Improved Gaussian Process Regression Solar Output Forecast with Pre-clustering Techniques

Fatemeh Najibi
Artificial Intelligence
Research Centre (CitAI)
City, University of London
London, EC1V 0HB, UK
fatemeh.najibi@city.ac.uk

Dimitra Apostolopoulou
Department of Electrical
and Electronic Engineering
City, University of London
London, EC1V 0HB, UK
dimitra.apostolopoulou@city.ac.uk

Eduardo Alonso
Artificial Intelligence
Research Centre (CitAI)
City, University of London
London, EC1V 0HB, UK
e.alonso@city.ac.uk

***Abstract*—Power system operations are becoming more challenging with the increasing penetration of renewable-based resources such as photovoltaic (PV) generation. In this regard, obtaining accurate solar power output forecasts allows a deepening penetration of renewable-based resources in a secure and reliable way. In this paper, we propose a probabilistic framework to predict short-term PV output taking into account the uncertainty of weather data as well as the variability of PV output over time. To this end, we use datasets comprising of meteorological weather data such as temperature, irradiance, zenith, and azimuth and solar power output. We cluster these data in categories and train a Matérn 5/2 Gaussian Process Regression model for each cluster. More specifically, we cluster the data into one to eight different partitions by making use of the k-means algorithm. In order to identify the optimal number of clusters we use the Elbow and Gap methods. We compare the results obtained for the different number of clusters with the (i) 5-fold cross-validation; and (ii) holding out 30 representative days as test data. The results showed that the optimal number of clusters is four, since in comparison to higher number of clusters the increase in the forecast error was marginal.**

I. Introduction

Nowadays power systems are facing significant challenges from the increasing integration of renewable-based energy resources. The use of renewable-based generation to meet the load contributes to a sustainable future (see, e.g., [1]). As such, many countries are trying to meet the majority of their demand with environment friendly generation (see, e.g., [2], [3]). However, doing so is challenging due to the inherent intermittency and uncertainty of weather characteristics on which the output of such resources heavily relies on (see, e.g., [4], [5]). In this regard, building an efficient forecasting model is of vital importance.

Photovoltaic (PV) output forecast models are built based on different techniques: (i) statistical methods; (ii) Artificial Intelligence (AI); (iii) physical models; and (iv) hybrid approaches [6]. The statistical methods are based on analysing historical data while AI methods focus on the nonlinear relation between historical weather data and solar output to construct a probabilistic model [7]. Since the results that belong in the second group are assessed by error metrics which are based on statistics, they can also be categorised in the first group [8]. Physical models are mainly established based on the monitoring of satellite images and numerical weather forecasts to forecast the solar power output. Hybrid models are the combination of all aforementioned methods.

In this paper, we propose a framework that predicts the short-term solar power output based on the weather input data: temperature, zenith, azimuth, direct solar irradiance, diffused solar irradiance, and horizontal solar irradiance. We first categorise the data into distinct groups based on time of day and solar generation output using the k-means algorithm, which is a prominent clustering technique. The rationale behind this clustering is to gather similar data in clusters and determine a forecast model for each cluster. The clustering technique is used as a way to cope with the inherent variability and sparsity of PV output over time at different days and seasons. We use the Gaussian Process Regression (GPR) with a Matérn 5/2 kernel function (see, e.g., [9], [10]) to determine the nonlinear relationship between the input weather data and solar power output for each cluster. This methodology is suitable for modelling uncertainty sources of weather while being flexible to be implemented in time-series with a wide range of variations over time [11]. GPR takes the advantages of a kernel function to map the weather input data to solar power output data. The selection of the kernel function plays a crucial role in modelling the nonlinear relationship between the input data and the solar output [12]. By using GPR, the uncertainty of the input data is reflected to the output forecasts since this technique assumes each input as a random variable with an unknown distribution. This is true due to the Bayesian nature of GPR. Details on the proposed method may be found in [13].

The efficacy of the proposed approach is a function of the number of GPR models that are trained for a given dataset. In order to determine the optimal number of clusters that a dataset needs to be categorised, i.e., the number of GPR models used for a given dataset, we use the Elbow and Gap tests (see, e.g., [14], [15]). The Elbow and Gap tests show that the optimal number of clusters is four. To demonstrate the effect of different number of clusters in the performance of the proposed forecasting methodology, we present the results for one to eight number of clusters. For each cluster we use the input weather data to train our GPR model and determine the nonlinear relationship between the solar output and the weather data. To validate our methodology we use two datasets, which are based on two different locations at Denver and St. Lucia, and the 5-fold cross validation and holding-out data techniques. More specifically, to take into account different days in different seasons we choose 30 random days as hold-out test dataset.

II. Data Clustering

Clustering is a machine learning technique used as an unsupervised pattern classification learning method to partition the similar data in the same group based on distance or dissimilarity function [16]. In this work, we use the k-means clustering algorithm to group the data based on time of day and power output. Consider a set $\mathscr{X} = \{x_1, x_2, \ldots, x_N\}$ with N elements, where $x_i \in \mathbb{R}^n$ for all of $i = 1, \ldots, N$; the data point cluster number $C(i) \in \{1, \ldots, K\}$, $i \in \{1, \ldots, N\}$; the cluster centroid for cluster k $c_k \in \mathbb{R}^n$, $k = 1, \ldots, K$; and the Euclidean distance $d(x_i, c_k) = ||x_i - c_k||$, which is the distance between x_i and cluster centroid c_k. Then k-means clustering tries to minimise the following squared error function:

$$\underset{\{c_k\}_{k=1}^{K}}{\text{minimize}} \sum_{k=1}^{K} N_k \sum_{C(i)=k} d^2(x_i, c_k), \tag{1}$$

where N_k is the number of points assigned to cluster k. The performance of the k-means algorithm is greatly affected by the number of clusters. To determine the optimal number of clusters, we use two popular statistical algorithms, namely, the Gap and the Elbow (see, .e.g, [15], [17]).

The basic idea of Gap Statistic is to introduce reference datasets, which are generated with independent Monte Carlo simulations sampling from an empirical distribution and to calculate the sum of the squares of the Euclidean distance between two measurements in each cluster. To describe the Gap methodology we define the summation of all pairwise euclidean distances for all datapoints in cluster k to be $D_k = \sum_{i,i' \in C_k} d(x_i, x_i')$ and the normalized sum of intra-cluster distances to be $W_k = \sum_{k=1}^{K} \frac{1}{2N_k} D_i$. Then, we use the following function to measure the Gap value [15]:

$$\text{Gap}_n(k) = \mathbb{E}_n \left[\log(W_k)\right] - \log(W_k), \tag{2}$$

where $\mathbb{E}_n[\cdot]$ denotes the expectation operator under a sample of size n from the empirical distribution of the data. The optimal number of clusters based on the Gap statistic is the smallest number k that satisfies the following expression:

$$\text{Gap}_n(k) \geq \text{Gap}_n(k+1) - s_{k+1}, \tag{3}$$

where $s_k = \sqrt{1 + 1/B}\, s_d(k)$ is the simulation error and is calculated using the standard deviation $s_d(k)$ of B Monte Carlo replicates, in this study $B = 500$, drawn from the empirical distribution.

The Elbow technique uses the sum of squared errors (SSE), which is the sum of the distances between the sample points in each cluster and the centroid of the cluster as a performance indicator for a set number of clusters [18]. More specifically, the SSE is calculated over a series number of clusters. If small SSE values are obtained then that is an indication that each cluster is more convergent. When the number of clusters is set to approach the optimal number of clusters K, SSE shows a rapid decline. When the number of clusters exceeds K, SSE continues to decline but with a slower rate. Usually the optimal number of clusters K is obtained graphically at the point that looks like an "elbow", i.e., at the largest inflection point down. Once K is determined then if the selected number of clusters is less than K, the SSE will be greatly reduced for every 1 increase of the number of clusters. On the other hand, when the selected number of clusters is greater than K the change of the SSE will not be so obvious for every 1 increase of the selected number of clusters.

III. Proposed Gaussian Process Regression Framework

Once the data are clustered into K groups we train a GPR with a Matérn 5/2 kernel function to determine the nonlinear relationship between the direct solar irradiance, diffused solar irradiance, horizontal solar irradiance, temperature, zenith, and azimuth, which are the six weather input data and the solar output. GPR maps the input data into the solar output by defining a covariance function, which plays a crucial role in the process.

Let the training set $\mathscr{S} = \{(x^{(t)}, y^{(t)})\}_{t=1}^{T}$ be a set of random variables from some unknown distribution, where T is the period of available data with one hour resolution; $x^{(t)} \in \mathbb{R}^6$ is the vector containing all input data at time t; and $y^{(t)} \in \mathbb{R}$ the PV output at observation t. With the use of a Gaussian model we may relate the input with the output terms by:

$$y^{(t)} = f(x^{(t)}) + h(x^{(t)})^{\top} \beta + \epsilon^{(t)}, \text{ for } t = 1, \ldots, T, \tag{4}$$

where $\epsilon^{(t)}$ are i.i.d. "noise" variables with independent $\mathscr{N}(0, \sigma^2)$ distributions, $f(x^{(t)})$ is the mapping function $\mathbb{R}^6 \to \mathbb{R}$ and $h(x^{(t)})$ is a set of a fixed basis function. The explicit use of basis functions is a way to specify a non-zero mean over $f(x^{(t)})$. In this work we assume that $h(x^{(t)})$ is a 6×1 vector whose all entries are equal to the constant value of one, and β is the basis function coefficient 6×1 vector and is evaluated by maximising a likelihood function as described below. For notational convenience, we define:

$$X = \begin{bmatrix} (x^{(1)}) \\ \vdots \\ (x^{(T)}) \end{bmatrix} \in \mathbb{R}^{T \times 6}, y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(T)} \end{bmatrix} \in \mathbb{R}^{T}, \epsilon = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(T)} \end{bmatrix} \in \mathbb{R}^{T},$$

$$f = \begin{bmatrix} f(x^{(1)}) \\ \vdots \\ f(x^{(T)}) \end{bmatrix} \in \mathbb{R}^{T}, H = \left[h(x^{(1)}), \ldots, h(x^{(T)})\right] = \mathbb{1}_{6 \times T},$$

where $\mathbb{1}_{6 \times T}$ is a 6 by T matrix whose all elements are one. In matrix form we may rewrite (4) as

$$y = f(X) + H^{\top} \beta + \epsilon. \tag{5}$$

We assume a prior distribution over functions $f(X)$ as

$$f(X) \sim \mathscr{N}(0, K(X, X)), \tag{6}$$

where 0 is the mean value; $K(X, X)$ is the covariance matrix:

$$K(X, X) = \begin{bmatrix} k(x^{(1)}, x^{(1)}) & \ldots & k(x^{(1)}, x^{(T)}) \\ \vdots & \ddots & \vdots \\ k(x^{(T)}, x^{(1)}) & \ldots & k(x^{(T)}, x^{(T)}) \end{bmatrix},$$

where $k(\cdot,\cdot)$ is the kernel function. By using the kernel function we aim to actively model the unknown relationship between the input and the output variables. The kernel function is defined based on the likely pattern that we can observe in the data. One assumption to model the kernel may be that the correlation between any two points in the input set, i.e., $x^{(t)}, x^{(t')} \in \mathscr{S}$, with $t, t' = 1, \ldots, T, t \neq t'$, decreases with increasing the euclidean distance between them. This means that points with similar features behave similarly. Under this assumption, in this work we use the Matérn 5/2 as a kernel function, which is parameterised as follows:

$$\begin{aligned} k(x^{(t)}, x^{(t')}) \quad &= \sigma_f^2 \left(1 + \frac{\sqrt{5}d(x^{(t)}, x^{(t')})}{\sigma_l} + \frac{5d^2(x^{(t)}, x^{(t')})}{3\sigma_l^2}\right) \\ &\qquad e^{-\frac{\sqrt{5}d(x^{(t)}, x^{(t')})}{\sigma_l}}, \end{aligned} \tag{7}$$

where $d(x^{(t)}, x^{(t')})$ is the euclidean distance between any two input observations $x^{(t)}, x^{(t')} \in \mathscr{S}$ as defined in Section II; σ_l and σ_f, are two other kernel parameters which show respectively the characteristic length scale and the signal standard deviation that both belong in $\mathbb{R}^6$. The characteristic length scale σ_l defines how far the output $y^{(t)}$ needs to be away from the input features $x^{(t)}$ to become uncorrelated. These two parameters are greater than zero and are formulated as follows:

$$\sigma_l = 10^{\theta_l}, \sigma_f = 10^{\theta_f}. \tag{8}$$

We now define a new parameter θ to be:

$$\theta = \begin{bmatrix} \theta_l \\ \theta_f \end{bmatrix} = \begin{bmatrix} \log(\sigma_l) \\ \log(\sigma_f) \end{bmatrix} \in \mathbb{R}^{6\times 2}. \tag{9}$$

From (5) we may write that

$$y|f(X), X \sim \mathscr{N}(H^\top \beta, \sigma^2 I + K(X, X)), \tag{10}$$

since both $f(X)$ and ϵ have zero means. In order to determine the distribution that y follows, we need to determine three parameters, i.e., β, σ^2 and θ. $K(X, X)$ is a function of θ as may be seen in (7)-(9). β, σ^2, and θ are also known as the hyperparameters of the kernel function. In order to estimate the parameters we maximise the following marginal log-likelihood function

$$\log P(y|f(X), X) = \log P(y|X, \beta, \theta, \sigma^2). \tag{11}$$

Thus, the estimates of β, θ, and σ^2 denoted by $\hat{\beta}$, $\hat{\theta}$ and $\hat{\sigma}^2$ are given by

$$\hat{\beta}, \hat{\theta}, \hat{\sigma}^2 = \underset{\beta,\theta,\sigma^2}{\operatorname{argmax}} \log P(y|X, \beta, \theta, \sigma^2). \tag{12}$$

We may write from (10) and (11) that

$$P(y|X) = P(y|X, \beta, \theta, \sigma^2) = \mathscr{N}(H^T \beta, K(X, X) + \sigma^2 I). \tag{13}$$

Thus, the marginal log-likelihood function is

$$\begin{aligned} \log P(y|X, \beta, \theta, \sigma^2) = -\frac{1}{2}(y - H^\top \beta)^T [K(X, X) + \sigma^2 I]^{\,-1} \\ (y - H^\top \beta) - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|K(X, X) + \sigma^2 I|. \end{aligned} \tag{14}$$

We rewrite the likelihood function for the subset of parameters, σ^2 and θ, by expressing β as a function of the parameters of interest and replacing them in the likelihood function. Thus, we have that the estimate of β for given θ and σ^2 is:

$$\begin{aligned} \hat{\beta}(\theta, \sigma^2) = [H^\top [K(X, X|\theta) + \sigma^2 I]^{\,-1} H]^{\,-1} \\ H^\top [K(X, X|\theta) + \sigma^2 I]^{\,-1} y. \end{aligned} \tag{15}$$

By substituting (15) in (14) we have

$$\begin{aligned} \log P(y|X, \hat{\beta}(\theta, \sigma^2), \theta, \sigma^2) = -\frac{1}{2}(y - H\hat{\beta}(\theta, \sigma^2))^T \\ [K(X, X|\theta) + \sigma^2 I]^{\,-1}(y - H\hat{\beta}(\theta, \sigma^2)) \\ -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|K(X, X|\theta) + \sigma^2 I|. \end{aligned} \tag{16}$$

We now may determine the hyperparameters as the output of the above optimisation problem.

Once the hyperparameters are evaluated we may use (10) to predict the output of solar generation based on the input parameters.

IV. Numerical results

The proposed methodology presented in Section III is implemented in two datasets from different sites based on available lagged historical data we gathered from National Solar Radiation, Iowa Environmental Mesonet (IEM) and National Renewable Energy Laboratory and the University of Queensland. We combined all the data from these resources and built a consistent dataset. The two sites' details are given in Table I. To quantify the effect of the number of clusters on the forecasts obtained by the proposed framework, we modify the number of clusters, which are used in the development of the GPR models, from one to eight and compare the forecast error metrics for the various numbers of clusters for Site A.

To obtain meaningful comparison metrics we use 5-fold cross-validation and hold out validation as two of the most prevalent test methods used in recent studies [20]. In 5-fold cross validation the whole data is split into 5 folds: at each time, 4 folds are used as a training set and a one-fold as a testing set, until all folds are used to build the forecast model. We randomly select 30 days of a year as hold-out data while the remaining data are used for training and testing using 5-fold cross-validation.

A. Optimal number of clusters

We apply the Gap and Elbow algorithms on the datasets as described in Section II and find that the optimal number of clusters is four. As depicted in Fig. 1, four is the smallest number of clusters where the Gap value is higher than the precedent and the subsequent value, and satisfies (3). In Fig. 2, the "elbow" of the curve which happens at the optimal number of clusters is found in $K = 4$. As such it was shown with the Elbow method that the choice of four clusters means that if the

Site	Location	Size [MW]	Latitude [°]	Longitude [°]
A	Denver Intl Airport	30	39.8561 N	104.6737 W
B	St Lucia	0.433	27.498 S	153.013 E

Table I: Site description.

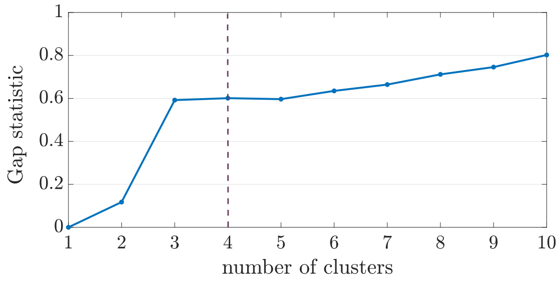

Figure 1: Gap optimal number of clusters

selected number of clusters is less than four, the SSE value will be greatly reduced for every 1 increase of the number of clusters. On the other hand, when the selected number of clusters is greater than four the change of the SSE value will not be so obvious for every 1 increase of the selected number of clusters.

B. Framework Implementation on Site A

To show how the solar power forecast is affected by the number of clusters we apply the proposed framework on Site A for one to eight clusters. More specifically, in the case of one cluster we only train one GPR model for the entire dataset, in the case of two clusters we train two GPR models one for each cluster; and so on until we have eight clusters and eight GPR models. We use the 5-fold cross validation and hold out validation techniques to obtain the forecast errors and be able to analyse the clustering effect on the accuracy of the solar power forecasting.

For Site A the available historical data comprise of hourly input weather data: diffused solar irradiance, horizontal solar irradiance, direct solar irradiance, temperature, zenith, and azimuth from 2006, i.e., we have 6 × 8760 data points for weather input data and the solar generation output. We implement the proposed framework in one to eight number of clusters; and select 30 random days as hold-out data as representative of different days of the year during different seasons. Each cluster is trained by using Matérn 5/2 GPR and tested by 5-fold cross-validation and hold-out techniques. The error metrics used are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T_\star}\sum_{t=1}^{T_\star}\left(\tilde{y}^{(t)} - y_\star^{(t)}\right)^2}, \tag{17}$$

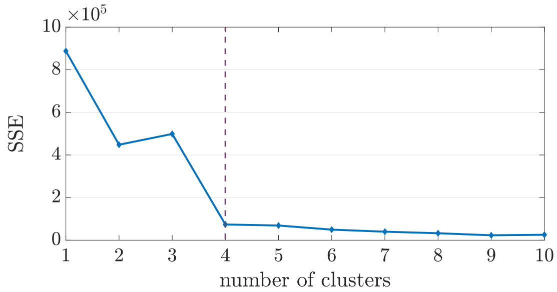

Figure 2: Elbow optimal number of clusters

no. of clusters	RMSE [MW]	MAE [MW]	RMSE [%]	MAE [%]
8	0.90	0.34	2.72	1.02
7	0.80	0.26	2.43	0.78
6	1.02	0.40	3.10	1.20
5	1.18	0.49	3.58	1.48
4	1.29	0.36	3.91	1.08
3	1.53	0.47	4.63	1.43
2	1.64	0.66	4.98	2.00
1	2.94	0.58	8.91	1.77

Table II: Training set error metrics for various number of clusters.

$$\mathrm{MAE} = \frac{1}{T_\star}\sum_{t=1}^{T_\star}\left|\tilde{y}^{(t)} - y_\star^{(t)}\right|, \tag{18}$$

$$\mathrm{MSE} = \frac{1}{T_\star}\sum_{t=1}^{T_\star}\left(\tilde{y}^{(t)} - y_\star^{(t)}\right)^2, \tag{19}$$

where $y_\star^{(t)}$, is the prediction value for solar generation at time t, and by $\tilde{y}^{(t)}$ the actual value at time t; $T_\star$ is the number of hourly intervals we are forecasting the solar output. We may also normalise values of the above metrics with respect to peak value.

The results of the forecast error metrics for the training and test sets for one to eight number of clusters are given in Tables II, III. The error metrics of the training data between the actual and the predicted values are based on the average error of all 5 folds for the training set. It should be noted that the test results are expected to be different from the training set results, since 30 hold-out days are not shown to the model during the training process. However, the results with any test set should be approximately the same as those obtained with the training set, as it may be seen in Tables II, III. We notice that the error metrics are usually improved as we increase the number of clusters. However, at the same time a choice of a large number of clusters increases the computational complexity of the model since for each cluster we build a GPR model. The number of clusters needs to balance the trade-off between two different objectives of minimum forecast error and minimum number of clusters due to the computational complexity.

In this regard, we further study the effect of the number of clusters in the forecast error and depict in Figs. 3, 4, 5 the forecasts for the training set along with the actual values. As seen in these figures the different patterns of solar generation are better captured and modelled in the case of eight clusters. However, partitioning the data into four clusters also leads to

no. of clusters	RMSE [MW]	MAE [MW]	RMSE [%]	MAE [%]
8	0.80	0.38	2.41	1.14
7	1.01	0.48	3.05	1.45
6	0.95	0.41	2.87	1.25
5	1.00	0.47	3.02	1.43
4	1.08	0.50	3.26	1.52
3	1.46	0.65	4.43	1.97
2	1.44	0.68	4.36	2.05
1	2.75	1.16	8.35	3.53

Table III: Test set error metrics for various number of clusters.

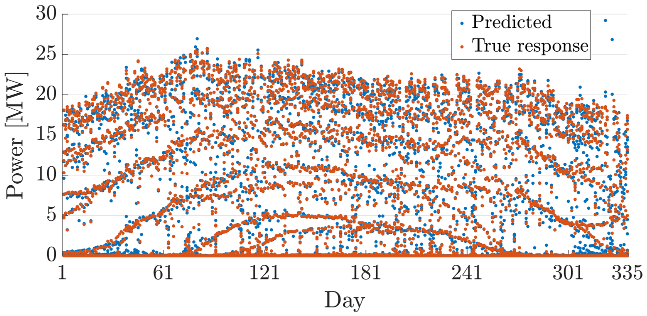

Figure 3: Proposed framework predictions of the training data set for one cluster.

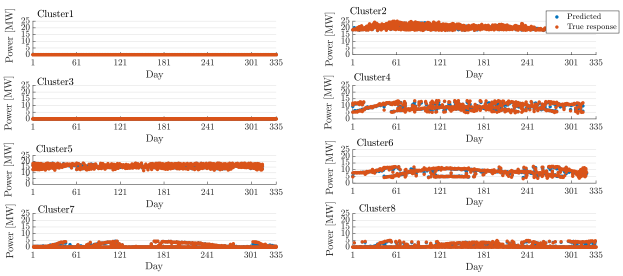

Figure 5: Proposed framework predictions of the training data set for eight clusters.

		RMSE[%]	MAE[%]
Proposed framework		**3.48**	**1.85**
[11]	Fall	13.85	8.48
	Winter	7.67	4.16
	Spring	13.6	8.08
	Summer	16.43	10.73
[42]	ELM	12.84	6.68
	FFBPG	13.33	7.53

Table IV: Forecast error metrics based on different methodologies for Site B.

good results in comparison to eight based on the results we can see in Fig. 6, where the sensitivity on the number of clusters to different normalised error metrics values is depicted. As such, we partition the data into four clusters and as seen in Fig. 7, clusters two and three, represent the seasonal variations while clusters one and four represent early morning and night times.

C. Comparison with existing methodologies

To compare our results with the existing methodologies, we use the same data as in [21], [22] which are available from University of Queensland. The temporal resolution of the data in [21] is 1-minute; however since we are interested in hourly values we select historical data with hourly resolution. We used 2012 data for training and 2013 data for testing. The authors of [21], categorized the data into four different seasons, i.e., fall, winter, spring and summer. Also, in [22], ELM method and the traditional feed-forward back propagation neural network (FFBPG) are used for forecast model. The results in Table IV, clearly show that the results of prediction for one year is better than the results in [21], [22].

V. Conclusion

In this work, we proposed a framework that predicts the short-term solar output based on weather input data: temperature, zenith, azimuth, direct solar irradiance, diffused solar irradiance, and horizontal solar irradiance. We clustered the data in a given number of groups based on time of day. We then trained a model for each cluster using GPR in order to learn the relationship between the input weather data and the PV generation. GPR is a kernel based nonlinear nonparametric regression technique, in which the covariance function plays a crucial role. In this work, we selected the Matérn 5/2 as a covariance or kernel function. This function was selected under the assumption that the correlation between any two points in the input feature set decreases with increasing the euclidean distance between them. We analysed the effect on the performance of the proposed framework of the number of chosen clusters. More specifically, we implemented two statistical methods, namely Gap and Elbow, to identify the optimal number of clusters. The methods showed that four is the optimal number of clusters. In the numerical results' section we used k-means algorithm to cluster the data based on solar output and time of day into one to eight clusters and calculated error forecast metrics. This sensitivity study also demonstrated the improved framework performance when four clusters are chosen in terms of balancing model complexity and accuracy.

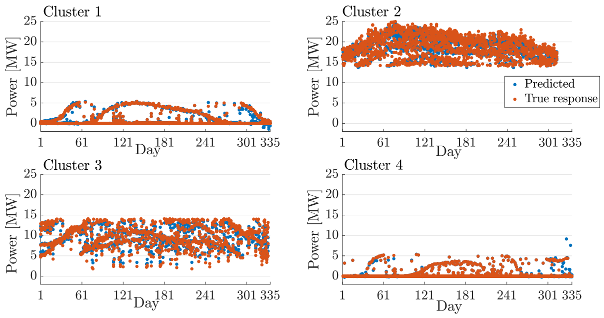

Figure 4: Proposed framework predictions of the training data set for four clusters.

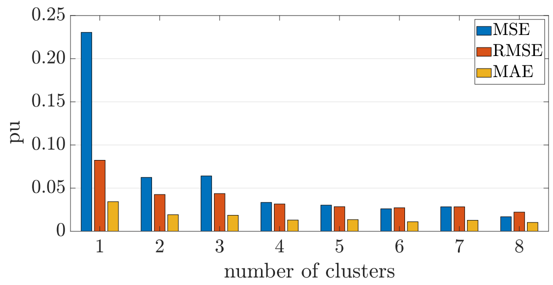

Figure 6: Sensitivity study on the number of clusters by comparing different normalised error metrics values.

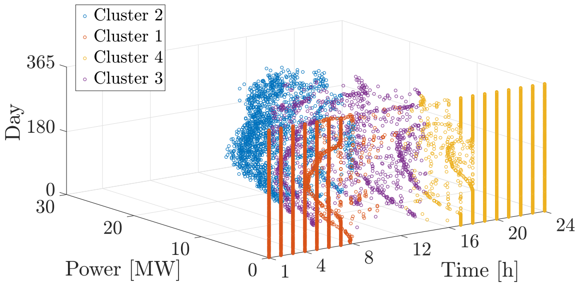

Figure 7: 3D graph of four clusters. Different colours represent different clusters.

References

[1] M. Gul, Y. Kotak, and T. Muneer, "Review on recent trend of solar photovoltaic technology," *Energy Exploration & Exploitation*, vol. 34, no. 4, pp. 485–526, 2016. [Online]. Available: https://doi.org/10.1177/0144598716650552

[2] R. J. Bessa, J. Dowell, and P. Pinson, *Renewable Energy Forecasting*. American Cancer Society, 2016, pp. 1–21. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118755471.sgd050

[3] D. Apostolopoulou and M. McCulloch, "Optimal short-term operation of a cascaded hydro-solar hybrid system: A case study in kenya," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 4, pp. 1878–1889, 2019.

[4] R. Shah, N. Mithulananthan, R. Bansal, and V. Ramachandaramurthy, "A review of key power system stability challenges for large-scale pv integration," *Renewable and Sustainable Energy Reviews*, vol. 41, pp. 1423 – 1436, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032114008004

[5] R. G. Wandhare and V. Agarwal, "Reactive power capacity enhancement of a pv-grid system to increase pv penetration level in smart grid scenario," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1845–1854, July 2014.

[6] A. Bracale, G. Carpinelli, and P. De Falco, "A probabilistic competitive ensemble method for short-term photovoltaic power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 551–560, April 2017.

[7] M. Rana, I. Koprinska, and V. G. Agelidis, "Forecasting solar power generated by grid connected pv systems using ensembles of neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–8.

[8] M. K. Behera, I. Majumder, and N. Nayak, "Solar photovoltaic power forecasting using optimized modified extreme learning machine technique," *Engineering Science and Technology, an International Journal*, vol. 21, no. 3, pp. 428 – 438, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2215098617316063

[9] C. Xu, Z. Feng, and Z. Meng, "Affective experience modeling based on interactive synergetic dependence in big data," *Future Generation Computer Systems*, vol. 54, pp. 507 – 517, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X15000527

[10] G. I. Nagy, G. Barta, S. Kazi, G. Borbly, and G. Simon, "Gefcom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1087 – 1093, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169207015001521

[11] Z. Li, Y. Han, and P. Xu, "Methods for benchmarking building energy consumption against its past or intended performance: An overview," *Applied Energy*, vol. 124, pp. 325 – 334, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306261914002505

[12] Y. Heo and V. M. Zavala, "Gaussian process modeling for measurement and verification of building energy savings," *Energy and Buildings*, vol. 53, pp. 7 – 18, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037877881200312X

[13] F. Najibi, D. Apostolopoulou, and E. Alonso, "Gaussian process regression for probabilistic short-term solar output forecast," 2020.

[14] M. A. Masud, J. Z. Huang, C. Wei, J. Wang, I. Khan, and M. Zhong, "I-nice: A new approach for identifying the number of clusters and initial cluster centres," *Information Sciences*, vol. 466, pp. 129 – 151, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025517301135

[15] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[16] G. Liu and X. Wu, "Time series clustering and evaluation of unknown working conditions of mismatched photovoltaic array systems," in *2018 IEEE 4th International Conference on Control Science and Systems Engineering (ICCSSE)*, Aug 2018, pp. 164–167.

[17] M. Antunes, D. Gomes, and R. L. Aguiar, "Knee/elbow estimation based on first derivative threshold," in *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2018, pp. 237–240.

[18] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[19] M. Aakroum, A. Ahogho, A. Aaqir, and A. A. Ahajjam, "Deep learning for inferring the surface solar irradiance from sky imagery," in *2017 International Renewable and Sustainable Energy Conference (IRSEC)*, Dec 2017, pp. 1–4.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[21] F. Golestaneh and Hoay Beng Gooi, "Batch and sequential forecast models for photovoltaic generation," in *2015 IEEE Power Energy Society General Meeting*, July 2015, pp. 1–5.

[22] F. Golestaneh, P. Pinson, and H. B. Gooi, "Very short-term nonparametric probabilistic forecasting of renewable energy generation with application to solar energy," *IEEE Transactions on Power Systems*, vol. 31, pp. 3850–3863, 2016.