Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Solar irradiance forecasting models without on-site training measurements

Andres Felipe Zambrano [*], Luis Felipe Giraldo

Department of Electrical and Electronic Engineering, Universidad de los Andes, Carrera 1 Este No. 19A-40, Edificio Mario Laserna, Piso 7, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 25 October 2019
Received in revised form
13 January 2020
Accepted 21 January 2020
Available online 25 January 2020

Keywords:
Solar irradiance forecasting
Satellite measurements
Metric learning
PV power systems
Machine learning

ABSTRACT

Much effort has been made to increase the integration of solar photovoltaic (PV) systems to reduce the environmental impacts of fossil fuels. An essential process in PV systems is the forecasting of solar irradiance to avoid safety and stability problems due to its intermittent nature. Most of the research has been focused on improving the prediction accuracy based on the assumption that enough on-site training data are available. However, in many situations, it is required for the implementation of PV systems in locations where not enough solar irradiance measurements have been collected. Our hypothesis is that measurements from other sites can be used to train accurate forecasting models, given an appropriate definition of site similarity. We propose a methodology that takes information from exogenous variables that are correlated to on-site solar irradiance and constructs a multidimensional space equipped with a metric. Each site is a point in this space, and the learned metric is used to select those sites that can provide measurements to train an accurate forecasting model on an unobserved site. We show through experiments with real data that using the learned metric provides better predictions than using the measurements collected from the whole set of available sites.

1. Introduction

The adverse effects of power generation with fossil fuels such as global warming, air pollution, and health issues have made nonconventional renewable energy sources (NRES) an attractive alternative that not only uses renewable resources for the generation but also has minimal environmental and social impacts [1–3]. Despite the advantages of NRES, their incorporation into power systems comes with several challenges due to the intermittency in the source availability caused by its high dependence on weather conditions [1]. This could cause a mismatch between the power supply and energy demand, generating issues on stability, safety, reliability and frequency response of the grid [4,5]. To solve these issues related to the intermittency of NRES, most of the solutions involve the implementation of storage [6,7] and demand response [6,8,9] systems.Both of these systems need to forecast the source availability to have appropriate management of storage energy and price changes [6,10,11]. In this context, there are several prediction time horizons to be considered depending on the task to be conducted. For example, very short-term horizons (less than 30 min) are used in electricity market clearing, short (30m-6h) and medium-term (6h-2d) horizons are useful in economic load dispatching and operational security of the system, and long term (more than 1 day) horizons are used in maintenance scheduling [12].

In the case of photovoltaic (PV) power systems, which is one of the most studied NRES, a great deal of research has been devoted to proposing methods to forecast solar irradiance to do short- and medium-term horizon analysis, which is important when addressing most of the grid issues [10,11]. For instance, Autoregressive Integrated Moving Average (ARIMA) and ARIMA with Exogenous Variables are time series-based statistical models that are typically used in PV power systems [13]. Also, linear and nonlinear data-driven approaches such as support vector regression, random forests, k-nearest neighbors, and artificial neural networks have been largely studied [11,13,14]. A key aspect that defines the performance of these methods for solar irradiance forecasting is the amount of on-site historical data to train the prediction model. Large amounts of solar irradiance measurements on the target site are required to capture the radiation patterns and therefore to provide precise predictions [13]. This represents an issue for those places where on-site measurements are limited or

[*] Corresponding author.
E-mail addresses: af.zambrano10@uniandes.edu.co (A.F. Zambrano), lf.giraldo404@uniandes.edu.co (L.F. Giraldo).

https://doi.org/10.1016/j.renene.2020.01.092

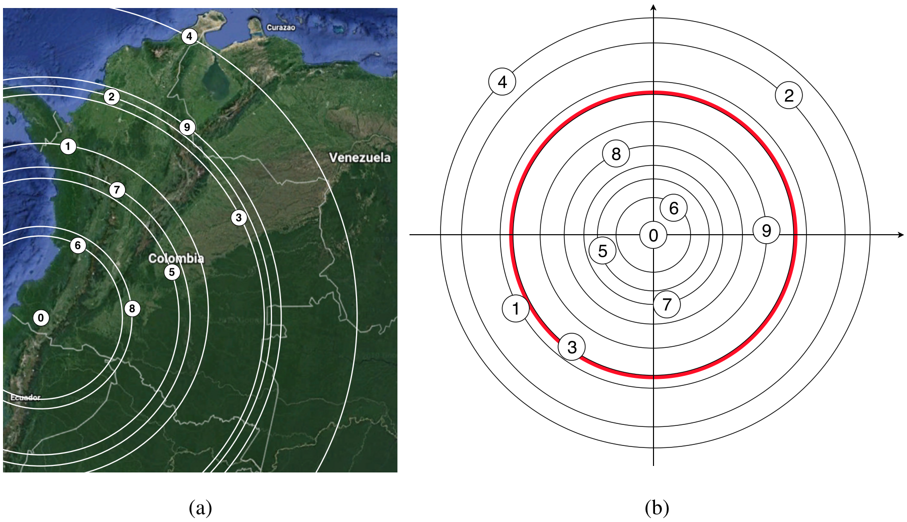


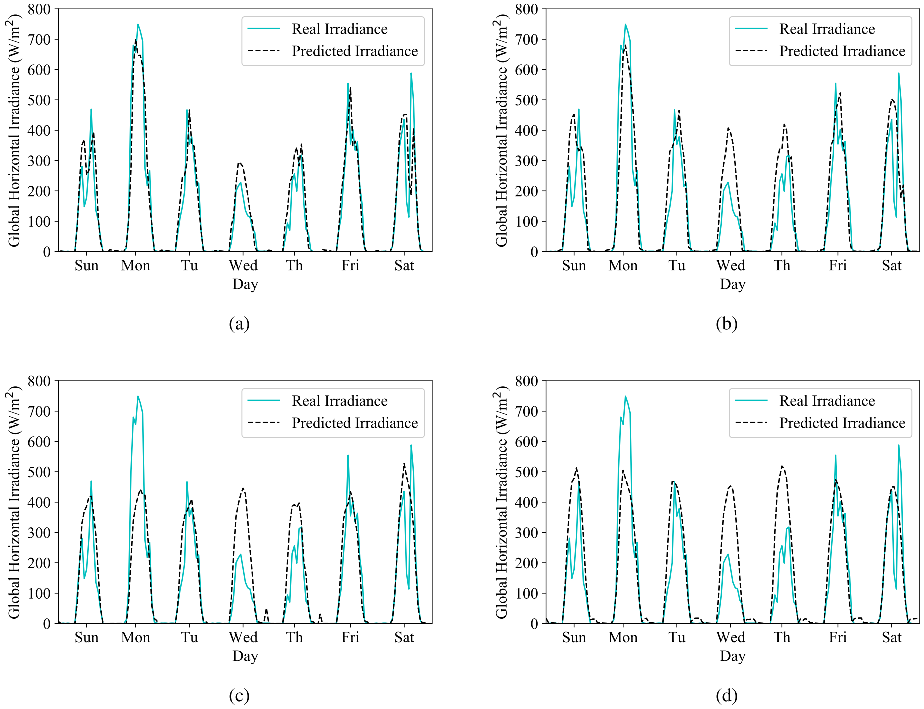
Fig. 1. (a) Geographical location of 10 sites, where sites 1 to 9 have solar irradiance measurements, and site 0 does not. (b) Sites located in a metric space where sites closer to site 0 are the ones that could have more similar radiation patterns. Note that the geographic location does not necessarily coincide with the location in the new metric space. Moreover, the methodology indicates the amount of nearest sites that have to be used for training the prediction model on the target site. In this example, for site 0, the 6 nearest sites have to be considered on the training set, which are 6, 5, 7, 8, 9 and 3.

not available. In these cases, an accurate solar radiation forecasting is not possible until enough on-site measurements are taken, implying months or even years of implementation delay. Most of the work has been focused on finding strategies that improve the model prediction accuracy based on the assumption that enough on-site training data are available [13]. However, little has been done to study the problem of finding forecasting models when there are not on-site irradiance measurements available for training.

Some work has studied this problem. For example, the work in Ref. [15] uses several interpolation methods to conduct very short-term solar irradiance forecasting at unobserved locations placed less than 1 km away from any station from a dense 1 km × 1.2 km network of 17 stations. Also, statistical models have shown to produce accurate forecast results at unobserved sites when they are close enough to an observed location [16]. Interpolation and statistical-model based approaches are useful to forecast solar irradiance at unobserved locations when geospatial closeness to observed sites is guaranteed. However, we can find many scenarios where a site where the nearest available measurements are more than 10 km away. To our knowledge, this issue has not been studied yet. There is a need to develop methodologies to obtain accurate forecasting models when there are not on-site irradiance measurements available for training and when there are not spatially close sites that can provide measurements. We do this in this paper.

We hypothesize that solar irradiance measurements from other sites, located more than 10 km away from the target place, can be used to train accurate forecasting models, given an appropriate definition of site similarity. This implies the construction of a multidimensional space equipped with a metric, where each site corresponds to a point in this space, and the metric allows for determining which of the sites can provide measurements to train a model to forecast irradiance on a site that does not have training data. In this paper, we propose a methodology that takes information from exogenous variables that are correlated to on-site solar irradiance and constructs that space where a metric is learned. This metric is used as a tool to select those sites that can provide solar irradiation measurements to train an accurate forecasting model on a previously unknown site without training data. These exogenous variables include elevation, latitude/longitude, clear sky models, and satellite measurements from the target site (where the PV system will be implemented and radiation have not been measured) and from other sites where radiation measurements are available. Fig. 1 illustrates an example of applying the proposed methodology to a specific site to select which set of measurements has to be used for training of the prediction model. Fig. 1a shows the real geographical location of 10 sites, where sites 1 to 9 have solar irradiance measurements, and site 0 does not. Fig. 1b shows each site in a metric space where sites closer to site 0 are the ones that could provide data to train an accurate forecasting model. Note that the geographic location does not necessarily coincide with the location in the new metric space. The methodology that we propose indicates the amount of nearest sites that have to be used for training a forecasting model on the target site. In this example, for site 0, the 6 nearest sites have to be considered on the training set, which is 6, 5, 7, 8, 9 and 3.

This paper is structured as follows. First, artificial neural networks are trained to forecast solar irradiance with a horizon changing from 1 to 48 h using full on-site information (Section 2). The input vector is constructed using on-site global horizontal irradiance (GHI), satellite, and clear sky model (CSM) [17] measurements following a 2-D time-series configuration, with date and hour as coordinate axes. Using this model, we quantify the feature importance for GHI forecasting. We show that the CSM variables are some of the most predictive features, particularly for long-term forecasting horizons. The goal here is to have a reference model before we introduce the methodology for finding a forecasting model without on-site training data. Then, we present different approaches to construct the metric space used to determine those sites that can provide irradiation measurements to train models for prediction in sites without historical data (Section 3). These approaches are based on principal component analysis [18], learning to rank [19], and kernel machines [20]. The goal of these approaches is to learn a Weighted Mahalanobis distance [21] and several nearest sites that determine which of the sites can provide training data for an accurate prediction on a previously unknown target site. Through experiments on real data, we show that the proposed methodology improves the prediction accuracy compared to a strategy that uses all the available data or that uses the geographical distance as a metric. We finish this paper with a



(a)

(b)

(c)

(d)

Fig. 2. Global Horizontal Irradiance (GHI) prediction for different time horizons: a) 1 h, b) 2 h, c) 6 h, d) 48 h.

discussion about the usefulness and the reach of the proposed methodology to the challenge of solar forecasting in PV power systems (Section 4).

2. Reference irradiance forecasting model trained with full-information

In this section, we train a prediction model using all the available information at different locations to compute performance indicators and to calculate the feature importance for prediction. This provides a reference to evaluate the proposed methodology for forecasting irradiance without on-site training data.

2.1. Predictor variables

Global horizontal irradiance (GHI) is the variable to be predicted. Studies on solar irradiance forecasting typically use, in addition to the past samples of GHI, temperature and humidity time series as predictor variables [11]. However, it is known that atmospheric dynamics and cloudiness can considerably affect on-site solar irradiance. Therefore, we include satellite measurements such as surface pressure, wind-speed, ozone column and precipitation as additional predictor variables of the model. Also, hour, date, latitude, longitude, and elevation of each location where the measurements are collected are used as features of the ANN.

Clear Sky Model. Clear sky models (CSMs) are used to estimate the GHI under a sky with no clouds as a function of site altitude, solar elevation angle, aerosol concentration, water vapor, air turbidity and other variables related to solar geometry and atmospheric conditions [11,17]. Even though CSMs have been used to model the general maximum output of a solar photovoltaic system, as far as we know, it has never been used as a feature of a data-driven solar irradiance prediction model along with satellite and

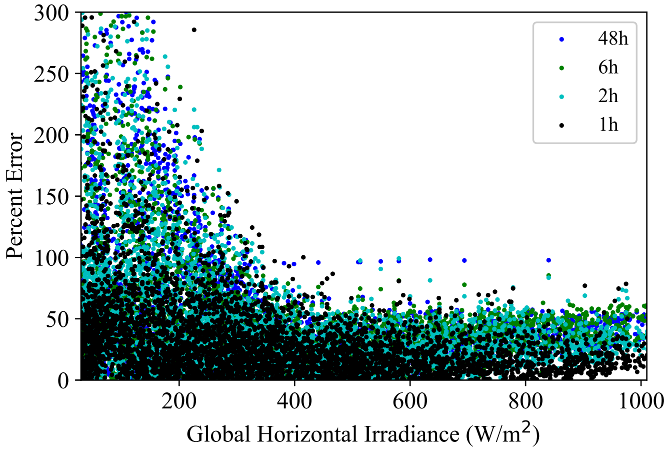


Fig. 3. Percent error of the prediction, calculated as $\left|y - y_{pred}\right|/y$, versus the GHI to be predicted for horizons of 1, 2, 6 and 48 h.

Table 1
Performance indicators for different forecasting horizons.

Horizon (h)	1	2	3	4	5	6	12	24	48
MAE (*W*/ *m*^{2})	44.63	55.94	58.90	61.35	61.50	61.58	63.41	63.13	63.83
MBE (*W*/ *m*^{2})	1.18	−1.07	0.94	2.84	−0.24	1.52	4.42	−2.33	4.72
RMSE (*W*/ *m*^{2})	89.16	107.35	112.79	115.84	117.22	117.26	119.07	119.39	120.88
nRMSE (%)	47.94	57.72	60.64	62.28	63.02	63.04	64.02	64.19	65.02
R^2	0.8921	0.8436	0.8274	0.8179	0.8136	0.8134	0.8076	0.8166	0.8017

Table 2
Feature Importance calculated using Equation (6). Features at time (t-k) represent the last measurement of that variable for a forecasting horizon of k hours. Features at time (t-24n) are variables measured 24, 48 or 72 h before the moment of prediction.

Feature	*FI* (%)
Hour	13.52
Latitude	7.50
Clear Sky Model (t)	6.36
Temperature (t-k)	6.07
Elevation	6.00
Radiation (t-k)	5.90
Surface Pressure (t-k)	5.36
Surface Pressure (t-24n)	4.88
Longitude	4.31
Radiation (t-24n)	3.77
Eastward Wind (t-k)	3.38
Temperature (t-24n)	3.28
Specific Humidity (t-k)	2.94
Month	2.65
Northward Wind (t-k)	2.46
Precipitable Water Vapor (t-k)	2.44
Eastward Wind (t-24n)	2.18
Specific Humidity (t-24n)	2.17
Ozone Column (t-24n)	1.92
Ozone Column (t-k)	1.88
Precipitable Water Vapor (t-24n)	1.88
Year	1.86
Precipitable Liquid Water (t-k)	1.49
Northward Wind (t-24n)	1.47
Precipitable Liquid Water (t-k)	1.32
Precipitable Ice Water (t-k)	1.16
Precipitable Ice Water (t-24n)	1.10
Day	0.72

on-site measurements. In this work we incorporate the estimation of GHI provided by the CSM in Ref. [22] into the set of variables used to predict solar irradiance.

2-D Time Series Analysis. Weather-related variables, especially solar radiation, have seasonal characteristics. For example, radiation at 10 a.m. could be estimated by its measurements at 8 and 9 a.m. and also by the measurement at the same hour (10 a.m.) on the previous days. For this reason, the work in Ref. [23] proposed a time series 2-D configuration, being hour and day the two-axis where the time series evolve. If the prediction model uses measurements from the last hours and the previous days at the same hour of the desired prediction, the predictor model can learn behaviors related to the seasonal characteristics of the time series. In our work, we use the 2-D time series configuration of the predictor variables for solar irradiance prediction.

2.2. Performance indicators

To measure the performance of a predictor model, the following five indicators, described in Refs. [10,12], are used. Mean Absolute Error (*MAE*), measures the differences between the real and predicted values. Mean Bias Error (*MBE*) calculates the average bias in the prediction. Root Mean Square Error (*RMS E*) and its normalized version (*nRMS E*) quantify the squared error in the prediction, giving more weight to bigger prediction mistakes. Finally, the coefficient of determination (R^2) estimates the variance of the solar radiation that can be predicted with the model. Equations (1)–(5) show the expressions for this indicators, with y and y_{pred} being the real and predicted measurements:

$$MAE = \frac{1}{N}\sum \left| y - y_{pred} \right| \tag{1}$$

Table 3
Performance indicators for forecasting models using a cross-validation analysis for a horizon of 1 h. The row associated with station *i* contains information from the performance of a model trained with data from all the other stations and tested with data from station *i*.

Station	*MAE* (W/m^2)	*MBE* (W/m^2)	*RMSE* (W/m^2)	*nRMSE* (%)	R^2
0	62.07	30.52	109.65	69.94	0.7777
1	48.71	16.13	93.31	52.86	0.8803
2	43.45	9.14	83.89	40.93	0.9077
3	60.83	−8.61	97.39	47.52	0.8791
4	47.28	27.54	81.54	34.99	0.9394
5	41.01	17.06	86.34	84.89	0.7837
6	60.61	−29.80	114.22	60.72	0.8266
7	55.95	13.45	109.75	54.30	0.8531
8	93.82	−30.11	111.15	67.65	0.8139
9	61.10	−15.86	114.44	56.49	0.8576

$$MBE = \frac{1}{N}\sum y - y_{pred} \tag{2}$$

$$RMS\ E = \sqrt{\frac{1}{N}\sum \left(y - y_{pred}\right)^2} \tag{3}$$

$$nRMS\ E = \frac{\sqrt{\frac{1}{N}\sum \left(y - y_{pred}\right)^2}}{y_{mean}} \tag{4}$$

$$R^2 = 1 - \sum \frac{\left(y - y_{pred}\right)^2}{\left(y - y_{mean}\right)^2} \tag{5}$$

2.3. Experimental results

We choose an artificial neural network (ANN) as a prediction model since it has been widely used for nonlinear approximation [24] and specifically for solar radiation forecasting [10,12]. The dataset used here to evaluate the solar radiation forecasting model comes from two free-access databases with on-site and satellite measurements. On-site measurements of Global Horizontal Irradiance (GHI), temperature, and relative humidity were provided by the Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). These measurements are time series with the same acquisition times coming from 10 weather stations at locations with different geographical and weather conditions in Colombia (see Fig. 1a). Satellite measurements are obtained from NASA's modern-era retrospective analysis for research and applications MERRA-2 [25]. The features extracted from this database are hourly wind speed, precipitation, atmospheric pressure, and ozone column centered at the location of the weather stations with a spatial resolution of $0.625^{o} \times 0.5^{o}$. To consider as much as possible seasonal effects of each variable on the forecasting, we used more than 10 years of measurements (2005–2017) on all the results presented for both approaches, with and without on-site measurements. Usually, a distinction between diffuse and direct irradiance should be made. However, in the Colombian case, IDEAM only provides GHI measurements. As we consider this parameter important enough, GHI is the only variable to be predicted in this study.

2.3.1. Forecasting results

The prediction of solar radiation was conducted for seven forecasting horizons ranging from 1 to 48 h using 90% of the dataset

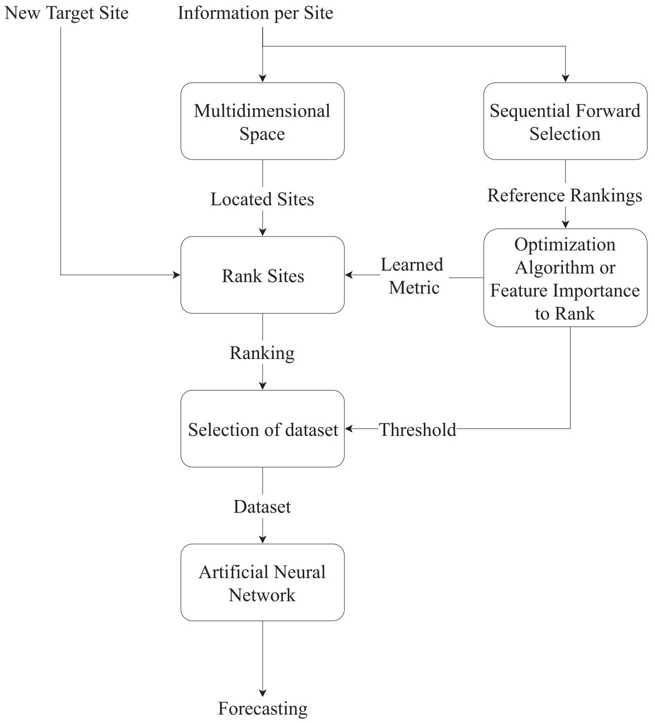


Fig. 4. Methodology for training forecasting models without on-site measurements with a metric based on reference rankings for each weather station.

at each location to train the ANN, and using the remaining 10% to validate each model. Fig. 2 illustrates the performance of the model showing the real and predicted radiation time series for a randomly selected week of the validation set. With a prediction horizon of 1 h, the model knows enough information to accurately predict the atmospheric dynamics and cloudiness (Fig. 2a). However, as it is expected, when the forecasting horizon increases, the prediction accuracy decreases. Note that for long time horizons, the predicted irradiance tends to follow the baseline pattern given by the clear sky model estimation.

Table 1 presents the performance indicators for different time horizons. In general, these results are comparable with the ones presented in current literature for solar irradiation forecasting [11], especially for horizons lower than 4 h. However, the behavior of our model for longer time horizons tends to be better than the ones previously reported in current literature [11,26]. Observe that each performance indicator in Table 1 tends to keep an approximately constant value as the time horizon increases, as opposed to results previously reported in other works that show an increasing prediction error when the time horizon increases. The reason for this situation is that the forecasting model can capture the patterns given by the CSM estimations. It means that the CSM estimations provide a baseline pattern that is useful for medium and long term predictions, ensuring a lower bound for the error in the predicted irradiance. This is evident in Fig. 2d, where the predicted irradiance is similar to the characteristic curve provided by the CSM. These results show the importance of considering the CSM estimation as a predictor variable of the ANN. Some small non-zero irradiance values can be seen in Fig. 2c and d in days with low irradiance measurements, as the result of increasing the prediction horizon. To understand this situation (days with low irradiance), Fig. 3 shows the percent error of the prediction, calculated as $\left|y - y_{pred}\right|/y$, versus the GHI to be predicted, to see a possible correlation between them. This figure shows that for the lowest values of GHI the percent error of the prediction tends to be higher. Therefore, GHI at cloudy days, or specific moments when GHI decreases sharply, is more difficult to predict especially for higher time horizons.

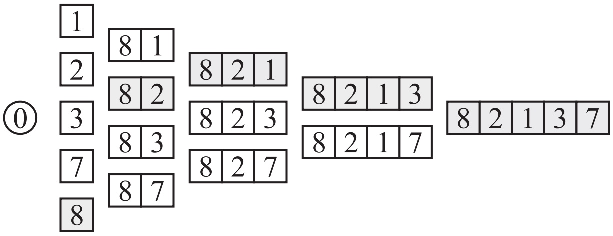


Fig. 5. Part of sequence forward selection process to find the ideal ranking for site 0. Site 8 is the one that provides data to have the most accurate predictions on site 0. Then, site 2 is the one that provides data that combined with the data from site 8 has more accurate predictions on site 0. Following this sequential process, the partial ideal site ranking for prediction in site 0 corresponds to 8, 2, 1, 3, 7.

2.3.2. Feature importance

We compute the feature importance (*FI*) to identify the most important variables for solar irradiance. Using the trained ANN and the method proposed in Ref. [27], the *FI* of feature *i* is computed as the sum of the weights of its connections with the first layer of the ANN over the total weights of that layer:

$$FI_i = \frac{\sum_{j=1}^{N_{neurons}} \left|w_{ij}\right|}{\sum_{k=1}^{N_{features}} \sum_{j=1}^{N_{neurons}} \left|w_{kj}\right|} \tag{6}$$

The idea behind this equation is to use the weights assigned by the first layer of the trained ANN to each feature as an indicator of its importance in the prediction task. Table 2 shows that hour and latitude are some of the most important features, meaning that the specific moment of prediction and the distance with the equator are important for irradiance forecasting. Also, as expected, the CSM estimations have high importance to predict solar irradiance. Moreover, some features measured 24 h before the prediction also appear in the top 10 of the most important variables, justifying the importance of considering the seasonal behavior of time series to predict solar irradiance using ANN.

3. Training forecasting models without on-site measurements

The methodology previously described needs years of measurements to train an accurate prediction model. This requirement could imply excessive delays in the start-up of a power supply plant. To avoid these delays, it is necessary to find a methodology that allows for an accurate prediction without on-site historical measurements before they become available. The first methodology considered to solve this problem is training the ANN with all available measurements obtained at other stations. To study this approach, we implemented a cross-validation analysis for each station with a forecasting horizon of 1 h. It means that the model for station *i* is trained with data from all the other stations and tested with data from station *i*, without using measurements of station *i* on the training, to simulate that scenario where a new site without measurements is the target of the prediction model. Different time horizons could be tested, but this represents a big combinatorial problem. To study several methodologies to predict irradiance on a place where data is not available, we focus this analysis on a prediction horizon of 1 h for all methodologies. Table 3 shows that, using this methodology, only 2 of the 10 stations

Table 4
Reference ranking and performance indicators when applying SFS for site 0.

Ranking SFS	*MAE* (W/m^2)	*MBE* (W/m^2)	*RMSE* (W/m^2)	*nRMSE* (%)	R^2
8	77.59	−7.24	117.17	74.75	0.7461
2	67.78	−42.60	114.59	73.10	0.7572
1	80.95	−66.90	141.52	90.27	0.6296
3	61.75	−7.63	100.30	63.98	0.8139
7	68.25	29.27	107.25	68.42	0.7873
6	62.27	30.83	108.23	69.04	0.7833
9	**57.27**	**20.31**	**103.48**	**66.01**	**0.8020**
4	58.86	28.43	108.07	68.94	0.7840
5	60.33	28.47	108.23	69.04	0.7834

Table 5
Amount of sites selected to obtain the best prediction result for each station.

Station	Amount of selected stations
0	7
1	8
2	8
3	8
4	9
5	5
7	5
8	6
9	4
Mean	6.0

exceeded 0.892 for the R^2 indicator, which is the reference indicator obtained in Table 1. These results suggest that selecting sites that provide measurements to train the forecasting model could potentially provide more accurate forecasting models to be used on the target location. We hypothesize that defining a multidimensional space equipped with a metric will allows us to choose the appropriate measurements for training. Following, we will present several methodologies to select those sites that can potentially provide measurements for training an accurate forecasting model that can be used to predict solar irradiance on a particular site without training measurements.

Multidimensional space. We construct a space where each site corresponds to a point. Let x_i be the vector associated with site i in this space where each entry corresponds to the average of the historical data measured from the following variables: temperature and humidity; satellite measurements available on free access databases that are surface pressure, ozone column, wind speed and precipitation; and location-specific features that are latitude, longitude, elevation, and CSM average irradiance estimation. These variables are known to be correlated to solar irradiance and can provide useful information related to a site's radiation patterns.

Metric. A metric is a function that defines a distance between two points [28,29]. In the context of solar irradiance forecasting, we need to define a distance function $D(x_i, x_j) = D_{ij}$ that quantifies how close or similar are the sites i and j. Given this distance function and a threshold, we can choose those sites that are closer to the target site. It means that we can determine which sites are the ones that can provide historical measurements to train a model to forecast irradiance on a site that does not have them. In the following subsections, we introduce several approaches to learn the distance function D_{ij} and to determine the threshold.

3.1. Learning to rank

We can learn a distance function to rank sites according to their closeness to a previously unknown target site. The first approach that we propose to learn a metric involves two steps using solar irradiance measurements from training sites: a first step where for each training site an optimal ranking of the remaining sites is discovered according to their measurements prediction accuracy; and a second step where, given the optimal rankings for each site, an optimization process is conducted to find a distance function that is able to produce the same reference rankings obtained in the first step. The metric that results from the optimization process can be used to determine those sites that can provide measurements for training given a new unknown location where there are not irradiance measurements available. This methodology is illustrated in Fig. 4.

The first step involves searching the ideal combination of sites that maximizes the prediction accuracy. An exhaustive search represents a combinatorial problem that has excessive computational costs. For the feasibility of implementation, we conduct this search using a Sequential Forward Selection (SFS) algorithm [30]. This is a greedy algorithm that sequentially adds data to the training dataset from different sites in a way that the mean absolute error (Equation (1)) on the target site is minimized. Fig. 5 shows the result of partially conducting SFS process using an ANN as a prediction model to find the ideal site ranking when the target site is site 0. According to this figure, site 8 is the one that provides data to have the most accurate predictions on site 0. Then, site 2 is the one that provides data that combined with the data from site 8 has more accurate predictions on site 0. Following this sequential process, the partial ideal site ranking for prediction in site 0 corresponds to 8, 2, 1, 3, 7.

Table 4 shows the ranking provided by the SFS algorithm for site 0. Note that a worse performance was obtained using the datasets provided by all the other sites than using just a subset of that data. According to all the performance indicators, we can say that the best prediction is obtained when the ANN is trained using data from only seven out of nine sites. In this study, the best prediction is the one that minimizes the sum of *MAE* and *RMSE*. We use this criterion to determine the best prediction because both errors are measured on the same scale and units, and other indicators such as *NRMSE* and R^2 are highly correlated to *RMSE*. Interestingly, the sites that can be discarded in the ranking have the biggest differences in geographical and weather conditions with the location where the prediction is made, justifying the need for a metric to determine which sites provide data for training. Applying the same procedure to every site, we determine that for most of the sites the use of a subset of data gives a better prediction. Table 5 shows the number of sites selected to get the best prediction result for each station.

Having the ideal site ranking, we formulate a solution to the second step in the proposed approach for metric learning. In this case, we use an optimization algorithm to find a metric such that it produces the ideal ranking of sites.

Table 6
Average performance indicators when determining the threshold for a forecasting model on station 0. Note that the lowest *MAE* and *RMSE* are obtained on average when the first 6 sites are selected, and using the measurements for all other sites worsen the prediction accuracy. Therefore the threshold selected for station 0 is 6.

Amount of selected stations	*MAE* (W/m^2)	*MBE* (W/m^2)	*RMSE* (W/m^2)	*nRMSE* (%)	R^2
1	68.32	−15.19	109.91	61.26	0.8301
2	60.70	−6.69	100.22	56.02	0.8577
3	55.23	−3.34	97.27	54.53	0.8651
4	52.73	−1.24	95.28	53.37	0.8707
5	52.18	2.88	95.82	53.66	0.8697
6	**51.98**	**−0.90**	**95.46**	**53.41**	**0.8708**
7	52.25	1.24	95.64	53.59	0.8686
8	53.25	0.14	96.95	54.26	0.8663
9	54.23	−6.48	101.04	56.58	0.8542

3.1.1. *Optimization problem*

To formulate the optimization problem we first introduce the definition of Mahalanobis distance (WMD) [21]. This distance function is defined using a quadratic form with a positive semi-definite matrix *W* that determines the geometry of the function:

$$D_{ij}^2 = (x_i - x_j)^T W (x_i - x_j). \tag{7}$$

In this case, we assume that *W* is a diagonal matrix. It means that the elements of *W* weight each entry of the vector $x_i - x_j$. Entries with larger weights are the most relevant ones to calculate the distance. Equation (7) calculates the square of the distance function over points that are in the multidimensional space that we previously defined. The goal here is to find *W* such that the site ranking obtained using this distance function coincides with the reference site ranking.

The optimization algorithm is formulated such that it finds the *W* matrix that minimizes the difference between these two rankings for station *k*. Let s_i be the position on the ranking given by the metric of the site that should be at position *i* in the reference ranking obtained through SFS. Let D_{ik} be the distance between site at position *i* in the reference ranking and the target site *k*. Knowing that the site at position *i* in the reference ranking has to be closer to *k* than the site at position *j*, the rank s_i determined by the metric should be lower than the rank s_j. Similarly, distance D_{ik} should be lower than D_{jk}. Therefore, the optimization process should seek to minimize $s_i - s_j$ and $D_{ik} - D_{jk}$ so that these differences should be negative for every pair (i,j). In this way, this optimization process will conduct a search that reduces the difference between the reference ranking and the ranking developed by the metric. Using this idea and the cost function proposed in Ref. [19] for ranking in information retrieval problems, we formulate the minimization problem for a specific site *k*:

$$\begin{aligned} \min J_k(W) &= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[s_i - s_j + \ln(1 + e^{s_j - s_i})\right] \\ &+ \lambda \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[D_{ik} - D_{jk} + \ln\left(1 + e^{D_{jk} - D_{ik}}\right)\right] \end{aligned} \tag{8}$$

The first term of the cost function accounts for the difference in ranking positions between the reference ranking and the one induced by the metric. The second term can be seen as a regularization term that accounts for the difference in distances between the sites according to their ranking positions. This regularization term allows for avoiding those candidate solutions that result in the same value of the difference between rankings. Moreover, according to Ref. [19], using the term with the logarithm function makes the cost function smoother, facilitating the search process of the optimization algorithm.

3.1.2. *Incorporating principal component analysis*

After solving the optimization problem in Equation (8), the algorithm converges to different local minima for random initialization due to correlations between the variables that define each vector x_i. To deal with this issue, principal component analysis (PCA) was applied before the optimization algorithm to map each point into a new space where its variables have a minimum correlation. This is a linear transformation that is also called a whitening transformation. Defining a matrix *U*, which applies the PCA transformation $y = U^T x$, Equation (7) can be rewritten as Equation (9):

$$D_{ij}^2 = (x_i - x_j)^T U W U^T (x_i - x_j) \tag{9}$$

In this way, the optimization algorithm looks for local minima in a new context where the correlations between each feature are approximately zero.

3.1.3. *Kernelized Mahalanobis Distance*

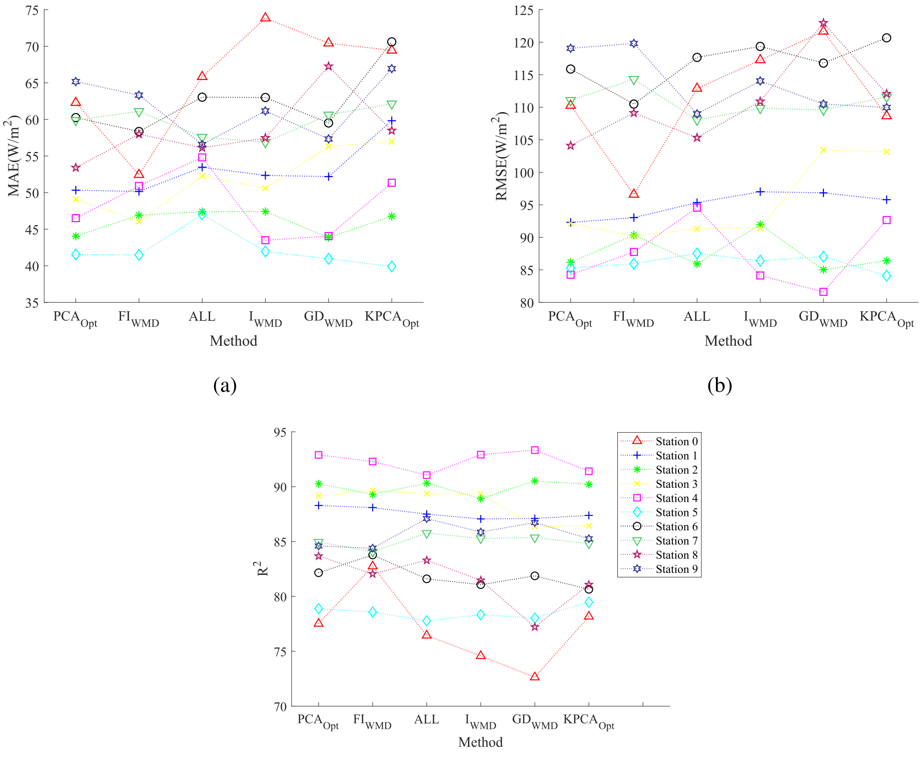
It is possible to learn the metric in a space that results from a nonlinear transformation of the data. Learning a Kernelized Mahalanobis Distance (kWMD) is an option to define nonlinear transformations of the metric space to potentially improve the ranking results. According to the definition of Mahalanobis distance, Equation (7) can be rewritten as

$$\begin{aligned} D_{ij}^2 &= x_i^T W x_i - 2x_i^T W x_j + x_j^T W x_j \\ &= x_i^T W^{1/2} W^{1/2} x_i - 2x_i^T W^{1/2} W^{1/2} x_j + x_j^T W^{1/2} W^{1/2} x_j \\ &= \hat{x}_i^T \hat{x}_i - 2\hat{x}_i^T \hat{x}_j + \hat{x}_i^T \hat{x}_i, \end{aligned} \tag{10}$$

where $\hat{x}_i = W^{1/2} x_i$. As it is shown in Refs. [29,31], we can use the so-called kernel trick for the computation of the inner products involving vectors $\hat{x}_i$ and $\hat{x}_j$ in a higher dimensional space. Equation (10) can be redefined using the kernel trick as

$$D_{ij}^2 = k(\hat{x}_i, \hat{x}_i) - 2k(\hat{x}_i, \hat{x}_j) + k(\hat{x}_j, \hat{x}_j), \tag{11}$$

where $k(\hat{x}_i, \hat{x}_j)$ denotes the dot product in the higher dimensional space and is called kernel. Three types of kernels (polynomial, radial base, and sigmoid functions), linear combination of kernels, and kernel composition [32] were tested to find the kernel that constructs a metric whose ranking is similar to the reference ranking. The best result was obtained with a composition of sigmoid kernels as shown in Equation (12), where parameters γ and r in Equation (12) are found in the same optimization problem for metric learning.



(a)

(b)

(c)

Fig. 6. Performance indicators a) Mean Absolute Error *MAE*, b) Root Mean Square Error *RMSE*, and c) R^2 on each methodology and weather station. All: historical measurements were used from all the other sites to train the model; PCA-Opt: metric developed by the optimization algorithm and PCA; KPCA-Opt: Metric adding the kernelized version of WMD; I-WMD: Euclidean distance (no optimization process); FI-WMD: Mahalanobis distance with weights given by the feature importance; GD-WMD: Euclidean distance based on geographic location.

Table 7
Selected threshold for each station at each step of the crossvalidation test based on a leave-one-station-out scheme.

Station	Threshold
0	6
1	6
2	6
3	7
4	6
5	7
7	7
8	8
9	7
Mean	6.6

$$k(\widehat{x}_i, \widehat{x}_j) = \tanh\left(\gamma \tanh\left(\widehat{x}_i^{\,T} \widehat{x}_j\right) + r\right) \tag{12}$$

3.2. Feature importance to rank

Another approach for metric learning results from using the feature importance (*FI*) computed in Section 3.2 directly as the weights of the matrix *W* in Equation (7) for those variables used to construct the space where the metric is defined. Variables that are more important for prediction have larger weights in the definition of the metric. The *FI* for variables measured at two different times are added to get one weight for each variable. For example, Surface Pressure$(t-k)$ and Surface Pressure $(t-24n)$ in Table 2 had *FI* 5.36 and 4.88, respectively. Thus, the weight associated with the variable Surface Pressure in the space where the metric is defined is 10.24. The feature importance of on-site solar radiation variables is not used to define the metric because space where the metric is defined does not include this information.

3.3. Determining the threshold

Remember that the goal in this work is to determine those sites that can potentially provide measurements for training a predictive model on a previously unknown site. It means that not only the metric has to be learned but also the number of neighboring sites that are considered to provide measurements for training. To do that, we compute the mean absolute error (Equation (1)) at each step of the SFS algorithm to determine the average number of sites for training that have to be added to produce the best prediction and round it to the nearest integer. Table 6 shows the average performance indicators when determining the threshold for a forecasting model on station 0. When it is assumed that there are not irradiance measurements of this site, first, we apply the first step of SFS to stations 1 to 9 (sites with measurements) and average the indicators to get the first row of this table. After that, we apply the second step of SFS to stations 1 to 9 and average the indicators to get the second row. We repeat this process for all the 9 steps of SFS. Finally, we select the amount of sites that give us the best prediction result for stations 1 to 9 as the threshold for the station 0. Note that the lowest *MAE* and *RMSE* are obtained on average when the first 6 sites are selected, and that using the measurements for all other sites worsen the prediction accuracy. Therefore, the threshold selected for station 0 is 6.

3.4. Comparing the proposed approaches

We compare the performance of six methodologies to select the data to train a prediction model: i) one based on the Euclidean distance on the metric space (no optimization), ii) the Mahalanobis distance after the optimization process in Equation (8) and principal component analysis (PCA) (Section 3.1.2), iii) the kernelized Mahalanobis distance (Section 3.1.3), iv) a weighted Euclidean distance with weights proportional to feature importance (Section 3.2), v) an Euclidean distance based on the geographic position of the stations, and vi) training the prediction model using the measurements from all available sites (i.e., selecting all sites). To compare the performance of each methodology we applied a cross-validation test for each one of the ten available weather stations in a leave-one-station-out scheme, from the real-world dataset described in Section 2.3. For each test station, each methodology was used to learn the metric and to select those stations that provide measurements for the training step. For each weather station, performance indicators were computed to compare the proposed methodologies.

Fig. 6 shows the performance of each proposed methodology using the performance indicators *MAE*, *RMSE*, and R^2 at each station (Equation (1), (3), and (5)). Each station is shown with a specific color and marker. These results show that the methodologies based on the optimization algorithm and PCA (PCA-Opt), and Feature Importance Weighted Mahalanobis Distance (FI-WMD) tend to have better results concerning the performance indicators *MAE*, R^2 and *RMSE* than using all the available data for training. Even a selection process based on a simple Euclidean distance on the metric space provides better results than using measurements from all available sites if the results from station 0 are ignored. On the other hand, interestingly, when that selection is made using an Euclidean distance based on just geographical location, the performance of the prediction model is worse. These results suggest that the proposed multidimensional space and the learned metric are key to define a subset of the measurement selection process. Also, Table 7 shows that at each step in the cross-validation test, the selected threshold is around six.

These results support our hypothesis that selecting a subset of measurements rather than using all the available data can lead to more accurate prediction models, and that a multidimensional space constructed using information based on weather and environmental variables is useful to discover those sites that can potentially provide measurements for training. As far as we know, this type of study using data from sites located tens or hundreds of kilometers away from the target place has not been done yet. We consider this study a reference for future works on predicting irradiance for places without available measurements. The proposed methodology can potentially increase the performance of prediction for any predictive model. We used ANN but the same procedure can be applied to any other model such as ARIMAX, SVM or Random Forest.

4. Conclusions

This paper addressed the problem of solar irradiance forecasting when there are no historical on-site irradiance measurements for training a prediction model. We hypothesized that historical irradiance measurements from other sites with similar radiation patterns could be used to train a prediction model. We proposed a methodology based on metric learning to solve this issue. In this methodology, a multidimensional space was constructed based on measurements that included temperature, humidity, and satellite measurements, where each site corresponded to a point in this space. Then, a metric was defined to quantify the similarity between sites concerning their radiation patterns. This allowed for determining which sites can provide data to train a prediction model that will be used on a target site with similar patterns. We used real data and performance indicators to evaluate the proposed approaches for metric learning.

In the first part of this work, we showed that clear sky model (CSM) estimations represent an important source of information when they are included as predictor variables in a solar irradiance forecasting model such as Artificial Neural Network. We reported that the model can learn the pattern given by the CSMs, providing a baseline pattern for long term predictions. This significantly improves the accuracy for long-term predictions compared to other models presented in the current literature. Also, we showed that a 2-D time series structure was important to improve the forecast performance by considering the seasonal behavior of some variables.

In the second part of this work, we presented several approaches to learn a metric based on concepts such as Mahalanobis distance, learning to rank, kernel machines, principal component analysis, sequential forward data selection, and feature importance for prediction. We showed that learning a metric to select sites with similar patterns outperforms a methodology that uses data from all the available sites.

We visualize this work as an important step for solar PV power systems implementation in locations of difficult access. Future research in this direction include using deep artificial neural networks as prediction models, and different kernel machines to characterize a variety of nonlinear transformations in the definition of the metric. Also, an alternative strategy to sequential forward selection strategy can be used to find site rankings that are closer to the optimal one.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Andres Felipe Zambrano: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Visualization. **Luis Felipe Giraldo:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision.

References

[1] S. Sen, S. Ganguly, Opportunities, barriers and issues with renewable energy development - a discussion, Renew. Sustain. Energy Rev. 69 (2017) 1170–1181, https://doi.org/10.1016/j.rser.2016.09.137. http://www.sciencedirect.com/science/article/pii/S1364032116306487.

[2] M. Mohammadi, Y. Noorollahi, B. Mohammadi-ivatloo, H. Yousefi, Energy hub: from a model to a concept - a review, Renew. Sustain. Energy Rev. 80 (2017) 1512–1527, https://doi.org/10.1016/j.rser.2017.07.030. http://www.sciencedirect.com/science/article/pii/S1364032117310985.

[3] M.C. Lott, S. Pye, P.E. Dodds, Quantifying the co-impacts of energy sector decarbonisation on outdoor air pollution in the United Kingdom, Energy Pol. 101 (2017) 42–51, https://doi.org/10.1016/j.enpol.2016.11.028. https://www.sciencedirect.com/science/article/pii/S030142151630622X.

[4] T. Sharma, P. Balachandra, Model based approach for planning dynamic integration of renewable energy in a transitioning electricity system, Int. J. Electr. Power Energy Syst. 105 (2019) 642–659, https://doi.org/10.1016/j.ijepes.2018.09.007. http://www.sciencedirect.com/science/article/pii/S0142061518306938.

[5] C.B. Martinez-Anido, B. Botor, A.R. Florita, C. Draxl, S. Lu, H.F. Hamann, B.-M. Hodge, The value of day-ahead solar power forecasting improvement, Sol. Energy 129 (2016) 192–203, https://doi.org/10.1016/j.solener.2016.01.049. http://www.sciencedirect.com/science/article/pii/S0038092X16000736.

[6] A. Ghasemi, M. Enayatzare, Optimal energy management of a renewable-based isolated microgrid with pumped-storage unit and demand response, Renew. Energy 123 (2018) 460–474, https://doi.org/10.1016/j.renene.2018.02.072. http://www.sciencedirect.com/science/article/pii/S0960148118302167.

[7] I. Ranaweera, O.-M. Midtgård, M. Korps, Distributed control scheme for residential battery energy storage units coupled with pv systems, Renew. Energy 113 (2017) 1099–1110, https://doi.org/10.1016/j.renene.2017.06.084. http://www.sciencedirect.com/science/article/pii/S0960148117305888.

[8] D. Neves, A. Pina, C.A. Silva, Assessment of the potential use of demand response in dhw systems on isolated microgrids, Renew. Energy 115 (2018) 989–998, https://doi.org/10.1016/j.renene.2017.09.027. http://www.sciencedirect.com/science/article/pii/S0960148117308893.

[9] E. Nyholm, M. Odenberger, F. Johnsson, An economic assessment of distributed solar pv generation in Sweden from a consumer perspective - the impact of demand response, Renew. Energy 108 (2017) 169–178, https://doi.org/10.1016/j.renene.2017.02.050. http://www.sciencedirect.com/science/article/pii/S0960148117301362.

[10] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: a review, Renew. Energy 105 (2017) 569–582, https://doi.org/10.1016/j.renene.2016.12.095. http://www.sciencedirect.com/science/article/pii/S0960148116311648.

[11] R.H. Inman, H.T. Pedro, C.F. Coimbra, Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci. 39 (6) (2013) 535–576, https://doi.org/10.1016/j.pecs.2013.06.002. http://www.sciencedirect.com/science/article/pii/S0360128513000294.

[12] F. Barbieri, S. Rajakaruna, A. Ghosh, Very short-term photovoltaic power forecasting with cloud modeling: a review, Renew. Sustain. Energy Rev. 75 (2017) 242–263, https://doi.org/10.1016/j.rser.2016.10.068. http://www.sciencedirect.com/science/article/pii/S136403211630733X.

[13] D. Yang, J. Kleissl, C.A. Gueymard, H.T. Pedro, C.F. Coimbra, History and trends in solar irradiance and pv power forecasting: a preliminary assessment and review using text mining, Sol. Energy 168 (2018) 60–101, https://doi.org/10.1016/j.solener.2017.11.023, advances in Solar Resource Assessment and Forecasting, http://www.sciencedirect.com/science/article/pii/S0038092X17310022.

[14] S. Sobri, S. Koohi-Kamali, N.A. Rahim, Solar photovoltaic generation forecasting methods: a review, Energy Convers. Manag. 156 (2018) 459–497, https://doi.org/10.1016/j.enconman.2017.11.019. http://www.sciencedirect.com/science/article/pii/S0196890417310622.

[15] A.W. Aryaputera, D. Yang, L. Zhao, W.M. Walsh, Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging, Sol. Energy 122 (2015) 1266–1278, https://doi.org/10.1016/j.solener.2015.10.023. http://www.sciencedirect.com/science/article/pii/S0038092X15005745.

[16] D.J. Gagne, A. McGovern, S.E. Haupt, J.K. Williams, Evaluation of statistical learning configurations for gridded solar irradiance forecasting, Sol. Energy 150 (2017) 383–393, https://doi.org/10.1016/j.solener.2017.04.031. http://www.sciencedirect.com/science/article/pii/S0038092X17303158.

[17] M. J. Reno, C. W. Hansen, J. S. Stein, Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis, Sandia National Laboratories Report.

[18] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.

[19] C.J. Burges, R. Ragno, Q.V. Le, Learning to rank with nonsmooth cost functions, in: Advances in Neural Information Processing Systems, 2007, pp. 193–200.

[20] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond, MIT press, 2001.

[21] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, Chemometr. Intell. Lab. Syst. 50 (1) (2000) 1–18.

[22] W. Holmgren, C. Hansen, M. Mikofski, Pvlib python: a python package for modeling solar energy systems, Journal of Open Source Software 3 (2018) 884, https://doi.org/10.21105/joss.00884.

[23] F.O. Hocaoğlu, Ömer N. Gerek, M. Kurban, Hourly solar radiation forecasting using optimal coefficient 2-d linear filters and feed-forward neural networks, Sol. Energy 82 (8) (2008) 714–726, https://doi.org/10.1016/j.solener.2008.02.003. http://www.sciencedirect.com/science/article/pii/S0038092X08000376.

[24] J. Qiu, K. Sun, I.J. Rudas, H. Gao, Command filter-based adaptive nn control for mimo nonlinear systems with full-state constraints and actuator hysteresis, IEEE Transactions on Cybernetics (2019) 1–11, https://doi.org/10.1109/TCYB.2019.2944761.

[25] M.M. Rienecker, M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.-K. Kim, et al., Merra: nasa's modern-era retrospective analysis for research and applications, J. Clim. 24 (14) (2011) 3624–3648.

[26] A. Fouilloy, C. Voyant, G. Notton, F. Motte, C. Paoli, M.-L. Nivet, E. Guillot, J.-L. Duchaud, Solar irradiation prediction with machine learning: forecasting models selection method depending on weather variability, Energy 165 (2018) 620–629, https://doi.org/10.1016/j.energy.2018.09.116. http://www.sciencedirect.com/science/article/pii/S0360544218318802.

[27] J. Heaton, S. McElwee, J. Fraley, J. Cannady, Early stabilizing feature importance for tensorflow deep neural networks, in: 2017 International Joint Conference on Neural Networks, 2017, pp. 4618–4624.

[28] P. Zezula, G. Amato, V. Dohnal, M. Batko, Similarity Search: the Metric Space Approach, vol. 32, Springer Science & Business Media, 2006.

[29] B. Kulis, Metric learning: a survey, Foundations and Trends in Machine Learning 5 (4) (2013) 287–364, https://doi.org/10.1561/2200000019. https://www.nowpublishers.com/article/Details/MAL-019.

[30] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recogn. 43 (1) (2010) 5–13, https://doi.org/10.1016/j.patcog.2009.06.009. http://www.sciencedirect.com/science/article/pii/S0031320309002520.

[31] B. Schölkopf, The kernel trick for distances, in: Advances in Neural Information Processing Systems, 2001, pp. 301–307.

[32] Z. Zhang, Customizing Kernels in Support Vector Machines, Master's thesis, University of Waterloo, 2007, http://hdl.handle.net/10012/3063.