IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 8, 2021, accepted April 30, 2021, date of publication May 4, 2021, date of current version June 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077475

Solar Radiation Intensity Probabilistic Forecasting Based on K-Means Time Series Clustering and Gaussian Process Regression

ZHENDONG ZHANG[1], CHAO WANG[2], XIAOSHENG PENG[3], (Member, IEEE), HUI QIN[1,2], HAO LV[1], JIALONG FU[1], AND HONGYU WANG[3]

[1]School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[2]State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China
[3]School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Hui Qin (hqin@hust.edu.cn)

This work was supported in part by the National Key Research and Development Program of China (Technology and application of wind power/photovoltaic power prediction for promoting renewable energy consumption) under Grant 2018YFB0904200, and in part by the Open Research Fund of State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, under Grant IWHR-SKL-KF201914.

ABSTRACT Solar radiation intensity is intermittent and uncertain under the influence of meteorological conditions. Clustering them and obtaining high-precision and reliable probabilistic forecasting results play a vital role in the planning and management of solar power. In this study, a novel K-means time series clustering (K-MTSC) algorithm is first proposed to cluster solar radiation intensity and compared with astronomy method and K-means. Then, different feature inputs for different categories of solar radiation intensity are screened. Afterwards, the different kernel functions of Gaussian process regression (GPR) are compared and optimal kernel function is selected in terms of deterministic forecasting and probabilistic forecasting for different categories. Finally, the case study in Tibet province, China are performed to verify the validity and practicability of this research model and method. In this experiment, the average accuracy of GPR is 44% higher than that of Artificial Neural Network ANN, and 17% higher than that of Support Vector Regression. The experiments show that (1) the clustering results obtained by the K-MTSC algorithm have a larger inter-group distance and a smaller intra-group distance, and at the same time, it will not destroy the continuity of the time series. (2) The probability forecast results obtained by GPR are reliable and high-accuracy.

INDEX TERMS Solar radiation intensity, K-means time series clustering, probabilistic forecasting, Gaussian process regression.

I. INTRODUCTION

With the increasing depletion of traditional fossil energy and the environmental problems, photovoltaics, as a renewable and clean energy, have received attention from all over the world [1]. However, solar power is affected by natural climate conditions, and its output power is unbalanced in space and unstable in time, showing strong randomness, volatility and intermittent characteristics [2]. The randomness and uncertainty of solar energy resources make it difficult for independent photovoltaic systems to continuously output stable power [3], which not only aggravates the pressure of peak and frequency modulation of the power grid, but also affects the safe and stable operation of the power system, thereby seriously restricting the power grid's ability to absorb solar power [4]. Therefore, obtaining solar radiation intensity deterministic forecasting results with high accuracy and probabilistic forecasting results with high reliability are very important for the application of solar power.

Solar radiation intensity prediction methods can be mainly divided into two categories: physical process driven method and data driven method [5]. Based on meteorological data and satellite images, the physical process driven method builds mathematical and physical equations to

The associate editor coordinating the review of this manuscript and approving it for publication was Grigore Stamatescu.

simulate the change process of solar radiation intensity at a certain regional or global scale [6], such as Numerical Weather Prediction (NWP) [7]. Typical NWP models include European Centre for Medium-Range Weather Forecasts (ECMWF) [8], Fifth-generation Mesoscale Model (MM5) [9] and Weather Research and Forecasting (WRF) [10]. Mathiesen and Kleissl [11] evaluated the numerical forecast of daytime solar radiation intensity in the United States. The study used SURF-RAD ground measurement data to verify the prediction performance of North American Model (NAM), Global Forecast System (GFS) and ECMWF [11]. The physical process driven method has the advantages of high accuracy and strong interpretability while it has the disadvantages of difficult data collection, complex modeling and time-consuming solution [12].

The data driven methods looks for relevant factors from historical data to predict solar radiation intensity, such as time series models, machine learning models and deep learning models [13]. Time series models mainly include Moving Average model (MA), Auto-regressive model (AR), Auto-regressive Moving Average model (ARMA) and their variants [14]. Data stationarity assumption is the precondition of these time series models [15]. Machine learning models such as Support Vector Regression (SVR) and Artificial Neural Network (ANN) has been used to predict solar radiation intensity [16]. Deo *et al.* [17] integrated SVR and discrete wavelet transformation algorithm for short- and long-term global solar radiation forecasting, and the case study in Australia verified the prediction performance of the hybrid model. Amrouche and Pivert [18] combined ANN and spatial modelling techniques for daily global solar radiation forecasting and the model's forecasting results were compared to measured data for the two locations. In recent years, deep learning methods [19] have shown excellent performance in image recognition and natural language processing. Deep learning methods such as Recurrent Neural Network (RNN) [20] and Convolutional Neural Network (CNN) [21] are gradually being used to predict solar radiation intensity. Ghimire *et al.* [22] proposed a hybrid model based on CNN and Long Short-term Memory (LSTM) network for solar radiation forecasting. In their model, CNN is used to extract features while LSTM is used for prediction, and the experiment shows the accuracy of the hybrid deep learning model. Deterministic forecasting model cannot quantify the uncertainty of forecasts.

The randomness and uncertainty of solar radiation intensity make it difficult to be fully forecasted accurately. Therefore, probabilistic forecasting can provide more abundant information for dispatching decision-makers. Estimating the solar radiation intensity prediction interval corresponding to a certain degree of confidence is an idea of quantifying uncertainty [23]. Huang and Wei [24] proposed a daily-ahead probabilistic photovoltaic power forecasting method based on an improved quantile CNN and obtained the upper and lower interval corresponding to 90% confidence levels. Probabilistic forecasting is a more comprehensive method than interval method, such as Bayesian theory and Gaussian process regression (GPR) [25]. Liu *et al.* obtained the spatiotemporal probabilistic forecasting results of solar radiation intensity based on deep learning method and Bayesian inference [26]. Yang *et al.* used GPR to obtain the probability density function (PDF) of solar power output, which can provide a reference for decision-makers to avoid future risks [27]. GPR has the theoretical derivation support and the advantage of strong reliability, which is widely used in probabilistic forecasting.

The solar intensity radiation is affected by meteorological conditions and has different characteristics in different seasons. The astronomy division method [28] uses the height of the sun at noon and the length of day and night as the basis for seasonal changes, but it is difficult to adopt a set of seasonal boundaries to finely divide the four seasons of the world. Therefore, for a specific area, how to use historical data for clustering is the first focus of this study. K-means [29] is a good clustering method, but its classification results will destroy the continuity of the time series. A new K-means time series clustering (K-MTSC) algorithm is proposed for solar radiation intensity clustering. On the other hand, GPR has many kernel functions. Comparing the performance differences of different kernel functions on different categories can provide reference for decision-makers to choose kernel functions, which is conducive to obtaining more reliable and accurate probability prediction results. This is the second focus of this research.

The main contributions are summarized as follows:

(1) A novel time series clustering algorithm called K-MTSC is proposed for solar radiation intensity clustering, whose clustering results have a larger inter-group distance and a smaller intra-group distance, and it doesn't destroy the continuity of the time series.

(2) Selecting the optimal kernel function for GPR are conducive to improving the probabilistic prediction accuracy and reliability.

The remainder of this paper is organized as follows. In Section 2, the implementation details of methods in this study are introduced. In Section 3, performance evaluation metrics are explained. In Section 4, K-MTSC and GPR are applied to solar radiation intensity case study in Tibet, China. In Section 5, the work of this paper is summarized and our conclusions are given.

II. METHODOLOGY

A. K-MEANS TIME SERIES CLUSTERING ALGORITHM

When K-means clustering algorithm is applied to time series problem, the continuity of data will be destroyed. Therefore, this research proposes the K-means time series clustering (K-MTSC) algorithm. The common clustering criteria for K-means and K-MTSC are as follows: (1) the number of category is K; (2) the distance between the centers of different categories should be as large as possible; (3) the average distance from each sample in the same category to the center is as

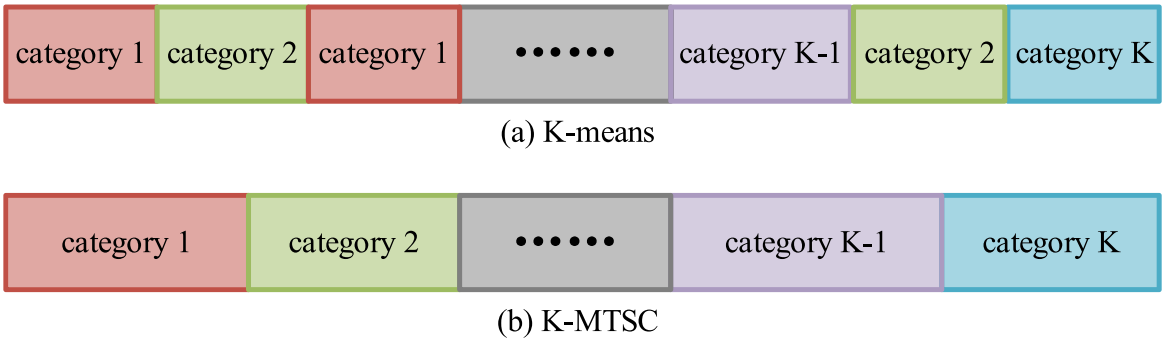

FIGURE 1. Diagram of K-means and K-MTSC.

small as possible. Compared with K-means, K-MTSC needs to satisfy the constraint that the sample points in the same category should be continuous. The diagram of K-means and K-MTSC is shown in FIGURE 1. The meaning of K-MTSC is to divide K categories for a year. Using different feature inputs for different categories is more conducive to improving accuracy than using one input for all datasets. K-MTSC does not filter the features, but only classifies the dataset. The feature selection is performed by correlation coefficients.

K-means time series clustering can be regarded as an optimization problem. The sequence to be clustered is represented by $X = [x_1, x_2, \cdots x_i, \cdots, x_N]$, where x_1 is the i-th sample vector and N is the number of samples. The objectives of K-means time series clustering are as follows:

$$f_{obj,1} = \max\{D_O\} = \max\{\sum_{i=1}^{K}\sum_{j=i+1}^{K} D_{O,i,j}\} \tag{1}$$

$$f_{obj,2} = \min\{D_I\} = \min\{\frac{1}{K}\sum_{i=1}^{K} D_{I,i}\} \tag{2}$$

where D_O and D_I are the distance between the centers of different categories and the average distance from each sample in the same category to the center, respectively. K is the number of categories.

$$D_{O,i,j} = d(\bar{x}_i, \bar{x}_j) \tag{3}$$

$$D_{I,i} = \frac{1}{m}\sum_{x_z \in C_i} d(\bar{x}_i, x_z) \tag{4}$$

where $D_{O,i,j}$ is the distance between the centers of i-th and j-th categories. $\bar{x}_i$ is the center vector of i-th category (C_i). $\bar{x}_i = \frac{1}{m}\sum_{x_z \in C_i} x_z$. x_z and m is a sample and the number of C_i, respectively. $d(\bar{x}_i, \bar{x}_j)$ is the function for calculating Euclidean distance. $D_{I,i}$ is the average distance from each sample (x_z) in i-th category (C_i) to the i-th center $(\bar{x}_i)$.

The constraints of K-means time series clustering are as follows:

$$L(C_i) = K \tag{5}$$

$$x_z \in C_i \text{ and } x_{z+1} \in C_{i+1} \quad \text{if } x_z \text{ is the last element of } C_i \tag{6}$$

where $L(C_i)$ is the number of all categories.

To solve this optimization problem, a K-means time series clustering algorithm based on Genetic Algorithm (GA) is proposed. The steps of K-MTSC algorithm are as follows:

Step 1: randomly generate K different integers on the interval [1, N] and sort them with ascending order;

Step 2: use these integers as the starting indexes for different categories; perform classification operations according to the index to obtain the clustering results and calculate the cluster centers;

Step 3: calculate the distance (D_O) between categories and the distance (D_I) in the same category; fitness = 0.5* D_I − 0.5* D_O;

Step 4: repeat steps 1 to 3 using genetic algorithm until the clustering result corresponding to the optimal fitness is selected. The pseudo code of K-MTSC is shown in FIGURE 2.

B. GAUSSIAN PROCESS REGRESSION

A series of continuous random variables subject to the Gaussian distribution constitute the Gaussian process. In the case of discrete Gaussian process, deriving the Gaussian distribution parameters of unknown samples based on known sample information is Gaussian process regression (GPR). Gaussian process regression [27] assumes that each sample obeys the Gaussian distribution, and any linear combination of samples obeys the joint Gaussian distribution. The schematic diagram of GPR is shown in FIGURE 3.

A common form of regression model is represent as follows:

$$Y = f(X) + \xi \tag{7}$$

where Y, X and ξ are observations, features and noisy ($\xi \sim N(0, \sigma_n^2)$). $N(0, \sigma_n^2)$ represents a Gaussian distribution with mean (μ) and standard deviation (σ_n^2).

According to the definition of GPR, prior distribution of observations Y is as follows:

$$Y \sim N(0, K(X, X) + \sigma_n^2 I_n) \tag{8}$$

Algorithm : K-means time series clustering algorithm
input:
X: the sequence to be clustered K: the number of categories
output:
categories: clustering result centers: centers of categories

```
N = size(X, 1); % the length of sequence X
for i = 1 : p % p is the number of Genetic algorithm population
  index = rand(1, N)*K; % generates K different rand integers on the interval [1, N]
  indexes(i, :) = index;
end
while (the iteration exit condition of genetic algorithm is not satisfied)
  for i = 1 : p
    index = indexes(i, :);
    index = sorted(index); % ascending order, a delimited index of different categories
    [categories, centers] = clustering(X, index); % divide categories by index and calculate category centers
    [D_O, D_I] = distance(categories, centers); % calculate the distance (D_O) between categories and the distance (D_I) in the same category
    fitness(i) = 0.5*D_I-0.5*D_O ; % weighting method; minimize
  end
  selection operation in GA;
  crossover operation in GA;
  mutation operation in GA;
  new population (indexes);
end
the optimal clustering result ([categories, centers]) is obtained according to fitness order;
```

FIGURE 2. Pseudo code of K-MTSC.

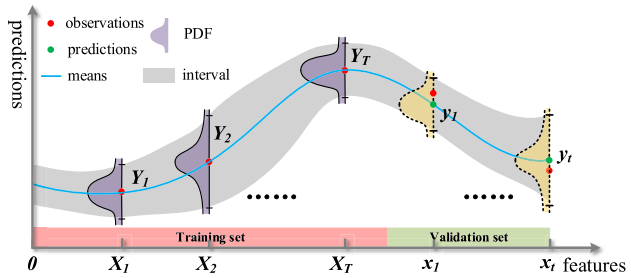

FIGURE 3. Diagram of Gaussian process regression.

And the joint prior distribution of observations Y and predictions y can be obtained:

$$\begin{bmatrix} Y \\ y \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I_n & K(X,x) \\ K(x,X) & K(x,x) \end{bmatrix}\right) = N\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{9}$$

where X, Y, x and y represent the training set features and observations, validation set features and predictions, respectively. $K(X,X) = (\kappa_{ij})$ is a symmetric positive definite covariance matrix, whose element κ_{ij} measures the correlation between X_i and X_j through a kernel function κ. $K(X,x) = K(x,X)^T$ is the covariance matrix between the validation set x and training set X, abbreviated as K_* and K_*^T. $K(x,x)$ is the covariance matrix of the validation set itself, abbreviated as K_{**}. I_n is an n-dimensional unit matrix.

The posterior conditional distribution of the validation set predictions y can be obtained as follows:

$$y|Y \sim N(\bar{y}, \sigma_y^2) \tag{10}$$
$$\bar{y} = K_* K^{-1} Y \tag{11}$$
$$\sigma_y^2 = K_{**} - K_* K^{-1} K_*^T \tag{12}$$

where $\bar{y}$ and σ_y^2 are validation set prediction mean and Gaussian distribution variance, respectively.

Therefore, the deterministic predictions of GPR are y and the interval predictions corresponding to 95% confidence level are $[\bar{y} - 1.96\sigma_y, \bar{y} + 1.96\sigma_y]$. The probability density function (PDF) of i-th predictions is as follows:

$$p(y_i) = \frac{1}{\sqrt{2\pi}\sigma_{y,i}} \exp\left(-\frac{(y_i - \bar{y}_i)^2}{2\sigma_{y,i}^2}\right) \tag{13}$$

C. KERNEL FUNCTION

There are many alternative kernel functions for GPR after a long period of development.

(1) Squared Exponential Kernel

$$k(x_i, x_j) = \sigma_f^2 \exp\left[-\frac{1}{2}\frac{r^2}{\sigma_l^2}\right] \tag{14}$$

where σ_l is the characteristic length scale, and σ_f is the signal standard deviation. $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$ is the Euclidean distance between x_i and x_j.

(2) Exponential Kernel

$$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right) \tag{15}$$

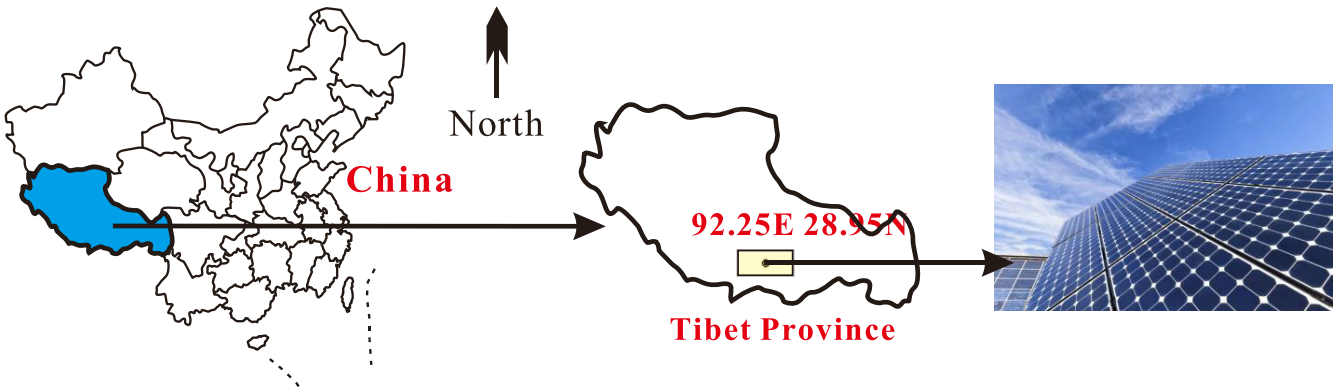

FIGURE 4. Research area.

TABLE 1. Algorithm parameters.

algorithm	symbol	meaning	value
K-MTSC	K	the number of all categories	4
	N	population size in GA	100
	p_c	crossover probability in GA	0.95
	η_c	crossover distribution index in GA	20
	p_m	mutation probability in GA	0.25
	η_m	mutation distribution index in GA	20
	M	iteration numbers in GA	500
K-means	K	the number of all categories	4
	M	iteration numbers	1000
ANN	n_i	number of input layer nodes	feature numbers
	n_h	number of hidden layer nodes	optimized by random search
	n_o	number of output layer nodes	1
	f	activation function	ReLU
SVR	θ	kernel parameters	optimized by function (fitrsvm) in Matlab
GPR	θ	kernel parameters	optimized by function (fitrgp) in Matlab

(3) Matern 3/2 Kernel

$$k(x_i, x_j) = \sigma_f^2 \exp(1 + \frac{\sqrt{3}r}{\sigma_l}) \exp(-\frac{\sqrt{3}r}{\sigma_l}) \quad (16)$$

(4) Matern 5/2 Kernel

$$k(x_i, x_j) = \sigma_f^2 \exp(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}) \exp(-\frac{\sqrt{5}r}{\sigma_l}) \quad (17)$$

(5) Rational Quadratic Kernel

$$k(x_i, x_j) = \sigma_f^2 (1 + \frac{r^2}{2\alpha\sigma_l^2})^{-\alpha} \quad (18)$$

where α is a kernel parameter, and other variables have the same meaning as before.

Exploring the differences of different categories of solar radiation intensity datasets on different kernel functions and selecting the optimal kernel function is one of the key points of this research, which is of vital importance to improve the prediction accuracy.

III. EVALUATION METRICS

A. DETERMINISTIC FORECASTING EVALUATION METRICS

Deterministic forecasting results are evaluated by mean absolute error (MAE) and root mean square error (RMSE) [30] in this study, as follows:

$$MAE = \frac{1}{Te} \sum_{i=1}^{Te} |y_i - Y_i| \quad (19)$$

$$RMSE = \sqrt{\frac{1}{Te} \sum_{i=1}^{Te} (y_i - Y_i)^2} \quad (20)$$

where y_i and Y_i are i-th prediction and observation, respectively. Te is the number of validation set. The smaller the MAE or the RMSE, the higher the prediction accuracy.

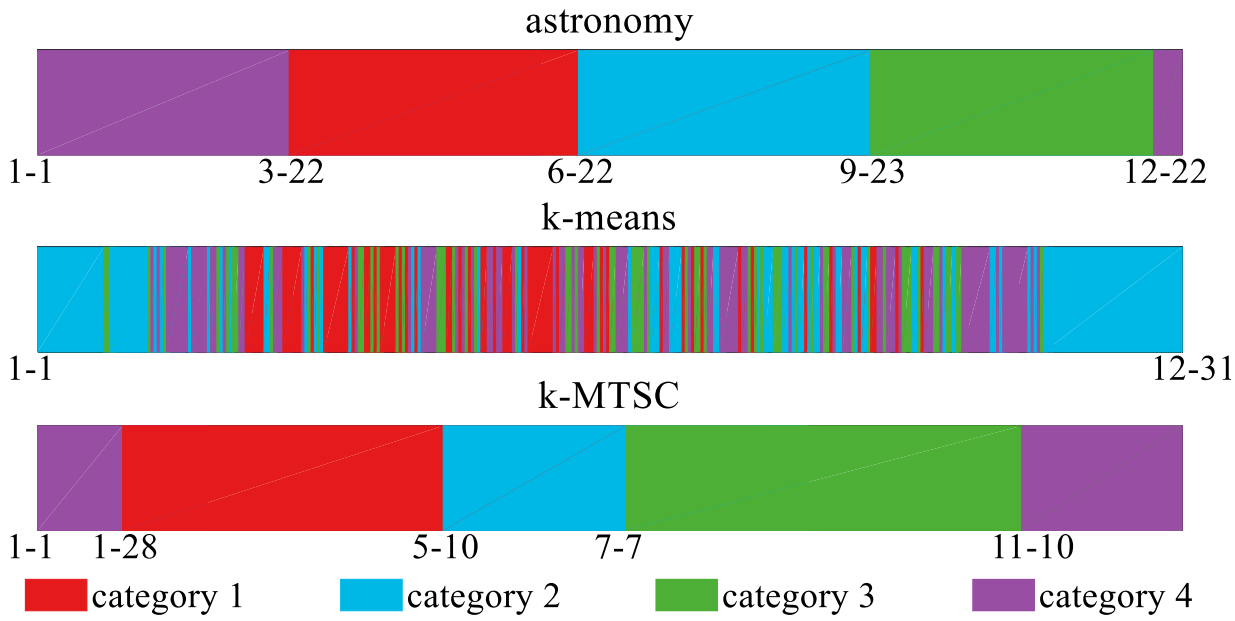

FIGURE 5. Clustering results.

TABLE 2. Average distance within the same category.

clustering algorithm	astronomy	K-means	K-MTSC
category 1	589.03	344.43	519.98
category 2	559.24	278.85	564.86
category 3	308.58	511.59	505.08
category 4	337.34	364.90	154.16
mean	448.55	374.94	436.02

TABLE 3. Average distance between categories.

clustering algorithm	astronomy	K-means	K-MTSC
category 1 and 2	321.49	787.93	247.39
category 1 and 3	405.89	971.35	205.47
category 1 and 4	438.84	410.16	353.25
category 2 and 3	158.23	473.99	366.56
category 2 and 4	207.31	399.23	566.24
category 3 and 4	186.05	609.89	225.52
mean	286.30	608.76	327.40

B. PROBABILISTIC FORECASTING EVALUATION METRIC

Probabilistic forecasting results are evaluated by continuous ranked probability score (CRPS) [31], as follows:

$$CRPS = \frac{1}{Te}\sum_{i=1}^{Te}\int_{-\infty}^{+\infty}[F(y_i) - H(y_i - Y_i)]^2 dy_i \quad (21)$$

$$F(y_i) = \int_{-\infty}^{y_i} p(x)dx \quad (22)$$

$$H(y_i - Y_i) = \begin{cases} 0 & y_i < Y_i \\ 1 & others \end{cases} \quad (23)$$

where $p(y_i)$ and $F(y_i)$ are probability density function and cumulative distribution function of i-th probabilistic prediction. The smaller the CRPS, the better the comprehensive performance.

C. INTERVAL FORECASTING EVALUATION METRICS

Interval forecasting results are evaluated by reliability evaluation (RE) criterion and sharpness evaluation (SE) criterion [23], as follows:

$$RE = (\frac{\xi^{(1-\alpha)}}{Te} - (1-\alpha)) \times 100\% \quad (24)$$

$$SE = \frac{1}{Te}\sum_{i=1}^{Te}(Ub_{1-\alpha} - Lb_{1-\alpha}) \quad (25)$$

where $\xi^{(1-\alpha)}$ is the number of times that observation points do indeed lie within the α level prediction intervals. $Ub_{1-\alpha}$ and $Lb_{1-\alpha}$ are the lower and upper bound of the α level prediction interval.

IV. CASE STUDY

A. CASE INTRODUCTION

Solar radiation intensity data used in this study is collected from the station (92.25°E, 28.95°N) in Tibet province, China, as shown in FIGURE 4. The total time span of the data is from January 1, 2010 to December 31, 2010. The step length of a period is one hour.

In this case, there are five tasks to be completed:

(1) Perform time series clustering and compare different clustering algorithm;

(2) Construct datasets and filter feature inputs for different categories;

(3) Deterministic forecasting results comparison;

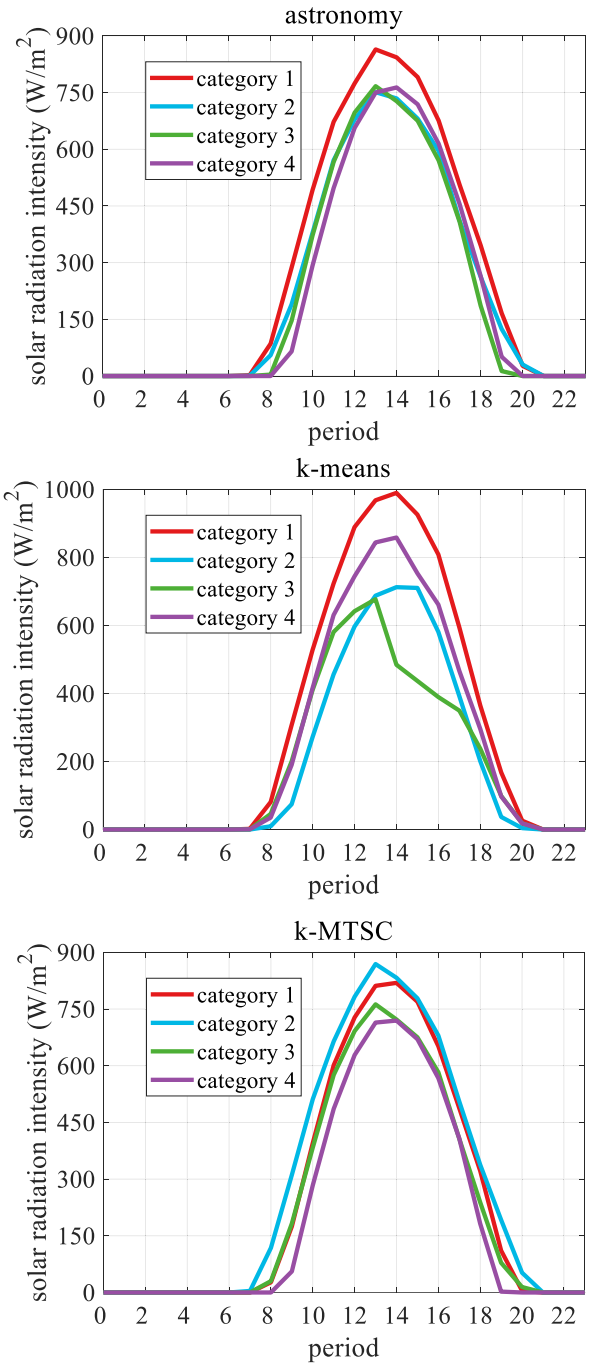

FIGURE 6. Centers of four categories.

(4) Probabilistic and interval forecasting results comparison;

(5) Display the probabilistic forecasting results.

B. EXPERIMENT RESULTS AND DISCUSSION

1) CLUSTERING COMPARISON

K-MTSC algorithm is proposed to perform time series clustering and compared with astronomy method and K-means. The parameters in K-MTSC are shown in TABLE 1. The time boundaries of the four seasons are defined as March 22, June 22, September 23 and December 22 in astronomy, which can be regarded as a time series clustering method (K = 4). Clustering results of three algorithms are shown in FIGURE 5. The four time boundaries of K-MTSC are January 28, May 10, July 7 and November 10, respectively. The K-means clustering result destroys the continuity of time series variables.

In order to further compare the differences between different clustering algorithms, centers of four categories are shown in FIGURE 6. The average distance within the same category and the average distance between categories are listed in TABLE 2 and TABLE 3, respectively. The results are analyzed as follows:

(1) Centers of category 2 to 4 of astronomy method are too close to distinguish these categories. Astronomy method are only a rough classification of the changes in the four seasons around the world, and cannot achieve very accurate classification of all specific areas.

(2) The four centers of K-means are obviously different, however it destroys the continuity of time and is not suitable for time series variables.

(3) K-MTSC combines the advantages of the two methods. On the one hand, the classification has continuity, and on the other hand, the cluster centers are obviously different.

(4) The average inner distances of astronomy and K-MTSC are 448.55 and 436.02, which shows that the latter samples are more concentrated in the same category. The average outer distances of astronomy and K-MTSC are 286.30 and 327.40, which shows that the latter has a better ability to distinguish different categories.

(5) The difference between astronomy and K-MTSC is that the former is a fixed classification for each year, but the latter can be based on dataset dynamic classification, the result is more refined.

2) CONSTRUCT DATASETS AND FEATURE SELECTION

According to the divided categories, four datasets are constructed, whose total length is 15 days, 15 days, 20 days and 20 days, respectively. Detailed statistical information is shown in TABLE 4, where T, Ta and Te represent the size of total samples, training set samples and validation set samples.

In order to improve the forecast accuracy, the correlation coefficient is used to filter the feature input of the four datasets. Historical features are used as candidate features [32]. The input features are historical solar radiation intensity. F23 represents the solar radiation intensity 23 periods ago. F24 represents the solar radiation intensity 24 periods ago. In this study, F1-F96 were selected as candidate features, and feature factors with correlation coefficients greater than 0.85 were left as feature inputs. Absolute value of feature correlation coefficient radar charts are plotted in FIGURE 7. Correlation coefficient varies from 0 to 1 from the center point to the external. The purple line represents the correlation coefficient of each feature, the orange line represents the 0.85 standard line, and the green dot represents the features that are left as input. Correlation coefficients of

TABLE 4. Statistical information of four datasets.

Datasets	category	time	T	Ta	Te	min	mean	max
unit		1 period = 1 h		period		W/m^2	W/m^2	W/m^2
Dataset 1	1	Feb/4/2010-- Feb/19/2010	360	312	48	0	203	889
Dataset 2	2	May/29/2010--Jun/13/2010	360	312	48	0	318	1055
Dataset 3	3	Oct/17/2010--Nov/6/2010	480	408	72	0	235	881
Dataset 4	4	Nov/25/2010--Dec/15/2010	480	408	72	0	189	744

TABLE 5. Correlation coefficients of feature inputs.

features	F1	F23	F24	F25	F47	F48	F49	F71	F72	F73	F95	F96
Dataset 1	0.907	0.851	0.903	--	--	0.895	--	--	0.894	--	--	0.879
Dataset 2	0.886	--	0.894	0.851	--	0.876	--	--	0.881	--	--	0.886
Dataset 3	0.884	--	0.889	--	--	0.875	--	--	0.875	--	--	0.879
Dataset 4	0.925	0.919	0.987	0.919	0.916	0.983	0.915	0.916	0.985	0.916	0.914	0.981

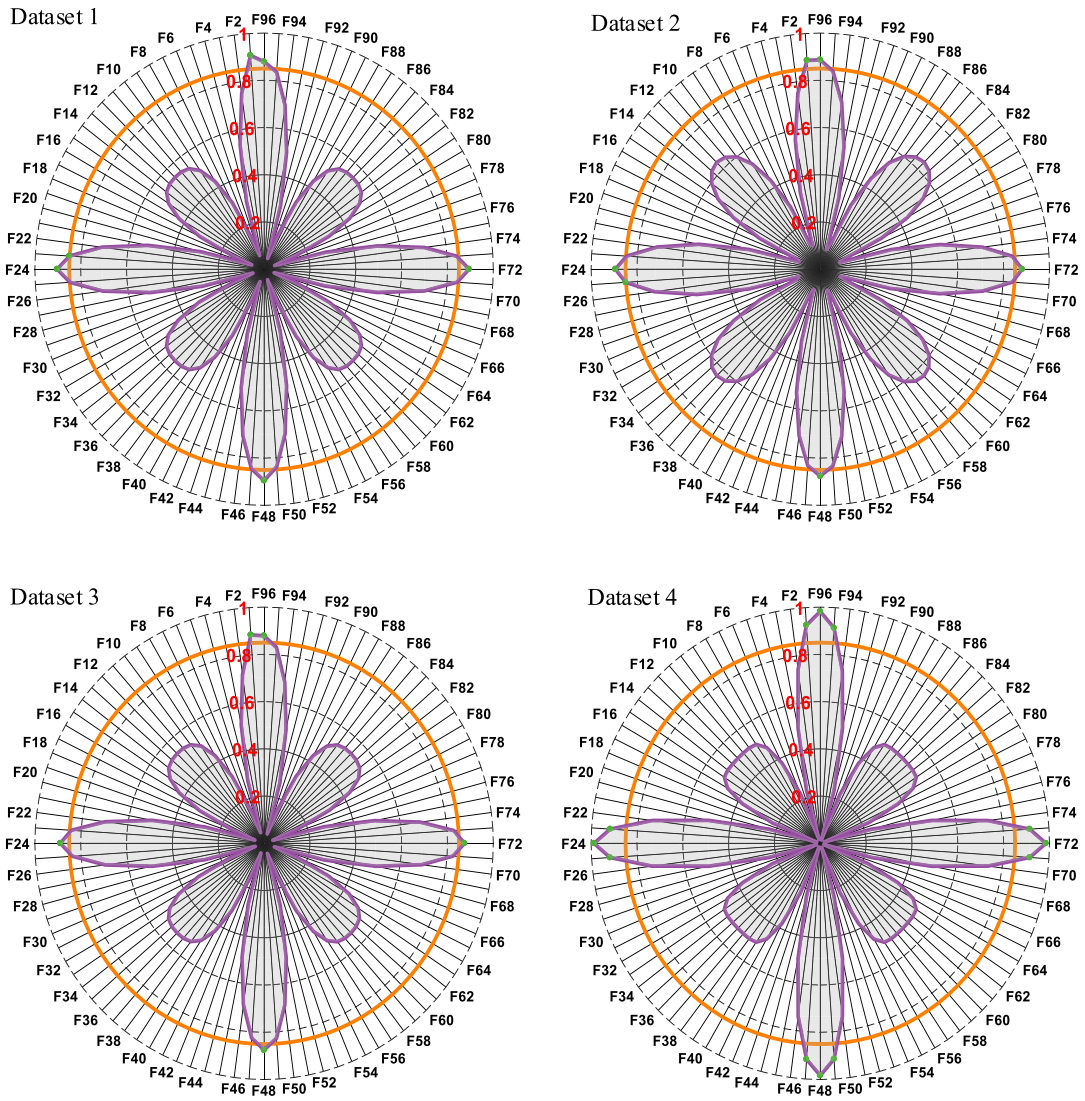

FIGURE 7. Feature correlation coefficient radar chart.

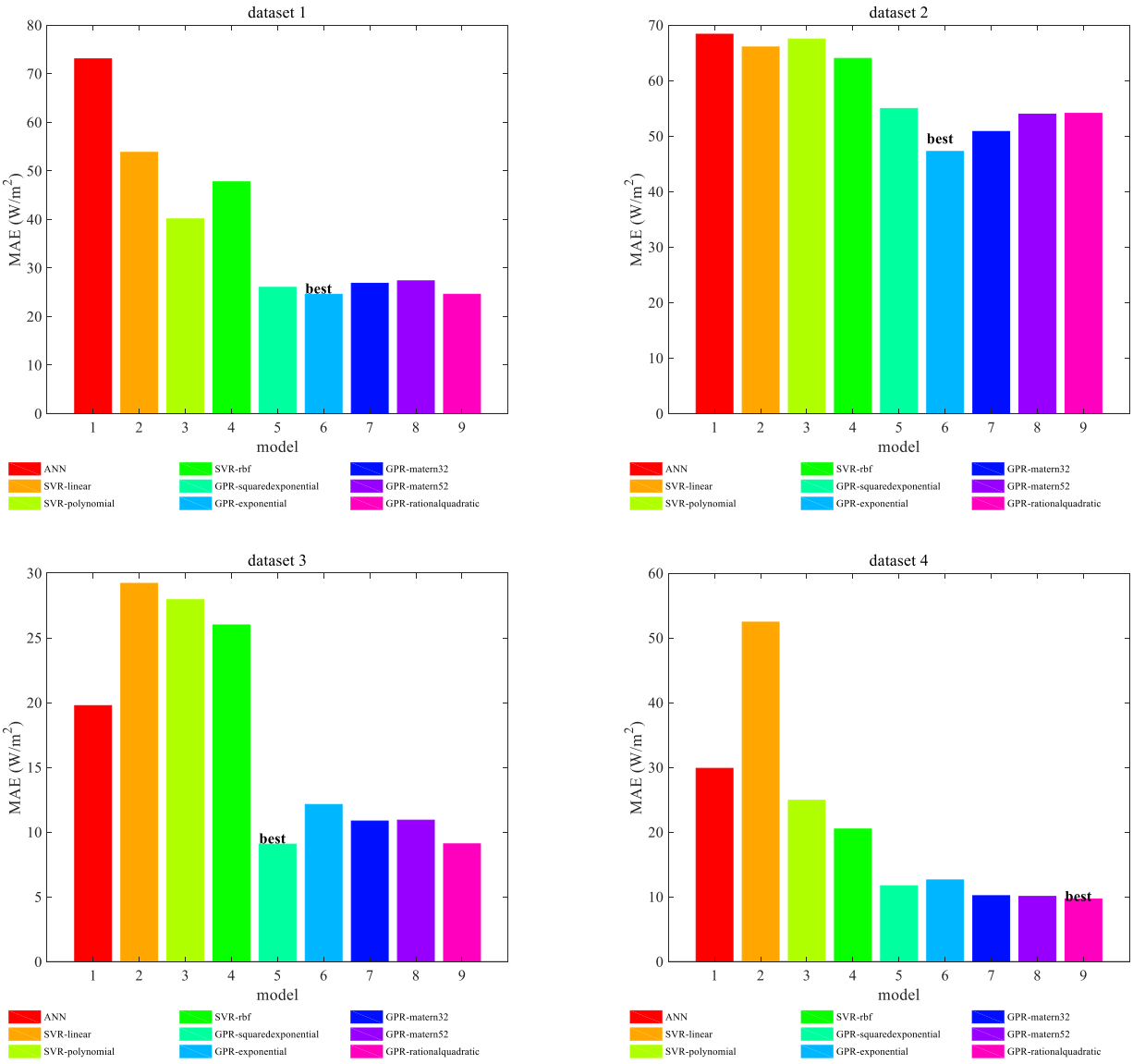

FIGURE 8. Deterministic forecasting metric (MAE) comparison.

feature inputs are listed in TABLE 5. Therefore, feature inputs of four datasets are [F1, F23, F24, F48, F72, F96], [F1, F24, F25, F48, F72, F96], [F1, F24, F48, F72, F96], [F1, F23, F24, F25, F47, F48, F49, F71, F72, F73, F95, F96], respectively.

3) DETERMINISTIC FORECASTING RESULTS COMPARISON

In order to verify the predictive performance of GPR, it was compared with ANN and SVR. The kernel functions in SVR are linear, polynomial and rbf kernels. The kernel functions in GPR are squared exponential, exponential, matern32, matern52 and rational quadratic kernels. The hyper-parameters of these models are shown in TABLE 1. The number of input layer nodes is equal to the number of feature inputs. The number of hidden layer nodes is optimized by random search. They are 16, 32, 64 and 32 on four datasets after optimization. The number of output layer nodes is 1. The activation function used in ANN is ReLU.

Deterministic forecasting metrics are listed in TABLE 6 and shown in FIGURE 8. The results can be analyzed as follows:

(1) In comparison of different models, the overall prediction accuracy of GPR is higher than that of SVR and ANN.

(2) In datasets 1 and 2, MAEs of GPR-exponential are 24.57W/m^2 and 47.24W/m^2, which is the smallest metric among all models, indicating that the exponential kernel function of GPR is the best on these two datasets. Similarly, the squared exponential and rational quadratic functions of GPR are the best performing kernel functions on datasets 3 and 4, respectively.

(3) The metrics MAE and RMSE are related to the mean of the dataset. The mean of dataset 2 is larger than other datasets, so the MAE and RMSE are larger than other datasets, which is normal.

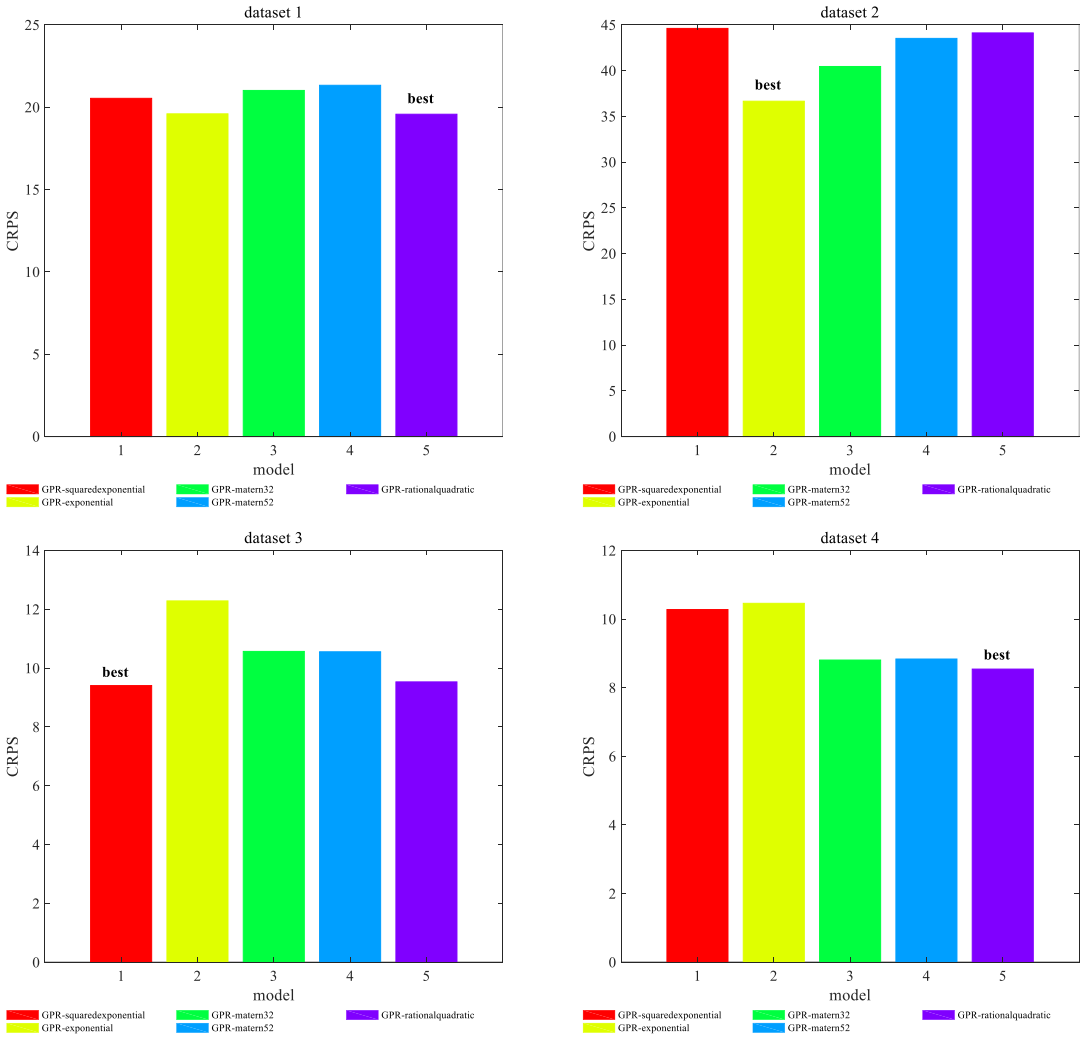

FIGURE 9. Probabilistic forecasting metric (CRPS) comparison.

TABLE 6. Deterministic forecasting metrics.

dataset		dataset 1		dataset 2		dataset 3		dataset 4	
metrics		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ANN		73.05	92.63	68.36	123.84	19.77	33.84	29.85	106.58
SVR	linear	53.81	88.77	66.06	103.07	29.21	37.14	52.44	59.26
	polynomial	40.10	62.04	67.50	102.88	27.95	31.29	24.91	33.46
	rbf	47.73	71.80	64.00	106.70	26.00	32.65	20.48	30.99
GPR	squared exponential	26.01	55.02	54.97	106.88	9.07	23.42	11.72	30.59
	exponential	24.57	49.78	47.24	93.31	12.14	29.08	12.62	30.46
	matern32	26.84	55.20	50.83	101.30	10.86	27.33	10.19	27.05
	matern52	27.34	56.09	53.97	107.55	10.93	28.09	10.08	27.53
	rational quadratic	24.58	52.63	54.10	109.09	9.10	24.69	9.66	27.00

(4) On the evaluation metric RMSE, compared with ANN model, GPR improved the accuracy of the four datasets by 46%, 25%, 31% and 75%, respectively. Compared with SVR model, GPR improved the accuracy of the four datasets by 20%, 9%, 25% and 13%, respectively.

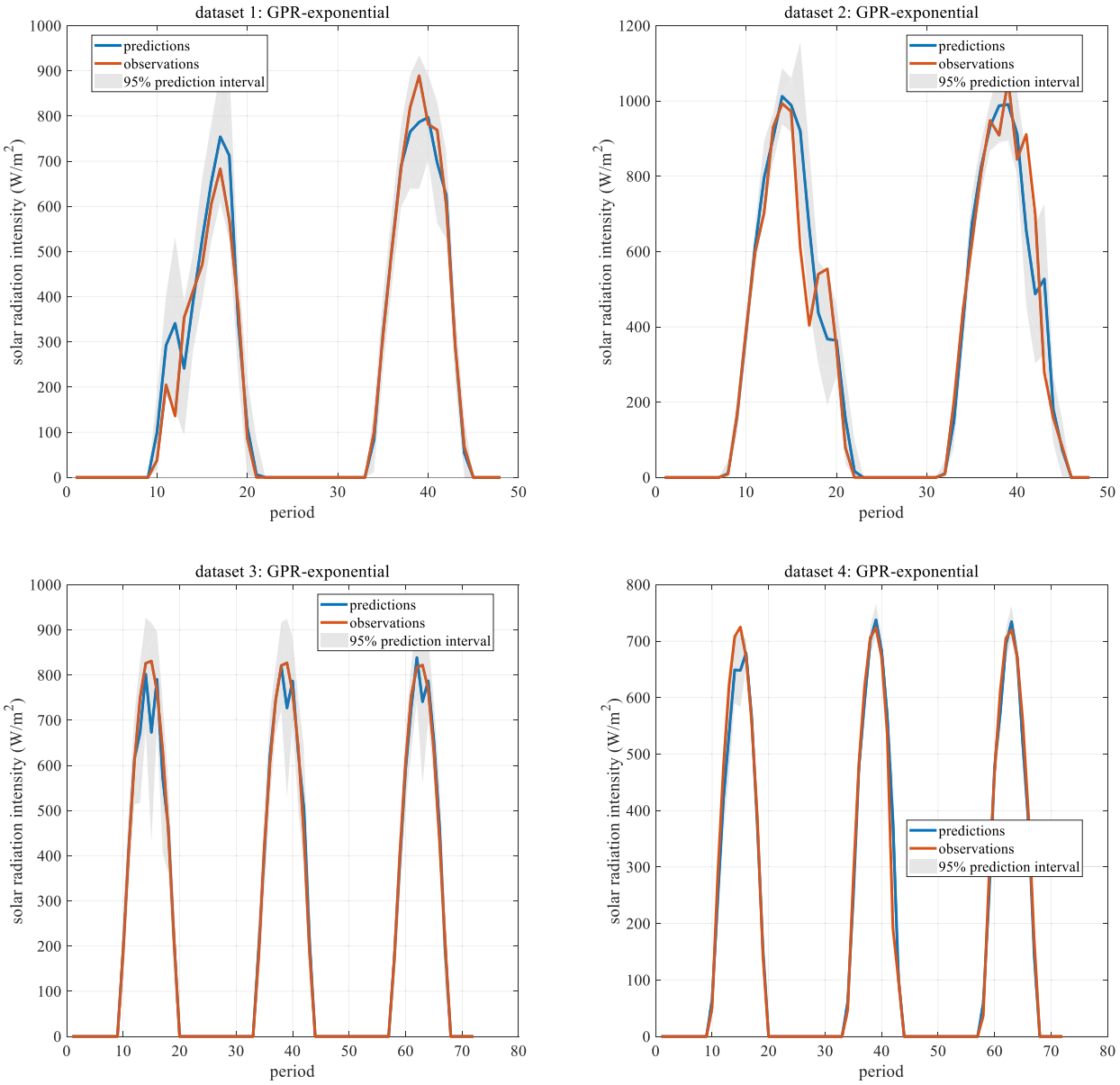

FIGURE 10. Interval prediction results of four validation sets.

4) PROBABILISTIC FORECASTING RESULTS COMPARISON

The probabilistic forecasting results of GPR with different kernel functions are compared. Probabilistic forecasting metrics are listed in TABLE 7 and shown in FIGURE 9. In datasets 1 and 4, the CRPS of rational quadratic kernel function are 19.58 and 8.54, which are the smallest metrics, indicating that it is the best on these two datasets. The kernel functions with best probabilistic prediction performance on datasets 2 and 3 are exponential and squared exponential functions, respectively.

The interval forecasting results are shown in TABLE 8. In dataset 1, the RE and SE of GPR with rational quadratic kernel function are 0.83% and 92.06W/m^2, respectively. It has a large coverage and a narrow prediction range, and is the most appropriate kernel function for interval prediction results in the experiment. The kernel functions with best

TABLE 7. Probabilistic forecasting metrics.

CRPS		dataset 1	dataset 2	dataset 3	dataset 4
GPR	squared exponential	20.54	44.62	9.41	10.28
	exponential	19.60	36.66	12.29	10.46
	matern32	21.02	40.45	10.57	8.81
	matern52	21.33	43.53	10.56	8.84
	rational quadratic	19.58	44.12	9.53	8.54

TABLE 8. Interval forecasting metrics.

dataset		dataset 1		dataset 2		dataset 3		dataset 4	
	metrics	RE (%)	SE(W/m^2)	RE (%)	SE(W/m^2)	RE (%)	SE(W/m^2)	RE (%)	SE(W/m^2)
GPR	squared exponential	-1.25	90.37	-24.17	105.78	5.00	79.05	-8.89	30.71
	exponential	2.92	128.54	-7.50	137.22	5.00	120.02	-1.94	41.11
	matern32	0.83	104.82	-11.67	117.54	5.00	88.87	-4.72	31.69
	matern52	0.83	99.38	-15.83	111.14	5.00	86.18	-7.50	30.67
	rational quadratic	0.83	92.06	-17.92	107.16	5.00	80.21	-6.11	29.59

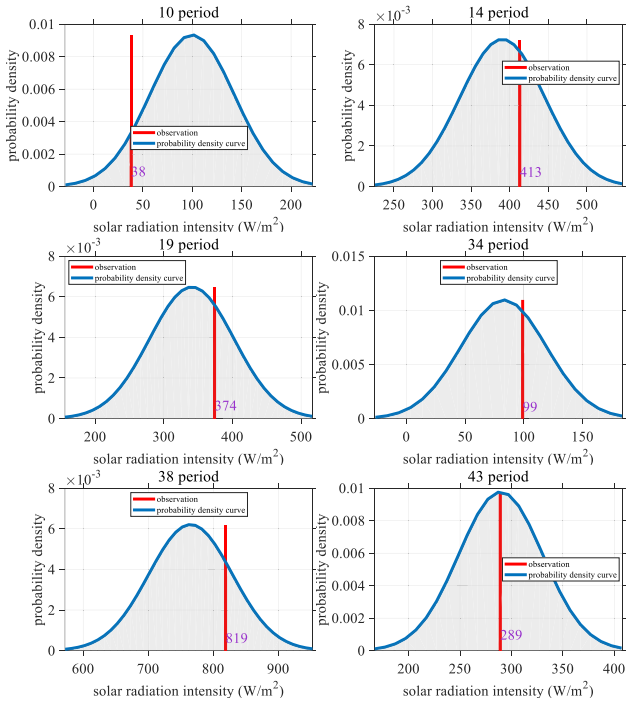

FIGURE 11. PDF of GPR- exponential on dataset 1.

interval prediction performance on datasets 2 to 4 are exponential, squared exponential functions and rational quadratic, respectively. The comparison results of different kernel functions in probabilistic prediction and interval prediction are similar.

5) DISPLAY THE PROBABILISTIC FORECASTING RESULTS

Taking GPR-exponential as an example, interval prediction results of four validation sets are shown in FIGURE 10. The prediction line is very close to the observation line, indicating that the prediction accuracy is high. Most observation points are in the 95% interval, indicating that the prediction is reliable. PDFs of GPR- exponential on dataset 1 are shown in FIGURE 11. These PDFs are full, and no curve is excessively high or low, wide or narrow, which indicates the obtained PDF is suitable. The observation lines in some periods (14, 34 and 43) are near the center of curve while others (10, 19, and 38) are a little far from the center, which just indicates that the probabilistic forecasting is reliable. The prediction accuracy of the period of sudden change will be lower, and it may be out of the confidence interval. In the probabilistic forecasting, there must be some observations that are not in the interval, which just shows that the forecast is reliable. If all the observations are within the interval, it means that the interval loses its significance in probability. The confidence interval of the period when the forecast accuracy is high is narrower, and vice versa, which is in line with the general law of probabilistic forecasting.

V. CONCLUSION

Obtaining reliable high-quality solar radiation intensity prediction results is very important for the planning and application of solar energy. Solar radiation intensity has different characteristics in different seasons. Clustering solar radiation intensity in a year is helpful to grasp the forecast characteristics in a targeted manner. Firstly, a novel time series clustering algorithm (K-MTSC) is proposed to cluster solar radiation intensity and compared with astronomy method and K-means. Next, Feature selection is performed for different categories. Finally, GPR is applied for solar radiation intensity probabilistic forecasting and compared with different models and different kernel functions. The experiment results of a case study in Tibet province, China show that:

(1) The clustering results obtained by the K-MTSC algorithm have a larger inter-group distance and a smaller intra-group distance, and at the same time, it will not destroy the continuity of the time series.

(2) Selecting different features for different categories of solar radiation intensity can obtain more accurate prediction results. The probability forecast results obtained by GPR are reliable.

(3) In this experiment, the average accuracy of GPR is 44% higher than that of Artificial Neural Network ANN, and 17% higher than that of Support Vector Regression.

The main contribution of this research is to propose the K-MTSC algorithm, which can complete the clustering of time series variables. At the same time, the GPR model can obtain reliable and high-precision solar radiation intensity forecasting results. It is one of the future research work to complete the multi-step ahead probabilistic forecasting of solar radiation intensity.

REFERENCES

[1] C. Feng, M. Cui, B.-M. Hodge, S. Lu, H. F. Hamann, and J. Zhang, "Unsupervised clustering-based short-term solar forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 4, pp. 2174–2185, Oct. 2019.

[2] J. Huang and M. Perry, "A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting," *Int. J. Forecasting*, vol. 32, no. 3, pp. 1081–1086, Jul. 2016.

[3] Z. Zhang, H. Qin, J. Li, Y. Liu, L. Yao, Y. Wang, C. Wang, S. Pei, and J. Zhou, "Short-term optimal operation of wind-solar-hydro hybrid system considering uncertainties," *Energy Convers. Manage.*, vol. 205, Feb. 2020, Art. no. 112405.

[4] Z. Zhang, H. Qin, L. Yao, Y. Liu, Z. Jiang, Z. Feng, and S. Ouyang, "Improved multi-objective moth-flame optimization algorithm based on R-domination for cascade reservoirs operation," *J. Hydrol.*, vol. 581, Feb. 2020, Art. no. 124431.

[5] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. M. Shah, "Review on forecasting of photovoltaic power generation based on machine learning and Metaheuristic techniques," *IET Renew. Power Gener.*, vol. 13, no. 7, pp. 1009–1023, May 2019.

[6] H. S. Jang, K. Y. Bae, H.-S. Park, and D. K. Sung, "Solar power prediction based on satellite images and support vector machine," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1255–1263, Jul. 2016.

[7] D. Yang, "On post-processing day-ahead NWP forecasts using Kalman filtering," *Sol. Energy*, vol. 182, pp. 179–181, Apr. 2019.

[8] P. Pinson and R. Hagedorn, "Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations," *Meteorol. Appl.*, vol. 19, no. 4, pp. 484–500, 2012.

[9] Y. Lei and W. Dongxiao, "Comparison of MM5 and satellite derived precipitation and wind speed during typhoon chanchu (2006) in the south China sea," in *Proc. 2nd IITA Int. Conf. Geosci. Remote Sens.*, Aug. 2010, pp. 344–347.

[10] T. Katopodis, I. Markantonis, N. Politi, D. Vlachogiannis, and A. Sfetsos, "High-resolution solar climate atlas for greece under climate change using the weather research and forecasting (WRF) model," *Atmosphere*, vol. 11, no. 7, p. 761, Jul. 2020.

[11] P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states," *Sol. Energy*, vol. 85, no. 5, pp. 967–977, May 2011.

[12] C. Feng, M. Cui, B.-M. Hodge, and J. Zhang, "A data-driven multi-model methodology with deep feature selection for short-term wind forecasting," *Appl. Energy*, vol. 190, pp. 1245–1257, Mar. 2017.

[13] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE J. Power Energy Syst.*, vol. 1, no. 4, pp. 38–46, Dec. 2015.

[14] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Sol. Energy*, vol. 83, no. 10, pp. 1772–1783, Oct. 2009.

[15] Z. Zhang, H. Qin, Y. Liu, Y. Wang, L. Yao, Q. Li, J. Li, and S. Pei, "Long short-term memory network based on neighborhood gates for processing complex causality in wind speed prediction," *Energy Convers. Manage.*, vol. 192, pp. 37–51, Jul. 2019.

[16] E. Eğrioğlu and R. Fildes, "A new bootstrapped hybrid artificial neural network approach for time series forecasting," *Comput. Econ.*, Nov. 2020. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10614-020-10073-7.pdf

[17] R. C. Deo, X. Wen, and F. Qi, "A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset," *Appl. Energy*, vol. 168, pp. 568–593, Apr. 2016.

[18] B. Amrouche and X. L. Pivert, "Artificial neural network based daily local forecasting for global solar radiation," *Appl. Energy*, vol. 130, pp. 333–341, Oct. 2014.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] Z. Zhang, H. Qin, J. Li, Y. Liu, L. Yao, Y. Wang, C. Wang, S. Pei, P. Li, and J. Zhou, "Operation rule extraction based on deep learning model with attention mechanism for wind-solar-hydro hybrid system under multiple uncertainties," *Renew Energ*, vol. 170, pp. 92–106, Jun. 2021.

[21] Z. Si, Y. Yu, M. Yang, and P. Li, "Hybrid solar forecasting method using satellite visible images and modified convolutional neural networks," *IEEE Trans. Ind. Appl.*, vol. 57, no. 1, pp. 5–16, Jan. 2021.

[22] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, "Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms," *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113541.

[23] U. Yolcu, E. Egrioglu, E. Bas, O. C. Yolcu, and A. Z. Dalar, "Probabilistic forecasting, linearity and nonlinearity hypothesis tests with bootstrapped linear and nonlinear artificial neural network," *J. Experim. Theor. Artif. Intell.*, vol. 33, no. 3, pp. 1–22, Apr. 2019.

[24] Q. Huang and S. Wei, "Improved quantile convolutional neural network with two-stage training for daily-ahead probabilistic forecasting of photovoltaic power," *Energy Convers. Manage.*, vol. 220, Sep. 2020, Art. no. 113085.

[25] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted Gaussian process regression," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 300–308, Jan. 2018.

[26] Y. Liu, H. Qin, Z. Zhang, S. Pei, C. Wang, X. Yu, Z. Jiang, and J. Zhou, "Ensemble spatiotemporal forecasting of solar irradiation using variational Bayesian convolutional gate recurrent unit network," *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113596.

[27] Y. Yang, S. Li, W. Li, and M. Qu, "Power load probability density forecasting using Gaussian process quantile regression," *Appl. Energy*, vol. 213, pp. 499–509, Mar. 2018.

[28] Q. Duan, Y. Feng, and J. Wang, "Clustering of visible and infrared solar irradiance for solar architecture design and analysis," *Renew. Energy*, vol. 165, pp. 668–677, Mar. 2021.

[29] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for K-means clustering," *IEEE Trans. Inf. Theory*, vol. 61, no. 2, pp. 1045–1062, Feb. 2015.

[30] Z. Zhang, H. Qin, Y. Liu, L. Yao, X. Yu, J. Lu, Z. Jiang, and Z. Feng, "Wind speed forecasting based on quantile regression minimal gated memory network and kernel density estimation," *Energy Convers. Manage.*, vol. 196, pp. 1395–1409, Sep. 2019.

[31] Z. Zhang, L. Ye, H. Qin, Y. Liu, C. Wang, X. Yu, X. Yin, and J. Li, "Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression," *Appl. Energy*, vol. 247, pp. 270–284, Aug. 2019.

[32] U. Yolcu, Y. Jin, and E. Egrioglu, "An ensemble of single multiplicative neuron models for probabilistic prediction," in *Proc. IEEE Symp. Comput. Intell.*, Dec. 2016, pp. 1–8.

ZHENDONG ZHANG was born in Yichang, China, in November 1994. He received the B.S. degree from China Three Gorges University, Yichang, in 2016. He is currently pursuing the Ph.D. degree in hydraulic and hydropower engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China.

His research interests include probabilistic forecasting, machine learning and deep learning, reservoir operation, multi-objective evolutionary algorithm, and wind, solar, and hydropower complementary system operation.

CHAO WANG was born in Enshi, China, in November 1989. He received the B.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2011 and 2016, respectively.

He is currently a Senior Engineer with the China Institute of Water Resources and Hydropower Research, Beijing, China. His research interests include forecasting, the modeling and operation theory in water resources management, and information management of water resources projects.

XIAOSHENG PENG (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Huazhong University of Science and Technology, China, in 2006 and 2009, respectively, and the Ph.D. degree in electrical engineering from Glasgow Caledonian University, in 2012. He has worked as a Postdoctoral Research Fellow with Glasgow Caledonian University, funded by EDF Energy. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Huazhong University of Science and Technology. His research interests include big data, artificial intelligence and its application in power systems, new and renewable energy forecasting, and condition monitoring of power plants. He is a member of IET and IEC SC8A WG2. His Ph.D. degree was funded by EPSRC.

HUI QIN was born in Yicheng, China, in September 1983. He received the B.S. and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2006 and 2011, respectively.

He is currently a Professor with the School of Civil and Hydraulic Engineering, HUST. His research interests include forecasting, reservoir (group) optimal dispatch, flood resource utilization, water resources optimal allocation, and power system optimal dispatch.

HAO LV was born in Chaoyang, China, in August 1996. He received the B.S. degree from Harbin Engineering University, Harbin, China, in 2019. He is currently pursuing the master's degree in hydraulic and hydropower engineering with the Huazhong University of Science and Technology (HUST), Wuhan, China.

His research interests include forecasting, hydropower energy system optimization operation, multi-objective optimization algorithm, and water resources planning and management.

JIALONG FU was born in Nanchang, China, in November 1997. He received the B.S. degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020, where he is currently pursuing the master's degree in hydraulic and hydropower engineering.

His research interests include forecasting, hydropower energy system optimization operation, multi-objective optimization algorithm, and water resources planning and management.

HONGYU WANG received the B.S. degree from the School of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2017, where he is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering. His research interests include condition monitoring, pattern recognition, and artificial neural networks.