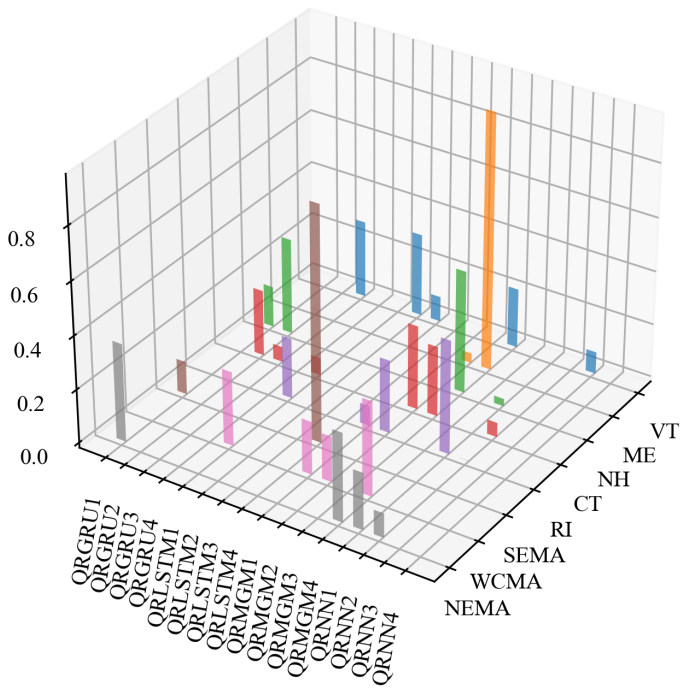

Fig. 5: Weights of the QR-models in KCGC in eight areas

The weight distributions of QR-models in KCGC in 8 areas are illustrated in Fig. 5. It can be seen that the number of QR-models included in KCGC is at least two and up to six, whose weights are significantly different. Table 4 compares the number of QR-models selected by different combined models in 8 areas. We can see that KCGC and CGC tend to select a larger number of QR-models in combination than MBM. This property may enable the former two models to gain more stability of combinations.

Table 4: The number of QR-models selected by different combined models

Models	VT	ME	NH	CT	RI	SEMA	WCMA	NEMA
KCGC	5	3	4	6	5	2	4	4
CGC	6	2	4	6	5	2	4	4
MBM	4	2	4	5	5	1	3	2
SA	5	5	5	5	5	5	5	5
PLWA	5	5	5	5	5	5	5	5

The results of each combined model and the best QR-model in each region are listed in Tables 5-7, where the minimum values are marked in bold and the "Best QR-model" column provides the value and the name of the corresponding QR-model. By comparing the three tables, it can be concluded that in different regions, the best QR-model varies. Even in the same region, the best individual model under different metrics also varies. In VT, for example, the QR-models with the minimum CRPS and MAPE are QRMGM3 and QRGRU3 respectively. However, KCGC is usually better than the best QR-model in each metric. In other words, KCGC combines the advantages of different models and shows better performance than any of individual model.

Table 5: CRPS of the best QR-model and combined models (MW)

Areas	Best QR-model	KCGC	CGC	MBM	SA	PLWA
VT	29.300 (QRMGM3)	**28.736**	28.977	29.132	29.133	29.119
ME	31.290 (QRMGM3)	**31.173**	31.360	31.365	40.096	39.155
NH	28.288 (QRGRU2)	**28.180**	28.617	29.356	29.438	29.405
CT	96.517 (QRMGM1)	**90.347**	90.508	90.892	102.521	104.003
RI	19.877 (QRMGM1)	**18.509**	18.616	18.754	21.020	21.001
SEMA	97.888 (QRLSTM4)	**97.234**	97.710	102.409	99.185	102.303
WCMA	55.836 (QRMGM2)	**50.081**	50.130	50.225	58.053	58.076
NEMA	59.676 (QRMGM1)	**57.230**	57.323	57.685	60.208	60.357

Note: Smaller values are the better. The best value in each row is in bold.

Figs. 6 and 7 illustrate the relative improvements of combined models in CRPS and MAPE respectively. By comparing the two graphs, we can find that KCGC, CGC and MBM exceed SA and PLWA in all areas. The reason is that SA and PLWA based on average weight only relies on the respective performance of the base model to assign weight, and does not consider the overall optimization when the base models combined with each other. As a result, models that perform well may not play well with others. By contrast, KCGC, CGC and MBM can find the cooperative relationship between the base models by solving the optimization

Table 6: MAPE of the best QR-model and combined models (%)

Areas	Best QR-model	KCGC	CGC	MBM	SA	PLWA
VT	7.703 (QRGRU3)	**7.648**	7.734	7.692	7.731	7.728
ME	3.531 (QRMGM3)	**3.515**	3.633	3.517	4.906	4.797
NH	3.339 (QRMGM3)	**3.307**	3.459	3.440	3.615	3.610
CT	4.660 (QRMGM2)	4.227	4.236	**4.136**	5.217	5.290
RI	3.621 (QRMGM1)	**3.342**	3.457	3.405	3.956	3.953
SEMA	**10.133 (QRMGM2)**	10.229	10.291	10.432	10.640	10.969
WCMA	4.569 (QRMGM2)	**3.840**	3.860	3.846	5.013	5.014
NEMA	3.372 (QRLSTM1)	3.341	3.377	**3.328**	3.575	3.601

Note: Smaller values are the better. The best value in each row is in bold.

Table 7: MAE of the best QR-model and combined models (MW)

Areas	Best QR-model	KCGC	CGC	MBM	SA	PLWA
VT	39.900 (QRGRU3)	39.587	39.891	39.808	39.502	**39.495**
ME	42.684 (QRMGM3)	**42.457**	43.740	42.482	55.560	54.374
NH	39.097 (QRMGM3)	**38.688**	40.382	39.936	39.920	39.873
CT	136.304 (QRMGM2)	122.589	122.735	**120.554**	148.300	150.156
RI	29.095 (QRMGM1)	**25.845**	26.670	26.380	31.007	30.976
SEMA	140.591 (QRLSTM4)	140.685	**140.239**	143.826	146.319	150.068
WCMA	77.617 (QRMGM2)	**65.351**	65.593	65.472	82.909	82.933
NEMA	84.367 (QRMGM4)	82.479	83.197	**82.399**	88.534	88.869

Note: Smaller values are the better. The best value in each row is in bold.

problem, leading to obtain more stable and effective combination results. They performed much better in CT, RI and WCMA regions than in other regions.

Although the improvement in some regions seems not very significant, any small optimization is worthy of attention, especially when the individual model has achieved good prediction results. In similar work [35], Li et al. also used the datasets from ISO-NE and their combination method can reach an improvement ratio of 1.90% to 7.36% compared with the best individual model and 0.13% to 1.01% compared with the best another combination model. In the case 1 of this paper, the proposed method has an improvement ratio of 0.37% to 10.31% compared with the best individual model and 0.10% to 1.53% compared the best another combination model (CGC). So the improvement degree of the proposed method is not trivial.

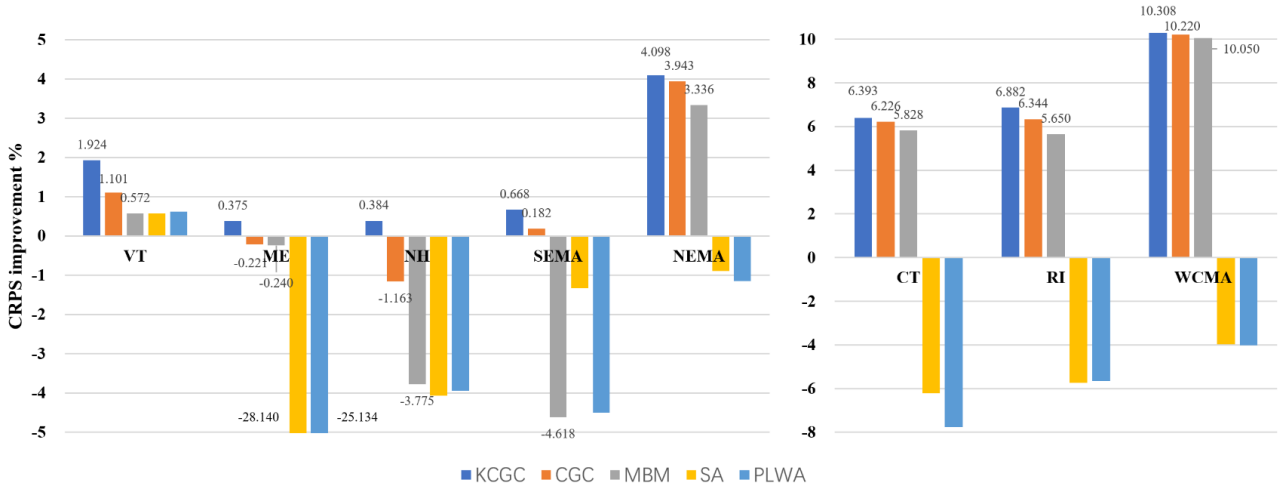

Fig. 6: Relative CRPS improvements of the combined models compared with the best local QR-models.

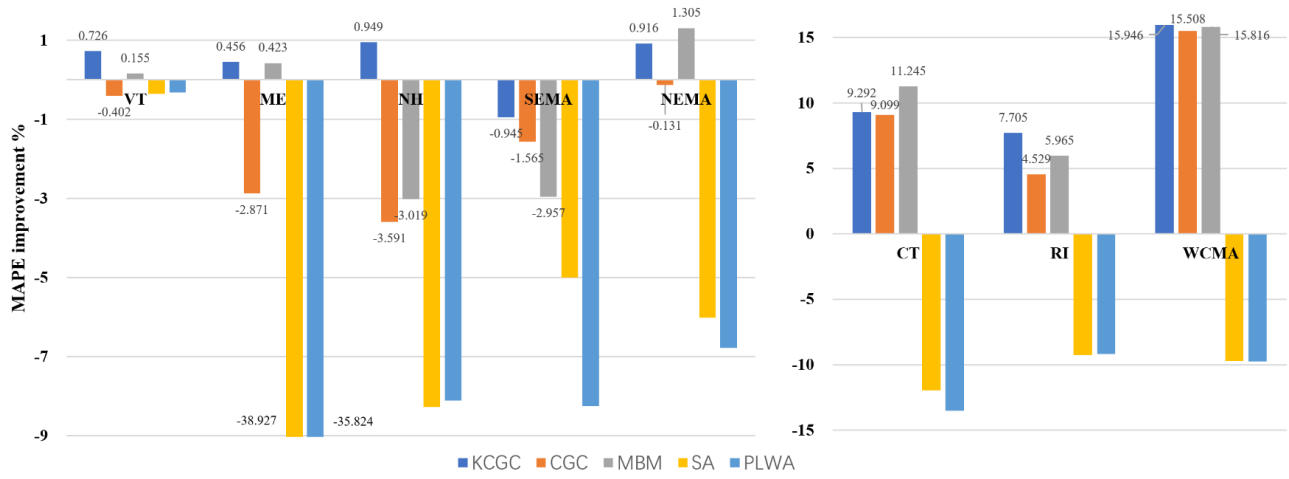

Fig. 7: Relative MAPE improvements of the combined models compared with the best local QR-models.

On average, KCGC gives the highest improvement rates among the five compared methods, with 3.879% in CRPS, 4.381% in MPAE and 5.196% in MAE. In addition, we can see that KCGC is more robust than CGC and MBM. In NH, CGC and MBM all fail, their CPRS and MAPE are inferior to the local best model, while KCGC has some improvement. Also, similar cases exist in other regions. Particularly, compared with CGC, KCGC can obtain lower MAPE under the premise of ensuring the maximum optimization of CRPS. As shown in Section 5.5.1, CGC requires the assumption that the base model obeys the Gaussian distribution. If this condition is not met, the results will be biased when CGC uses the GAQ to simulate the PDF of the base model, which in turn influences the process of combination and the final prediction. Although MBM outperforms KCGC in MAPE improvement in some regions, KCGC always achieves better CRPS in all regions.

6.3. Case study 2

Variations in load are strongly seasonal. To verify the applicability of the proposed method, case 2 is conducted to analyze the system load in summer and winter. The reason why we choose only summer and winter datasets is that they are more representative. Because temperature has a big impact on electricity demand, the peak load demand tends to occur during the two seasons of the year. Forecasting dataset containing the peak value is also more challenging and valuable. A total of 35040 data in 2013 from EMCM are used as the D_1. Like case 1, 8400 data points are randomly sampled from D_1 as a sub-training set and 16 QR-models are trained. The time spans covered by D_2, D_3 and D4 in summer and winter are shown in Table 8, and each contains 840 data points. The alternative features of load and time for IMAB are the same as case 1, while the meteorological features include maximum temperature, minimum temperature, average temperature, humidity and rainfall.

Table 8: The time spans of D_2, D_3, D_4 in summer and winter

Datasets	Span	D_2	D_3	D_4
Summer	Starting	2014/7/2 19:00	2014/7/11 13:00	2014/7/20 7:00
	Ending	2014/7/11 12:45	2014/7/20 6:45	2014/7/30 0:45
Winter	Starting	2014/11/27 13:45	2014/12/6 7:45	2014/12/15 1:45
	Ending	2014/12/6 7:30	2014/12/15 1:30	2014/12/23 19:30

Table 9 compares the performance metrics of the best QR-model and combined models in D_4. The "Number" row indicates the number of QR-models selected by different combined method. In D_4, KCGC has the best CRPS among them, and the improvement rates are 5.159% in summer data and 2.483% in winter data compared with the best QR-model. In MAPE, KCGC is second only to MBM in summer, but it reaches the best in winter. In contrast, the MAPE of CGC is not optimized in summer and is worse than KCGC in both seasons. SA and PLWA show some improvement, but far less than KCGC.

Table 9: Performance metrics of the best QR-model and combined models in D_4.

Datasets	Metrics	Best QR-model	KCGC	CGC	MBM	SA	PLWA
Summer	CRPS(MW)	409.806(QRNN4)	**388.664**	390.308	393.283	397.040	400.143
	MAPE(%)	5.910(QRNN4)	5.836	6.131	**5.828**	6.343	6.419
	MAE(MW)	527.673(QRNN4)	524.856	552.738	**523.303**	575.742	581.316
	Number	-	2	2	2	5	5
Winter	CRPS(MW)	99.589(QRMGM3)	**97.116**	97.982	99.761	99.391	98.411
	MAPE(%)	1.917(QRMGM3)	**1.852**	1.885	1.881	1.938	1.916
	MAE(MW)	**125.970(QRMGM3)**	126.292	127.797	127.963	130.810	129.329
	Number	-	6	6	4	5	5

Note: Smaller values are the better. The best value in each row is in bold.

Fig. 8 shows the PDFs of the QR-models, CGC and KCGC at four typical times in summer. The CGC and KCGC include the same base models, which are QRMGM3 and QRNN4. However, the PDFs of the same QR-model are different by estimating with different methods. Thereinto, when the predicted results of QRMGM3 are not Gaussian distribution, GAQ skews its PDF away from the real value and further leads to poor combinations in CGC. Whereas, since KDE can obtain the fine estimation regardless of the type of the distribution function, the combination given by KCGC is closer to the real. Fig. 9 depicts the prediction results of KCGC in winter dataset. It is clear that KCGC can precisely delineate the real load curve and obtain the appropriate interval forecasting results. At the same time, the prediction errors are evenly distributed on both sides of the zero value rather than continuously increasing, indicating that KCGC has robust prediction performance.

6.4. Robustness analysis

The prediction effectiveness of the model may be affected when training data present to be noisy. It is necessary to verify the robustness of the model in the case of data pollution. One common way to simulate data contamination in a real-world environment is to add perturbed data [56]. First, the dataset is normalized in the range (0,1). Second, 5% stochastic disturbances with uniform random numbers in [-0.05,0.05] are added to the normalized data.

The analysis experiments are conducted on four datasets, which are selected form CT and SEMC regions of ISO-NE and two datasets of ECMC. CT and SEMC are representative in case study 1 because they have a relative low and high percentage of CRPS optimization in KCGC respectively. Table 10 shows that the performance metrics between the best QR-model and combined models after adding perturbed data in four datasets. KCGC obtains the best CRPS in all four datasets and the best MAPE and MAE in three other

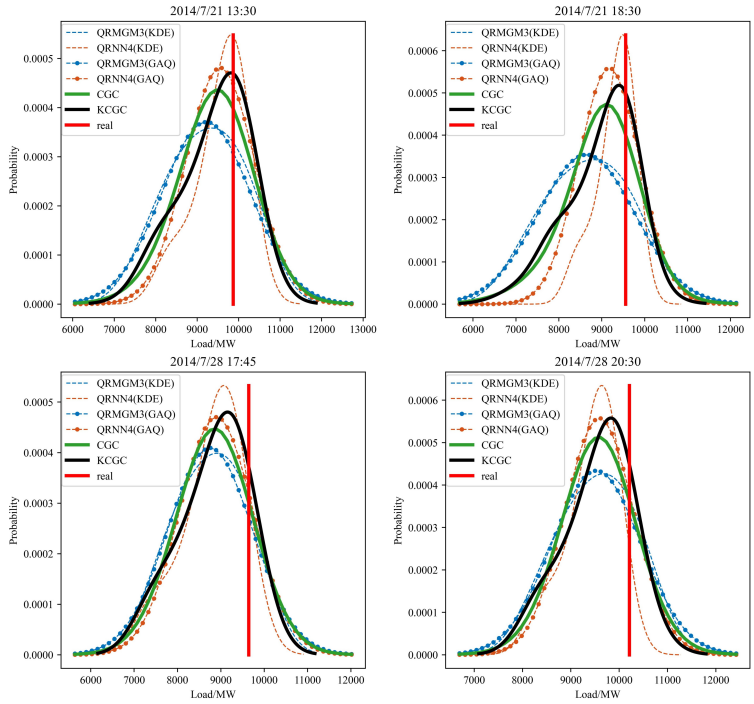

Fig. 8: The PDF graphs of loads from summer dataset for QR-models, CGC and KCGC at four different times. *Note*: the red line is real load and the other two solid lines are the PDFs predicted by CGC and KCGC respectively. Two types of dotted lines represent the PDFs obtained by the base model using different estimation methods, of which the one with circular mark is obtained by GAQ and the other is obtained by KDE.

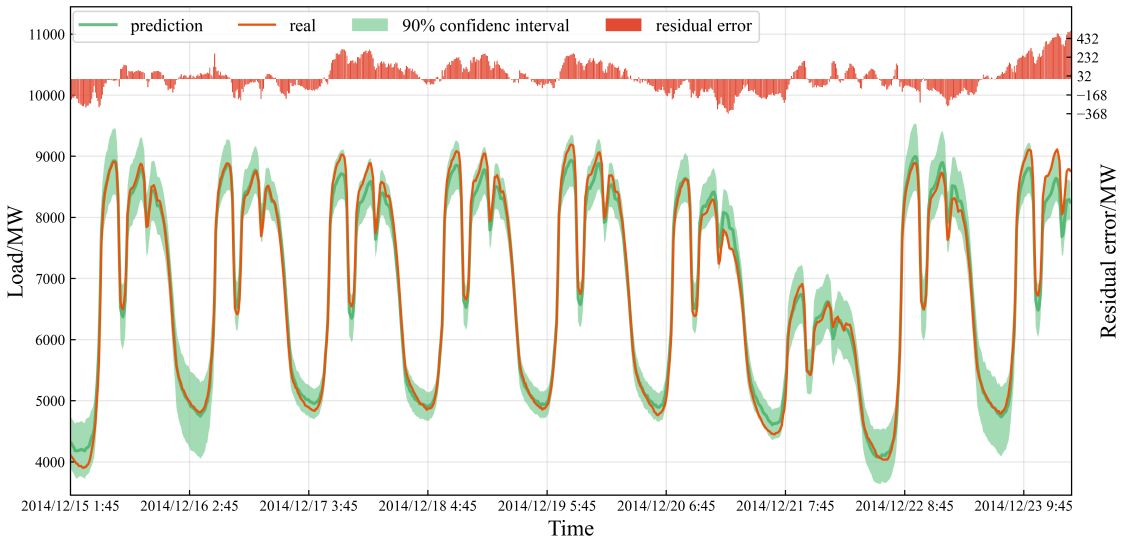

Fig. 9: Prediction results of KCGC in winter dataset.

datasets except CT region. In CT region, the MAPE and MAE of KCGC are only inferior to MBM. Thus, the KCGC model has good robustness and low sensitivity to perturbed data.

Table 10: Comparison of performance metrics after adding perturbed data.

Datasets	Metrics	Best QR-model	KCGC	CGC	MBM	SA	PLWA
CT (ISO-NE)	CRPS(MW)	152.804 (QRMGM1)	**150.211**	150.586	150.634	154.823	153.627
	MAPE(%)	7.785 (QRGRU2)	7.735	7.764	**7.704**	8.072	7.968
	MAE(MW)	219.526 (QRGRU2)	216.347	216.928	**215.599**	218.139	219.518
SEMA (ISO-NE)	CRPS(MW)	79.214 (QRMGM2)	**78.063**	78.248	78.450	82.607	81.941
	MAPE(%)	8.410 (QRGRU3)	**8.343**	8.394	8.381	9.015	8.957
	MAE(MW)	113.604 (QRMGM2)	**111.560**	111.887	112.035	113.301	113.253
Summer (EMCM)	CRPS(MW)	342.452 (QRNN1)	**338.627**	339.492	342.817	536.323	490.682
	MAPE(%)	5.464 (QRNN1)	**5.410**	5.441	5.427	7.991	7.378
	MAE(MW)	472.508 (QRNN1)	**471.334**	474.503	475.553	599.237	583.478
Winter (EMCM)	CRPS(MW)	224.821 (QRGRU1)	**222.489**	223.182	224.638	247.681	242.865
	MAPE(%)	5.105 (QRGRU1)	**5.067**	5.077	5.097	5.481	5.421
	MAE(MW)	324.364 (QRGRU1)	**322.526**	322.952	323.232	324.956	324.879

Note: Smaller values are the better. The best value in each row is in bold.

7. Conclusion

A novel combined model, named KCGC, is proposed for probabilistic forecasting in this paper, and is tailored to the nonparametric situation. KDE is introduced into the quantile conversion stage to obtain the best fit in the face of different quantiles. In order to combine the converted results, the CRPS integrated with KDE is selected as the objective function in KCGC. A complete theoretical deduction helps KCGC transform into a quadratic programming problem. Optimal weights are obtained by solving the quadratic programming problem. The whole combination process does not need distribution assumption and additional parameter estimation, which means that KCGC extends the combining probabilistic forecast problem to non-parametric environment.

Detailed empirical comparisons illustrate that no individual model is the best for all datasets. Even in the same data set, the best base model is not the same under different evaluation metrics. However, KCGC can integrate the strengths of base models and outperform any individual model. In addition, when common combination methods such as SA and PLWA fail, KCGC can still work well. Comparing other advanced combination methods (CGC and MBM), KCGC also can show its unique advantages: 1) Due to

the removal of distribution constraints on the predicted results of the base model, KCGC is able to achieve better performance in both probabilistic and deterministic predictions. 2) KCGC is robust to noisy data.

Future works will focus on the following three directions: (1) Consider other types of quantile forecasting models. Base models with good accuracy and diversity may further improve the combined model. (2) Use more elaborate combinations. Our study on combination stage is to assign weights for different base models. It is also worth investigating to determine weights for different quantiles or even each time point. (3) Apply to more energy forecasting problems. Since the combined model is performed on quantile results, it is not limited to only probabilistic load forecasting and can be extended to different forecasting problems, such as renewable energy forecasting involving wind power or solar irradiance.

Acknowledgements

This paper is funded by the National Natural Science Foundation (Nos.72171068, 71771073), and the Anhui Provincial Natural Science Foundation for Distinguished Young Scholars (2108085J36).

References

[1] Ding J, Wang M, Ping Z, Fu D, Vassiliadis VS. An integrated method based on relevance vector machine for short-term load forecasting. European Journal of Operational Research 2020;287(2):497–510.

[2] Guo C, Ye C, Ding Y, Wang P. A multi-state model for transmission system resilience enhancement against short-circuit faults caused by extreme weather events. IEEE Transactions on Power Delivery 2021;36(4):2374–85.

[3] Yang Y, Peng JCH, Ye C, Ye ZS, Ding Y. A criterion and stochastic unit commitment towards frequency resilience of power systems. IEEE Transactions on Power Systems 2022;37(1):640–52.

[4] Zhang W, Maleki A, Rosen MA. A heuristic-based approach for optimizing a small independent solar and wind hybrid power scheme incorporating load forecasting. Journal of Cleaner Production 2019;241:117920.

[5] Xie G, Chen X, Weng Y. Enhance load forecastability: Optimize data sampling policy by reinforcing user behaviors. European Journal of Operational Research 2021;.

[6] He Y, Liu R, Li H, Wang S, Lu X. Short-term power load probability density forecasting method using kernel-based support vector quantile regression and copula theory. Applied energy 2017;185:254–66.

[7] Afrasiabi M, Mohammadi M, Rastegar M, Afrasiabi S. Deep learning architecture for direct probability density prediction of small-scale solar generation. IET Generation, Transmission & Distribution 2020;14(11):2017–25.

[8] Yang X, Fu G, Zhang Y, Kang N, Gao F. A naive bayesian wind power interval prediction approach based on rough set attribute reduction and weight optimization. Energies 2017;10(11).

[9] Khosravi A, Nahavandi S, Creighton D. Construction of optimal prediction intervals for load forecasting problems. IEEE Transactions on Power Systems 2010;25(3):1496–503.

[10] Beyaztas U, Arikan BB, Beyaztas BH, Kahya E. Construction of prediction intervals for palmer drought severity index using bootstrap. Journal of Hydrology 2018;559:461–70.

[11] Zhang W, He Y, Yang S. Day-ahead load probability density forecasting using monotone composite quantile regression neural network and kernel density estimation. Electric Power Systems Research 2021;201:107551.

[12] Zhang S, Wang Y, Zhang Y, Wang D, Zhang N. Load probability density forecasting by transforming and combining quantile forecasts. Applied Energy 2020;277:115600.

[13] He Y, Xu Q, Wan J, Yang S. Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function. Energy 2016;114:498–512.

[14] Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. 2016.

[15] Nagy GI, Barta G, Kazi S, Borbély G, Simon G. Gefcom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. International Journal of Forecasting 2016;32(3):1087–93.

[16] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017;30:3146–54.

[17] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, p. 785–94.

[18] Zhang W, Maleki A, Rosen MA, Liu J. Sizing a stand-alone solar-wind-hydrogen energy system using weather forecasting and a hybrid search optimization algorithm. Energy conversion and management 2019;180:609–21.

[19] Gholipour Khajeh M, Maleki A, Rosen MA, Ahmadi MH. Electricity price forecasting using neural networks with an improved iterative training algorithm. International Journal of Ambient Energy 2018;39(2):147–58.

[20] He Y, Wang Y. Short-term wind power prediction based on eemd–lasso–qrnn model. Applied Soft Computing 2021;105:107288.

[21] Oreshkin BN, Dudek G, Pełka P, Turkina E. N-beats neural network for mid-term electricity load forecasting. Applied Energy 2021;293:116918.

[22] Xu X, Chen Y, Goude Y, Yao Q. Day-ahead probabilistic forecasting for french half-hourly electricity loads and quantiles for curve-to-curve regression. Applied Energy 2021;301:117465.

[23] He Y, Qin Y, Wang S, Wang X, Wang C. Electricity consumption probability density forecasting method based on lasso-quantile regression neural network. Applied energy 2019;233:565–75.

[24] Wang Y, Gan D, Sun M, Zhang N, Lu Z, Kang C. Probabilistic individual load forecasting using pinball loss guided lstm. Applied Energy 2019;235:10–20.

[25] Yang Y, Hong W, Li S. Deep ensemble learning based probabilistic load forecasting in smart grids. Energy 2019;189:116324.

[26] Zhang Z, Qin H, Liu Y, Yao L, Yu X, Lu J, et al. Wind speed forecasting based on quantile regression minimal gated memory network and kernel density estimation. Energy conversion and management 2019;196:1395–409.

[27] Mendes-Moreira J, Soares C, Jorge AM, Sousa JFD. Ensemble approaches for regression: A survey. Acm computing surveys (csur) 2012;45(1):1–40.

[28] Lahouar A, Slama JBH. Day-ahead load forecast using random forest and expert input selection. Energy Conversion and Management 2015;103:1040–51.

[29] Barrow DK, Crone SF. A comparison of adaboost algorithms for time series forecast combination. International Journal of Forecasting 2016;32(4):1103–19.

[30] Massaoudi M, Refaat SS, Chihi I, Trabelsi M, Oueslati FS, Abu-Rub H. A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting. Energy 2021;214:118874.

[31] Dudek G. Heterogeneous ensembles for short-term electricity demand forecasting. In: 2016 17th International Scientific Conference on Electric Power Engineering (EPE). IEEE; 2016, p. 1–6.

[32] Nowotarski J, Liu B, Weron R, Hong T. Improving short term load forecast accuracy via combining sister forecasts. Energy 2016;98:40–9.

[33] Hall SG, Mitchell J. Combining density forecasts. International Journal of Forecasting 2007;23(1):1–13.

[34] Bracale A, Carpinelli G, De Falco P. A probabilistic competitive ensemble method for short-term photovoltaic power forecasting. IEEE Transactions on Sustainable Energy 2017;8(2):551–60.

[35] Li T, Wang Y, Zhang N. Combining probability density forecasts for power electrical loads. IEEE Transactions on Smart

Grid 2019;11(2):1679–90.

[36] Taylor JW, Jeon J. Probabilistic forecasting of wave height for offshore wind turbine maintenance. European Journal of Operational Research 2018;267(3):877–90.

[37] Székely GJ, Rizzo ML. A new test for multivariate normality. Journal of Multivariate Analysis 2005;93(1):58–80.

[38] He Y, Zheng Y. Short-term power load probability density forecasting based on yeo-johnson transformation quantile regression and gaussian kernel function. Energy 2018;154:143–56.

[39] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 2007;102(477):359–78.

[40] Limaye D, Whitmore C. Selected statistical methods for analysis of load research data. final report. Tech. Rep.; Synergic Resources Corp., Bala-Cynwyd, PA (USA); 1984.

[41] Jin H, Shi L, Chen X, Qian B, Yang B, Jin H. Probabilistic wind power forecasting using selective ensemble of finite mixture gaussian process regression models. Renewable Energy 2021;174:1–18.

[42] Taylor JW. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of Forecasting 2000;19(4):299–311.

[43] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9(8):1735–80.

[44] Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with lstm recurrent networks. Journal of machine learning research 2002;3(Aug):115–43.

[45] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078 2014;.

[46] Reshef D, Reshef Y, Mitzenmacher M, Sabeti P. Equitability analysis of the maximal information coefficient, with comparisons. arXiv preprint arXiv:13016314 2013;.

[47] Zhang Y, Zhang Z, Liu K, Qian G. An improved iamb algorithm for markov blanket discovery. J Comput 2010;5(11):1755–61.

[48] Koller D, Sahami M. Toward optimal feature selection. Tech. Rep.; Stanford InfoLab; 1996.

[49] Cao Z, Wan C, Zhang Z, Li F, Song Y. Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting. IEEE Transactions on Power Systems 2019;35(3):1881–97.

[50] Tang Y. Deep learning using linear support vector machines. arXiv preprint arXiv:13060239 2013;.

[51] Jain P, Kakade S, Kidambi R, Netrapalli P, Sidford A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. Journal of Machine Learning Research 2018;18.

[52] Lee HW, Kim Nr, Lee JH. Deep neural network self-training based on unsupervised learning and dropout. International Journal of Fuzzy Logic and Intelligent Systems 2017;17(1):1–9.

[53] Silverman BW. Density estimation for statistics and data analysis. Routledge; 2018.

[54] Wang Y, Chen Q, Sun M, Kang C, Xia Q. An ensemble forecasting method for the aggregated load with subprofiles. IEEE Transactions on Smart Grid 2018;9(4):3906–8.

[55] Iso new england. Website; 2022. `https://www.iso-ne.com/isoexpress/web/reports/load-and-demand`.

[56] Luo L, Li H, Wang J, Hu J. Design of a combined wind speed forecasting system based on decomposition-ensemble and multi-objective optimization approach. Applied Mathematical Modelling 2021;89:49–72.