Assessing the Impact of Spatial Proximity Data on the Solar Insolation Prediction

Sunghwan Bae, *Member, IEEE* and Saeed D. Manshadi, *Member, IEEE*

***Abstract*—Improving the prediction of the availability of solar energy resources became a necessary component in the operation of utilities with a high penetration level of renewable energy resources. In this paper, the solar insolation data in spatial proximity is leveraged to investigate the error in the prediction of solar insolation using multiple learning algorithms. Different error measures are utilized to evaluate the accuracy of the presented linear and non-linear learning algorithms. Essential data pre-processing steps are conducted on the solar insolation data available from multiple meteorological stations in spatial proximity. The impact of utilizing the spatio-temporal data compared to temporal data is analyzed. A comprehensive analysis based on multiple error measures is presented to compare the prediction error while employing multiple learning algorithms. It is shown that it is possible to identify the particular station and the particular learning algorithm that contribute the most in improving the solar insolation prediction of a specific location.**

***Index Terms*—Spatio-temporal data, solar insolation, learning algorithms, spatial proximity, weather reported data.**

I. Introduction

RENEWABLE energy resources including solar energy have brought several opportunities and challenges to the operation of the power system. Over recent years, the share of solar energy resources in the electricity generation mix has rapidly increased because of the aim to reduce the carbon footprint of electricity generation as well as the decrease in their installment cost. Nevertheless, power system operators still need to accommodate the intrinsic intermittency in the availability of solar energy resources with advanced prediction techniques [1]. Since the availability of solar energy resources is directly related to solar insolation, solar insolation prediction is critical to ensure a smooth operation of the power system with a high penetration level of solar energy resources. This work aims to leverage the spatial proximity data to enhance the prediction of solar insolation prediction by finding answers to the following questions.

1) What would be the best way to describe the accuracy of a learning algorithm for solar insolation prediction?
2) Is it possible to identify a single set of error measure, learning algorithm, and predetermined spatio-temporal relationship for solar insolation prediction?
3) Is it possible to define a specific set of spatial proximity data to enhance solar insolation prediction?
4) Is it possible to provide the best combination of spatial proximity data and learning algorithm in prediction performance perspective?

In [1], satellite and numerical weather prediction were used as the best tools for the hour-ahead and day-ahead solar insolation prediction because it was an adequate prediction technique for time horizons of more than 5 hours. To estimate and measure weather information, various algorithms such as metaheuristic, machine learning, and neural network have been proposed for solar insolation, wind speed, and raindrops, respectively [2]–[4]. However, modern solar insolation prediction is more complex. The sky image is leveraged in [5], [6], where short-term solar insolation is predicted with multiple total sky images. A ground-based sky imaging system is applied to solar insolation prediction [6]. Various satellite images are used to predict solar insolation and power in [7]–[9]. In [7], a solar power prediction model is proposed based on various satellite images and a support vector machine (SVM) learning scheme. To improve the results obtained with ground data, satellite's global horizontal irradiance (GHI) data, as well as total cloud cover data, are used as additional inputs for the artificial neural network model in [8]. Solar insolation prediction using satellite images based on cloud motion vectors is proposed in [9]. PV module characteristics are also considered in the prediction model where historical data includes measurements of the PV currents, voltages, and the module temperature, information normally available to the PV plant operator [10].

Among the recent regression models, a weighted Gaussian process regression approach is provided in [11] that data samples with higher outlier potential have a low weight. In terms of Markov properties, the model of [12] predicted the probability distribution function of power generation of PV systems based on the higher order Markov chain. The authors of [13] presented Hidden Markov Model with Pearson R model which is utilized for the extraction of shape based clusters from the input meteorological parameters. In [14], K-means clustering algorithm was applied to collect meteorological data and one-hour ahead prediction of solar insolation is performed based on meteorological factors including the cloud cover and SVM. In [15], at the early stage, the author used one year of three publicly available numerical weather prediction models, the North American mesoscale forecast system, the global forecast system, and the short-range ensemble forecast. Then, after the analog ensemble/blending procedure, forecast error could be reduced in PV power prediction compared with persistence model, system advisor model, and SVM respectively. The authors of [2] proposed a shallow neural network with an embedded sensor system to achieve a cost-effective solar insolation estimate of large PV plants. In [16], the authors reviewed various prediction methodologies including simple second-order regression, shallow neural network, quantile ran-

The authors are with the Department of Electrical and Computer Engineering, San Diego State University, San Dieg, CA, 92182, USA email:(sbae@sdsu.edu; smanshadi@sdsu.edu)

dom forest, k-nearest neighbors, and support vector regression. Then these models are averaged out as an ensemble model with optimal weights. Although 32 PV datasets and solar insolation information are utilized for predictive models, it only focused on the dedicated analysis of individual PV plant without characterizing the interdependence between multiple PV sites. According to [16], since the accuracy of solar power prediction heavily depends on the accuracy of available solar insolation, it is necessary to verify the spatio-temporal relationship to enhance the prediction of solar insolation in multiple meteorological stations.

To capture spatio-temporal relations of the solar insolation, a generative model with convolutional graph auto-encoder is proposed in [17] by verifying the reliability, sharpness, and continuous ranked probability score. Another attempt presented in [18] was to simulate the solar power generation by a wavelet-based variability model which requires a single sensor's time-series measurements and distance-based correlation function. In [19] the probabilistic forecast model was proposed to predict densities of solar generation, where quantile regression with L_1 penalization is adapted to deal with the high dimensional input of numerous solar sites. The authors of [20] point out that the performance of prediction highly depends on the time resolution (the sampling rate), especially for partly cloudy and partly sunny weather, which in turns causes significant errors. Considering temporal and spatial dynamics [20], the proposed hidden Markov models of [21] estimated solar power distributions and geographic auto-correlation based on high temporal and spatial resolution data from closely distributed solar sites., e.g., 1–15mins resolution data, and subdivided 2km^2 regions from one of two 256km^2 areas. It is shown in [22] that a spatial-temporal model could successfully reduce prediction errors. This paper aims to further assess the importance of spatio-temporal relationships on improving the solar insolation prediction given various learning algorithms based on various error metrics. The contributions of this paper are summarized as follows:

1) The real data for solar insolation in spatial proximity is exploited to assess the impact of utilizing the spatio-temporal correlation within the data for improving the solar insolation prediction. A data cleansing process is implemented on the available data set to enhance the accuracy of the learning algorithm. Besides, the most important features for short-term solar insolation prediction are selected which verifies the importance of the solar insolation data in spatial proximity.
2) The performance of various linear and non-linear learning algorithms based on temporal data and spatio-temporal data are investigated. The merits of each algorithm are discussed in detail to customize the best strategies to enhance solar insolation prediction.
3) The impact of the spatio-temporal data from adjacent stations on the solar insolation of a station is illustrated. Also, the most influential station to improve the performance of solar insolation prediction for each station is identified.
4) A comprehensive analysis is executed over the performance of each of the implemented learning algorithm, where multiple error metrics are presented to measure the accuracy of insolation prediction based on various measures. We summarize the best combination of learning methods and important adjacent stations for each station, respectively.

II. Learning Algorithms and Methodologies

Suppose that there is an ideal and unknown target function $f\colon \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is input space whose elements are $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y}$ is output space whose elements are $y \in \mathcal{Y} = \mathbb{R}$. The data set $\mathcal{D} = (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$ consisting of a vector of elements $(\mathbf{x}_i, y_i)$. Since the target f is unknown, the function is learned from the given data set. Through a learning algorithm $\mathcal{A}$, a hypothesis g is selected from hypothesis set $\mathcal{H}$, and g is used for a learned prediction function when $g \approx f$. Fig. 1 shows a summary of the learning setup for predictive models [23].

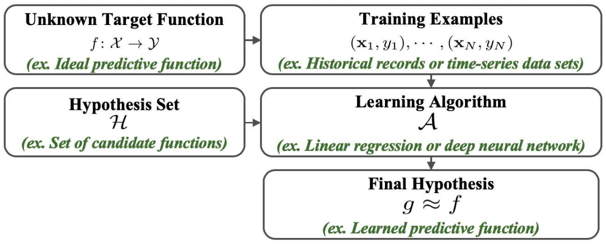
Unknown Target Function
$f\colon \mathcal{X} \to \mathcal{Y}$
(ex. Ideal predictive function)
Training Examples
$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$
(ex. Historical records or time-series data sets)
Hypothesis Set
$\mathcal{H}$
(ex. Set of candidate functions)
Learning Algorithm
$\mathcal{A}$
(ex. Linear regression or deep neural network)
Final Hypothesis
$g \approx f$
(ex. Learned predictive function)

Fig. 1. Summary of the general learning setup.

A. Linear regression (LR) with multiple features

The LR method is a linear learning algorithm. Suppose n-multiple features as input space for learning algorithms with m-training examples. For i^{th} training example $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^n$, $\mathbf{x}_i^T = [x_i^{(1)}, \cdots, x_i^{(j)}, \cdots, x_i^{(n)}]$, where $x_i^{(j)}$ is the value of the feature j in i^{th} training example. For convenience of notation, $x_i^{(0)} = 1$, then $\mathbf{x}_i \in \mathbb{R}^{n+1}$ are defined in linear models. For multiple features input $\mathbf{x}$, a hypothesis g of the LR method is

$$g_{_{\text{LR}}}(\mathbf{x}) = \theta^T \mathbf{x}, \tag{1}$$

where $\theta \in \mathbb{R}^{n+1}$ is a vector of parameters (weights) parameterizing the space of the LR method function $g\colon \mathcal{X} \to \mathcal{Y}$. As a conventional method, θ is adjusted by using the cost function with training examples. For m-training examples, the cost function J of the LR method is as follows:

$$J_{_{\text{LR}}}(\theta) = \frac{1}{2}\sum_{i=1}^{m}(g_{_{\text{LR}}}(\mathbf{x}_i) - y_i)^2. \tag{2}$$

Then θ is updated to minimize the sum of the squared residuals in the cost function by using the gradient descent algorithm and the learning rate.

B. Least absolute shrinkage and selection operator (LASSO)

The LASSO method is also a linear learning algorithm that adds a penalty term (L_1 regularization) in the cost function of the linear regression which is the absolute value of the magnitude of the parameters, e.g., the penalty term is $\lambda\|\theta\|_1$, where λ is a tuning parameter that controls the speed of the improvement in error by adjusting the penalty effect. Similar to the LR method, a hypothesis g of the LASSO method is

$$g_{\text{LASSO}}(\mathbf{x}) = \theta^T \mathbf{x}. \tag{3}$$

and the cost function J of the LASSO method is

$$J_{\text{LASSO}}(\theta) = \frac{1}{2}\sum_{i=1}^{m}(g_{\text{LASSO}}(\mathbf{x}_i) - y_i)^2 + \lambda\|\theta\|_1. \tag{4}$$

when λ is zero, the LASSO method is simply equivalent to the LR method.

C. Support vector regression (SVR)

The SVR method is a non-linear learning algorithm. The ϵ-support vector regression (ϵ-SVR) is represented here. For m-training examples, the standard form of SVR is

$$\min_{\mathbf{w},b,\xi,\xi^*} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*), \tag{5a}$$

$$\text{subject to} \quad \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \epsilon + \xi_i, \tag{5b}$$

$$y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i^*, \tag{5c}$$

$$\xi_i, \xi_i^* \ge 0, i = 1, \cdots, m, \tag{5d}$$

where given hyper-parameters $C > 0$ and $\epsilon > 0$, $\phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ into a higher dimensional space, $\mathbf{w}$ and b are coefficients, and two positive slack variables ξ and ξ^* are introduced to represent the distance from the actual values to the corresponding boundary values of the ϵ-tube with support vectors. Based on the Karush-Kuhn-Tucker (KKT) conditions, the optimization problem is solved by transforming into the dual problem with Lagrange multipliers. The dual problem is

$$\min_{\alpha_i,\alpha_i^*} \quad \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T Q(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)$$

$$+ \epsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*), \tag{6a}$$

$$\text{subject to} \quad \mathbf{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0, \tag{6b}$$

$$0 \le \alpha_i, \alpha_i^* \le C, i = 1, \cdots, m, \tag{6c}$$

where α_i, α_i^* are Lagrange multipliers, $\mathbf{e}^T$ is the all ones vector, and Q is an m by m positive semi-definite matrix. $Q_{k,l} = K(\mathbf{x}_k, \mathbf{x}_l)$, where $K(\mathbf{x}_k, \mathbf{x}_l)$ is the kernel function, e.g., $K(\mathbf{x}_k, \mathbf{x}_l) = \phi(\mathbf{x}_k)^T\phi(\mathbf{x}_l)$. The kernel function is used as the radial basis function (RBF) kernel, so $K(\mathbf{x}_k, \mathbf{x}_l) = \exp(-\gamma\|\mathbf{x}_k - \mathbf{x}_l\|^2)$, where γ is kernel parameter. Once the dual problem is solved, a hypothesis g of SVR with m-training examples is

$$g_{\text{SVR}}(\mathbf{x}) = \sum_{i=1}^{m}(-\alpha_i + \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b. \tag{7}$$

D. Deep neural network (DNN)

An artificial neural network (ANN) combines multiple processing layers that use simple and interconnected non-linear functions in parallel. Specifically, components of an ANN have an input layer, multiple hidden layers, and an output layer. The layers are interconnected via nodes, or neurons, and each layer gets the outputs of the previous layer as its input. A multilayer perceptron (MLP) which is a class of feedforward neural network is used here. If there are many hidden layers with large-scale neurons in an ANN, its structure can be considered as a deep neural network (DNN). Fig. 2 shows an example of DNNs structure with input, hidden, and output layers. The DNN method is a non-linear learning algorithm. When the DNN method has L total layers (hidden and output layers), a hypothesis g is

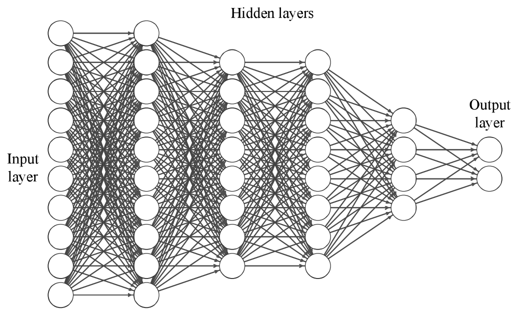

Fig. 2. A structure example of a feedforward deep neural network (DNN).

$$g_{\text{DNN}} = \varphi^{[L]}\left(\mathbf{W}^{[L]}\nu^{[L-1]} + b^{[L]}\right), \tag{8}$$

where for the output layer L, $\varphi^{[L]}$ is an activation function, $\mathbf{W}^{[L]}$ is an $n^{[L]}$ by $n^{[L-1]}$ weight matrix ($n^{[l]}$ is the number of neurons in l^{th} layer), $\nu^{[L-1]}$ is the outputs of the previous $L-1$ hidden layer, and $b^{[L]}$ is an $n^{[L]}$ by 1 bias vector. Similarly, as $\nu^{[L-1]} = \varphi^{[L-1]}\left(\mathbf{W}^{[L-1]}\nu^{[L-2]} + b^{[L-1]}\right)$, in the first hidden layer, $\nu^{[1]} = \varphi^{[1]}\left(\mathbf{W}^{[1]}\mathbf{x} + b^{[1]}\right)$ with multiple features input $\mathbf{x}$. And for m-training examples, the cost function J of the DNN method is

$$J_{\text{DNN}}(\mathbf{W}^{[L]}, b^{[L]}, \cdots, \mathbf{W}^{[1]}, b^{[1]}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(g_{\text{DNN}}(\mathbf{x}_i), y_i), \tag{9}$$

where the function $\mathcal{L}$ can be various loss functions such as the cross-entropy loss, the mean absolute error and the mean square error for different predictive purposes. Then all parameters of the cost function can be updated by various gradient descent algorithms with proper learning rates.

III. Prediction Performance Evaluation

The following metrics are presented to measure the performance of learning algorithm presented in the previous section.

1) Mean-absolute error (MAE): MAE reflects the average magnitude of the errors.

$$\frac{1}{s}\sum_{t=1}^{s}|y_t - \hat{y}_t|, \tag{10}$$

where s is the number of samples of test data, y_t is actual value, and $\hat{y}_t$ is the predicted value.

2) Relative (or normalized) mean-absolute error (rMAE or nMAE): rMAE is the normalized MAE by the mean of actual values.

$$\left[\frac{1}{s}\sum_{t=1}^{s}|y_t - \hat{y}_t|\right] \bigg/ \left[\frac{1}{s}\sum_{t=1}^{s}y_t\right]. \tag{11}$$

3) Root-mean-square error (RMSE): RMSE is the square root of the quadratic mean of errors.

$$\sqrt{\frac{1}{s}\sum_{t=1}^{s}(y_t - \hat{y}_t)^2}. \tag{12}$$

4) Relative (or normalized) root-mean-square error with the mean (rRMSE_mean): rRMSE_mean is the normalized RMSE by the mean of actual values.

$$\left[\sqrt{\frac{1}{s}\sum_{t=1}^{s}(y_t-\hat{y}_t)^2}\right] \Bigg/ \left[\frac{1}{s}\sum_{t=1}^{s}y_t\right]. \tag{13}$$

5) Relative (or normalized) root-mean-square error with the maxmin (rRMSE_maxmin): rRMSE_maxmin is the normalized RMSE by the difference between the maximum and the minimum values of actual values.

$$\left[\sqrt{\frac{1}{s}\sum_{t=1}^{s}(y_t-\hat{y}_t)^2}\right] \Bigg/ [y_{\max}-y_{\min}]. \tag{14}$$

IV. Data Preparation

A. Geographical information and data set

The data set for solar insolation from four weather stations are utilized here [24]. Fig. 3 shows the geographical distribution of these four weather stations in San Diego County, California. The aerial distance between stations is about 10 to 20 miles. As the minimum resolution, hourly solar insolation [Wh/m^2] is currently available at each station from 00:00 to 23:00 (PST) [24]. Note that different time resolution can significantly affect on the prediction performance due to the weather dynamics [20], [25]. In addition to the time resolution, the prediction of solar insolation depends on the geographical location (spatial data) and ambient conditions (air temperature, wind, and gust speed) [25]. The main motivation is to verify the effectiveness of the geographical information, and the proposed learning models are also scalable regardless of the time resolution. Here, the time resolution is set to be an hour to verify the impact of spatial proximity data on the solar insolation prediction. The detail information including station ID, latitude and longitude, as well as elevation are presented in Table I [24]. These stations are selected as not only they are geographically dispersed, but also they are in various elevations in the mountain area with diverse cloud coverage.

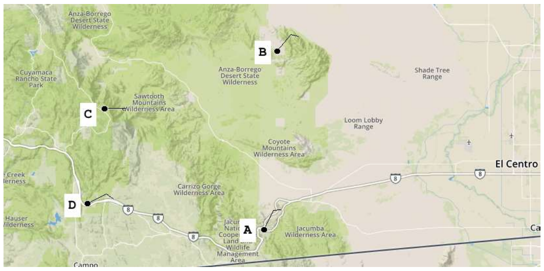
Anza-Borrego Desert State Wilderness
B
Anza-Borrego Desert State Wilderness
Shade Tree Range
Cuyamaca Rancho State Park
C
Sawtooth Mountains Wilderness Area
Loom Lobby Range
Coyote Mountains Wilderness Area
El Centro
Creek
lerness
8
8
Carrizo Gorge Wilderness Area
D
8
8
8
Hauser
/ilderness
Jacur
Natic
Cooper
Land and
Wildlife
Management
Area
A
Jacumba Wilderness Area
Ca
Campo

Fig. 3. Geographical location of weather stations in San Diego, California.

TABLE I
Information of weather stations.

Station_ID (Name)	TNSC1 (A)	FHCC1 (B)	MLGC1 (C)	CMNC1 (D)
Latitude	32.67	32.99	32.88	32.72
Longitude	-116.09	-116.07	-116.43	-116.46
Elevation [ft]	2044	781	5737	3268

B. Data cleansing

Here, the full records of hourly solar insolation data from March 13$_{\text{th}}$ 2017, 00:00 to June 30$_{\text{th}}$ 2017, 23:00 is utilized. The same resolution is employed for the short-term prediction of solar insolation. Since the data gathering process is not ideal, there are multiple series of missing and duplicated values in the raw data set. Therefore, imputation techniques are employed to replace these abnormal values by interpolating neighbor values in highly correlated time steps. After data cleansing, the total of 110 days of data is available in which are compatible with 2,640 hourly solar insolation samples. All samples of 110 days are used in approximately 99 days for training and verification examples and are used in approximately 11 days for test data.

C. Data pre-processing

Data pre-processing includes input data dimension (feature selection) and input data normalization (feature normalization). Feature selection is necessary to utilize an adequate time horizon as input features to capture daily status. Intuitively, the short-term prediction is mainly affected by the daily periodicity of local daylight hours as well as the temporal tendency of highly correlated time intervals. In addition to one station's temporal data, spatio-temporal data from adjacent stations might be meaningful to improve the prediction of solar insolation. As a preliminary step, we need to quantify the importance of available input information to investigate the potential improvement of utilizing the spatio-temporal data from an adjacent station. To do so, the Gini index of decision tree algorithms is leveraged as feature importance values. In Fig. 4, the top 10 most important features are represented in station A perspective, when both station A's temporal data A$[t]$ and station B's spatio-temporal data B$[t]$ are utilized together, where $[t]$ means a t-hour ahead temporal information. Note that all of the importance values are relatively scaled to the maximum importance value. It is not surprising that the most effective time features are a few last hours on that day as well as a few hours around on previous days. This is why highly correlated previous time steps such as 1, 25, and 48 show a dominant importance value. Even though the result

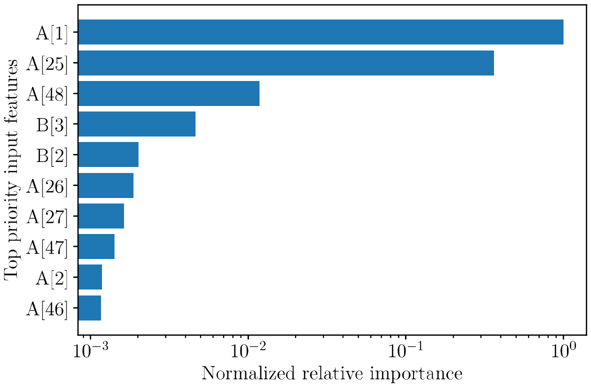

Fig. 4. For station A, the top important input features are illustrated when both its own temporal data A$[t]$ and station B's spatio-temporal data B$[t]$ are utilized together, where $[t]$ means a t-hour ahead temporal information.

of feature importance cannot capture all relationships between output and input data, it is reasonable to select a proper amount of significant features to capture daily periodicity and daily tendency. Based on a heuristic search, a 30-hour time horizon is selected as input features for our simulation.

Feature normalization can be useful when features use different scales, ranges, and units. The presented learning algorithms might converge during training without feature normalization because features of previous time horizon use the same unit as the input. However, in case that models are difficult for training, features should be normalized by removing the mean and scaling to unit variance as a standard scale. Here, the statistics are extracted only from training examples because learning algorithms do not have any information about test data.

V. Simulation Results

Once the essential steps of data cleansing and preprocessing are verified, the input data is ready to train the learning algorithms. To perform hour-ahead prediction, the parameter λ for the LASSO method is assumed to be 2.0×10^2, while this parameter is zero for the LR method. For the SVR method, γ of RBF is 10^{-7}, C is 10^3, and ϵ is 10^{-1}. The DNN method has a total of five layers with neurons where four hidden layers have 240, 120, 60, and 30 activation functions of rectified linear units (ReLU), respectively, and an output layer has a linear activation function.
In the training data, target hours of prediction are set from 06:00 to 20:00 as meaningful daylight periods. The rest of the time intervals are dropped to eliminate the unnecessary night-time periods. This will enhance the prediction accuracy by exposed to a misleading great performance when the prediction of solar insolation is obvious to be zero. The performance of the presented learning algorithms is investigated based on the proposed error metrics given spatio-temporal data from four adjacent stations. For each station, the performance of learning algorithms is compared given the temporal data. Then, the possibility of improvement in the performance of each learning algorithm with the spatio-temporal data from adjacent stations is analyzed. For each station, the goal is to determine the best learning algorithm and find the interdependencies to the insolation of adjacent stations.

A. Station A

The performance of the presented learning algorithms when utilizing the temporal data for station A is presented in Table II, where multiple forecast errors measures are listed. First, in terms of the two linear algorithms, the regularization of the LASSO method demonstrates an improvement for all error measures compared to the LR method. For example, the LASSO method shows 6.4% improvement in both the MAE and the rMAE (nMAE). Since the rMAE (nMAE) is a normalized value of the MAE, it has the same error improvement as the MAE. Second, the prediction errors substantially decreased from linear algorithms to non-linear algorithms. Similar to the DNN method, non-linearity of the SVR method indicates compatible improvement in both the MAE and the RMSE, compared with two linear algorithms. In terms of comparing the two non-linear algorithms, the DNN method has a slightly smaller error for all metrics compared to the SVR method. Thus, given temporal data, the observation is that non-linear algorithms are generally dominant considering all error metrics and the DNN method outperforms other learning algorithms.

TABLE II
Prediction error results are shown when Station A uses temporal data with four different learning algorithms.

Error Measure	LR	LASSO	SVR	DNN
MAE [Wh/m^2]	30.7	28.74	23.62	22.87
rMAE (nMAE) [%]	5.55	5.19	4.27	4.13
RMSE [Wh/m^2]	58.75	55.78	52.73	52.6
rRMSE_mean [%]	10.61	10.08	9.53	9.5
rRMSE_maxmin [%]	5.68	5.39	5.1	5.09

The opportunity for improving the performance of solar insolation prediction while using the temporal data is limited. To further improve the accuracy of solar insolation prediction, learning algorithms can also utilize the spatio-temporal data of the adjacent stations. Such improvements in insolation prediction of station A in terms of the decrease in the MAE and the RMSE are illustrated in Fig. 5. This two metrics are selected because they are consistent in their normalized values presented as the rMAE (nMAE), the rRMSE_mean, and the rRMSE_maxmin. It is interesting to note that regardless of the choice of the learning algorithms, the prediction performance of station A is improved by utilizing spatio-temporal data from adjacent stations. For example, given the LR method, the improvements in the MAE are 0.2%, 3.9%, and 5.4% with adding the spatio-temporal data from stations B, C, and D, respectively. These improvements are 0.6%, 2.8%, and 6.0% in terms of the RMSE by adding the same spatio-temporal data. The LASSO method also shows similar improvements to the LR method. However, the improvements in error metrics when spatio-temporal data is utilized are less significant compared to the LR method.

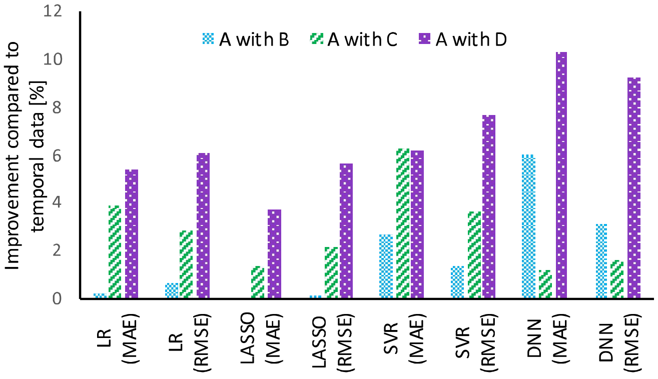

Fig. 5. Based on TABLE II, the error improvement percentage of using spatio-temporal data against temporal data is represented for Station A. The adjacent stations are B, C, and D, and various learning methods are implemented.

The improvement in predicting insolation while utilizing spatio-temporal data is more significant when non-linear learning algorithms are employed. For example, using the SVR method demonstrates a larger improvement compared to LR and LASSO methods. The improvements in the MAE are 2.6%, 6.2%, and 6.2% with the utilization of spatio-temporal data from station B, C, and D, respectively. These improvements are 6.0%, 1.2%, and 10.3% when the DNN method is employed. The improvements in the RMSE are similar.

It is interesting to figure out that the data of which station has the largest contribution on improving the insolation forecast of station A. Given all learning algorithms assessed by all error metrics, utilizing the spatio-temporal data for station D led to the largest improvement in the insolation prediction of station A. Thus, it is fair to argue that a proper combination of adjacent stations and learning algorithms results in a considerable synergy effect for solar insolation prediction performance. The role of utilizing spatio-temporal data for station B is not significant when LR, LASSO, and SVR methods are used. However, the improvement in the performance is notable when the DNN method is implemented. For station C, the largest improvement is achieved by when the SVR method is implemented. With the spatio-temporal data from station C, the SVR method with the rMAE of 4% outperforms the DNN method with the rMAE of 4.08%. Thus, the best learning algorithm might vary depending on the available data set. However, the best performance is achieved by utilizing the spatio-temporal data from station D.

B. Station B

The performance of the presented learning algorithms with utilizing the temporal data for station B are presented in Table III, where the prediction error is significantly improved compared to station A. For example, the rMAE using the LR method is decreased to 2.83% from 5.55% in station A. Here, the linear learning algorithms demonstrate similar performance. Implementing the LASSO method, the MAE is slightly larger than that in the LR, while the RSME is vice versa. The regularization of the LASSO method can successfully suppress large deviations and outliers from actual values, but it also sacrifices its bias by slightly increasing the average magnitude of the prediction error metrics. This bias-variance trade-off is a typical issue of existing predictive models. This not an issue for the non-linear learning algorithms. Compared to the LR method, both the MAE and the RMSE are improved using SVR and DNN methods, where the performance of DNN is better. Since the non-linear nature of SVR and DNN methods can enhance prediction performance, non-linear algorithms are more suitable to deal with temporal data in station B.

TABLE III
PREDICTION ERROR RESULTS ARE SHOWN WHEN STATION B USES TEMPORAL DATA WITH FOUR DIFFERENT LEARNING ALGORITHMS.

Error Measure	LR	LASSO	SVR	DNN
MAE [Wh/m^2]	16.3	16.37	12.77	12.55
rMAE (nMAE) [%]	2.83	2.85	2.22	2.18
RMSE [Wh/m^2]	24.2	23.89	20.32	20.21
rRMSE_mean [%]	4.21	4.15	3.53	3.51
rRMSE_maxmin [%]	2.31	2.28	1.94	1.93

The changes in the selected error metrics (i.e., the MAE and the RMSE) for insolation prediction of station B when utilizing the spatio-temporal data from adjacent stations are presented in Fig. 6. First, given the LR method, both the MAE and the RMSE are worse than the performance of utilizing temporal data only. Specifically, the MAE demonstrates -12.9%, -5.9%, and -8.8% performance degradation when adding spatio-temporal data for station A, C, and D, respectively. This is worse once the RMSE is considered as those degradation are -39.3%, -7.2%, and -13.7%. Implementing the LASSO method also unfolds similar degradation in the performance. Thus, for station B, utilizing the spatio-temporal data from adjacent stations fails to deliver any improvement by implementing the two linear algorithms.

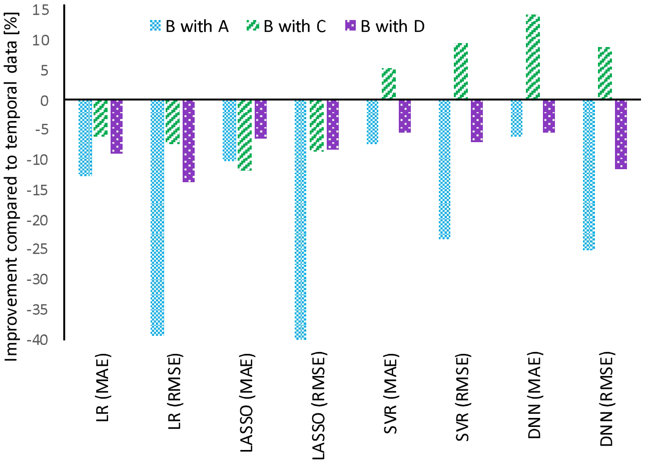

Fig. 6. Based on TABLE III, the error improvement percentage of using spatio-temporal data against temporal data is represented for Station B. The adjacent stations are A, C, and D, and various learning methods are implemented.

Next, the impact of employing non-linear learning algorithms to explore the potential improvements from utilizing the spatio-temporal data is investigated. For the SVR method, the MAE indicates -7.3% and -5.3% degradation in performance once utilizing the data from station A and D, but 5.5% improvement when the data from station C is utilized. Similarly, the RMSE represents -23.2% and -6.8% degradation in performance once utilizing the data from stations A and D, but 9.6% improvement in performance once utilizing the data from station C. Once the DNN method is implemented, the MAE is improved 14.1% once utilizing the data from station C, it indicates -6.2% and -5.3% degradation in performance once utilizing the data from station A and D. Results considering the RMSE are also similar. Unlike the linear learning algorithms, implementing SVR and DNN led to an increase in the performance of insolation prediction for station B once utilizing the data from station C. Hence, in terms of contributions from adjacent stations, station C contributes to improving the prediction error of station B when SVR and DNN methods are utilized. However, both station A and D caused a depreciation in the prediction performance of station B regardless of learning algorithms. Interestingly, by contrast with station A, adding spatio-temporal data does not always lead to performance improvement. The observation is that prediction errors depend on not only linear/non-linear algorithms but also the particular spatio-temporal data sets from adjacent stations. Therefore, it is necessary to determine the appropriate combinations of non-linear algorithms and spatio-temporal information to obtain performance gain in prediction errors. Here, the best insolation prediction is achieved by integrating the spatio-temporal data from station C. The best performing algorithm in terms of the RSME is the SVR method, while it is the DNN method in terms of the MAE.

C. Station C

The performance of learning algorithms based on various error metrics is presented in Table IV, where the temporal data of station C is used for solar insolation prediction. The interesting observation for the two linear models is the better performance of the LR method compared to the LASSO method. Thus, the bias imposed by the regularization in the LASSO method exacerbates the solar insolation prediction. This example demonstrates that although the LASSO method is generally suitable for feature selection and regularization purposes, it is not always the case that considering regularization will enhance the performance. This is because of the incorrect penalizing of substantively essential features. Similar to station A and station B, the performance of non-linear learning algorithms is better than linear ones when utilizing the temporal data.

TABLE IV
PREDICTION ERROR RESULTS ARE SHOWN WHEN STATION C USES TEMPORAL DATA WITH FOUR DIFFERENT LEARNING ALGORITHMS.

Error Measure	LR	LASSO	SVR	DNN
MAE [Wh/m^2]	35.28	39.41	24.5	22.02
rMAE (nMAE) [%]	6.93	7.74	4.81	4.33
RMSE [Wh/m^2]	53.21	55.05	43.93	41.36
rRMSE_mean [%]	10.45	10.81	8.63	8.12
rRMSE_maxmin [%]	5.32	5.5	4.39	4.13

The improvements in the performance of learning algorithms when station C utilizes spatio-temporal data from adjacent stations are illustrated in Fig. 7. Despite station B, the overall solar insolation forecast errors of station C are generally improved with the utilization of the spatio-temporal data from adjacent stations. Nevertheless, the one exception is when the LR method is implemented, where spatio-temporal data from station A leads to -0.27% degradation in terms of the RMSE, while according to the MAE it led to an improvement in insolation prediction. This is because of the difference like error metrics. While the spatio-temporal data can reduce the average magnitude of the errors in the prediction, it may also add a few outliers which had an adverse impact on the increase in the RMSE. For the insolation prediction of Station C, the spatio-temporal data from station A is an interesting example. This illustrates the benefit of regularization considering the bias-variance trade-off, where the LASSO method presents a larger improvement compared to the LR method. Moreover, the most significant improvement is a 13.5% improvement in the MAE when the LASSO method is employed with spatio-temporal data of station D. However, the improvement brings down the MAE to 34.87 which is still larger than 32.06 of that when the LR method is employed. Therefore, utilizing the spatio-temporal data mitigated the adverse impact of the bias imposed by the LASSO method compared to the LR method.

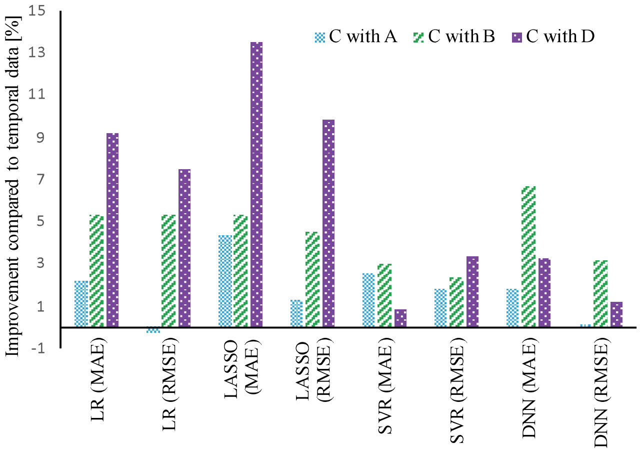

Fig. 7. Based on TABLE IV, the error improvement percentage of using spatio-temporal data against temporal data is represented for Station C. The adjacent stations are A, B, and D, and various learning methods are implemented.

Utilizing the spatio-temporal data led to an improvement in the performance of SVR and DNN methods. However, the contributions of the spatio-temporal data highly depend on learning algorithms, the adjacent station, and the selected error metrics. Consistent with linear models, utilizing the spatio-temporal data from station A has the smallest improvement compared to other adjacent stations. However, while the spatio-temporal data of station B has the largest impact on the improvement when the DNN method is implemented, the improvements in the SVR method highly depend on the choice of the error metric. If the MAE is taken into consideration, utilizing the spatio-temporal data from station B contributes the most to improve the insolation prediction, and the contribution of data from station A is more than station D. However, the most improvement is achieved by utilizing the spatio-temporal data from station D when the RMSE is considered. Therefore, it is not straightforward to call that the data from which station is contributing the most to improve the insolation prediction when the SVR method is implemented. However, the best performance is captured when spatio-temporal data from station B is utilized for insolation prediction in station C with the DNN method. Contrasting this with the best performance of station B reveals that sharing the spatio-temporal data between these stations is very useful for improving the insolation prediction.

D. Station D

Various metrics for prediction error of given learning algorithms with temporal data of station D is illustrated in Table V. The error metrics for LR and LASSO methods are very close. For example, with employing the LASSO method, there is a

TABLE V
PREDICTION ERROR RESULTS ARE SHOWN WHEN STATION D USES TEMPORAL DATA WITH FOUR DIFFERENT LEARNING ALGORITHMS.

Error Measure	LR	LASSO	SVR	DNN
MAE [Wh/m^2]	30.8	31.37	20.02	19.71
rMAE (nMAE) [%]	5.42	5.52	3.53	3.47
RMSE [Wh/m^2]	48.28	47.9	42.4	41.34
rRMSE_mean [%]	8.5	8.43	7.46	7.28
rRMSE_maxmin [%]	4.5	4.46	3.95	3.85

slight error increase of 0.57 in terms of the MAE and the error decrease of 0.38 in terms of the RMSE. This is because of the difference in LASSO and LR methods given the regularization in the LASSO method, particular features on temporal data are penalized to eliminates outliers or large deviations from actual values, where this leads to the increase in the overall magnitude of the errors. To deal with bias-variance trade-off, non-linear models are also examined here. The error metrics presents smaller prediction error when implementing SVR and DNN methods compared to the implementation of LR and LASSO methods. Those error metrics are also very close. However, employing the DNN method presents a slightly smaller error.

The improvements in insolation prediction of station D utilizing the spatio-temporal data from the adjacent station are described in Fig. 8. Here, the spatio-temporal data of station C has the largest contribution to improve the performance of insolation prediction given all algorithm and all error metrics. It is observed that the spatio-temporal data from station A is leading to a decreased in the performance of the LR method as well as the DNN and SVR methods considering the RMSE. It is not decreased when the LASSO method is employed due to the tendency of the LASSO method to remove outlier that will contribute to a larger value of the RMSE. It is very interesting to note that the spatio-temporal data from station D was the largest contributor to improve the performance of station A, while data from station A is not helpful for the insolation forecast of station D. The improvement in insolation prediction is more significant when implementing SVR and DNN methods compared to implementing LR and LASSO methods. For example, implementing the SVR method, the MAE indicates prediction improvements of 1.0%, 2.0% and 6.1% when the spatio-temporal data from stations A, B, and C are utilized, respectively. Similarly, when the DNN method is employed, the MAE indicates prediction improvements of 1.8%, 0.7% and 6.0% when the spatio-temporal data from stations A, B, and C are utilized, respectively. Thus, the best performance is achieved once the DNN method is implemented and spatio-temporal data from station C is utilized. The role of station C points out that combinations of specific information and particular algorithms can produce a significant effect by exploring potential improvements.

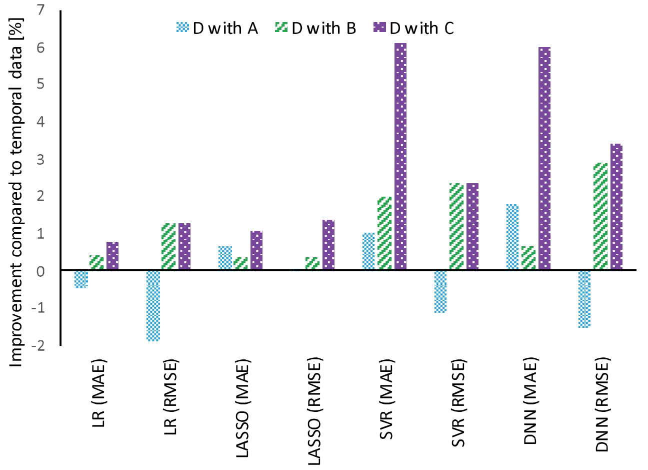

Fig. 8. Based on TABLE V, the error improvement percentage of using spatio-temporal data against temporal data is represented for Station D. The adjacent stations are A, B, and C, and various learning methods are implemented.

E. Summary

The summary of the best-performing algorithms given temporal and spatio-temporal data as well as the most important adjacent station and the algorithm with the most improvement in performance with the utilization if spatio-temporal data is given in Table VI. Although the implementation of the DNN method would generally outperform other learning algorithms, there are interesting exceptions to explore. For example, for station B, the best choice of learning algorithms utilizing the spatio-temporal data could be the DNN method or the SVR method depending on the choice of error metric. In addition, the contribution of the spatio-temporal data in improving the performance of the learning algorithm is different. Therefore, when implementing a specific algorithm, the largest improvement gain does not necessarily mean the best performance among learning algorithms. Also, the superiority of one algorithm by utilizing the spatio-temporal data of one station does not necessarily mean that it would be the best algorithm for all stations.

TABLE VI
SUMMARY OF THE BEST COMBINATION OF LEARNING METHODS AND IMPORTANT ADJACENT STATIONS FOR EACH STATION, RESPECTIVELY.

Combination	A	B	C	D
The best method with temporal data only	DNN	DNN	DNN	DNN
The largest positive gain of method and specific station	DNN & D	DNN & C	LASSO & D	SVR & C
The best combination of method and important station	DNN & D	DNN & C	DNN & B	DNN & C

To increase the prediction performance further, one may choose the ensemble (hybridization) methods that combine multiple learning methods to generate better predictive performance than the results obtained from any of learning algorithms alone [15], [16]. Nevertheless, if the spatial information, per se, is not beneficial to a specific weather station, the ensemble methods would not be effective. This happens when station B utilizes station A and station D which led to worse prediction errors. Finally, one may reduce the prediction errors by incorporating learning algorithms into the solar insolation stochastic process, e.g., the probability distribution of solar power can be modeled by the Gaussian distribution and beta distribution [26]–[28]. While a typical distribution can be constructed by collecting the realized samples over some period of time, either minutes, hours, days or weeks, a conditional distribution can be modeled by predicted value of solar insolation, calendar date, and ambient conditions [28]. Thus, combining distribution sets and learning algorithms potentially enables the proposed prediction techniques to be more accurate, which in turns can predict unbiased solar insolation values.

VI. Conclusions

The impact of various learning algorithms for solar insolation prediction is assessed by utilizing the available data in spatial proximity. Temporal and spatio-temporal information is utilized for linear and non-linear learning algorithms for solar insolation prediction. Multiple error metrics highlighted that the improvement achieved by utilizing the spatio-temporal data depends on the specific location and the employed algorithms. The best way to describe the accuracy of a solar prediction depends on the combination of spatio-temporal data and learning algorithms. The largest positive gain of a certain algorithm for spatio-temporal of the target station did not necessarily mean that it would be the best algorithm for other stations. Therefore, for each location, a proper combination of non-linear algorithms and spatio-temporal information should be obtained to ensure the best performance improvement in prediction errors by utilizing spatial proximity data. By doing so, it is possible to determine a specific set of spatial proximity data to enhance the solar insolation prediction. Since our proposed methods are scalable for different time resolution (sampling rate), the future research is to further improve the prediction performance by using multiple data sets such as high temporal and spatial resolution information and ambient conditions (air temperature, wind, and gust speed).

References

[1] W. Glassley, J. Kleissl, C. C. van Dam, H. Shiu, and J. Huang, "California Renewable Energy Forecasting, Resource Data, and Mapping." California Energy Commission, Tech. Rep., 2011.

[2] F. Mancilla-David, F. Riganti-Fulginei, A. Laudani, and A. Salvini, "A neural network-based low-cost solar irradiance sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 3, pp. 583–591, March 2014.

[3] Yen-Wei Chen, N. E. Mendoza, Z. Nakao, and T. Adachi, "Estimating wind speed in the lower atmosphere wind profiler based on a genetic algorithm," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 4, pp. 593–597, Aug 2002.

[4] B. Denby, J. . Prevotet, P. Garda, B. Granado, L. Barthes, P. Gole, J. Lavergnat, and J. . Delahaye, "Combining signal processing and machine learning techniques for real time measurement of raindrops," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 6, pp. 1717–1724, Dec 2001.

[5] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, and P. Kalb, "3D cloud detection and tracking system for solar forecast using multiple sky imagers," *Solar Energy*, vol. 118, pp. 496–519, 2015.

[6] H. Yang, B. Kurtz, D. Nguyen, B. Urquhart, C. W. Chow, M. Ghonima, and J. Kleissl, "Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego," *Solar Energy*, vol. 103, pp. 502–524, 2014.

[7] 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 Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.

[8] L. M. Aguiar, B. Pereira, P. Lauret, F. Díaz, and M. David, "Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting," *Renewable Energy*, vol. 97, pp. 599–610, 2016.

[9] A. Hammer, D. Heinemann, E. Lorenz, and B. Lückehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, vol. 67, no. 1-3, pp. 139–150, 1999.

[10] E. Scolari, L. Reyes-Chamorro, F. Sossan, and M. Paolone, "A Comprehensive Assessment of the Short-Term Uncertainty of Grid-Connected PV Systems," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 3, pp. 1458–1467, 2018.

[11] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 1, pp. 300–308, 2018.

[12] M. J. Sanjari and H. B. Gooi, "Probabilistic Forecast of PV Power Generation Based on Higher Order Markov Chain," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2942–2952, 2017.

[13] S. Bhardwaj, V. Sharma, S. Srivastava, O. S. Sastry, B. Bandyopadhyay, S. S. Chandel, and J. R. Gupta, "Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model," *Solar Energy*, vol. 93, pp. 43–54, 2013.

[14] K. Y. Bae, H. S. Jang, and D. K. Sung, "Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 935–945, 2017.

[15] X. Zhang, Y. Li, S. Lu, H. Hamann, B. M. S. Hodge, and B. Lehman, "A Solar Time-based Analog Ensemble Method for Regional Solar Power Forecasting," *IEEE Transactions on Sustainable Energy*, vol. 3029, no. c, 2018.

[16] L. Gigoni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-Ahead Hourly Forecasting of Power Generation from Photovoltaic Plants," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 831–842, 2018.

[17] M. Khodayar, S. Mohammadi, M. Khodayar, J. Wang, and G. Liu, "Convolutional Graph Auto-encoder: A Deep Generative Neural Architecture for Probabilistic Spatio-temporal Solar Irradiance Forecasting." *arXiv Machine Learning*, pp. 1–8, 2018. [Online]. Available: http://arxiv.org/abs/1809.03538

[18] M. Lave, J. Kleissl, and J. S. Stein, "A wavelet-based variability model (WVM) for solar PV power plants," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 2, pp. 501–509, 2013.

[19] X. G. Agoua, R. Girard, and G. Kariniotakis, "Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting," *IEEE Transactions on Sustainable Energy*, pp. 1–9, 2018.

[20] C. Schuss, B. Eichberger, and T. Rahkonen, "Impact of sampling interval on the accuracy of estimating the amount of solar energy," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, May 2016, pp. 1–6.

[21] M. D. Tabone and D. S. Callaway, "Modeling Variability and Uncertainty of Photovoltaic Generation: A Hidden State Spatial Statistical Approach," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 2965–2973, 2015.

[22] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 538–546, 2018.

[23] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.

[24] Mesowest, *Data Download Interface Help Page*. [Online]. Available: https://mesowest.utah.edu/html/help/download.html

[25] C. Schuss, T. Fabritius, B. Eichberger, and T. Rahkonen, "Moving photovoltaic installations: Impacts of the sampling rate on maximum power point tracking algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 5, pp. 1485–1493, May 2019.

[26] Bei Zhang, P. Dehghanian, and M. Kezunovic, "Spatial-temporal solar power forecast through use of gaussian conditional random fields," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, July 2016, pp. 1–5.

[27] F. Y. Ettoumi, A. Mefti, A. Adane, and M. Bouroubi, "Statistical analysis of solar measurements in algeria using beta distributions," *Renewable Energy*, vol. 26, no. 1, pp. 47 – 67, 2002.

[28] S. Ryu, S. Bae, J. Lee, and H. Kim, "Gaussian residual bidding based coalition for two-settlement renewable energy market," *IEEE Access*, vol. 6, pp. 43 029–43 038, 2018.