Typical Solar Radiation Year Construction Using *k*-means Clustering and Discrete-time Markov Chain

Shuai Li[a] , Hongjie Ma[b*] ,Weiyi Li[a]

[a] Key Laboratory of Efficient Utilization of Low and Medium Grade Energy, MOE, Tianjin University, Tianjin 300072, China

[b] Institute of Industrial Research, University of Portsmouth, Portsmouth, Hampshire PO1 2EG, United Kingdom

Highlights

A Typical Solar Radiation Year synthesis method reflecting fluctuation is proposed.

Clear-sky ratio based parameters are used to present the solar radiation fluctuation.

K-means clustering and discrete-time Markov chain are employed to synthesize.

The distribution and transition rule of solar radiation fluctuation are analysed.

Abstract

Daily solar radiation (DSR) fluctuation and transition rules affect the design of the energy storage system and online control strategy of solar energy utilisation systems. However, the current synthesis methods for the typical meteorological year do not emphasise such features of DSR. To overcome this shortcoming, this study presents an innovative synthesis method for a typical solar radiation year (TSRY) based on k-means clustering and discrete-time Markov chain (DTMC). The historical distributions of clear-sky ratio (CSR) in four representative regions were analysed, and a six-dimensional feature vector that represents the DSR fluctuations based on CSR was defined. Then, based on the feature vector, k-means clustering was used to cluster the historical DSR into four types. Subsequently, a DTMC-based model was built for transition rule estimation among the four types of solar radiation. Finally, the TSRY was established based on the clustering categories and transition rules among them. The innovative synthesis method was also verified in this study. Results for the four regions showed that the average error of the synthesised TSRY has maximum and minimum values of 10% and 6% in all seasons, respectively, compared with historical data. The proposed method could represent DSR fluctuation and transition characteristics of certain regions and could also be extended to other regions.

Corresponding author: Hongjie Ma. Tel: +44 7421307789

Email: hongjie.ma@port.ac.uk

Keywords:
DSR, Typical Solar Radiation Year, Clear-sky ratio, k-means clustering, Discrete-time Markov chain, Typical Meteorological Year

Nomenclature	
CSR	Clear-sky Ratio
SEUS	Solar energy utilization system
PV	Photovoltaic solar cells
TSRY	Typical Solar Radiation Year
TMY	Typical Meteorological Year
DTMC	Discrete-time Markov chain
CSP	Concentrated Solar Power
SURFRAD	Surface Radiation budget network
UTC	Universal Time Coordinated
DSWR	Downwelling short wave radiation
CDSWR	Clear-sky downwelling short wave radiation

1 Introduction

Solar radiation as a renewable energy has been extensively applied in recent years because of energy and environmental restrictions. Solar energy utilisation systems (SEUS) have different variations, such as photovoltaic (PV) solar cells and solar organic Rankine cycle power generation systems. The design of SEUS is highly dependent on the law of solar radiation, such as the capacity of capacitance for the PV system, the thermal energy storage size for the solar organic Rankine cycle power generation systems[1] or the solar field size for the solar hybrid power plant[2]. However, solar radiation as a non-stable energy is affected by multiple factors, including latitude, district, season and cloud distribution and so on [3]. Thus, DSR fluctuation and transition rules should be explored to guide the design of a high-efficiency SEUS [4]. Belkaid et al. proposed a new PV maximum power point tracking (MPPT) strategy for rapid solar radiation fluctuations to optimise the operation of the PV system. The test results showed that the new MPPT strategy, which considers the solar radiation fluctuations, can improve efficiency by approximately 5% [5]. Kaplani et al. presented a stochastic simulation model to determine the minimum installed peak power and storage capacity considering the

DSR fluctuation to optimise the PV system design [6].

At present, research specific to DSR can be conducted using three methods, including the empirical/statistical model of solar radiation, the solar radiation prediction model and the typical DSR database.

The principle for the empirical/statistical model of solar radiation is based on the long term measured data analysis of a particular region. In other words, according to the statistical relationships among parameters, such as latitude, solar radiation, total solar radiation, direct radiation, scattered radiation and radiation peak, the corresponding statistical model could be established[7,8] or the parameters related to the typical empirical model such as a sine wave mode or a trigonometric model in conjunction with a sine and cosine wave could be optimised[9,10]. Fariba et al. from Yazd University in Iran summarised 78 typical empirical models of solar radiation [11]. In line with the inputs of these models, they classified these models into four classes. The first class covers 35 empirical models using solar radiation as their input; the second class consists of 6 empirical models utilising cloud distribution as their input; the third class includes 16 empirical models adopting temperature as their input; and the fourth class incorporates 21 empirical models that use meteorological parameters, such as precipitation, relative humidity, dew point temperature, soil temperature and vaporisation temperature as their inputs. Hassan et al. established, validated and compared their empirical model with 17 ambient-temperature-based models for estimating global solar radiation based on 20 years of historical solar radiation data [12]. Janjai et al. proposed a semi-empirical model for estimating clear sky global and direct normal irradiances, which express global and direct normal irradiances as empirical functions of aerosol parameters, precipitable water, total column ozone, air mass and solar zenith angle [13] .

In contrast to the empirical model of solar radiation, the solar radiation prediction model is used to predict solar radiation within a short period. This model uses real-time environmental information as input and solar radiation as output, thereby highlighting the relationship between the two. Relevant environmental information includes sunrise and sunset times, temperature, relative humidity, wind speed, air pressure, longitude and latitude and clear-sky index (CSI) [14]. Given that these parameters have a typical non-linear relationship with solar radiation, a forecast model for such is often constructed based on methods using artificial intelligence [15]. For instance, Landeras et al. attempted to use three temperature-based models to estimate the DSR, including gene expression programming, artificial neural network (ANN) and adaptive neuro-fuzzy inference system [16]. For these two models, Hussain et al. proposed a frequency coherence and phase synchronization based method to evaluate their predictive performance[17]. As a branch of the stochastic method, the Markov model is also applied[18], Saurabh et al. combined the hidden Markov model and the generalised fuzzy model to estimate solar radiation [19]. Voyant et al. used Bayesian rules to select a hybrid stochastic model consisting of multi-layer perceptron and auto-regressive and moving average to improve the prediction accuracy of short-term solar radiation[20]. Meanwhile, Deo et al. utilised the support vector machine to estimate DSR based on sunshine hours, evaporation, precipitation, wind speed and so on[21]. The solar radiation prediction model based on semi-empirical model and stochastic method, represented by Kaplani and Kaplanis, is another kind of common method. These models takes into account either 1[22], or 2, or 3[23] morning measurements to predict the hourly solar radiation profile for the remaining hours of the day[24]. These models can be utilised to predict solar radiation rules in the scale of days, hours or even minutes.

The comparison of the methods described previously shows that the empirical model of solar radiation stresses in-depth abstraction of the historical data in a region, as it serves as a guide for industrial and agricultural productions. For example, Marcel et al. utilised a model called r.sun to estimate the solar radiation potential of PV systems in Central and Eastern Europe [25]. According to the empirical model, the average tendency of DSR is reflected by each region. Meanwhile, the solar radiation prediction model is adopted to express the relationship between solar radiation and easily measured meteorological data. As such, solar radiation in the scale of days, hours or even minutes could be forecasted by the current meteorological parameter. For instance, ANN is used to estimate hourly global irradiation for the online optimisation of the tilt angle of a solar collector [26] or to predict the generating capacity of PV in the following hours or days to optimise energy management of the corresponding PV system [6,27]. In summary, the two kinds of models describe solar radiation variations in two different time scale.

In addition to the empirical and predictive models of solar radiation, the typical DSR database-based method is used to provide a reference solar radiation for SEUS design. For example, Lou et al. developed typical meteorological year in Hong Kong using machine learning and multivariable regression [28]. A representative solar radiation database for the duration of one year is known as the typical meteorological year (TMY). The TMY consists of months that are selected from individual historical years and concatenated to form a complete year [29]. In contrast to the two previous models, the application target of TMY is to provide a standard reference for SEUS design in accordance with a certain area. In addition, the TMY data are derived from real typical segments of historical solar radiation, which make it more suitable for the design of SEUS compared with other models [30].

As the United States has a meteorological data collection network that covers its entire territory, it leads the world in conducting research on TMY. The first TMY data set for the U.S. was produced by Sandia National Laboratories in 1978 for 248 regions using long-term weather and solar data from the 1952–1975 SOLMET/ERSATZ database [31]. With the addition of new historical data sets and improvements on the algorithm, the National Renewable Energy Laboratory (NREL) successively released the second- and third-generation TMY. The third-generation TMY was obtained through synthesis by Finkelstein–Schafer statistics [32]. The corresponding differences were observed in the operational details, such as the selection and weights of meteorological parameters, and the exclusion of candidate samples [33,34]. This method is also broadly applied to TMY synthesis in other districts of the world, such as Hong Kong [35], Cyprus [36] and Turkey [37]. Considering that meteorological parameters, such as humidity and wind speed, that have little effect on the design of SEUS, including concentrated solar power (CSP) and PV, were introduced by the NREL at the time of TMY synthesis, Cebecauer et al. designated only solar radiation and average dry-bulb temperature as composite characteristics, based on which a solar geographical information system synthesis approach focusing on the application of CSP and PV was presented [38].

Although these TMY synthesis methods have the capability to reflect the overall change trend of solar radiation for a given region, these methods ignore the DSR fluctuations and transition rules. In fact, these two features of DSR can provide important design guidance for the energy storage system and online control strategy for SEUS. Evidence shows that these features are important and could be modelled. Based on the historical data from 2004 to 2013, Hussain et al. determined that the transfer process of different mean DSR has a certain periodic variation rule by introducing time–frequency joint analysis [17]. Pearce et al. reported that rapid variations of solar energy largely affect the output of SEUS because of their short response time [39]. Therefore, in power distribution grids with high-density PV, fluctuations in the produced electrical power may occur, leading to unpredictable variations of node voltage and power in electrical networks. In small grids, such as those that exist on islands, these fluctuations can cause instabilities [40]. Meyer et al. observed that a 10-min solar radiation simulation step may result in an energy production error of 2% to 3% relative to a 1-min step [41].

In this study, four sites located in diverse time zones in the U.S. are selected to analyse the

CSR distributions. Afterward, a six-dimensional feature vector (containing six independent characteristic parameters) that represents the DSR fluctuations in each area based on CSR is defined. Then, k-means clustering is used to conduct clustering analysis for dividing historical DSR data into four classes. Afterward, the discrete-time Markov chain (DTMC) is adopted to build a model illustrating the transition rules, which include the distribution and transition probabilities, among the four types of DSR. The typical solar radiation year (TSRY) can be ultimately obtained based on the clustering categories and transition rules.

2 Data sources

The data source used in this research is derived from the Surface Radiation Budget Network (SURFRAD) database. SURFRAD consists of data from seven sites, and for convenience of presentation, four model sites are selected as the research objects, as shown in Fig. 1. The sites are located in four different time zones from UTC −6:00 to −8:00, and their altitudes range from 230 m to 1,007 m. The figure shows that the average global horizontal radiations of the sites have unique characteristics of representativeness. Among them, the site situated at Desert Rock has the highest average horizontal radiation value, which reached 5.7 kWh/m^2/day. By contrast, the site situated at Fort Peck only reached a minimum value of 3.76 kWh/m^2/day.

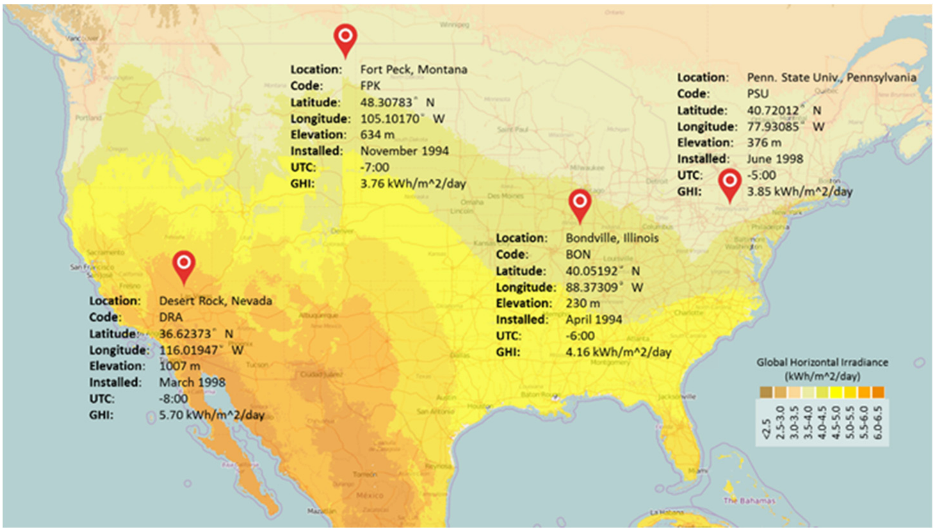

Fig. 1. Locations and information of selected four sites in SURFRAD database.

Compared with the solar radiation data provided by the National Solar Radiation Data Base with a sampling interval of 0.5 h or 1 h, SURFRAD exhibits a higher data acquisition frequency. SURFRAD data had an interval of 3 min before January 2009, after which it was changed to 1 min. The data collected with high sampling frequency provides the foundation for the research on DSR fluctuation.

3 Fluctuation patterns of solar radiation

Many studies have defined standardised factors that characterise solar radiation fluctuations because they can exclude the effects of absolute radiation differences on the results due to geographical differences. These standardised factors are essentially the same, but have different definitions for different research purposes. One of the factors, called clearness index, is defined as the ratio of the global horizontal radiation to the corresponding radiation available outside the atmosphere and is calculated using an empirical formula [42]. Muselli et al. utilised the clearness index as the characteristic parameter of solar radiation to classify the typical meteorological days from global irradiation records [43]. Maffi et al. combined the clearness index and daily fractals to build a daily solar irradiance classification model [44]. Marty et al. separated clear- and cloudy-sky situations for climate research by using the clear-sky index, which is defined as real apparent emittance divided by clear-sky situation [45].

In this paper, a factor called clear-sky ratio (CSR) is proposed. The CSR is defined as the downwelling short wave radiation (DSWR) divided by the clear-sky downwelling short wave radiation (CDSWR) and is estimated based on sensor data [46,47]. Fig. 2 illustrates the transformation of DSWR into CSR. Fig. 2(A) presents the DSWR and CDSWR on 1 June 2015 in BON region, and Fig. 2(B) shows the fluctuation of CSR corresponding to Fig. 2(A). CSR > 1 will only occur when the direct sunlight is not blocked and the clouds around the direct light path appear with strong scattering. Given that this condition rarely occurs and the duration is short, this study defines CSR as 1 for DSWR/CDSWR > 1 to facilitate subsequent analysis. Moreover, the fluctuation amplitude of radiation can reach 265.3 W/m^2/minute and that of the CSR can reach 0.315/minute. Drastic fluctuation on CSR can significantly influence the output and the operation of SEUS.

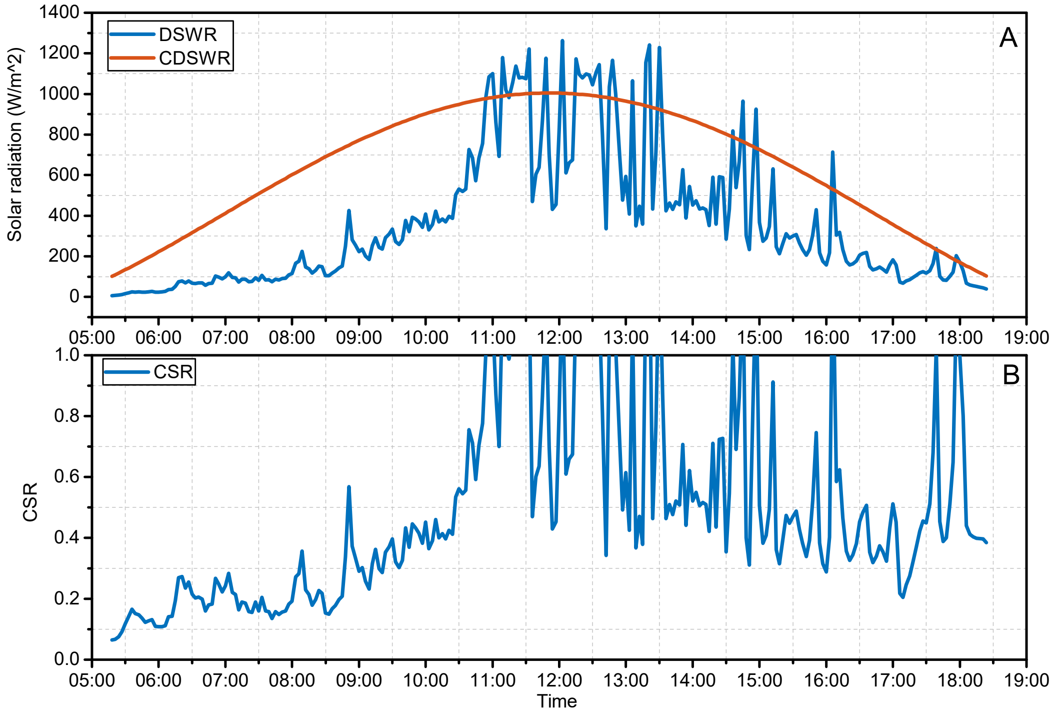

Fig. 2 An example of DSWR and CSR fluctuations.

Fig. 3 displays the CSR distributions of the four sites. The resolution of each image is 180 × 180, and the colour scale for each pixel is dimensionless. The colour scale denotes the probability of CSR at that pixel. The figure shows that the CSR in the DRA region is primarily distributed at approximately 1 at different timeframes, signifying that it is dominantly sunny. In comparison, the fluctuations of DSR in the other three sites are rather significant; the fluctuation in the FPK region is principally distributed around noon, whereas the fluctuations in the BON and PSU regions occur even during the daytime. The CSR of the FPK region fluctuates between 0.33 and 0.67, which is a smaller range than those of the BON and PSU regions. The comparison of the radiation rules of the BON and PSU regions showed that a CSR separation zone exists between 0.56 and 0.89 in PSU, which suggests that the weather in that area has two possibilities. One possibility is dominantly sunny and dusky, and rarely cloudy, and the other possibility is rapid cloud change rate. From the previously presented analysis, the radiation fluctuation of each of the four sites has their own unique features.

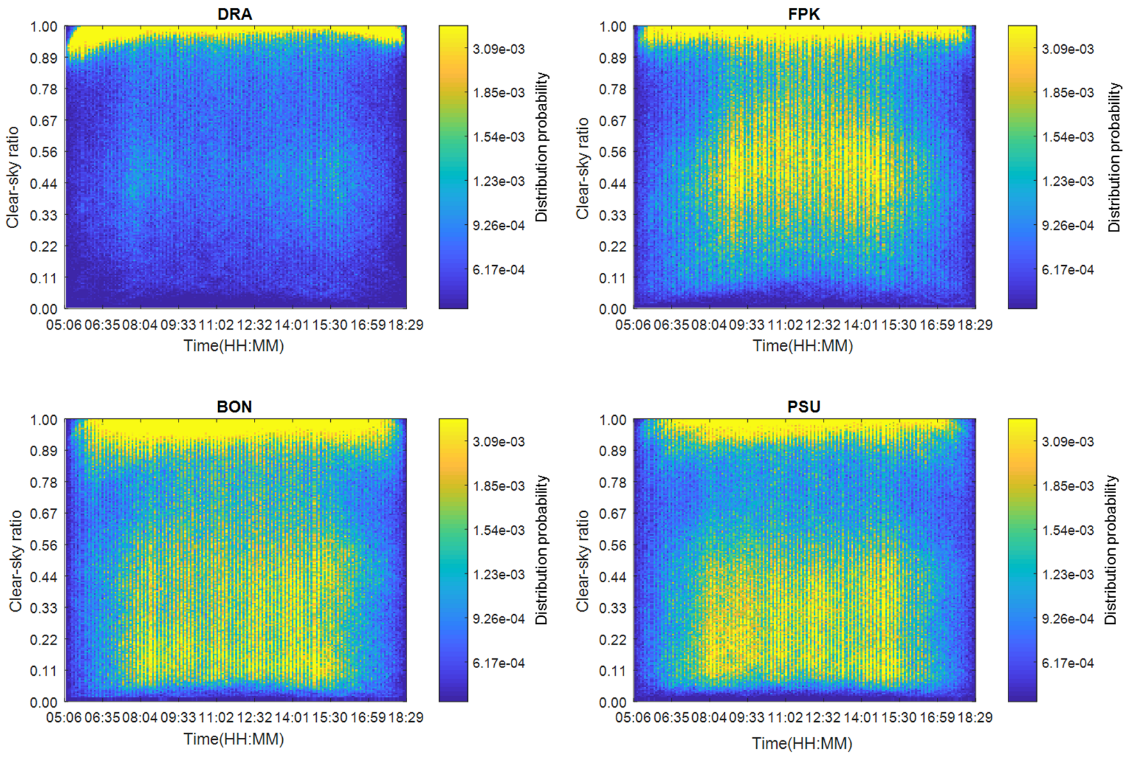

Fig. 3 Distributions of CSR for the four selected sites

4 Methods

As presented in Fig. 4, a typical solar radiation model that comprehensively reflects the fluctuation patterns of DSR is constructed in line with the following process. First, the CSR is derived from the measured value provided by SURFRAD and the clear-sky radiation value estimated by Long et al [47]. Second, a six-dimensional feature vector (containing six independent characteristic parameters) is extracted to present the solar fluctuation rules. Then, based on the feature vector and k-means algorithm, clustering is conducted during individual days specific to the CSR. Afterward, DTMC is used to model the transition rules among the individual classes of clustering results. By selecting samples for every clustering class centre combined with the transition sequences generated by DTMC, a typical CSR sequence can be obtained. Finally, in combination with the CDSWR, the TSRY can be synthesised. Among them, the number of clustering classes and the length of radiation can be designated as needed. In this study, the radiation fluctuations of the selected regions are divided into four classes. Meanwhile, solar radiation is divided on a quarterly basis (three months) for analysis and discussion.

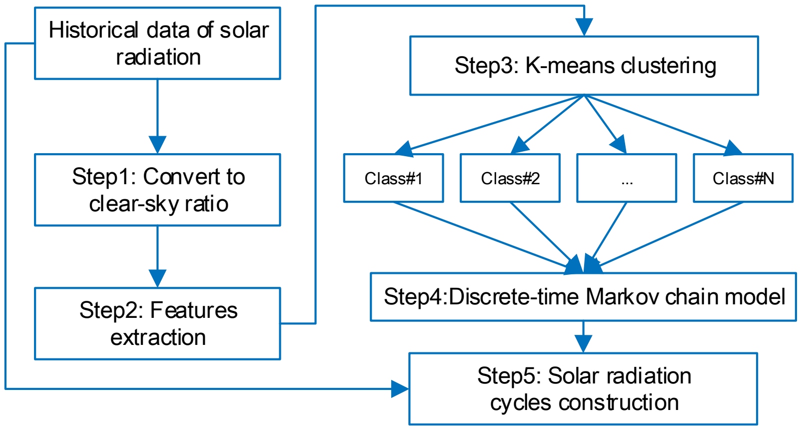

Fig. 4 The construction processes of Typical Solar Radiation Year

The differences between TSRY and TMY in the construction method are as follows:

1. Inputs are different: TSRY only takes the solar radiation features as inputs. In addition to the solar radiation, TMY takes some other meteorological parameters, such as humidity and wind speed as well. This allows TMY to more fully reflect the meteorological characteristics of a region, but considering these parameters have little effects on the design of SEUS, including solar power (CSP) and PV, the introduction of these inputs may be negative for typical DSR selection influences[38].
2. Selection criteria are different: TMY is usually synthesized using Finkelstein-Schafer (FS) statistics method. This method is an empirical methodology for selecting individual months from different years over the available period based on the comparison between the long-term cumulative frequency distribution Function (CDF) of each month and the CDF for each individual year of the month. The problem with CDF based method is that it can only reflect the distribution of radiation and does not reflect the time-dependent features of DSR such as the rate of change of radiation. TSRY introduced a multidimensional feature classification method (k-means) to solve this problem.
3. Time scales are different: The TMY synthesis scale is one month, which means that all days of the selected month are treated as typical DSRs into the final meteorological year. And it is difficult to guarantee for a region with less historical data. TSRY selects typical DSRs on a daily scale and models the sample sequence with the DTMC. So, TSRY could be used to build a typical solar radiation year even at a location with less historical data.

4.1 Definition of fluctuation features

A six-dimensional feature vector is defined to represent the variation of CSR. As shown in

Fig. 5, the weather is classified into sunny, dusky and cloudy days based on the CSR. The extraction of the feature vector is shown in the following pseudo-code:

```
If CSR >0.95 and duration > 30min
    The fragment is sunny
Else if CSR <0.3 and duration >30min
    The fragment is dusky
Else
    The fragment is cloudy
End
F1 = Dusky duration/sunshine time
F2 = Cloudy duration/sunshine time
F3 = Average CSR under cloudy
F4 = Standard deviation of CSR under cloudy
F5 = Maximum change rate of CSR under cloudy
F6 = Average change rate of CSR under cloudy
Feature vector = [F1,F2,F3,F4,F5,F6]
```

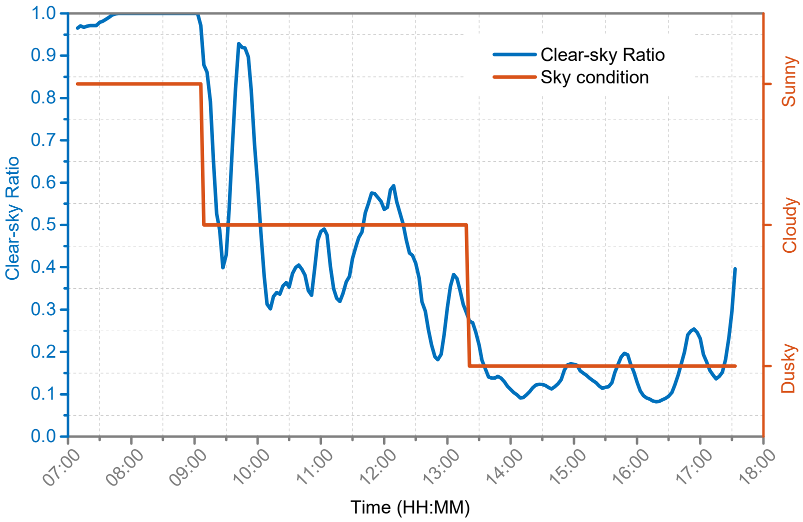

Fig. 5 Features definition of clear-sky index

The correlation analysis results, which were verified by the significance test (p = 0.001), between the six features showed that the correlation between F1 and F3 is relatively high. It can reach -0.77, -0.76, -0.88, -0.87 for DRA, FPK, BON and PSU respectively. It indicates that the removal of F1 or F3 does not have a significant effect on the clustering results theoretically. However, it needs to be noticed that the final synthesis accuracy in the three of all the four regions has been reduced after removing the F3 from the feature vector. The average error is increased from 6% to 13% for DRA region, from 7% to 9% for FPK region, from 9% to 13% for BON region, only the average error in PSU region is decreased slightly from 10% to 9 %.

Considering that the algorithm in this study is not very sensitive to the computational load, six features are all adopted in the clustering model of DSR.

As discussed previously, based on the six-dimensional feature vector and historical data set, the DSR can be classified and a typical representation of each class can be obtained using k-means clustering.

4.2 *k*-means for clustering

The k-means clustering algorithm is used to conduct clustering analysis for DSR. The algorithm could be used to divide samples into k disjoint clusters based on their feature vectors [48]. Objects that are classified into the same cluster have similar feature values. k is a positive integer number specifying the number of clusters and has to be provided in advance. For a given set of observations $\{X_1, X_2, \dots, X_n\}$, where each observation X is a d-dimensional feature vector $[F_1, F_2, \dots, F_d]$, k-means clustering aims to partition the n objects into k $(\leq n)$ sets $S = \{S_1, S_2, \dots, S_k\}$ to minimise the within-cluster sum of squares (sum of the distance functions of each point in the cluster to the k centroid). The mathematical description is shown in Equation (1), where U_i is the centroid of S_i, which denote a typical feature vector of S_i. A distance function is required to compute the distance (i.e., similarity) between two objects. Many distance function, such as Euclidean and cosine distances, are commonly used [49]. DSR clustering considers the differences in the overall trend of all features rather than the absolute value of each feature. For this reason, as shown in Equation (2), cosine distance is chosen as the distance function, where X_s, X_t refer to two feature vectors, whereas d_{st} refers to the distance between the two feature vectors. The steps of the k-means clustering algorithm are shown in Table 1 [50].

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \left\| X - U_i \right\|^2 \tag{1}$$

$$d_{\mathrm{st}} = 1 - \frac{X_{\mathrm{s}} X_t'}{\sqrt{\left(X_{\mathrm{s}} X_s'\right)\left(X_t X_t'\right)}} \tag{2}$$

Table 1 *k*-means algorithm for DSR clustering

Index	Actions	Remarks
Step 1	Define the number of clusters k.	DSRs were divided into four classes ($k = 4$) in this study.
Step 2	Initialise the k cluster centroids by arbitrarily dividing all	In this study, the k initial cluster centres are initialised by random

	objects into k clusters, computing their centroids and verifying that all centroids are different from each other.	instances from the solar historical data set, which are composed of the feature vector of the CSR.
Step 3	Iterate over all objects and compute the distances to the centroids of all clusters. Assign each object to the cluster with the nearest centroid.	Each instance of DSR is assigned to its closest cluster centre based on the cosine distance.
Step 4	Recalculate the centroids of the modified clusters.	Each cluster centre is updated to be the mean of its constituent instances.
Step 5	Repeat Step 3 until the centroids do not change any more.	The final centroids are the typical DSR instances for each class. Each typical instance could represent the average level of the corresponding class.

4.3 Discrete-time Markov chain for transition rule estimation

DTMC could be used to further process the clustering results obtained by the k-means algorithm. DTMC is a random process that undergoes transition from one state to another state on a state space, and it must possess a property that is usually characterised as "memorylessness": the probability distribution of the subsequent state depends only on the current state, and not on the sequence of events that preceded it. This specific kind of "memorylessness" is called the Markov property. A Markov chain is specified by the following components: $Q = \{q_1, q_2, \dots, q_N\}$ denotes a set of N states, as shown in Equation (3), and A is a transition probability matrix between N states where each a_{ij} represents the probability of moving from states i to j [51]. Therefore, in a first-order Markov chain, the probability of a particular state depends only on the previous state. This assumption is shown in Equation (4). Notably, given that each a_{ij} expresses the probability $P(q_j|q_i)$, as shown in Equation (5), the laws of probability require that the output values of a given state must have a sum of 1. In this study, the number of daily solar types is 4, such that A is a 4-by-4 transition probability matrix. For example, a_{12} represents the probability that the next day's weather belongs to class 2 if the current weather belongs to class 1.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \tag{3}$$

$$P(q_i \mid q_1 \cdots q_{i-1}) = P(q_i \mid q_{i-1}) \tag{4}$$

$$\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i \tag{5}$$

The daily weather is assumed to be independent of each other. Thus, the DTMC can estimate the transition probability of this random process. DTMC processing starts by calculating the probability distribution of each DSR class without considering the order between them and then estimating the transition probabilities according to the transfer order between each class. As such, the DTMC can model the transition rules of DSR.

5 Results and discussion

As discussed previously, the two important steps in the synthesis process are k-means-based clustering and DTMC-based transition rule estimation. Given the space constraints, a typical synthesis process of DSR for the four regions only in spring was described in detail. All the data of the other seasons are given in Tables 2 to 4.

5.1 Analysis of the clustering results

The k-means clustering results of the four regions in spring are given in Fig. 6. The six-dimensional fluctuation feature vector $[F_1, \ldots, F_6]$ that comprise each class is described in Section 4.1. The four coloured blocks denote the feature distributions of four classes, namely, Class 1 (C_1), Class 2 (C_2), Class 3 (C_3) and Class 4 (C_4).

From Fig. 6, information can be analysed from the difference between different cluster classes in four regions, as follows:

[1] $\boldsymbol{C_1}$: Sunny days are dominant in this class. The DSR in this class has a similar distribution, with a lower proportion of dusky (F_1) and cloudy (F_2) and a higher mean value of CSR for cloudy (F_3) among the four sites.

[2] $\boldsymbol{C_2}$: Dusky days has the largest proportion in this class. The DSR in this class has a similar distribution, with a higher proportion of dusky (F_1), a lower proportion of cloudy (F_2) and a lower mean value of CSR for cloudy (F_3) among the four sites.

[3] $\boldsymbol{C_3}$: Cloudy has the principal percentage in this class. The DSR in this class has a similar distribution, with a lower proportion of dusky (F_1) and a higher proportion of cloudy (F_2) among the four sites.

[4] $\boldsymbol{C_4}$: Mixed weather dominates the majority of this class. Mixed weather (C_4) is mainly constituted by sunny and cloudy in the DRA and FPK regions, but is composed of dusky and cloudy in the BON and PSU regions.

The results showed that the six-dimensional fluctuation feature vector defined in this study and the adopted clustering algorithm can be used to obtain four classes of typical DSR. Based on the clustering results, the transfer rules between different classes based on DTMC are analysed in the subsequent section.

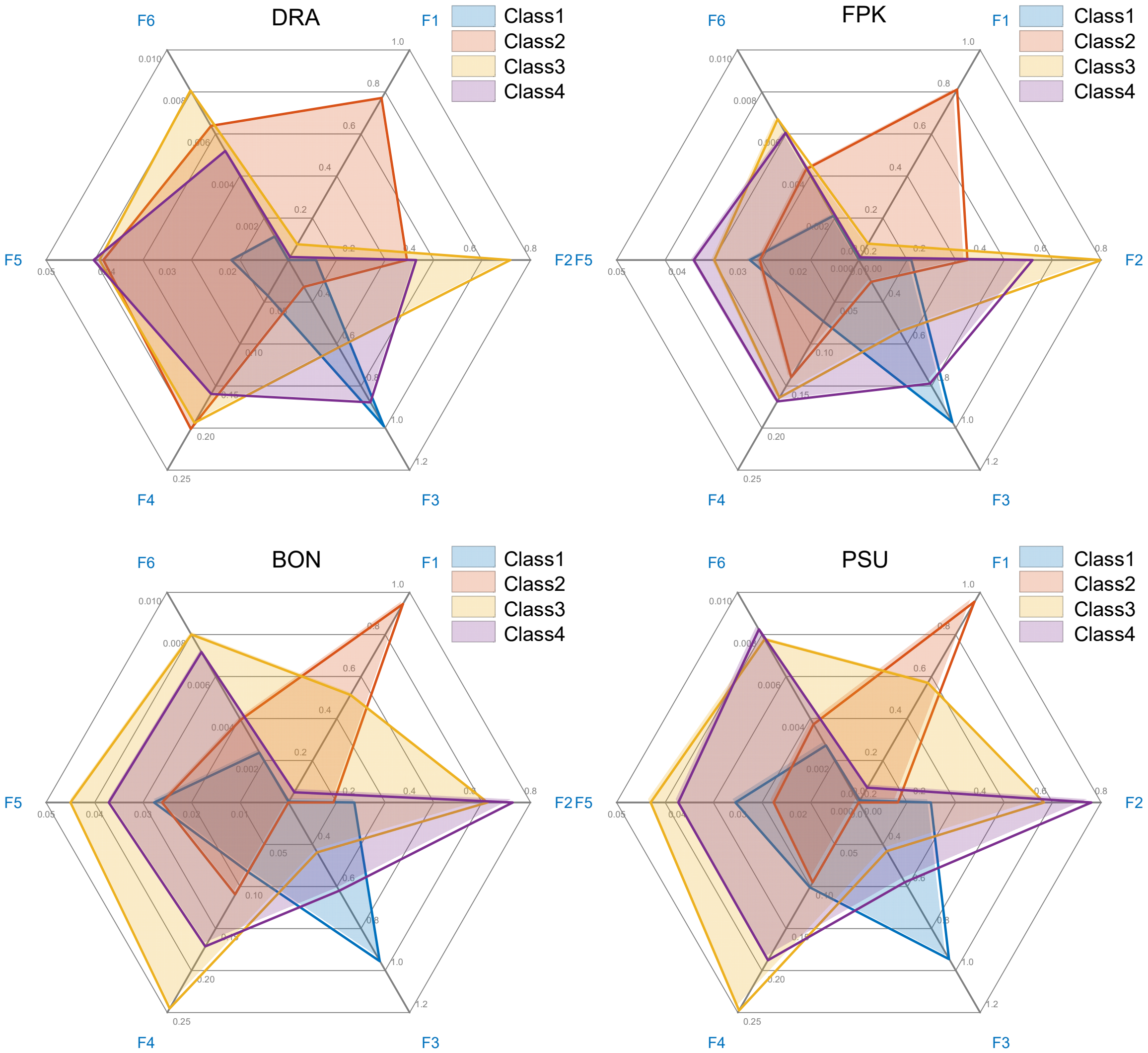

Fig. 6 Features distribution during spring for individual class at four different regions

Table 2 Features distribution for individual class at four different regions

		DRA				FPK				BON				PSU			
		C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
Spring	F1	0.00	0.77	0.07	0.01	0.00	0.81	0.08	0.01	0.00	0.95	0.51	0.05	0.01	0.95	0.57	0.07
	F2	0.09	0.39	0.73	0.42	0.17	0.36	0.80	0.57	0.22	0.15	0.66	0.74	0.24	0.13	0.61	0.77
	F3	0.99	0.33	0.62	0.88	0.97	0.30	0.54	0.79	0.95	0.19	0.44	0.62	0.94	0.19	0.43	0.58
	F4	0.04	0.20	0.19	0.16	0.07	0.14	0.16	0.17	0.08	0.11	0.24	0.17	0.10	0.09	0.25	0.19
	F5	0.02	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.03	0.03	0.04	0.04	0.03	0.02	0.04	0.04
	F6	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
Summer	F1	0.00	0.01	0.04	0.45	0.01	0.79	0.12	0.03	0.87	0.00	0.38	0.03	0.88	0.01	0.35	0.04
	F2	0.07	0.37	0.68	0.55	0.17	0.36	0.74	0.52	0.29	0.29	0.69	0.72	0.28	0.27	0.72	0.72
	F3	0.99	0.90	0.68	0.55	0.97	0.32	0.59	0.82	0.27	0.93	0.51	0.65	0.26	0.94	0.50	0.65
	F4	0.03	0.15	0.21	0.34	0.09	0.20	0.23	0.19	0.18	0.10	0.30	0.17	0.16	0.10	0.28	0.18
	F5	0.02	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.04	0.04	0.04	0.05	0.04
	F6	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.01
Autumn	F1	0.00	0.02	0.06	0.61	0.00	0.80	0.12	0.02	0.00	0.93	0.47	0.04	0.00	0.94	0.48	0.04
	F2	0.06	0.37	0.72	0.48	0.13	0.35	0.76	0.51	0.19	0.18	0.65	0.72	0.23	0.17	0.67	0.74
	F3	0.99	0.90	0.64	0.44	0.98	0.31	0.56	0.83	0.96	0.21	0.47	0.65	0.95	0.21	0.45	0.62
	F4	0.03	0.17	0.21	0.28	0.06	0.18	0.20	0.18	0.07	0.12	0.28	0.17	0.10	0.11	0.27	0.18
	F5	0.01	0.05	0.05	0.05	0.03	0.03	0.04	0.05	0.03	0.03	0.05	0.04	0.03	0.03	0.05	0.04
	F6	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
Winter	F1	0.00	0.86	0.06	0.01	0.00	0.83	0.05	0.01	0.01	0.96	0.53	0.04	0.00	0.95	0.58	0.06
	F2	0.07	0.28	0.77	0.45	0.11	0.33	0.82	0.53	0.21	0.10	0.66	0.78	0.24	0.13	0.62	0.80
	F3	0.99	0.27	0.59	0.86	0.98	0.30	0.52	0.81	0.95	0.18	0.43	0.58	0.94	0.19	0.42	0.55
	F4	0.04	0.16	0.18	0.16	0.06	0.11	0.13	0.17	0.08	0.08	0.21	0.15	0.10	0.09	0.22	0.17
	F5	0.02	0.03	0.04	0.04	0.02	0.02	0.03	0.04	0.03	0.02	0.04	0.03	0.04	0.02	0.04	0.04
	F6	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01

5.2 Analysis of the transition rules

Based on the clustering results, DTMC is utilised to calculate the transition rules, which include the distribution proportion and the transition probability of individual classes of DSR. These two parameters are used to determine the sequence of DSR that comprises TSRY.

5.2.1 Distribution proportion

As shown in Fig. 7, the distribution proportions of the four classes in the four regions in spring have the following rules:

- In the DRA region, sunny days are dominant and dusky days account for only a small percentage. The proportion of *sunny days* (C_1) is 0.44, whereas that of *dusky days* (C_2) is only 0.04. This means that the potential of solar energy in the DRA region is high and SEUS has no harsh requirements for the capacity of the energy storage equipment, such as thermal energy storage or capacitors, used to suppress output fluctuations.

- In the FPK region, cloudy days and mixed weather are predominant. The proportion of *cloudy days* (C_3) is 0.42, whereas that of *mixed weather* (C_4) is 0.29. The proportion of dusky days is low at 0.07. This indicates that the solar-only power generation system in the area needs to have a large enough energy storage equipment or the solar field size in a hybrid power generation system needs to be well designed to suppress the impact of DSR fluctuations on the output.
- In the BON and PSU regions, mixed weather is predominant. The proportion of *mixed weather* (C_4) is 0.46 and 0.45 at these two sites.

5.2.2 Transition probability

The transition probability among individual classes of the four regions in spring is also shown in Fig. 7. The figure shows the following transfer laws among different classes:

- In the DRA region, sunny weather likely occurs for many continuous days because the transition from *sunny* (C_1) to *sunny* (C_1) has the highest probability (0.68). Moreover, *sunny* (C_1) in the DRA region has a high proportion based on the distribution analysis.
- In the FPK region, each type of DSR is rarely presented consecutively because the self-transfer probability of these four types of solar radiation is low (the maximum is not more than 0.39).
- In the BON and PSU regions, continuously mixed weather has a high probability. As the probability of *mixed weather* (C_4) to transfer to itself is high (up to 0.45 and 0.50 for BON and PSU, respectively). When the weather type falls into *dusky* (C_2) and *cloudy* (C_3), the probability of shifting to *mixed weather* (C_4) during the next day is high.

The relatively stable solar radiation characteristics indicate that the DRA area is suitable for deploying multiple types of SEUS, such as organic Rankine cycle-based combined heat and power systems or PV systems. In contrast, the FPK area is dominated by cloudy weather and day-to-day variation is significant, indicating that the area is not suitable for deployments the SEUS that require long start or stabilize time. BON and PSU regions are dominated by mixed weather and the day-to-day variation is small, so multiple types of SEUS can be deployed with sufficient energy storage system.

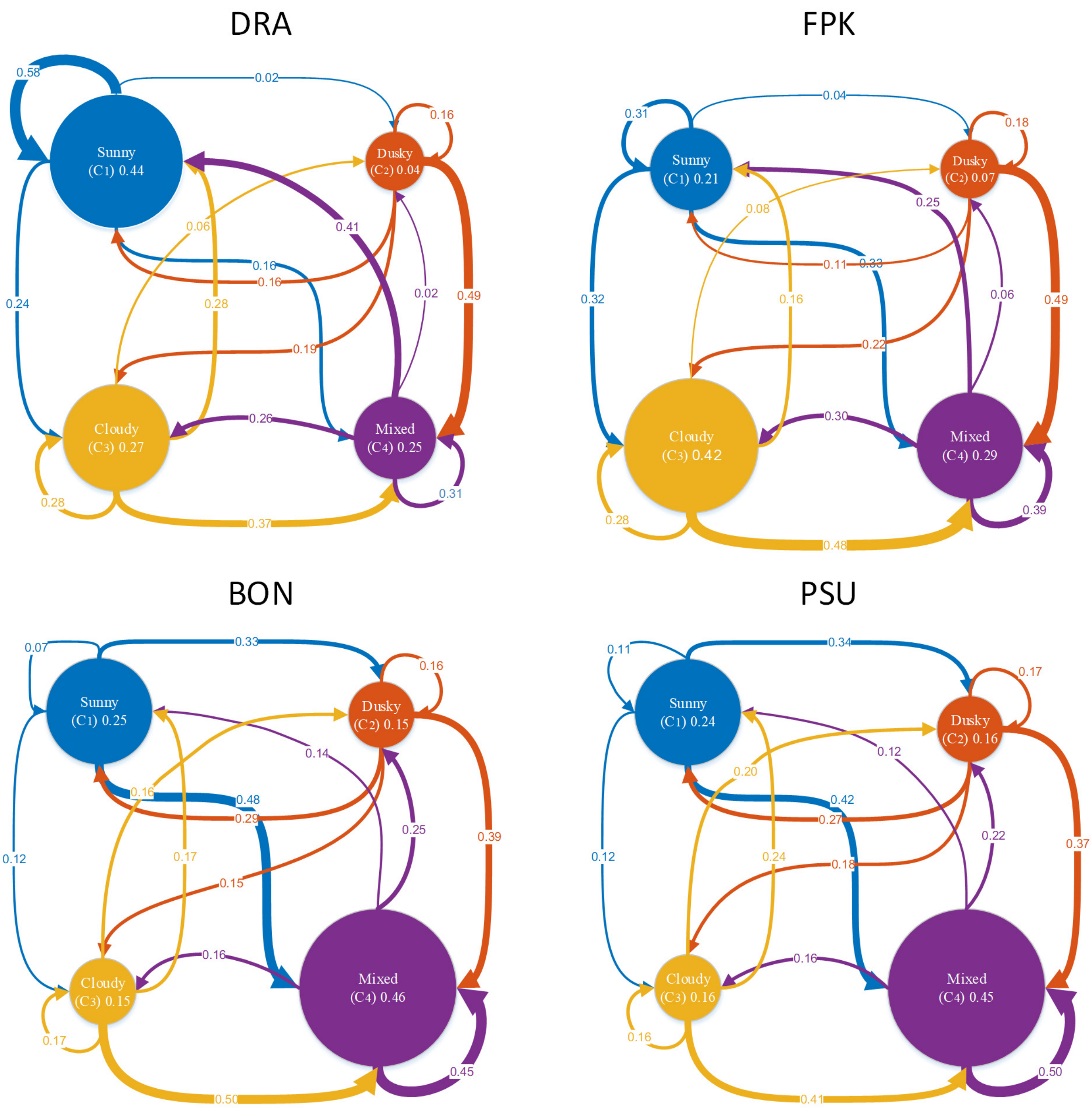

Fig. 7 Distribution proportion and transition probability of the four daily solar classes

Table 3 Distribution proportion for individual classes in four seasons

	DRA				FPK				BON				PSU			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
Spring	0.44	0.04	0.27	0.25	0.21	0.07	0.42	0.29	0.25	0.15	0.15	0.46	0.09	0.17	0.22	0.52
Summer	0.60	0.20	0.18	0.03	0.23	0.08	0.36	0.32	0.07	0.24	0.18	0.51	0.23	0.11	0.16	0.50
Autumn	0.60	0.23	0.15	0.03	0.30	0.08	0.33	0.29	0.34	0.07	0.13	0.46	0.12	0.18	0.16	0.16
Winter	0.44	0.04	0.27	0.25	0.20	0.06	0.50	0.24	0.27	0.20	0.14	0.38	0.19	0.20	0.15	0.46

Table 4 Transition probabilities for individual classes in four seasons

		DRA				FPK				BON				PSU			
		S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
Spring	E1	0.58	0.16	0.28	0.41	0.31	0.11	0.16	0.25	0.07	0.29	0.17	0.14	0.11	0.27	0.24	0.12
	E2	0.02	0.16	0.06	0.02	0.04	0.18	0.08	0.06	0.33	0.16	0.16	0.25	0.34	0.17	0.20	0.22
	E3	0.24	0.19	0.28	0.26	0.32	0.22	0.28	0.30	0.12	0.15	0.17	0.16	0.12	0.18	0.16	0.16
	E4	0.16	0.49	0.37	0.31	0.33	0.49	0.48	0.39	0.48	0.39	0.50	0.45	0.42	0.37	0.41	0.50
Summer	E1	0.74	0.48	0.29	0.30	0.41	0.09	0.14	0.25	0.14	0.43	0.18	0.19	0.04	0.32	0.12	0.16
	E2	0.16	0.25	0.27	0.30	0.03	0.22	0.11	0.06	0.18	0.04	0.11	0.05	0.21	0.05	0.12	0.07
	E3	0.01	0.03	0.09	0.09	0.31	0.22	0.31	0.36	0.29	0.11	0.22	0.17	0.40	0.12	0.28	0.20
	E4	0.09	0.24	0.36	0.32	0.25	0.48	0.44	0.33	0.38	0.42	0.49	0.58	0.35	0.51	0.49	0.57
Autumn	E1	0.74	0.45	0.33	0.26	0.45	0.13	0.23	0.27	0.04	0.30	0.07	0.06	0.03	0.33	0.14	0.09
	E2	0.18	0.31	0.30	0.30	0.03	0.23	0.12	0.06	0.50	0.18	0.22	0.28	0.41	0.08	0.17	0.21
	E3	0.01	0.03	0.09	0.15	0.27	0.19	0.29	0.34	0.09	0.17	0.21	0.13	0.10	0.23	0.19	0.15
	E4	0.08	0.22	0.28	0.28	0.25	0.45	0.36	0.34	0.37	0.35	0.49	0.54	0.47	0.35	0.50	0.55
Winter	E1	0.58	0.14	0.27	0.45	0.30	0.09	0.17	0.20	0.09	0.36	0.27	0.18	0.12	0.32	0.25	0.17
	E2	0.01	0.21	0.07	0.04	0.02	0.21	0.05	0.05	0.41	0.15	0.23	0.25	0.29	0.11	0.16	0.20
	E3	0.25	0.19	0.30	0.20	0.30	0.12	0.22	0.26	0.11	0.17	0.16	0.14	0.14	0.22	0.12	0.13
	E4	0.16	0.46	0.36	0.32	0.38	0.57	0.56	0.48	0.39	0.32	0.34	0.43	0.45	0.36	0.47	0.50

5.3 Verification of TSRY

TSRY is finally synthesised based on the results presented in Sections 5.1 and 5.2. The final synthesis results of TSRY and traditional TMY are compared to evaluate this method. Afterward, the characteristic parameters of TSRY and historical data are also compared.

5.3.1 Comparisons between TSRY and TMY

The comparison results shown in Fig. 8 indicate that the overall trend of the output of TSRY and TMY is the same. This finding shows that TSRY can also reflect the annual variation of solar radiation as TMY. However, as shown in the enlarged subgraph, the DSR features have several differences. The solar radiation feature of TMY and TSRY is basically the same for the DRA region, with less fluctuation. In the other regions, where solar radiation exhibits more fluctuation, significant differences between TMY and TSRY are observed. TMY uses a 1-hour sampling interval, which leads to the loss of most of the fluctuating features of solar radiation, whereas TSRY uses a 3-min sampling interval, which preserves the fluctuating features of solar radiation. In the subsequent section, the ability of TSRY to express the fluctuating patterns of solar radiation is further validated by comparing with historical data.

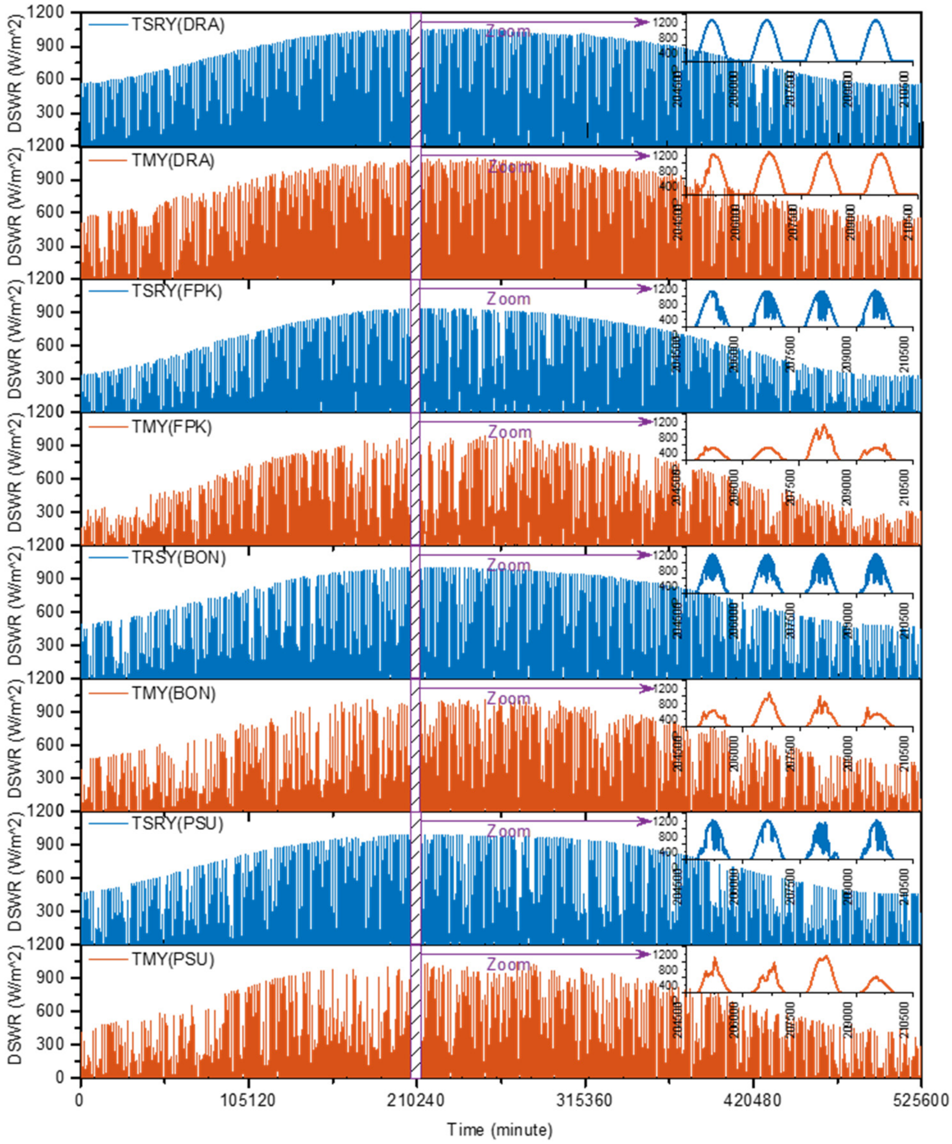

Fig. 8 Comparisons between TSRY and TMY

5.3.2 Comparisons between TSRY and historical data

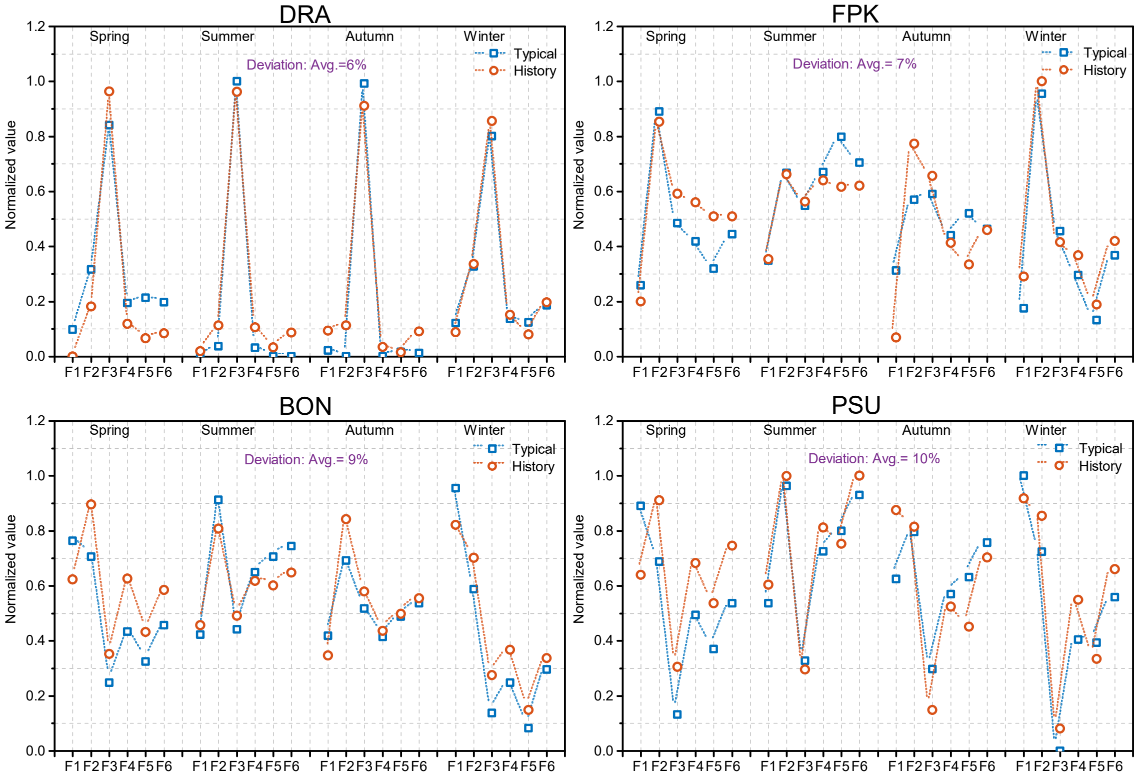

Fig. 9 Features distribution comparisons between synthesis results and historical data

Fig. 9 shows that TSRY can reflect the outstanding characteristics of the four regions and four seasons. The six characteristic parameters of TSRY have distributions similar to that of historical data. The average deviation is 6% for the DRA region, 7% for the FPK region, 9% for the BON region and 10% for the PSU region. This finding indicates that TSRY synthesis based on the method proposed in this study has the capability to represent the characteristics of DSR for a given region.

Fig. 9 also shows that the deviations in spring and autumn are higher than that in summer or winter because DSR is more prone to change during spring and autumn. For the same reason, the final synthesis result deviation is highest in the PSU region where solar radiation fluctuations are stronger than in the other regions. Increasing the solar clustering classes or reducing the length of clustering segments can be used to improve the TSRY synthesis precision with frequent fluctuation. However, these two improved methods would need more samples for modelling.

6 Conclusions

An innovative TSRY synthesis method is proposed to address the inability of the existing TMY synthesis approach in efficiently representing the DSR fluctuation and transition rules.

DSR data from four selected regions are selected as research basis, and the CSR distribution of these regions is analysed. The results showed that sunny weather is predominant in the DRA region, although solar radiation showed different fluctuation patterns. Fluctuation in the FPK region usually occurs around noon, resulting in greater changes in solar radiation values compared with morning or evening. In the BON and PSU regions, the fluctuation is evenly distributed throughout the entire daytime.

Based on the analysis, a six-dimensional feature vector is proposed to represent the solar radiation fluctuations. Then, k-means clustering is used to cluster the DSR into four classes. The clustering results of all cases show that $C_1 - C_4$ represent sunny, dusky and cloudy days and mixed weather, although individual regions have its own unique characteristics.

Based on the clustering results, DTMC was used to model the transition rules, which include the distribution and transition probabilities, of the four classes based on k-means clustering. Distribution analysis shows that sunny and cloudy days dominate the DRA and FPK regions, respectively, whereas mixed weather dominates the BON and PSU regions. Transition probability analysis shows that the DRA region has the highest probability of continuous sunny days. However, in the FPK region, each type of DSR is rarely presented consecutively. Meanwhile, in the BON and PSU regions, continuous mixed weather has a high probability.

The comparison results of TSRY and TMY show that TSRY can also accurately reflect the annual variation of solar radiation as TMY. In addition, the six-dimensional vectors of TSRY and historical DSR data are compared. The results of the comparison for the four regions show that the average error of synthesised TSRY has the maximum and minimum values of 10% and 6%, respectively, indicating that the synthesised TSRY adopted in this study can successfully represent the fluctuation and transition patterns of DSR in a certain region.

Subsequently, relevant studies from two perspectives will be conducted. First, the diverse characteristic parameter combinations of DSR for constructing the TSRY that are targeted in different applications, including the weighting aspect of the parameters, the average dry-bulb temperature and the radiation duration, will be investigated. Second, the influence of historical data sample sizes on the precision of the final synthesis outcomes will be explored to determine the minimum sample size of historical meteorological data that can satisfy the synthesis requirements.

Acknowledgements

The authors sincerely appreciate the help from the researchers affiliated with the SURFRAD network. The authors are grateful to the Earth System Research Laboratory, Global Monitoring Division of the National Oceanic and Atmospheric Administration for providing the radiation data.

References

[1] Zhang M, Xu C, Du Xet al., Off-design performance of concentrated solar heat and coal double-source boiler power generation with thermocline energy storage, Applied Energy, 2017, 189: 697~710.

[2] Zhao Y, Hong H, Jin H, Optimization of the solar field size for the solar－coal hybrid system, Applied Energy, 2017, 185: 1162~1172.

[3] Chang T, Liu F, Ko Het al., Oscillation characteristic study of wind speed, global solar radiation and air temperature using wavelet analysis, Applied Energy, 2017, 190: 650~657.

[4] Lappalainen K, Valkealahti S, Output power variation of different pv array configurations during irradiance transitions caused by moving clouds, Applied Energy, 2017, 190: 902~910.

[5] Belkaid A, Colak I, Isik O, Photovoltaic maximum power point tracking under fast varying of solar radiation, Applied Energy, 2016, 179: 523~530.

[6] Kaplani E, Kaplanis S, A stochastic simulation model for reliable pv system sizing providing for solar radiation fluctuations, Applied Energy, 2012, 97: 970~981.

[7] Pashiardis S, Kalogirou S A, Pelengaris A, Statistical analysis for the characterization of solar energy utilization and inter-comparison of solar radiation at two sites in cyprus, Applied Energy, 2017, 190: 1138~1158.

[8] Zarzo M, Martí P, Modeling the variability of solar radiation data among weather stations by means of principal components analysis, Applied Energy, 2011, 88 (8): 2775~2784.

[9] Bulut H, Büyükalaca O, Simple model for the generation of daily global solar-radiation data in turkey, Applied Energy, 2007, 84 (5): 477~491.

[10]Li H, Ma W, Lian Yet al., Estimating daily global solar radiation by day of year in china, Applied Energy, 2010, 87 (10): 3011~3017.

[11]Besharat F, Dehghan A A, Faghih A R, Empirical models for estimating global solar radiation: a review and case study, Renewable and Sustainable Energy Reviews, 2013, 21: 798~821.

[12]Hassan G E, Youssef M E, Mohamed Z Eet al., New temperature-based models for predicting global solar radiation, Applied Energy, 2016, 179: 437~450.

[13]Janjai S, Sricharoen K, Pattarapanitchai S, Semi-empirical models for the estimation of clear sky solar global and direct normal irradiances in the tropics, Applied Energy, 2011, 88 (12): 4749~4755.

[14]Kaushika N D, Tomar R K, Kaushik S C, Artificial neural network model based on interrelationship of

direct, diffuse and global solar radiations, Solar Energy, 2014, 103: 327~342.

[15]_enkal O, Kuleli T, Estimation of solar radiation over turkey using artificial neural network and satellite data, Applied Energy, 2009, 86 (7-8): 1222~1228.

[16]Landeras G, López J J, Kisi Oet al., Comparison of gene expression programming with neuro-fuzzy and neural network computing techniques in estimating daily incoming solar radiation in the basque country (northern spain), Energy Conversion and Management, 2012, 62: 1~13.

[17]Hussain S, Al-Alili A, A new approach for model validation in solar radiation using wavelet, phase and frequency coherence analysis, Applied Energy, 2016, 164: 639~649.

[18]Hocaoğlu F O, Stochastic approach for daily solar radiation modeling, Solar Energy, 2011, 85 (2): 278~287.

[19]Bhardwaj S, Sharma V, Srivastava Set al., Estimation of solar radiation using a combination of hidden markov model and generalized fuzzy model, Solar Energy, 2013, 93: 43~54.

[20]Voyant C, Darras C, Muselli Met al., Bayesian rules and stochastic models for high accuracy prediction of solar radiation, Applied Energy, 2014, 114: 218~226.

[21]Deo R C, Wen X, Qi F, A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset, Applied Energy, 2016, 168: 568~593.

[22]Kaplanis S, Kaplani E, A model to predict expected mean and stochastic hourly global solar radiation i(h;nj) values, Renewable Energy, 2007, 32 (8): 1414~1425.

[23]Kaplanis S, Kaplani E, Stochastic prediction of hourly global solar radiation for patra, greece, Applied Energy, 2010, 87 (12): 3748~3758.

[24]Kaplani E, Kaplanis S, Prediction of solar radiation intensity for cost-effective pv sizing and intelligent energy buildings: InTech, 2012,.

[25]Amp M, X, úriet al., A new gis-based solar radiation model and its application to photovoltaic assessments, TRANSACTIONS IN GIS, 2004, .

[26]Mehleri E D, Zervas P L, Sarimveis Het al., Determination of the optimal tilt angle and orientation for solar photovoltaic arrays, Renewable Energy, 2010, 35 (11): 2468~2475.

[27]Mellit A, Pavan A M, Performance prediction of 20kw p grid-connected photovoltaic plant at trieste (italy) using artificial neural network, Energy Conversion and Management, 2010, 51 (12): 2431~2441.

[28]Lou S, Li D H W, Lam J Cet al., Prediction of diffuse solar irradiance using machine learning and multivariable regression, Applied Energy, 2016, 181: 367~374.

[29]Bulut H, Generation of typical solar radiation data for istanbul, turkey, International Journal of Energy Research, 2003, 27 (9): 847~855.

[30]Yang L, Wan K K, Li D Het al., A new method to develop typical weather years in different climates for building energy use studies, Energy, 2011, 36 (10): 6121~6129.

[31]Hall I J, Prairie R R, Anderson H Eet al., Generation of a typical meteorological year: Sandia Labs., Albuquerque, NM (USA), 1978,.

[32]Finkelstein J M, Schafer R E, Improved goodness-of-fit tests, Biometrika, 1971, 58 (3): 641~645.

[33]Marion W, Urban K, User's manual for tmy2s: typical meteorological years: derived from the 1961-1990 national solar radiation data base: National Renewable Energy Laboratory, 1995,.

[34]Wilcox S, Anderberg M, George Ret al., Completing production of the updated national solar radiation database for the united states, Proceedings of the solar conference: American solar energy society; American Institute of Architects, 2007, 1.

[35]Chan A L, Chow T, Fong S Ket al., Generation of a typical meteorological year for hong kong, Energy Conversion and Management, 2006, 47 (1): 87~96.

[36]Kalogirou S A, Generation of Typical meteorological year (tmy-2) for nicosia, cyprus, Renewable Energy, 2003, 28 (15): 2317~2334.

[37]Pusat S, Ekmekçi I, Akkoyunlu M T, Generation of typical meteorological year for different climates of turkey, Renewable Energy, 2015, 75: 144~151.

[38]Cebecauer T, Suri M, Typical meteorological year data: solargis approach, Energy Procedia, 2015, 69: 1958~1969.

[39]Pearce J M, Expanding photovoltaic penetration with residential distributed generation from hybrid solar photovoltaic and combined heat and power systems, Energy, 2009, 34 (11): 1947~1954.

[40]Soubdhan T, Emilion R, Calif R, Classification of daily solar radiation distributions using a mixture of dirichlet distributions, Solar Energy, 2009, 83 (7): 1056~1063.

[41]Meyer R, Beyer H G, Fanslau Jet al., Towards standardization of csp yield prognosis, Proceedings of the SolarPACES 2009 conference, Berlin, Germany, 2009.

[42]Yu X, Wu Z, Jiang Wet al., Predicting daily photosynthetically active radiation from global solar radiation in the contiguous united states, Energy Conversion and Management, 2015, 89: 71~82.

[43]Muselli M, Poggi P, Notton Get al., Classification of typical meteorological days from global irradiation records and comparison between two mediterranean coastal sites in corsica island, Energy Conversion and Management, 2000, 41 (10): 1043~1063.

[44]Maafi A, Harrouni S, Preliminary results of the fractal classification of daily solar irradiances, Solar Energy, 2003, 75 (1): 53~61.

[45]Marty C, Philipona R, Clear-sky index to separate clear-sky from cloudy-sky situations in climate research, Geophysical Research Letters, 2000, 27 (17): 2649~2652.

[46]Long C N, Gaustad K L, The shortwave (sw) clear-sky detection and fitting algorithm: algorithm operational details and explanations: Pacific Northwest National Laboratory, 2004,.

[47]Long C N, Ackerman T P, Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects, Journal of Geophysical Research: Atmospheres, 2000, 105 (D12): 15609~15626.

[48]Lloyd S, Least squares quantization in pcm, IEEE Transactions on Information Theory, 1982, 28 (2): 129~137.

[49]Wagstaff K, Cardie C, Rogers Set al., Constrained k-means clustering with background knowledge, ICML, 2001, 577~584.

[50]Münz G, Li S, Carle G, Traffic anomaly detection using k-means clustering, GI/ITG Workshop MMBnet, 2007.

[51]Jurafsky D, Martin J H, Speech and language processing: Pearson, 2014.