Optimal Operation Strategy Analysis with Scenario Generation Method Based on Principal Component Analysis, Density Canopy, and *K*-medoids for Integrated Energy Systems

Bingtuan Gao, Yunyu Zhu, and Yuanmei Li

***Abstract*—The operation of integrated energy systems (IESs) is confronted with great challenges for increasing penetration rate of renewable energy and growing complexity of energy forms. Scenario generation is one of ordinary methods to alleviate the system uncertainties by extracting several typical scenarios to represent the original high-dimensional data. This paper proposes a novel representative scenario generation method based on the feature extraction of panel data. The original high-dimensional data are represented by an aggregated indicator matrix using principal component analysis to preserve temporal variation. Then, the aggregated indicator matrix is clustered by an algorithm combining density canopy and *K*-medoids. Together with the proposed scenario generation method, an optimal operation model of IES is established, where the objective is to minimize the annual operation costs considering carbon trading cost. Finally, case studies based on the data of Aachen, Germany in 2019 are performed. The results indicate that the adjusted rand index (ARI) and silhouette coefficient (SC) of the proposed method are 0.6153 and 0.6770, respectively, both higher than the traditional methods, namely *K*-medoids, *K*-means++, and density-based spatial clustering of applications with noise (DBSCAN), which means the proposed method has better accuracy. The error between optimal operation results of the IES obtained by the proposed method and all-year time series benchmark value is 0.1%, while the calculation time is reduced from 11029 s to 188 s, which verifies that the proposed method can be used to optimize operation strategy of IES with high efficiency without loss of accuracy.**

***Index Terms*—Scenario generation, principal component analysis (PCA), density canopy, *K*-medoids, integrated energy system.**

I. Introduction

WITH the increasing shortage of fossil energy and environmental pollution problems, the development of renewable energy has become an important means to solve the energy crisis. But the volatility and intermittency of renewable energy bring great challenges to the safe and stable operation of power systems, which also limit its large-scale consumption [1], [2]. Therefore, the integrated energy system (IES), which can promote the consumption of renewable energy and improve the energy efficiency, has been vigorously promoted and applied [3].

IES is a multi-energy system integrating unified planning and dispatch of electricity, gas, cooling, and heating [4]. Within the IES, there are distributed renewable energy sources such as photovoltaic (PV) and wind power, whose power output fluctuates randomly, and the load demand for cities and towns varies greatly from place to place and from time to time. Therefore, extracting features from historical power output and load data and generating typical scenarios can well reflect the complex operational characteristics of the IES [5]. And scenarios generated also have important practical application value in the scheduling, planning, and operation optimization of the IES [6], [7].

An IES historical operation dataset is a large collection of scenarios with various uncertainty factors, which needs to be streamlined into a small and representative set of scenarios. Using the typical scenarios for IES scheduling and optimal operation analysis can reduce the computational scale of the optimization model without affecting the accuracy. Typical scenario generation is mainly summarized as the following three methods [8], [9].

1) Typical day method, which selects a typical day as the scenario based on experience and operation object, and usually chooses the day with the largest peak-to-valley load difference as the typical day [10], [11].

2) Time series production method, which aims to simulate the time series of renewable energy output and load characteristics, and thus presents the actual grid operation scenarios [12], [13].

3) Clustering method, which can be divided into two steps: first, the scenario generation step is adopted to obtain a large number of scenarios and their corresponding probabilities, and then the scenario reduction step is used to obtain typical scenarios. In [14], a stochastic optimal operation

Manuscript received: October 18, 2022; revised: March 23, 2023; accepted: June 9, 2023. Date of CrossCheck: June 9, 2023. Date of online publication: August 8, 2023.

This work was supported by the State Grid Corporation of China "Research and Demonstration on Key Technologies of Distributed Energy Supply System with Complementary Renewable Energy" (No. 5230HQ19000J).

B. Gao (corresponding author), Y. Zhu, and Y. Li are with School of Electrical Engineering, Southeast University, Nanjing, China (e-mail: gaobingtuan@seu.edu.cn; 2797255229@qq.com; 230179744@seu.edu.cn).

DOI: 10.35833/MPCE.2022.000681

model based on multi-scenario simulation for IES was proposed, where the scenario generation method is based on a Latin hypercube sampling operation, and the scenario reduction method is based on the *K*-means clustering. In [15], Copula function was used to build a joint output model of time series from multiple wind farms. A large initial scenario set was generated by probabilistically sampling and splicing from the Copula model, and the *K*-means clustering algorithm was used to reduce the scenarios to generate typical joint output scenarios.

The clustering method described above is commonly used for reasons of computational efficiency and accuracy. If sufficient historical data are available, the first step of the method can be omitted. The second step is adopted to aggregate similar scenarios based on specific metrics such as probability, hourly magnitude, or cost for each scenario [16]. The typical scenarios can be used in the planning and optimal operation of IESs. Therefore, making typical scenarios contain the maximum amount of historical data information is the key issue in study of the scenario generation method. Reference [17] introduced the concept of correlation loss. A correlation loss weight of the proposed scenario reduction framework was used to balance the minimal correlation loss and the maximal similarity between the original scenario set and the reduced scenario set. Reference [18] aimed at minimizing the partial correlation loss and maximizing the probabilistic similarity degree before and after reduction. Reference [19] proposed a weighted clustering method to extract the extreme scenarios in a system, with the temporal variations and correlations between wind power and load considered. Clustering is a normal way to realize scenario reduction, which is mainly classified as hierarchy-based, division-based, grid-based, density-based, and model-based methods [20].

Currently, few researchers have focused on generating typical IES scenarios incorporating renewable energy and load uncertainty. In addition, statistical methods are commonly used to deal with high-dimensional historical data during scenario generation, which means that temporal changes are often ignored. This paper proposes a combined scenario generation method that aims at preserving temporal variation of the original scenario, which consists of five indicators, namely PV power, wind power, electric load, heat load, and gas load. The adjusted rand index (ARI) and silhouette coefficient (SC) are introduced to verify the effectiveness of the proposed method in this paper and the typical scenarios are applied in the optimization operation of the IES to demonstrate their practical application value. The main contributions of this paper can be briefly summarized as follows.

1) A novel typical scenario generation method based on the principal component analysis (PCA), density canopy, and *K*-medoids is proposed. PCA is first adopted to extract the feature of the high-dimensional data, then the density canopy is used to obtain the cluster number and cluster center, and finally the typical scenarios are implemented by *K*-medoids with the optimized cluster number and cluster center.

2) Together with typical scenarios considering the temporal variation obtained by the proposed method, an optimal operation model of an IES is established to minimize the annual operational cost including the purchase cost of electricity and gas, the operation and maintenance costs, the depreciation cost, and the cost of carbon trade.

3) Extensive case studies are conducted on the proposed method and the optimal operation strategy analysis of an IES based on the data of Aachen, Germany in 2019. Two cluster validity indexes are introduced to demonstrate the effectiveness of the proposed method. And overall simulation results verify the effectiveness and advantages of the proposed method.

The rest of the paper is organized as follows. Section II presents the proposed typical scenario generation method. Section III illustrates the mathematical model of the operation for IESs. Section IV presents the results of the studied cases. Finally, conclusions are described in Section V.

II. Proposed Typical Scenario Generation Method

A. Construction of Typical Scenarios

Panel data are a kind of multi-dimensional data that combines cross-sectional data and time-series data, which form a plane whenever arranged in the cross-sectional or time-series dimension, which looks like a panel as a whole [21]. Historical power output and load data of an IES over a year correspond to the feature of panel data since the data can be divided into two dimensions. From the cross-sectional dimension, the whole data are divided by the index value which includes wind power, PV power, electric demand, heat demand, and gas demand. From the time-series dimension, the sampling time is 1 hour. The historical daily data of an IES include 24 hourly data, while the hourly data are composed of index value corresponding to power output and energy demand.

Traditional clustering methods are difficult to implement for clustering this type of data. To facilitate clustering analysis, the dimensionality reduction of the original data is required. When constructing typical scenarios of an IES, a certain statistic of an index (e.g., electric demand, heat demand, gas demand) is commonly adopted to represent the all-day time-series values, which reduces the dimensionality from the perspective of time sequence and then clusters the statistic of all-day data using common clustering methods. In this section, in order to preserve the time-series characteristics of the historical data, dimensionality reduction is realized from the perspective of indexes rather than time series. PCA is adopted to construct an aggregated indicator for the scenario division to reduce the dimensionality of the original indexes. After the above processing, aggregated indicators of the all-year time series could be obtained and then the typical scenarios could be constructed using clustering methods.

B. Feature Extraction of Historical Data

PCA, also known as K-L transformation, is adopted for the feature extraction. First, the data are pre-processed to form the multi-indicator panel data of the IES, which in-

clude all-year data of wind power and PV output as well as electric, gas, and heat demands. Suppose that the original data are $x_{is}(t_n)$ $(i=1,2,\ldots,k;\ \ s=1,2,\ldots,m;\ \ t_n=t_1,t_2,\ldots,t_p)$, where i is the sequence number of the day (ranging from 1 to 365); s is the specific index (i.e., wind power output, PV output, electric demand, gas demand, and heat demand); and t_n is the time sequence (ranging from 1 to 24). Then, the original data set is formed as:

$$\boldsymbol{X}_{is}(t_n)=\begin{bmatrix} x_{11}(t_1) & x_{12}(t_1) & \cdots & x_{1m}(t_1) \\ x_{11}(t_2) & x_{12}(t_2) & \cdots & x_{1m}(t_2) \\ \vdots & \vdots & & \vdots \\ x_{11}(t_p) & x_{12}(t_p) & \cdots & x_{1m}(t_p) \\ x_{21}(t_1) & x_{22}(t_1) & \cdots & x_{2m}(t_1) \\ \vdots & \vdots & & \vdots \\ x_{k1}(t_p) & x_{k2}(t_p) & \cdots & x_{km}(t_p) \end{bmatrix} \tag{1}$$

The mean matrix of the original data set corresponding to the dimension of the index can be formed as:

$$\bar{\boldsymbol{X}}_{\cdot s}(t_n)=\begin{bmatrix} \bar{x}_{\cdot 1}(t_1) & \bar{x}_{\cdot 2}(t_1) & \cdots & \bar{x}_{\cdot m}(t_1) \\ \bar{x}_{\cdot 1}(t_2) & \bar{x}_{\cdot 2}(t_2) & \cdots & \bar{x}_{\cdot m}(t_2) \\ \vdots & \vdots & & \vdots \\ \bar{x}_{\cdot 1}(t_p) & \bar{x}_{\cdot 2}(t_p) & \cdots & \bar{x}_{\cdot m}(t_p) \end{bmatrix} \tag{2}$$

$$\bar{x}_{\cdot s}(t_n)=\frac{1}{m}\sum_{s=1}^{m}x_{is}\ (t_n) \tag{3}$$

Then, the correlation matrix $\boldsymbol{R}(t_n)$ corresponding to the time sequence is calculated as:

$$\boldsymbol{R}(t_n)=\begin{bmatrix} r_{11}(t_1) & r_{12}(t_1) & \cdots & r_{1m}(t_1) \\ r_{11}(t_2) & r_{12}(t_2) & \cdots & r_{1m}(t_2) \\ \vdots & \vdots & & \vdots \\ r_{11}(t_p) & r_{12}(t_p) & \cdots & r_{1m}(t_p) \\ r_{21}(t_1) & r_{22}(t_1) & \cdots & r_{2m}(t_1) \\ \vdots & \vdots & & \vdots \\ r_{m1}(t_p) & r_{m2}(t_p) & \cdots & r_{mm}(t_p) \end{bmatrix} \tag{4}$$

$$r_{ab}(t_n)=\frac{1}{k-1}\sum_{i=1}^{k}x_{ia}^{*}(t_n)x_{ib}^{*}(t_n) \tag{5}$$

$$x_{is}^{*}(t_n)=\frac{x_{is}(t_n)-\bar{x}_{\cdot s}(t_n)}{\sqrt{var_{\cdot s}(t_n)}} \tag{6}$$

where $var_{\cdot s}(t_n)$ is the variance of $x_{\cdot s}$.

The eigenvalues and eigenvectors of the correlation matrix $\boldsymbol{R}(t_n)$ as well as the contribution rate of each eigenvalue are calculated, and then the cumulative contribution rate is obtained by summing up each contribution rate. Thus, the linear expression of the principal components can be achieved, and the appropriate number of principal components are selected according to their cumulative contribution rates. The linear expressions of the selected l principal components are:

$$y_i(t_n)=\boldsymbol{x}_i^{\mathrm{T}}(t_n)\boldsymbol{\xi}_i(t_n)\quad i=1,2,\ldots,l;n=1,2,\ldots,p \tag{7}$$

where $\boldsymbol{x}_i$ is the original data of the i^{th} principal component; and $\boldsymbol{\xi}_i$ is the eigenvector of the i^{th} principal component.

An aggregated indicator of the specific time sequence can be obtained through the above process, and the aggregated indicator matrix of original scenarios is established. Suppose that $F_i(t_n)$ is the aggregated indicator of the t_n^{th} hour in the i^{th} day, which is expressed as:

$$F_i(t_n)=\alpha_1(t_n)y_1(t_n)+\alpha_2(t_n)y_2(t_n)+\ldots+\alpha_l(t_n)y_l(t_n) \tag{8}$$

$$\alpha_j(t_n)=\frac{\lambda_j(t_n)}{\sum_{j=1}^{l}\lambda_j(t_n)}\quad j=1,2,\ldots,l;n=1,2,\ldots,p \tag{9}$$

where $\alpha_j(t_n)$ is the contributed rate; and $\lambda_j(t_n)$ is the eigenvalue.

Then, the aggregated indicator matrix for the typical scenario generation of the IES is expressed as:

$$\boldsymbol{F}=\begin{bmatrix} F_1(t_1) & F_1(t_2) & \cdots & F_1(t_p) \\ F_2(t_1) & F_2(t_2) & \cdots & F_2(t_p) \\ \vdots & \vdots & & \vdots \\ F_k(t_1) & F_k(t_2) & \cdots & F_k(t_p) \end{bmatrix} \tag{10}$$

C. Clustering of Feature Data

An improved *K*-medoids algorithm coupled with density canopy is proposed to realize the clustering. The process of classic clustering methods like *K*-means and *K*-medoids is rather slow and the results are easily affected by the initial value. With the combination of canopy, which is an unsupervised and fast approximate "coarse" clustering algorithm proposed by Mccallum [22], the effect of the initial value selection on the clustering results is alleviated [23]. However, in classic canopy, two artificially set distance thresholds will affect the clustering results. If the distance threshold is set to-be too large, samples that belong to different clusters will be incorrectly classified into the same class, and if the distance threshold is set to be too small, samples that belong to the same cluster will be classified into different clusters. However, when the density canopy algorithm is adopted, setting the values of predefined distance threshold is unnecessary.

Several principals need to be defined beforehand. First, the Euclidean distance between sample F_i and sample F_j is expressed as:

$$d(F_i,F_j)=\sqrt{\sum_{n=1}^{p}(F_i(t_n)-F_j(t_n))^2} \tag{11}$$

Then, the average distance of all the samples in the data set is expressed as:

$$d_{\text{mean}}=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}d(F_i,F_j) \tag{12}$$

The density of sample F_i refers to the number of all samples, whose distance away from F_i is smaller than d_{mean}, which is expressed as:

$$\rho(F_i)=\sum_{j=1}^{n}f(d(F_i-F_j)-d_{\text{mean}}) \tag{13}$$

$$f(x)=\begin{cases}1 & x<0\\ 0 & x\geq 0\end{cases} \tag{14}$$

The improved *K*-medoids algorithm based on density canopy can be conducted based on the above definitions. The spe-

cific steps of density canopy are as follows.

Step 1: calculate the density of all the samples in the scenario set F, and select the sample with the largest density value to be the first cluster center, which is expressed as c_1. Then, c_1 is added to the set of cluster centers, $C=\{c_1\}$. All the samples in set F whose distance away from c_1 is smaller than d_{mean} will be removed.

The average distance of all the samples, whose center is c_1, is:

$$\partial(F_i)=\frac{2}{\rho(F_i)(\rho(F_i)-1)}\sum_{i=1}^{\rho(F_i)}\sum_{j=i+1}^{\rho(F_i)}d(F_i,F_j) \tag{15}$$

where a smaller value of $\partial(F_i)$ means that the samples within the cluster are closer and more similar.

Intraclass distance $s(F_i)$ is the distance between samples F_i and F_j with a higher local density, which is expressed as:

$$s(F_i)=\begin{cases}\min\{d(F_i,F_j)\} & \exists F_j,\rho(F_j)>\rho(F_i)\\ \max\{d(F_i,F_j)\} & \forall F_j,\rho(F_j)<\rho(F_i)\end{cases} \tag{16}$$

Then, with the above definitions, the density weight of a sample F_i can be defined as:

$$\omega(F_i)=\frac{\rho(F_i)s(F_i)}{\partial(F_i)} \tag{17}$$

Step 2: the selection of the next cluster center is based on the product weight, and the second cluster center c_2 is the remaining sample that has the biggest product weight. Add c_2 into set C, thus $C=\{c_1,c_2\}$. Remove the samples whose distance away from c_2 is smaller than d_{mean} afterward.

Step 3: after the above processing, calculate $\rho(F_i)$, $\partial(F_i)$, and $s(F_i)$ of all the remaining samples in data set F. Then, the product weights could be calculated, which are $\omega(F_i,c_1)$ and $\omega(F_i,c_2)$. The third cluster center c_3 will be the sample that has the biggest value of $\omega(F_i,c_1)\omega(F_i,c_2)$. Add c_3 into set C, thus $C=\{c_1,c_2,c_3\}$. Similarly, remove the samples whose distance away from c_3 is smaller than d_{mean} afterward.

Step 4: among the remaining samples, if there exists a sample F_j satisfying $\max\{\omega(F_j,c_1)\omega(F_j,c_2)\cdot\ldots\cdot\omega(F_j,c_{k-1})\}$, then F_j will be set as the k^{th} cluster center c_k. Add c_k into set C, thus $C=\{c_1,c_2,\ldots,c_k\}$. Meanwhile, remove the samples whose distance away from c_k is smaller than d_{mean}.

Step 5: repeat the above steps to find all the remaining cluster centers that satisfy the conditions in *Step 4*, and add them one by one to the centroid set C. Then, remove the remaining samples from the data set F that have a distance less than d_{mean} between the corresponding cluster centers. Repeat the process until the data set is empty.

With the above steps, the pre-clustering is completed. And the optimal value k and the initial cluster center could be obtained. Then, the K-medoids [24] is carried out to obtain the final clusters, whose specific steps are as follows.

Step 1: the K-medoids clustering is based on the number of cluster centers and clusters obtained from the density canopy algorithm. Calculate Euclidean distance between sample F_i and each cluster center, then the sample will be assigned to the cluster.

Step 2: for an obtained cluster i, calculate the sum of the Euclidean distances between each sample and the other samples separately. And the sample with the smallest sum of distances is selected as the new centroid.

Step 3: repeat *Steps 1* and *2* until the cluster centers no longer change.

The flow chart of the above process is shown in Fig. 1.

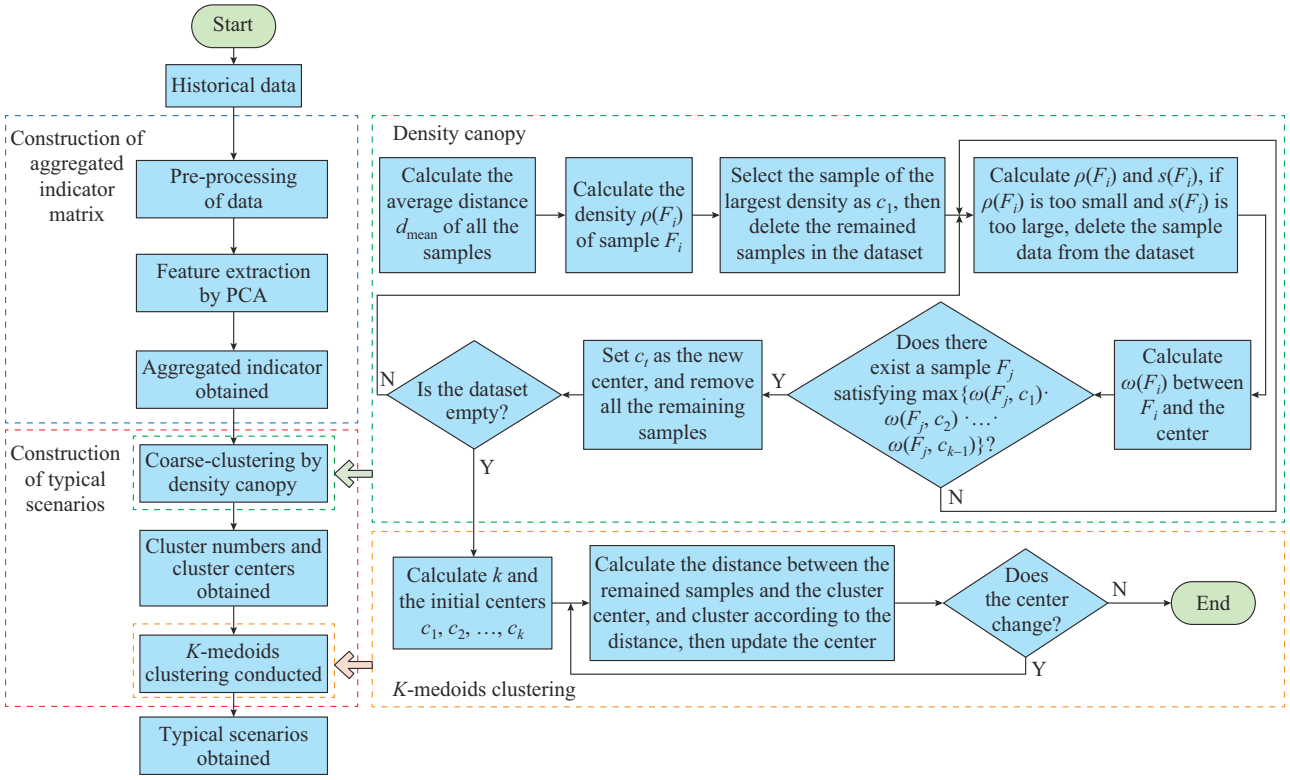

Fig. 1. Flow chart of scenario generation method.

III. Mathematical Model of Operation for IES

To illustrate the proposed scenario generation method and the optimal operation strategy, an IES of a single energy hub is adopted, whose structure is shown in Fig. 2. The original scenarios and typical scenarios are input separately to perform the optimal operation of the energy hub [25], with the time scale being one year. And the optimal operation model of the energy hub is presented in this section.

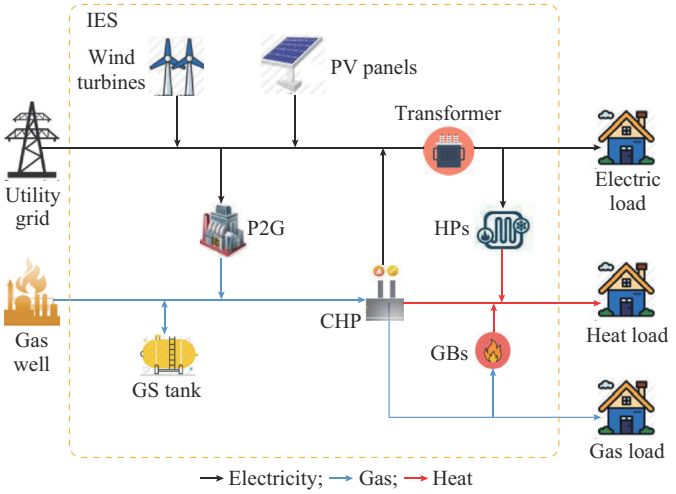

Fig. 2. Framework of an IES.

A. Objectives

The operation objectives of this multiple-energy system include the cost of purchased electricity, the cost of purchased gas, the operation and maintenance costs of wind turbines, PV panels, power-to-grid (P2G) equipment, gas storage (GS) tanks, gas boilers (GBs), heat pumps (HPs), combined heat and power (CHP) equipment, and unit depreciation costs, which is expressed as:

$$\min C_{sys}=\sum_{i=1}^{a}\left[C_e^{net}(t)+C_g^{net}(t)+\sum_{j=1}^{n}(C_{op}^{j}(t)+C_{de}^{j}(t))+C_T(t)\right] \tag{18}$$

where C_{sys} is the total operation cost of this system over a given operation cycle; a is the operation cycle; C_e^{net} and C_g^{net} are the costs of purchased electricity and purchased gas at moment t, respectively; C_{op}^{j} and C_{de}^{j} are the operation costs and depreciation of installed costs for equipment j, respectively; and C_T is the carbon trading cost.

$$C_e^{net}=P_e^{net}S_e\Delta T \tag{19}$$

$$C_g^{net}=P_g^{net}S_g\Delta T/Q_{LHV} \tag{20}$$

$$C_{op}^{j}=C_o^{j}P_j\Delta T \quad j\in N_e \tag{21}$$

$$C_{de}^{j}=\frac{C_{ins}^{j}\lambda_j P_j\Delta T}{P_N^{j}T_{year}} \tag{22}$$

$$\lambda_j=\frac{d_j(1+d_j^{L_j})}{(1+d_j)^{L_j}-1} \tag{23}$$

where P_e^{net} and P_g^{net} are the purchased electric and gas power, respectively; S_e and S_g are the cubic meter prices of electric and gas power, respectively; ΔT is the time period; Q_{LHV} is the calorific value; C_o^{j} is the operation and maintenance cost of equipment j; N_e is the amount of equipment; P_j is the power of equipment j; C_{ins}^{j} is the cost of per unit capacity for equipment j; λ_j is the capital recovery factor of equipment j; P_N^{j} is the capacity factor of equipment j; T_{year} is 8760; and d_j and L_j are the annual interest rate and depreciable life of equipment j, respectively.

Carbon trading is a way to encourage enterprises with high carbon reduction capacity and low costs to increase their reduction efforts and sell the remaining carbon emission allowances to those with higher carbon reduction costs, thereby achieving the targeted total carbon emissions [26]. In IES, the carbon trading for different devices is considered in the costs and the carbon trading market is set for all carbon-emitting equipment, and the total amount of carbon emissions is minimized through the market regulation mechanism. The equipment involved in the carbon trading market in this paper are gas-fired units and P2G equipment, and their carbon emissions are modeled as follows.

Gas-fired units are the source of carbon emissions in the carbon trading market, which would be fined for exceeding the carbon credits. The carbon trading cost of gas-fired units at moment t is expressed as:

$$C_{t,cd}^{G}=c_G^{cd}(e_G-e_m)P_t^{G} \tag{24}$$

where P_t^{G} is the amount of power generated by the gas-fired units; c_G^{cd} is the carbon emission price factor of gas-fired units; e_G is the carbon intensity of gas-fired units; and e_m is the carbon credit of the gas-fired units when involved in carbon trading.

Utilizing P2G, hydrogen and methane are yielded by electrolysis of water according to different market demands [26]. The specific reaction process is:

$$\begin{cases}2H_2O\rightarrow 2H_2+O_2\\ CO_2+4H_2\rightarrow CH_4+2H_2O\end{cases} \tag{25}$$

The amount of methane being produced by P2G when consuming per unit electric power is calculated as:

$$V_{t,CH_4}^{P2G}=\frac{3.6\eta_{P2G}P_t^{P2G}}{L_{CH_4}} \tag{26}$$

where P_t^{P2G} is the electric power consumed by P2G; η_{P2G} is the conversion efficiency of P2G; and L_{CH_4} is the calorific value of combusting per unit volume of natural gas, which is 36 MJ/m^3.

CO_2 is involved in the operation process of P2G, which needs to be purchased. That's why P2G is a participant in the carbon trading market. The cost of purchasing CO_2 at moment t is:

$$C_{t,buy}^{CO_2}=s_{CO_2}^{buy}M_t \tag{27}$$

where $s_{CO_2}^{buy}$ is the price coefficient of consuming per unit CO_2; and M_t is the mass of CO_2 required to participate in the reaction. Furthermore, the carbon credit of P2G is set to be zero since P2G belongs to carbon reduction devices. Therefore, the carbon trading cost of P2G is normally negative, which means obtaining profits.

B. Constraints

The constraints of the network consist of electric power balance, gas flow balance, and heat power balance.

$$P_e^{net}+P_w^{used}+P_p^{used}+P_{CHP}^{e}-P_{P2G}^{e}-P_{hp}^{e}=L_e \tag{28}$$

$$P_g^{net}+P_{P2G}^{g}+P_g^{release}-P_g^{store}-P_{CHP}^{g}-P_{gb}^{g}=L_g \tag{29}$$

$$P_{CHP}^{h}+P_{gb}^{h}+P_{hp}^{h}=L_h \tag{30}$$

where L_e, L_g, and L_h are the electric, gas, and heat demands, respectively; P_w^{used}, P_p^{used}, and P_{CHP}^{e} are the power generated by wind turbines, PV, and CHP, respectively; P_{P2G}^{e} and P_{hp}^{e} are the power consumed by P2G and HPs, respectively; P_{P2G}^{g} is the gas power generated by P2G; $P_g^{release}$ and P_g^{store} are the gas power released and stored by GS tanks, respectively; P_{CHP}^{g} and P_{gb}^{g} are the gas power consumed by the CHP and GB, respectively; and P_{CHP}^{h}, P_{gb}^{h}, and P_{hp}^{h} are the heat power generated by the CHP, GB, and HP, respectively.

The mathematical models of CHP, GB, HP, and P2G are expressed as:

$$c_f=\frac{Q_{CHP}}{P_{CHP}} \tag{31}$$

$$c_v=\frac{Q_{CHP}}{\eta_{CHP}F_{in}-P_{CHP}} \tag{32}$$

$$P_{gb,t}^{e}=\eta_{gb}P_{gb,t}^{g} \tag{33}$$

$$P_{hp,t}^{e}=\eta_{hp}P_{hp,t}^{h} \tag{34}$$

$$P_{t,gas}^{P2G}=\eta_{P2G}P_t^{P2G} \tag{35}$$

where c_f, Q_{CHP}, and P_{CHP} are the thermoelectric ratio, heat output, and electric output of CHP at a constant thermoelectric ratio, respectively; c_v is the thermoelectric ratio of the CHP when the thermoelectric ratio is variable; η_{CHP} is the conversion efficiency of the CHP; F_{in} is the gas flow input of the CHP; $P_{gb,t}^{g}$ is the amount of natural gas consumed by GBs; $P_{gb,t}^{e}$ is the heat power output of GBs; η_{gb} is the efficiency of electric heating for GBs; $P_{hp,t}^{e}$ is the electric power input into an HP; $P_{hp,t}^{h}$ is the heat power output of an HP; η_{hp} is the efficiency of electric heating for an HP; and $P_{t,gas}^{P2G}$ is the gas power output of P2G.

The constraints of CHP, GB, HP, and P2G are expressed as:

$$0\le P_{CHP}^{g}\le P_{CHP}^{g,\max} \tag{36}$$

$$0\le P_{gb}^{g}\le P_{gb}^{g,\max} \tag{37}$$

$$0\le P_{hp}^{e}\le P_{hp}^{e,\max} \tag{38}$$

$$0\le P_{P2G}^{e}\le P_{P2G}^{e,\max} \tag{39}$$

where $P_{CHP}^{g,\max}$ is the upper limit of gas power input into CHP; $P_{gb}^{g,\max}$ is the upper limit of gas power input into the GB; $P_{hp}^{e,\max}$ is the upper limit of electric power input into HP; and $P_{P2G}^{e,\max}$ is the upper limit of electric power input into P2G.

The GS tank is adopted here as the storage device for CH_4. The state of the tank and the charged/discharged gas power at moment t are related to the state of the moment before. The upper and lower bounds of the GS tank should also be satisfied. The constraints are expressed as:

$$S_t=S_{t-1}+\frac{P_{gas,store}^{t}\eta_{ch}\Delta t}{\beta_k}-\frac{P_{gas,release}^{t}\Delta t}{\beta_k\eta_{dch}} \tag{40}$$

$$S_{\min}\le S_t\le S_{\max} \tag{41}$$

where η_{ch} and η_{dch} are the charging and discharging rates of the GS tank, respectively; $P_{gas,store}^{t}$ is the gas power stored by the GS tank; $P_{gas,release}^{t}$ is the gas power released from the GS tank; Δt is the time scale of operation; $S_{\min}$ and $S_{\max}$ are the minimum and maximum values of the GS state S_t, respectively; and β_k is the nominal capacity of the GS tank.

Given that the GS tank could not realize charging and discharging at the same time, the following constraint should be introduced:

$$P_{gas,store}^{t}P_{gas,release}^{t}=0 \tag{42}$$

Suppose that the amount of released gas and stored gas could reach a balance, meaning that the amount of released gas and stored gas are the same, which is expressed as:

$$\sum_{t=1}^{T}P_{gas,store}^{t}=\sum_{t=1}^{T}P_{gas,release}^{t} \tag{43}$$

where T is the operation time.

To ensure the safe and stable operation of the upper grid and to reduce the pressure on its regulation, only energy purchases are considered here:

$$0\le P_{net}^{e}\le P_{net}^{e,\max} \tag{44}$$

$$0\le P_{net}^{g}\le P_{net}^{g,\max} \tag{45}$$

where $P_{net}^{e,\max}$ and $P_{net}^{g,\max}$ are the maximum power purchased from the upper electric and gas grids, respectively.

The constraints of the grid include nodal power constraints, DC power flow constraints, and node pressure constraints.

$$\sum_{j(i,j\in E)}P_{ji}-\sum_{j(i,j\in E)}P_{ij}-P_i=0 \tag{46}$$

$$F_l^{e}(t)=\frac{\theta_l^{i}(t)-\theta_l^{j}(t)}{x_l} \tag{47}$$

$$U_{\min,i}\le U_i\le U_{\max,i} \tag{48}$$

where E is the set of nodes; P_{ij} is the node power; F_l^{e} is the active power flow of line l; θ_l^{i} and θ_l^{j} are the phase angles of node i and node j, respectively; x_l is the reactance of line l; U_i is the pressure of node i; and $U_{\min,i}$ and $U_{\max,i}$ are the lower and upper pressure bounds of node i, respectively.

The transmission constraints are:

$$-P_{l,\max}\le P_{ij}\le P_{l,\max} \tag{49}$$

where $P_{l,\max}$ is the bound of the power transmitted through line l.

The nodal flow constraints of the gas network are [27]:

$$F_S^{i}-F_{hub}^{i}-\sum_{j\in E,j\ne i}F_l^{ij}=0\quad i\in N_G \tag{50}$$

where N_G is the set of gas nodes; F_S^{i} is the gas flow transmitted from the gas source to node i; F_{hub}^{i} is the gas flow exchanged between node i and the energy hub; and F_l^{ij} is the gas flow transmitted from gas source to nodes i and j.

The gas flow transmitted through the pipeline is influenced by the air pressure of the nodes located at both sides

and the transmission coefficient of the pipeline itself. Then, the steady-state equations are expressed as:

$$F_l^{ij} = S(i,j)K_{ij}\sqrt{\left|p_i^2 - p_j^2\right|} \tag{51}$$

$$S(i,j) = \begin{cases} 1 & p_i \geq p_j \\ -1 & p_i < p_j \end{cases} \tag{52}$$

$$p_{\min} \leq p_i \leq p_{\max} \tag{53}$$

where K_{ij} is the transmission coefficient of the pipeline between node i and node j, which is affected by the temperature, diameter, and length of the pipeline, as well as the friction coefficient; p_i and p_j are the gas pressures of node i and node j, respectively; and $p_{\min}$ and $p_{\max}$ are the lower and upper bounds of node pressure, respectively, which are both positive. When $S(i,j) < 0$, the pressure of node i is lower than that of node j, which means the gas flows from node j to node i.

Substituting K_{ij} into (51), it can be obtained that:

$$F_l^{ij} = S(i,j)C\frac{T_b}{P_b}\sqrt{\frac{\left|p_i^2 - p_j^2\right|D^5}{GT_fLZf}} \tag{54}$$

where C is a constant, which is 1.1494×10^{-3}; T_b is the base temperature; P_b is the base pressure; G is the gravity of per unit gas; T_f is the average temperature of the gas; L is the length of the pipeline; Z is the compression coefficient of the gas at the corresponding temperature; D is the diameter of the pipeline; and f is the fraction coefficient.

The conversion between the natural gas flow and power is through the calorific value, the relationship is as follows:

$$P_l^{ij} = H_{GV}F_l^{ij} \tag{55}$$

where P_l^{ij} is the power through pipeline l; and H_{GV} is the high calorific value of the natural gas.

C. Linearization Method

The constraints of the gas network considered in this paper are steady-state gas network constraints, which are difficult to solve directly using commercial software due to the presence of non-convex constraints [28]. Hence, the model needs to be linearized. In the steady-state equations of the natural gas network, the node pressure p_i appears in the form of p_i^2, and G_i is adopted to represent p_i^2, then:

$$G_i = p_i^2 \tag{56}$$

$$G_{ij} = p_i^2 - p_j^2 \tag{57}$$

Square both sides of (51) and combine (56) and (57):

$$(F_l^{ij})^2 = K_{ij}^2\left|G_{ij}\right| \tag{58}$$

Introduce auxiliary variables φ_{ij} into (51), then:

$$\varphi_{ij} = S(i,j)F_{ij}^2 \tag{59}$$

$$\varphi_{ij} = K_{ij}^2 G_{ij} \tag{60}$$

The range of node pressure is expressed in (53), and then the range of G_{ij} is expressed as:

$$0 \leq \left|G_{ij}\right| \leq p_{\max}^2 - p_{\min}^2 \tag{61}$$

Substitute (61) into (60), then:

$$-K_{ij}^2(p_{\max}^2 - p_{\min}^2) \leq \varphi_{ij} \leq K_{ij}^2(p_{\max}^2 - p_{\min}^2) \tag{62}$$

However, after the above equivalence, some non-linear terms still exist. Assume that $\boldsymbol{X}$ is the matrix of all the variables involved in the optimization, which includes continuous variables. Then, the optimization model can be expressed in compact form as:

$$\begin{cases} \min \boldsymbol{CX} \\ \text{s.t. } \boldsymbol{A}_1\boldsymbol{X} - \boldsymbol{b}_1 \leq \boldsymbol{0} \\ \quad \boldsymbol{A}_2\boldsymbol{X} - \boldsymbol{b}_2 = \boldsymbol{0} \\ \quad \boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT}) = \boldsymbol{0} \\ \quad \boldsymbol{X}_{CT}^{\mathrm{T}} \in \boldsymbol{\Theta} \end{cases} \tag{63}$$

where $\boldsymbol{C}$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ are the coefficient matrices; $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the constant vectors; $\boldsymbol{\Theta}$ is the possible domains of $\boldsymbol{X}$; and $\boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT}) = \boldsymbol{0}$ is used to represent the non-linear constraints. The linear model above includes both inequality and equality constraints, which are represented by $\boldsymbol{A}_1\boldsymbol{X} - \boldsymbol{b}_1 \leq \boldsymbol{0}$ and $\boldsymbol{A}_2\boldsymbol{X} - \boldsymbol{b}_2 = \boldsymbol{0}$, respectively.

The above optimization model is processed using an improved step-by-step linearization method. Based on the step-by-step linearization method, the step is corrected at each iteration, thus achieving accelerated convergence of the iterations. The non-linear constraints is transformed based on the Taylor formula. For the non-linear constraint $\boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT}) = \boldsymbol{0}$, suppose that the run point at the k^{th} iteration is $\boldsymbol{X}_{CT}^{(k)}$, and a first-order Taylor expansion is performed at $\boldsymbol{X}_{CT}^{(k)}$, (63) can be transformed into:

$$\begin{cases} \min \boldsymbol{CX} \\ \text{s.t. } \boldsymbol{A}_1\boldsymbol{X} - \boldsymbol{b}_1 \leq \boldsymbol{0} \\ \quad \boldsymbol{A}_2\boldsymbol{X} - \boldsymbol{b}_2 = \boldsymbol{0} \\ \quad \nabla x_{CT}\boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT} - \boldsymbol{X}_{CT}^{(k)}) = \boldsymbol{0} \\ \quad \boldsymbol{X}_{CT}^{\mathrm{T}} \in \boldsymbol{\Theta} \end{cases} \tag{64}$$

where ∇x_{CT} is the Newton's basic step.

During the iterative process, if the newly obtained solution $\boldsymbol{X}_{CT}$ is directly used as the run point of the next iteration, oscillations will occur, leading to a slower convergence rate. To shorten the convergence time, the step length of the k^{th} iteration is defined as $\Delta\boldsymbol{X}_{CT}^{(k)}$, which is calculated by $\boldsymbol{X}_{CT} - \boldsymbol{X}_{CT}^{(k)}$, and a step length correction factor λ is introduced, so that the running point of the next iteration is corrected to $\boldsymbol{X}_{CT}^{(k)} + \lambda\Delta\boldsymbol{X}_{CT}^{(k)}$. The factor λ is determined by the following linear model:

$$\begin{cases} \min \left\| \boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT}^{(k)} + \lambda\Delta\boldsymbol{X}_{CT}^{(k)}) \right\|_2 \\ \text{s.t. } \lambda \in [0, 1] \end{cases} \tag{65}$$

In iterative algorithms, the selection of the initial value is very important, and if it is not chosen well, the system can be unsolvable or fall into a local optimum. In this paper, the initial point of the iteration is the optimal solution of the optimization model without non-linear constraints, and the specific steps are as follows.

Step 1: after removing the non-linear constraints in (63), this linear programming model is solved using the commercial solver Cplex, and $\boldsymbol{X}_{CT}$ is used as initial run point $\boldsymbol{X}_{CT}^{(k)}$.

Step 2: solve (64), and achieve the solution $\boldsymbol{X}_{CT}$ of this iteration and the basic step $\Delta\boldsymbol{X}_{CT}^{(k)}$. The first-order Taylor expansion is performed for the non-linear constraints in this optimization model, and the specific expression is:

$$(F_l^{(k)})^2 \cdot sgn(F_l^{(k)}) - K_{ij}^2 (G_i^{(k)} - G_j^{(k)}) + 2F_l^{(k)} \cdot sgn(F_l^{(k)}) \cdot (F_l - F_l^{(k)}) - K_{ij}^2 (G_i - G_i^{(k)}) + K_{ij}^2 (G_j - G_j^{(k)}) = 0 \quad (66)$$

where $F_l^{(k)}$ is the natural gas flow; and $G_i^{(k)}$ and $G_j^{(k)}$ are the node pressures.

Step 3: based on $\boldsymbol{X}_{CT}^{(k)}$ and $\Delta\boldsymbol{X}_{CT}^{(k)}$, λ is obtained by iterating through (65), then $k = k + 1$.

Step 4: using the updated run point $\boldsymbol{X}_{CT}^{(k)}$ as the new run point, repeat *Steps 2* and *3* until the convergence residuals satisfy the convergence accuracy ε:

$$\left\| \boldsymbol{H}_{non\text{-}linear}(\boldsymbol{X}_{CT}^{(k)}) \right\|_2 \leq \varepsilon \quad (67)$$

IV. Case Studies

A. Results and Discussion

The proposed scenario generation method is applied to an IES including wind and PV power outputs as well as electric, heat, and gas demands in this subsection. The case is set in Aachen, Germany, with the time scale starting from January 1, 2019 to December 31, 2019 [29]. The data of renewable power are obtained from [30], which are presented in Fig. 3 and Fig. 4. And the load data are presented in Fig. 5.

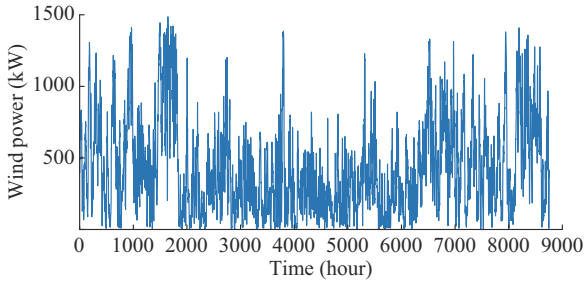

Fig. 3. Hourly wind power output.

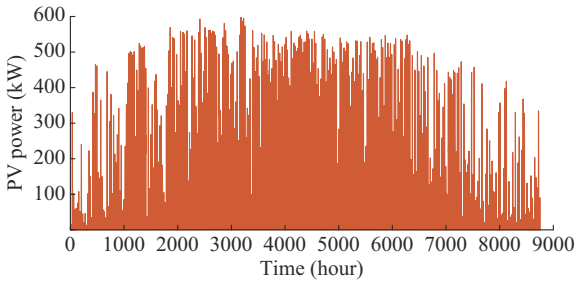

Fig. 4. Hourly PV power output.

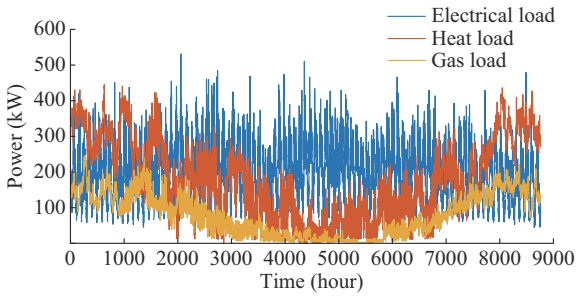

Fig. 5. Hourly demand in a year.

Aachen is rainy all year round. The duration of simulation can be divided into a heating season and a non-heating season, with the heating season extending from September 10 to April 10. Due to the regular heat demands from the residents during the heating season, the situation of these two periods is discussed separately. There are 8760 sequences all through the year, in which 5088 sequences are of the heating season and 3672 sequences are of the non-heating season. In this subsection, the proposed scenario generation method is applied to the historical scenarios of wind/PV power output and electric/gas/heat demand, respectively. Typical scenarios corresponding to the heating and non-heating seasons are obtained, respectively, which are presented in Fig. 6.

From the trends of curves in Fig. 6, we can conclude that:

1) Except for the scenarios with no obvious wind power output, the daily wind power output shows the characteristic of peak and valley, and the fluctuation range is relatively large. And the trend of wind power output in a typical scenario is relative to the trend of electric load, showing certain reverse-peak regulation characteristics.

2) The daily PV power output is influenced by solar energy resources, which are regular and fluctuating. The daily PV power output curve shows an obvious "sine wave" pattern, i.e., the PV power output is higher at noon, lower in the early morning and evening, and zero at night, and the output occurs mainly between 09:00 and 17:00. At the same time, the PV power output is closely related to the seasons, e.g., during the heating season, the shorter daylight hours and weaker illumination intensity lead to less PV power output compared with that during the non-heating season. In a typical scenario, during the non-heating season, the PV power output lasts from 05:00 to 19:00, and the peak-to-valley difference is greater.

3) The seasonal nature of the load is evident, with the total value of the load being higher during the heating season due to the higher heat demand. The gas load during the non-heating season is almost zero since the main purpose of natural gas is heating. The overall heat demand is higher during the heating season than during the non-heating season and fluctuates less. In a typical scenario, the heat and gas demand variation trends are similar throughout the day. For example, in scenario 2 for the heating season, the gas and heat demands both appear a simultaneous upward trend before 07:00, with a general decrease trend from 07:00 to 18:00 and a slow increase trend after 18:00.

4) In typical scenarios of the non-heating season, the gas and heat demands show the same trend, reflecting the energy consumption characteristics of consumers. In all typical scenarios, the electric load has more obvious peaks in the morning and evening. The clustering effect of typical scenarios obtained by the proposed method is significant. The scenario generation method replaces the original year-round historical data with several typical scenarios.

B. Validation of Proposed Scenario Generation Method

1) Accuracy Comparison

The quality of the clustering algorithm is closely related to the effectiveness of the typical scenario generation method.

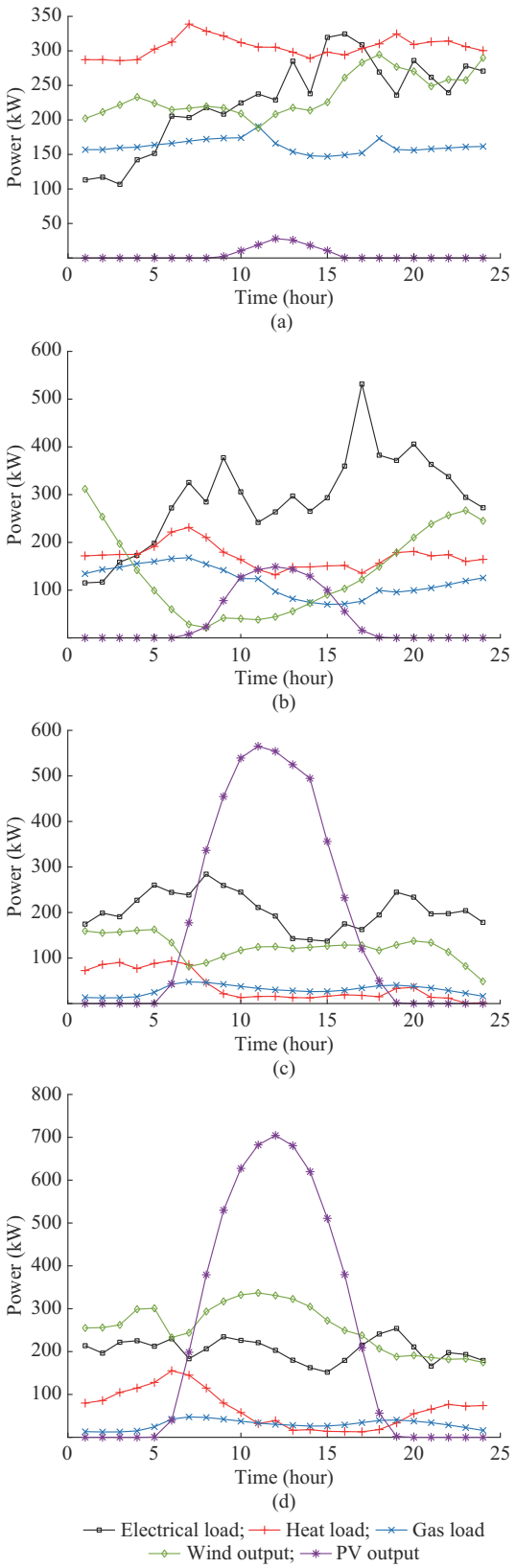

Fig. 6. Load and output curves of typical scenarios. (a) Scenario 1 in heating season. (b) Scenario 2 in heating season. (c) Scenario 1 in non-heating season. (d) Scenario 2 in non-heating season.

Therefore, the clustering effectiveness of the proposed method will be assessed by comparing the cluster validity indexes of the proposed method with those of the comparison method. Typical cluster validity indexes are external and internal cluster validation indexes, where ARI and SC are introduced to verify the accuracy of the proposed method [30]. The ARI and SC of the proposed method and comparative methods are presented in Table I, where DBSCAN stands for density-based spatial clustering of applications with noise.

TABLE I
RESULTS OF ARI AND SC

Method	ARI	SC
Proposed method	0.6153	0.6770
K-medoids	0.5816	0.6592
K-means++	0.6180	0.6754
DBSCAN	0.6102	0.6703

From the two indexes above, higher values of the ARI and SC are obtained when the proposed method is applied. The quality of simple *K*-medoids clustering method is relatively worse since it is affected by the initial cluster center. While *K*-means++ is an improved *K*-means clustering method, the clustering results are less dependent on the initial cluster center, and the results are slightly better than the *K*-medoids clustering method, which is yet inferior to the proposed method. The internal and external validation indexes obtained using the DBSCAN clustering method are better than the *K*-medoids and *K*-means++ clustering methods, but still inferior to the proposed method. The proposed method retains the time series of output and load variation over the day, which is superior to the method representing the renewable power output and load over the day by statistics.

2) Effectiveness Analysis in Optimal Operation

To verify the rationality and effectiveness of the proposed method, typical scenarios obtained in Section III are applied to the optimal operation of the system presented in Fig. 2. The problem is performed by the YALMIP optimization toolbox in MATLAB and optimized by the Cplex solver. The time step is set to be 1 hour, while the optimization scale lasts 365 days, i.e., 8760 hours. The energy hub is connected to the grid and the natural gas network, with the natural gas price set to be 2.5 ¥/m^3 and the low calorific value of natural gas set to be 9.95 kWh/m^3. The unit investment cost and operation and maintenance (O&M) costs of instruments in the energy hub are shown in Table II. Moreover, the O&M cost of the GS tank is 0.01 ¥/m^3. The parameters of related instruments are shown in Table III. And the cost of purchased electricity is calculated by the time-of-use tariff.

The all-year time series, typical day selection, *K*-means++ , and SOM methods are chosen as comparative methods to validate the proposed method. The day with the largest peak-to-valley difference is selected as a typical day. The optimization result based on all-year historical data is set as the criterion to analyze the validation. The optimization results are presented in Table IV.

It can be observed from Table IV that the operation cost difference between the proposed method and the all-year time series method is the smallest.

TABLE II
O&M COSTS OF INSTRUMENTS IN ENERGY HUB

Instrument	O&M (¥/kWh)
Wind turbine	0.296
PV	0.358
HP	0.200
GB	0.230
P2G	0.150

TABLE III
PARAMETERS OF RELATED INSTRUMENTS IN ENERGY HUB

Parameter	Value	Parameter	Value
$P_{CHP}^{g,\max}$	800 kW	$S_{\max}$	1
$P_{P2G}^{e,\max}$	600 kW	η_{P2G}	0.6
$P_{HP}^{e,\max}$	400 kW	η_{CHP}	0.45
$P_{gb}^{e,\max}$	400 kW	η_{gb}	0.75
$S_{\min}$	0.2	η_{hp}	0.8

TABLE IV
OPTIMIZATION RESULTS

Method	Operation cost (¥)	Calculation time (s)	Scenario number (day)
All-year time series	2675112.0	11028.8	365
Typical day selection	3001475.7	89.4	2
K-means++	2753644.3	187.3	4
SOM	2725939.1	162.4	4
Proposed method	2677817.8	188.0	4

The proposed method only applies four wind/PV/load historical scenarios, so the running time is much shorter compared to the all-year time series method. In contrast, the results obtained using the typical day selection method and the *K*-means++ method are less accurate, although both methods are of shorter running time compared to the proposed method.

As shown in Table V, the error of the annual operation cost using the typical day selection method is relatively large. It could be concluded that the typical scenarios selected by this method cannot comprehensively characterize the temporal feature of the annual data, which leads to a large error. Meanwhile, the error of the annual operation cost adopting the SOM method is rather smaller than the errors obtained by the *K*-means++ and typical day selection methods. The error obtained by the proposed method is only 0.1%, which means better accuracy.

TABLE V
ERROR OF ANNUAL OPERATION COST OF DIFFERENT METHODS

Method	Error (%)
All-year time series	0
Typical day selection	12.20
K-means++	4.29
SOM	1.90
Proposed method	0.10

The above analysis shows that the proposed method is more comprehensive than the typical day selection and *K*-means++ methods in characterizing the annual data of the multi-energy system. Hence, the annual operation of this IES could be reflected using the typical scenarios obtained by the proposed method.

C. Application in IES with Multiple Energy Hubs

To further realize the application of the proposed method in the IES, based on the typical scenarios obtained before, a sampling interval of 1 hour is taken for the optimal operation. The simulation is performed on the IES containing multiple energy hubs, which is a modification of the four-node energy hub test system in [31]. The four energy hubs are denoted as EH1, EH2, EH3, and EH4, which are connected by the grid and natural gas pipeline. NE denotes the gas source of the system, and NG denotes the electric power supplied to the IES by the external grid. The parameters of the electric line and gas pipeline refer to [32]. The operation cost of the system in different scenarios is shown in Table VI.

TABLE VI
OPERATION COST IN DIFFERENT SCENARIOS

Scenario	C_e^{net} (¥)	C_g^{net} (¥)	C_{op} (¥)	C_T (¥)	C_{svs} (¥)
Scenario 1 in heating season	10518.486	16317.041	12556.821	187.604	39204.744
Scenario 2 in heating season	17235.395	15114.839	14833.174	275.968	46907.440
Scenario 1 in non-heating season	9634.018	2049.993	3027.062	102.771	14608.302
Scenario 2 in non-heating season	9491.841	2467.858	4261.329	178.937	16042.091

Besides, in typical scenario 1 in the heating season, the costs of purchased electricity and purchased gas account for about 68.450% of the total operation cost, while the O&M cost accounts for about 32.029% of the total operation cost. While in typical scenario 2 during the heating season, the costs of purchased electricity and purchased gas account for about 68.966% of the total operation cost, and the O&M cost accounts for about 31.622% of the total operation cost. It is essential for the IES to maintain real-time interaction with the external grid and gas network during operation, which guarantees the stable supply and demand of the system.

On the basis of the above analysis, it can be concluded that the energy hub configured in this paper is capable of

meeting the demand for various forms of energy. However, the cost of purchased gas and electricity exceeds 50% of the total operation cost in all typical scenarios, indicating that the system is highly dependent on the outside energy supplies and needs to maintain real-time interaction with the external grid and external gas network to realize the system stability.

V. Conclusion

This paper presents a novel scenario generation method for an IES based on panel data feature extraction. PCA is first adopted to compress the multi-indicator data at a certain moment to form an aggregated indicator matrix for the scenario generation of an IES. Then, for the aggregated indicator matrix obtained, the improved canopy clustering method considering the density of the samples is used to perform coarse clustering, while the *K*-medoids clustering method is performed based on the obtained number of clusters and cluster centers to construct typical scenarios. In case studies, the proposed method is conducted utilizing the historical data from Aachen, Germany, which consist of wind/PV power output, electric load, heat load, and gas load. Two cluster validity indexes, ARI and SC, are introduced to verify the accuracy of the proposed method and the corresponding calculation index value is compared with other three typical clustering methods.

By analyzing the characteristics of the renewable energy output, multi-energy load curves obtained from the proposed method as well as the results of comparing the two cluster validity indexes, it is revealed that the proposed method is more accurate for practical applications. To validate the effectiveness of the proposed method when applying the obtained typical scenarios to analyze IESs, an optimal operation model considering carbon trading cost is established. The results show that the system is more dependent on the external network during this period than during the heating season.

Finally, by comparing the results obtained from the proposed method with those obtained from the all-year time series and other two comparative methods, the proposed method is shown to be more effective and efficient.

References

[1] H. Zhao, S. Miao, C. Li *et al*., "Research on optimal operation strategy for park-level integrated energy system considering cold-heat-electric demand coupling response characteristics," *Proceedings of the CSEE*, vol. 42, no. 2, pp. 573-589, Aug. 2022.

[2] P. Li, F. Zhang, X. Ma *et al*., "Operation cost optimization method of regional integrated energy system in electricity market environment considering uncertainty," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 1, pp. 368-380, Jan. 2023.

[3] L. Chen, Q. Xu, Y. Yang *et al*., "Community integrated energy system trading: a comprehensive review," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 6, pp. 1445-1458, Nov. 2022.

[4] X. Yu, X. Xu, S. Chen *et al*., "A brief review to integrated energy system and energy internet," *Transactions of China Electrotechnical Society*, vol. 31, no. 1, pp. 1-13, Jan. 2016.

[5] L. Liu, D. Wang, K. Hou *et al*., "Region model and application of regional integrated energy system security analysis," *Applied Energy*, vol. 260, p. 114268, Feb. 2020.

[6] J. Tan, P. Yang, and X. Zeng, "The improved CVaR day-ahead dispatch model of micro-energy-grid based on scenario reduction optimization guiding confidence value selection," *High Voltage Engineering*, vol. 48, no. 12, pp. 1-11, Oct. 2022.

[7] L. Yang, D. Wang, H. Jia *et al*., "Multi-objective stochastic expansion planning based on multi-dimensional correlation scenario generation method for regional integrated energy system integrated renewable energy," *Applied Energy*, vol. 276, pp. 1-32, Oct. 2020.

[8] J. Pei, J. Wang, Z. Wang *et al*., "Precise recovery of corrupted synchrophasors based on autoregressive Bayesian low-rank factorization and adaptive *K*-medoids clustering," *IEEE Transactions on Power Systems*. doi: 10.1109/TPWRS.2022.3221291

[9] M. Ding, J. Xie, X. Liu *et al*., "The generation method and application of wind resources/load typical scenario set for evaluation of wind power grid integration," *Proceedings of the CSEE*, vol. 36, no. 15, pp. 4064-4071, Jul. 2016.

[10] W. G. Zong, "Size optimization for a hybrid photovoltaic-wind energy system," *International Journal of Electrical Power and Energy Systems*, vol. 42, no. 1, pp. 448-451, Jun. 2012.

[11] L. Guo, R. Hou, Y. Liu *et al*., "A novel typical day selection method for the robust planning of stand-alone wind-photovoltaic-diesel-battery microgrid," *Applied Energy*, vol. 263, p. 114606, Apr. 2020.

[12] X. Peng, G. Gao, G. Hu *et al*., "Research on inter-regional renewable energy accommodation assessment method based on time series production simulation," in *Proceedings of 2019 IEEE Sustainable Power and Energy Conference (iSPEC)*, Beijing, China, Nov. 2019, pp. 2031-2036.

[13] C. Dong, M. Li, G. Fan *et al*., "Research and application of renewable energy accommodation capability evaluation based on time series production simulation," *Electric Power*, vol. 48, no. 12, pp. 166-172, Dec. 2015.

[14] F. Mei, J. Zhang, J. Lu *et al*., "Stochastic optimal operation model for a distributed integrated energy system based on multiple-scenario simulations," *Energy*, vol. 219, p. 119629, Mar. 2021.

[15] X. Tang, Q. Li, L. Hou *et al*., "Generation of typical sequential joint output scenarios of wind power based on Copula function," *Electric Power Engineering Technology*, vol. 39, no. 5, pp. 152-168, Sept. 2020.

[16] Y. Dvorkin, Y. Wang, H. Pandzic *et al*., "Comparison of scenario reduction techniques for the stochastic unit commitment," in *Proceedings of 2014 IEEE PES General Meeting Conference & Exposition*, National Harbor, USA, Jul. 2014, pp. 1-5.

[17] J. Hu and H. Li, "A new clustering approach for scenario reduction in multi-stochastic variable programming," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3813-3825, Feb. 2019.

[18] J. Hu, H. Li, and Z. Liu, "Scenario reduction based on correlation sensitivity and its application in microgrid optimization," *International Transactions on Electrical Energy Systems*, vol. 31, no. 3, p. e12747, Jan. 2021.

[19] K. Luo, W. Shi, and W. Wang, "Extreme scenario extraction of a grid with large scale wind power integration by combined entropy-weighted clustering method," *Global Energy Interconnection*, vol. 3, no. 2, pp. 140-148, Apr. 2020.

[20] M. Ahmed, R. Seraj, and S. M. S. Islam, "The *K*-means algorithm: a comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020.

[21] A. Özlem and Y. Güzin, "Hierarchical clustering of mixed variable panel data based on new distance," *Communications in Statistics – Simulation and Computation*, vol. 50, no. 6, pp. 1695-1710, Apr. 2021.

[22] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of Knowledge Discovery and Data Mining*, Boston, USA, Aug. 2000, pp. 169-178.

[23] G. Zhou, "Improved optimization of canopy-*K*means clustering algorithm based on Hadoop platform," in *Proceedings of the International Conference on Information Technology and Electrical Engineering*, Xiamen, China, Dec. 2018, pp. 1-6.

[24] S. A. Abbas, A. Aslam, A. U. Rehman *et al*., "*K*-means and *K*-medoids: cluster analysis on birth data collected in city Muzaffarabad, Kashmir," *IEEE Access*, vol. 8, pp. 151847-151855, Aug. 2020.

[25] Z. Yuan, S. He, A. Alizadeh *et al*., "Probabilistic scheduling of power-to-gas storage system in renewable energy hub integrated with demand response program," *Journal of Energy Storage*, vol. 29, p. 101393, Jun. 2020.

[26] F. Wen, N. Wu, and X. Gong, "China's carbon emissions trading and stock returns," *Energy Economics*, vol. 86, p. 104627, Feb. 2019.

[27] X. Xing, J. Lin, Y. Song *et al*., "Modeling and operation of the power-to-gas system for renewables integration: a review," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 2, pp. 168-178, Jun. 2018.

[28] J. Liu, W. Sun, and G. P. Harrison, "Optimal low-carbon economic environmental dispatch of hybrid electricity-natural gas energy systems considering P2G," *Energies*, vol. 12, no. 7, p. 1355, Apr. 2019.
[29] Aachen. (2019, Apr.). E2Watch. [Online]. Available: https://stadt-aachen.e2watch.de
[30] Ninja. (2020, Jul.). Renewables.ninja. [Online]. Available: https://www.renewables.Ninja
[31] Y. Bao, M. Wu, X. Zhou *et al*., "Piecewise linear approximation of gas flow function for the optimization of integrated electricity and natural gas system," *IEEE Access*, vol. 7, pp. 91819-91826, Jul. 2019.
[32] S. I. Hassan, A. Samad, O. Ahmad *et al*., "Partitioning and hierarchical based clustering: a comparative empirical assessment on internal and external indices, accuracy, and time," *International Journal of Information Technology*, vol. 12, no. 4, pp. 1377-1384, Dec. 2020.
[33] T. Krause, G. Andersson, K. Fröhlich *et al*., "Multiple-energy carriers: modeling of production, delivery, and consumption," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 15-27, Dec. 2010.

Bingtuan Gao received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in power electronics and electrical drives, all from Harbin Institute of Technology, Harbin, China, in 2002, 2004, and 2007, respectively. From 2008 to 2010, he was a Post Doctor with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, USA. He is currently a Professor with the School of Electrical Engineering, Southeast University, Nanjing, China. His research interests include robotics, renewable energy generation, and demand-side management.

Yunyu Zhu received the B.S. degree in electrical engineering from Shandong University, Jinan, China, in 2018, and the M.S. degree in electrical engineering from Southeast University, Nanjing, China, in 2021. Her research interest includes the analysis of energy supply and consumption methods and generation of operation scenarios for urban integrated energy systems.

Yuanmei Li received the B.S. degree in automation and the M.S. degrees in control science and engineering from Xinjiang University, Urumqi, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree in electrical engineering at Southeast University, Nanjing, China. Her research interests include integrated energy system optimization and distributed energy system optimization.