Renewable and Sustainable Energy Reviews 135 (2021) 110450

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: http://www.elsevier.com/locate/rser

Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting

M.O. Moreira [a,c], P.P. Balestrassi [b,c,*], A.P. Paiva [b], P.F. Ribeiro [c], B.D. Bonatto [c]

[a] Federal Institute of Education, Science and Technology – South of Minas Gerais, Carmo de Minas, MG, Brazil
[b] Institute of Production Engineering and Management, Federal University of Itajubá, Itajubá, MG, Brazil
[c] Institute of Electrical Systems and Energy, Federal University of Itajubá, Itajubá, MG, Brazil

ARTICLE INFO

Keywords:
Photovoltaic forecast
Design of experiments
Artificial neural networks
Ensemble

ABSTRACT

In recent years, renewable and sustainable energy sources have attracted the attention of various investors and stakeholders, such as energy sector agents and even consumers. It is perplexing to observe and anticipate the required levels of photovoltaic generation, which are inherent tasks for such rapid insertion into the electric grid. This distributed/renewable generation must be integrated in a coordinated way such that there is no negative impact on the electric performance of the grid, increasing in the complexity of energy management. In this article, a methodology for photovoltaic generation forecasting is addressed for a horizon of one week ahead, using a new approach based on an artificial neural network (ANN) ensemble. Two main questions will be explored with this approach: how to select the ANNs, and how to combine them in the ensemble. The design of experiments (DOE) approach is applied to the photovoltaic time series factors and ANN factors. Then, a cluster analysis is performed to select the best networks. From this point on, a mixture (MDE) is employed to determine the ideal weights for the ensemble formation. The methodology is detailed throughout the paper and, based on the combination of forecasts, the photovoltaic generation was estimated for a specific panel set located in the state of Minas Gerais, Brazil, reaching the value of 4.7% for the weekly mean absolute percentage error. The versatility of the proposed method allowed the change of the number of factors to be used in the experimental arrangement, the forecast model, and the desired forecast horizon, and consequently enhancing the forecasting determination.

1. Introduction

For many decades, the unidirectional flow of electric energy from generators to transmission, distribution and consumption at the end user has remained practically unchanged. From this perspective, distribution networks were not designed to support the insertion of generation units [1], but rather are based a unidirectional energy flow that carries energy from a substation to the final consumer [2]. Therefore, it is essential that there be a significant change in the paradigm of the energy supply system for the following reasons: (a) the shortage of fossil fuels leads to environmental and energy problems and (b) renewable energy has attracted the interest of many investors in different regions in recent years [3].

With the advent of smart grids, photovoltaic (PV) generation forecasts are key to managing distribution networks, micro-grids, or smart homes [4]. The system performance, as well as operational decisions, can be improved based on load or generation forecasts [5]. The main benefits of a more accurate forecast highlight the chances of avoiding over-voltages, which are observed when the PV generation is larger than the demand [6]. Some countries, such as Spain, provide incentives for better predictions of the next-day solar income level [7].

In comparison to load forecasting errors, which are generally approximately 1%–3% [8], generation errors (solar and wind) are significantly more substantial, reaching average values of 15%–20% [8]. Knowing that PV generation is highly uncertain and difficult to predict [9], it is of paramount importance to classify the forecast horizons according to Refs. [10]: long-term forecasts for periods longer than 1 month; medium-term forecasts for a range of 1–7 days; and short-term forecasts for a few hours ahead. The short- and medium-term horizons are the most studied, owing to the uncertainties related to climate dependence and the consequent decrease of reliability as the desired horizon increases. The importance of each horizon can be quantified

* Corresponding author. Institute of Production Engineering and Management,Federal University of Itajubá, Itajubá, MG, Brazil.
E-mail address: pedro@unifei.edu.br (P.P. Balestrassi).

https://doi.org/10.1016/j.rser.2020.110450
Received 3 December 2019; Received in revised form 8 September 2020; Accepted 5 October 2020
Available online 14 October 2020

according to the actions that must be taken (or avoided) at different moments of time. Short-term forecasts are useful for dealing with over-voltages [8], which can not only compromise the grid security in terms of over-loading of equipment, but can also cause permanent damage on motors, electronics, and electro-electronics [2]. Medium-term forecasts, which are the scope of this work, are used by operators to determine reserve requirements, storage dispatches and energy quality [11]. Long-term forecasts are useful for market decisions [12] and sector investments [13].

The artificial neural networks (ANN) was the technique chosen to be used in this work owing to its importance in the literature, as even studies that do not directly apply ANN mention that ANN plays an important role in PV generation forecasting [14].

As forecasting should provide both clarity and reliability [15], the forecasting method, presented in this work, is based on a design of experiments (DOE) approach, to assist in the choice of factors (both series and prediction models) that may result in more accurate estimates. DOE is a statistical tool in which each experimental run is a test, and it allows an investigator to discover some information regarding a process or system [16].

In the sequence, the best configurations observed in the DOE approach will be maintained, through a cluster analysis, to form a combined forecast. An ensemble tends to improve the results of the individual models [17]. The proposed combination considers that the definition of the ensemble weights is calculated by a mixture analysis.

Two issues are noteworthy in this scenario: a) Simulated processes often require a high number of combinations to test all possible solutions and b) forecasting models can generate more than one potential solution to the problem and it is not known, for sure, which is the best.

The originality of this methodology, applied to photovoltaic generation forecasting, aims to cover these gaps and is supported by two main pillars. The first one is, knowing that the processing architecture of current computers is limited, the application of DOE to reduce the number of combinations related to the ANN parameterization is

Table 1
Summary of related works.

Author and year	Forecasting method	Forecast horizon	Exogenous variables	Parametrization	Comment
Zhen et al. (2020) [18]	Hybrid model based on ANN, CNN, and LSTM	Minute	Sky image and solar irradiance	Number of layers and neurons	Some parameters of the forecast models have not been specified. Some of them are defined by scanning a range of values.
Theocharides et al. (2020) [19]	ANN	Hourly day-ahead	Incident global irradiance, ambient temperature, relative humidity, wind direction and speed, solar azimuth and elevation angles	Neurons, layers and epochs.	Parameters were defined by scanning a range of values.
Sangrody et al. (2020) [20]	Similarity-based forecasting models (SBFMs)	Minute day-ahead	Temperature, humidity, dew point, wind speed, irradiance and sky cover data	Based on k-nearest neighbors (KNN). Grid search method is applied to find optimal k.	At a certain point, the combination yielded larger errors. The parametric configuration of the model is not detailed.
Pan and Tan (2019) [21]	Based on cluster analysis and ensemble regression.	Hourly day-ahead	Hourly day-ahead	Hyperparameters selected using the grid search method.	Clusters based on climatic variables require greater computational effort
Wen et al. (2019) [22]	Deep recurrent neural network with long short-term memory units (DRNN-LSTM)	Hourly month-ahead	Global horizontal radiation, and diffuse horizontal radiation	Bayesian hyperparameters optimization	Automated parameterization is efficient, but still requires considerable processing effort when each forecast is performed. Combined forecasting is not explored.
Ozoegwu (2019) [23]	ANN	Month up to two years	Sunshine, temperature, cloudiness, precipitation, relative humidity, dew point, temperature, soil temperature, evaporation and pressure.	Fixed: One hidden layer with 20 hidden neurons. Levenberg-Marquardt training algorithm. Tangent sigmoid and linear transfer functions	The combination is performed between the models and not between the results of different predictors. Here, a new hybrid model is generated.
Qing and Niu (2018) [24]	Long Short-term memory (LSTM) networks and ANN using the classical backpropagation algorithm (BPNN)	Hourly day-ahead	Temperature, dew point, humidity, visibility, wind speed and descriptive weather summary	Vary, depending on the benchmark	Some parameters are randomly defined and others are found by combining many executions, which increases processing. Combined forecasting is not explored.
Bugala et al. (2018) [25]	ANN	Hourly day-ahead	Sunny hours, length of the day, air pressure, maximum air temperature, daily insolation and cloudiness	Parameters were estimated using a heuristic algorithm	For each forecasting round, the parameters must be estimated using heuristics, which increases the computational effort.
Ravinesh and Sahin (2017) [26]	ANN	Seasonal and Month	Satellite-derived data and land surface temperature	Several parameters were trailed. Model architecture may vary (55 for monthly forecasting and 9 for seasonal forecasting)	Although the model is parsimonious, there is a variation in the model architecture that stresses a variety of parameters, such as training algorithm and activation function. It is not indicated how the relationship between the parameters influences the result.
Sivaneasan and Goh (2017) [27]	ANN and fuzzy logic for pre-processing weather data	Month	Cloud cover, temperature, wind speed, and wind direction with irradiance value	Fixed: Levenberg-Marquardt training algorithm, 25 hidden neurons and tangent sigmoid as the activation function	There is no indication whether changing a given parameter improves or worsens the results. Combined forecasting is not explored
Cervone et al. (2017) [28]	ANN and Analog Ensemble	Hourly 3 day-ahead	Global horizontal irradiance, cloud cover, air temperature, solar azimuth and elevation	From 4 to 20 hidden neurons	ANN is initialized several times for the same station and the parameters are defined by searching a range of values
Lima et al. (2016) [29]	ANN	Hourly day-ahead	Relative humidity, temperature, wind speed, cloud cover, precipitable water	One hidden layer, hyperbolic tangent as activation function, Levenberg Marquardt as training algorithm	Parameters were defined by scanning a range of values. Some of them are not discussed or presented.

attractive, keeping statistical reliability. The second one is related to how to choose or combine the forecast results. In this case, it is proposed to apply the hierarchical cluster analysis to select the best networks and then use Mixture DOE to perform an ensemble.

The paper is divided as follows. Section 2 analyzes some related works, and discusses the main prediction methods, ANNs, and DOE. Section 3 describes the originality of the methodology for PV forecasting, which will be detailed throughout this work. Section 4 presents a case study for the forecast of a week ahead from recent generation data, obtained from a plant located in Minas Gerais, Brazil. Finally, Section 5 highlights the conclusions.

2. Background and literature review

This section investigated some of the most relevant related works as well as the gaps to be noted:

- Many of these works do not explore the potential of combined forecasting;
- In most cases, the parameters of the forecasting model are not formally defined, instead chosen empirically, fixedly, randomly or even by scanning a range of values. When this happens, it is not possible to identify which parameters influence the result, as well as it may require considerable computational effort when processing numerous possibilities.

It was also observed that some studies focus mainly on the forecast of solar radiation and not on the photovoltaic generation forecasting, as there may be a quite difference in the adherence of the forecast models in terms of uncontrollable factors, such as dust on the panels, damaged sensors, panel efficiency, etc. The following Table 1 summarizes these works, presenting the main characteristics and a critical comment on issues not addressed:

Zhen et al. [18] proposed a hybrid model, based on convolutional neural networks (CNN), Long Short-Term Memory (LSTM) and Artificial Neural Network (ANN), to forecast photovoltaic energy in real time. The normalization method was not discussed by the authors. First, the sky image pre-processing step requires high computational effort to extract features. Then, some parameters of the forecasting models are defined by scanning a range of values. As it is a real-time operation (15 min), it may not be an interesting operation if the model has to be recalibrated. The parameters adjusted for the neural network were the hidden layers and the number of neurons. It was not possible to see how the interaction between the model parameters affects the result.

Theocharides et al. [19] presented the photovoltaic energy forecast for a day ahead, with hourly resolution. The model is based on Artificial Neural Networks and uses the linear regressive correction method to adjust the forecast results using solar irradiance. No method of data normalization is discussed by the authors. The architecture of the forecasting model was configured based on the input data, which means that an empirical scan was performed to find the best parametric values in a data range. Post-processing combines K-means clustering and a linear regressive model.

Sangrody et al. [20] proposed the photovoltaic generation prediction through similar forecasting methods chosen through the k-nearest neighbors (KNN) method. The data normalization takes into account the feature scaling method, which considers the maximum and minimum values of the time series. The parametric configuration of the forecast model is not detailed and is defined by the grid search method. At a given moment, the authors report that the accuracy of the forecast is strictly related to the choice of climatic variables and that the process of combining some results did not promote improvements in the forecast.

The forecasting method proposed by Pan and Tan [21] makes use of cluster analysis to classify meteorological characteristics and then uses ridge regression to determine the weights of the ensemble. The forecast horizon for a day ahead was addressed, with hourly resolution. The normalization method used the maximum and minimum values of the time series. It was discussed that, in the first stage (cluster analysis), both generation data and only climatic data can be used. However, for the latter, the computational cost in terms of processing can be considerably increased. Hyperparameters were selected using the grid search method.

Wen et al. [22] performed an integrated forecast involving load and photovoltaic generation. The data was normalized using maximum and minimum values. The experiments were implemented in the Python programming language and, to define the parameters of the prediction model, an automated method is used that investigates the search space using Bayesian optimization. Of course, this type of search is more efficient than a manual or random scan, but it still consumes processing for each forecast that is performed. The photovoltaic forecast is estimated hourly for the period of one month ahead.

In Ref. [23], author developed a methodology aimed to forecast solar radiation for a year ahead horizon of monthly mean daily. This forecast is useful for the dimensioning of photovoltaic energy and also for agricultural activities. It is noticed that the case study, carried out in Nigeria, involved the comparison of several neural network architectures, implemented by the authors using the Matlab software, with fixed parameterization. The authors' idea focuses on the combination of the forecasting models and not on the results. This hybrid model uses the number of the month as part of the inputs in the long-term solar forecasting process.

Qing and Niu [24] performed a comparison between the Long Short-term memory (LSTM) networks and ANN models using the classical backpropagation algorithm (BPNN) for solar irradiance prediction. Depending on the benchmark, there was variation in the number of epochs, the number of hidden layers and the number of neurons per layer. Thus, there was a scan to test a range of parameters in each, which led to an increase in processing to find the ideal values. The linear scaling normalization method was used in the pre-processing of the inputs. The exogenous variables here are hourly weather forecasts for a specific day and will be used for photovoltaic forecast generation. The complete setup of the forecasting models was not presented and some parameters were defined through empirical execution to reach a certain value.

A selection of variables through the Pearson correlation coefficient was observed in Bugala et al. [25]. The authors analysed seven variables: number of sunny hours, maximum air temperature, daily sunshine, cloudiness, daytime duration and air pressure. The latter two were considered statistically irrelevant for the linear regression model and were not used. For the neural network, only the variable air pressure was ignored and a neural network of the radial basis function (RBF) type was generated for the prediction, with six neurons in the input layer, five neurons in the hidden layer and one neuron in the output layer. The error obtained by the RBF network 6: 6-5-1: 1 was considerably low, causing the authors to recommend their application for a day-ahead forecast. However, the study did not address the predictive impacts from a parametric variation in the neural network.

Ravinesh and Sahin [26] proposed a methodology for forecasting solar radiation for territories that have satellite data coverage. The normalization of the input data occurred through maximum and minimum values of the time series. The implementation took place through the Matlab software and several parameters were trialed, such as training algorithm and activation function. Network architecture may vary depending on the forecast context (monthly or seasonal). Although the neural network model is parsimonious, there was an intense parametric variation for each architecture that is generated and, consequently, increases the computational effort in terms of processing each time the forecast is performed. It is not indicated how the relationship between the parameters influences the result.

The work of Sivaneasan and Goh [27] applied ANN to evaluate the daily forecast for a period of one month ahead. The accuracy in the results was due to the application of fuzzy logic in the pre-processing of the inputs. The authors defined the parameterization of the neural

model in a fixed way, with 8 neurons in the input layer, 25 hidden neurons, tangent sigmoid as the activation function and Levenberg-Marquardt as training algorithm. A three-month period (January to March 2017) of climate data was chosen for training the model. It was not possible to identify in the work if these parameters were randomly defined or if they were found from the intensive processing of several combinations of executions. It is not discussed in the paper whether the change in certain parameters of the model may change the results or not.

Based on actual weather data and PV generation data from three power plants in Italy, Cervone et al. [28] proposed a forecast for 72 h ahead using an ANN and analog ensemble. The authors considered the solution scalable, and produced more reliable results when the methods were combined. However, the definition of the weights of this combination was done iteratively, generating approximately 1002 and 3004 combinations, and required parallel processing to obtain the results. ANN was initialized several times for the same station and the parameters were defined by searching a range of values.

Lima et al. [29], knowing that Brazil has enormous potential for producing PV energy, conducted a study to predict solar irradiance while considering a horizon of 24 h ahead. The authors used a cluster analysis to identify regions of the map with similar climatic characteristics. Several parameters were tested for each of the 110 stations in order to establish coherence between the results and the observed irradiation values, which demands high computational cost. The study did not consider a statistical analysis of the influence of varying the neural network parameters in the prediction, but revealed that the processing performed with this model increases the reliability of the forecast, and generates consistent results.

Some authors [30] have considered predicting the irradiation of a region based on the predictions of irradiation for neighbouring sites, as emitted by meteorological entities. In this case, the solution proposed by Ref. [30], based on neural networks, could be considered evolutionary, as the model is fed back with previous predictions as it advances chronologically in time. The authors only considered varying the number of network inputs, and fixed the values of parameters such as the number of hidden layers, activation function, and number of neurons.

In [31], the authors considered a systematically selected analysis of 38 articles. The searches were guided by keywords like “Big Data”, “Data Mining”, and “Machine Learning”, i.e., all related to the forecast of PV energy. In short, they found that neural networks have more accurate prediction algorithms than other models.

A comprehensive literature review was conducted by Das et al. [13], where the authors explored PV prediction from different perspectives: (a) the pre-processing of model inputs through normalization, (b) the correlation between input variables and generation data, (c) a short forecast horizon, (d) analysis of the performance of the methods, (e) details of the error evaluation criteria, and (f) analysis of the methods of forecasting, recent works from the perspectives of the techniques used, and the accuracy of the results. The authors concluded that neural networks and models based on a support vector machines (SVM) promote good execution, and have good adherence to the data. The most frequently used metric for error evaluation is the root mean square error (RMSE), and the most exploited horizon for forecasting is the short term. In this sense, knowing the scarcity of analyses for a medium-term period, our work intends to contribute with a forecast for a horizon of one week ahead.

The literature review examined by Sobri et al. [32] evaluated the different PV forecasting techniques and the recent progress achieved in this area. The analysis performed by the authors considered factors including the method of forecasting, time horizon, and error metric. In addition to finding that the application of the ANN and SVM methods is advantageous for the solution of nonlinear problems, they also showed that the ensemble-based models try to extract the precision and robustness of the individual methods.

Finally, the literature review presented by Barbieri et al. [33] aimed to list the methods used to forecast PV energy, as well as to review the statistical methods used for this purpose. In addition, the authors compared the different time horizons in terms of performance and ranking. The time setting is set according to the system operation. For long-term horizons, the suggestion is to use models based on numerical weather prediction (NWP).

Most studies do not cover a study horizon greater than 48 h [34], Aiming to fill this gap, this work will investigate the PV forecast for a horizon of one week ahead.

2.1. Forecasting methods

PV forecasting methods are generally classified as physical models or statistical models [36]. The physical models describe the physical state and characteristics of the generation plant, such as location, different meteorological variables [10], shadow effects, module type, and azimuth/tilt angle [11]. Statistical models consider data history, and attempt to extract knowledge from the past to forecast a time-series [32].

According to Sobri et al. [32], the statistical models include ANNs, SVMs, Chain Markov models, autoregressive models and regression models. The ANN has been extensively applied in many studies regarding non-linear time series problems (intrinsically associated with the prediction of PV generation) [37]. Even those studies that do not use neural networks in their analyses, mention that ANN is an important prediction technique [31].

2.2. Artificial neural network (ANN) ensemble

Neural networks are algorithms whose principle of functioning is inspired by the functioning of the brain [32]. Basically, the structure is composed of three layers: the input layer, hidden layer, and output layer [22]. Each layer has a specific number of nodes (or neurons) that interconnect one layer to another. The connection between the neurons is defined by weights, calculated iteratively in a training stage [28]. The networks have a random initialization in the training process [38], and learning is acquired based on the adjustment of weights, until an established criterion is reached [2], i.e., in a criterion to map the non-linearity between the inputs and the outputs [39].

A MLP network fully interconnects the layers, and is regulated based on supervised learning [40]. The modeling of a network with k inputs, only one output y and h hidden neurons is mapped by Equation (1) [33]:

$$y = y(x; w) = \sum_{j=0}^{h} \left[w_j f \left(\sum_{i=0}^{k} w_{ji} \cdot x_i \right) \right] \quad (1)$$

The weights and biases are represented by i and j, which interconnect the layers, and the network parameters are represented by the vector w. Considering that in most cases the neural network has a random initialization of weights and also has several parameters with a range of choices, we can obtain different results from each mapped configuration. The challenge is to determine which configuration (or set of them) promotes good model adjustments to the data.

An ensemble, initially proposed by Ref. [41], seeks to extract features from a model that, when combined, describes better results. In this work, the combination technique will be discussed in Sections 3.5 and 3.6.

2.3. Design of experiments (DOE) for ANN parameterization and time series factor selection

DOE is a tool that uses mathematical and statistical resources, and allows an analyst to understand certain phenomena that influence a desired process output [16]. This technique lets one build arrangements that allow the analyst to discover information regarding experimental setups that promote minor prediction errors.

At this point, the factors must be carefully selected with their respective levels. There are many possibilities to choose from, but observation is recommended as important in literature. The DOE approach allows for a small number of experiments to be performed in simulation process; therefore, its application becomes advantageous [42]. The bases for choosing each factor, as well as their respective levels, are detailed in the next section.

3. Proposed methodology

The proposed methodology is summarized in Fig. 1, and essentially uses DOE. The DOE factors are divided into two groups: ANN factors, and time series factors. The cluster analysis hierarchically selects forecast groups with the potential to compose a combined forecast. The mixture analysis applied to the cluster group tends to improve the prediction error decrease.

The next sections describe each step, starting from the historical generation series to the desired series of prediction. The forecast horizon considered was one week ahead.

3.1. Time Series Factors

A step that is commonly addressed in raw data pre-processing is a normalization method. The normalization process is an important step for machine learning algorithms. In general, this process aims to rescale or converts the original data set to a new standard. There is no consensus in the literature regarding the best normalization method, but some of them involve transforming the data using mean, standard deviation, maximum and minimum values. Some authors consider that, in addition to improving the accuracy of the forecasting algorithms, the normalization process also contributes positively to an improvement in performance [43]. In this case, the main goal is to reduce the magnitude, retain input correlations and keep the data on a scale that is close that of the neural network transfer functions. From this perspective, this work has adopted two methods that are commonly used in research of this nature.

The first method, usually called standardization, was applied in some studies focusing on PV forecasting [44], and scales data into a 0–1 interval by dividing every single observation "y_i" by the max value of time series "max(Y)". For each generation value "i", the corresponding normalized "$\widehat{y_i}$" is calculated using Equation (2):

$$\widehat{y_i} = \frac{y_i}{\max(Y)} \tag{2}$$

The second method, usually called feature scaling, is often applied for pre-processing input data [20]. It consists of, same as before, keeping data between range the 0 and 1 through the division of the observed value "y_i" minus the minimum value of the series "min(Y)", by the subtraction of the maximum value "max(Y)" and minimum value. The formula is shown in Equation (3):

$$\widehat{y_i} = \frac{y_i - \min(Y)}{\max(Y) - \min(Y)} \tag{3}$$

The use of external variables is a common method for improving the forecasting process, and should be considered when there is a strong correlation [13]. Meteorological data, such as temperature, irradiance, humidity, wind speed, hours of sunshine, cloud cover, and precipitation, are examples of these types of data. This measure can be positively or negatively correlated. The former indicates that, the series can increase or decrease proportionately, and, the second case indicates that one series can increase while another decreases (proportionally). This behaviour can be calculated using Pearson's correlation formula, and some other studies [19] has adopted this approach indeed:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Originally, data series are obtained with the observations ordered in time. For some authors [10], the use of similar days may contribute to improvements in the forecasting result. To investigate whether there is any relationship between PV productions on the same days of the week, this work propose to unstack the data considering each day of the week, resulting in seven separate data series. When the forecast is performed, if the data is unstacked, each time series is used as the input for the model, considering the specific day of the week. The unstacking process is summarized in Fig. 2.

The number of observations in the time series may vary depending on the measurement site, the condition of the equipment used, the storage devices, and the periodicity of the information collected. Considering this information, the methodology initially proposes the use of all available data for training the model, and at a second time, the use of a reduced part of this data history to evaluate the impacts on the forecast results.

3.2. ANN Factors

The layers that lie between the input layer and the output layer are

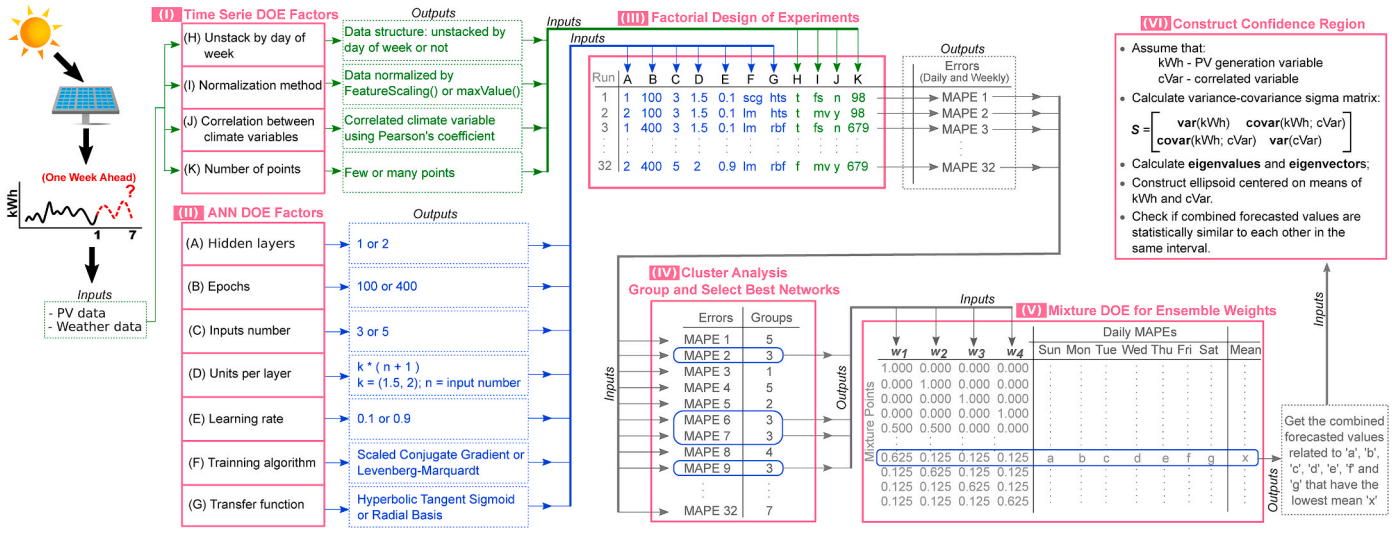


Fig. 1. Proposed methodology for photovoltaic (PV) generation forecast using artificial neural network (ANN), design of experiments (DOE), cluster analysis, and mixture DOE (MDE).

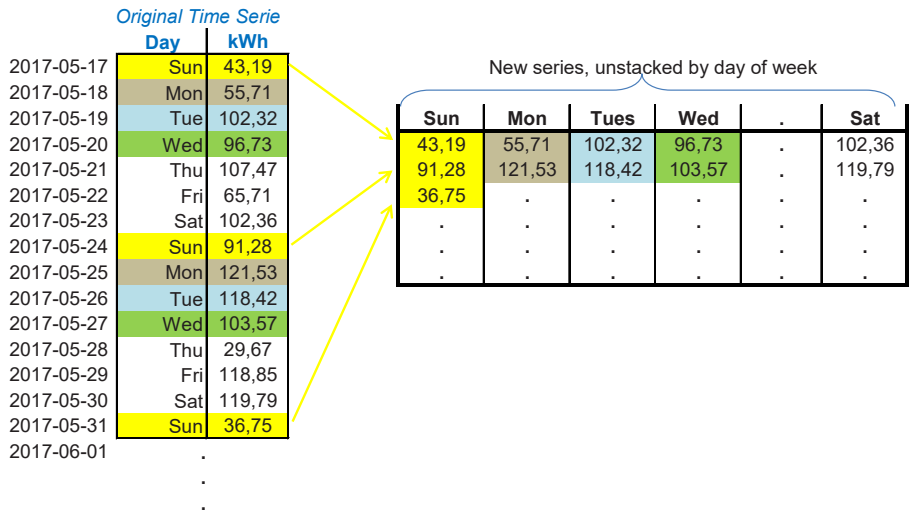


Fig. 2. Data Structure: sequential or unstacked by day of week.

known as the hidden layers. Each layer of an ANN has a specific number of neurons. Based on [45], the number of hidden layers chosen varies between one or two, with zero or three being rarely used.

Basically, each epoch represents a process of training the model with a set of data. The number of times the algorithm iterates over the same data to increase the accuracy of the output corresponds to the number of epochs. Evidently, increasing this number tends to increase the accuracy of the response, and consequently tends to increase the computational cost for processing the model.

This factor is related to the number of neurons in the input layer of the model. Depending on the data structure (as discussed before), training inputs and outputs can be organized from two different perspectives. The first one considers the data unstacked by the day of the week, while the other uses the data series in its original format.

For example, if a number of inputs is considered equal to three, the training configuration for the model will be as summarized in Fig. 3. In all cases, there will be only one output as the desired value.

As discussed in Ref. [46], the number of neurons in the intermediate layers plays an important role in prediction performance. Some authors consider defining the number of neurons randomly [47], whereas others consider it in a systematic way, such as [48].

Even knowing that there is no consensus in the literature for the definition of neural network architecture and that it depends on the nature of the problem being considered [49], this work considered the study of the nonlinear time series presented by Ref. [45], which used the following formula for the calculation of the number of neurons in the intermediate layers: $(K \times (N+1))$, where $\boldsymbol{N}$ is the number of inputs, and $K=1.5, 2$.

The learning rate can be considered as the step size for finding the solution of the problem and should be chosen carefully, as high values can lead to a fast convergence of the algorithm to sub-optimal solution points, and low values can lead to process stagnation in the search space.

As discussed in Ref. [45], the learning rate considered in this work may be 0.1 or 0.9.

Two algorithms were chosen to form the basis of the neural network training in the experiments carried out. The training algorithms are responsible for updating network weights and bias values [50]. The first algorithm, the scaled conjugate gradient (SCG), is often applied to the training process [51]. The second algorithm, Levenberg-Marquardt, was selected because it has been considered in the literature to have superior performance in for training feedforward neural networks [52].

Transfer (or activation) functions aim to compute the output of a layer from the data arriving from the immediately preceding layer. In this work, two functions were considered: radial basis [49] and hyperbolic tangent sigmoid [46]. Table 2 shows mathematical and graphical representations for each function.

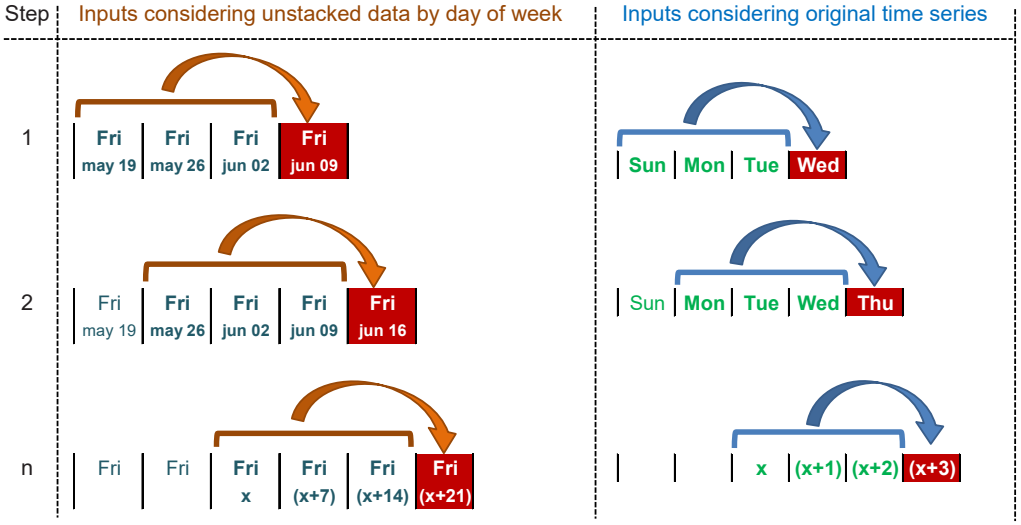
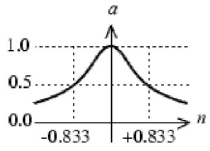
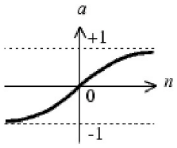


Fig. 3. ANN training steps considering unstacked and sequential data.

Table 2
Mathematical and graphical representation of transfer functions, according to Matlab® catalog.

Function Type	Mathematical Representation	Graphical Representation
Radial Basis	$f(x) = e^{-(\delta x)^2}$	
Hyperbolic Tangent Sigmoid	$f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	

3.3. Factorial DOE

The number of experimental runs increases exponentially as the number of factors increases. This can be a problem when there is a shortage of resources or time for examining all possible combinations.

The experimental matrix should be defined based on the factors and their respective levels. Thus, the amount of effort can be understood as the resolution to be employed in the experiment, and is defined by a predetermined design. Effort reduction refers to a fraction of the full factorial design, in cases where it is not possible to perform all combinations. In general, inferences can be made with a high level of confidence when performing fractional experiments. In Fig. 4, the red resolutions (identified by III) should be carefully considered, as the number of runs can be significantly too low.

In this simulated process, there are 11 factors (4 time series factors + 7 ANN factors) with two levels each. Thus, it would spend 2^k trials to test all combinations, where k corresponds to the number of factors, reaching a total of 2048 experimental runs for this analysis. This number is considerable for further analysis, and particularly if parameterization is implemented manually for each test.

To reduce the number of runs, DOE allows one to design the experiment with effort reduction without compromising the experimenter's inference analysis, through formula $2^{(k-p)}$, where p represents the effort reduction. Here the resolution IV is used, with 2^{11-6} effort reduction.

3.4. Cluster Analysis

The cluster analysis was performed to separate the best forecasting results (obtained by DOE) with similar characteristics, for further combination. It allows one to build a tree structure that interconnects information in groups through a linkage method, named the "dendrogram". The dendrogram is agglomerative, and is commonly designed with a bottom-up approach, where smaller clusters are grouped into larger clusters.

The grouping is based on Ward's linkage method and a Euclidean distance measurement. Ward's method seeks to minimize the sum of squared deviations internally in each cluster [29], from points to centroids (and therefore the variance between elements). Initially, each point is considered as a cluster and the sum of the squares of deviations is zero. As the clusters intersect, the sum of the squares of the deviations increases.

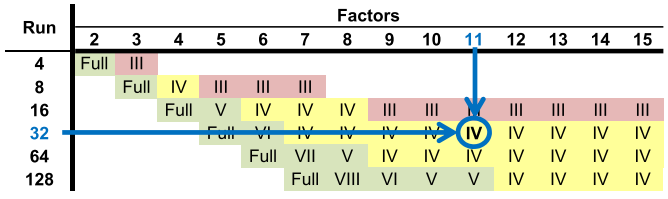


Fig. 4. Available factorial designs with respective resolution. Based on Minitab software.

The object of interest is the group with the smallest mean absolute percentage error (MAPE) values that have similarity. The significance of the differences between group pairs is statistically identified by a one-way analysis of variance, using Tukey's comparison as the primary procedure.

3.5. Mixture DOE

Combined forecasting tends to produce more accurate results. In this sense, this proposal seeks to achieve this combination through a mixture experiment. A mixture analysis considers the factors as "ingredients", and denotes that the proportions must be equal to 1 (one) [16]. The aim of the mixture is to find the weights that provide a prediction error smaller than the error obtained by the best individual prediction component.

For instance, w_1, w_2, ..., w_N, are the weights that make up the ensemble, where N is the number of components to be combined (previously selected through cluster analysis). Thus, $w_1 + w_2 + \ldots + w_N = 1$.

Fig. 5 (adapted from Ref. [16]) shows two simplex centroid designs. The first one (a) analyses the combination using three components, and $w_1 + w_2 + w_3 = 1$. The second one (b) considers four components, and $w_1 + w_2 + w_3 + w_4 = 1$. Each vertex of the triangle or tetrahedron is considered a "pure mixture", i.e., the ratio of the other components to that vertex is null. For example (considering the triangle), when $w_2 = 1$, w_1 and w_3 are automatically zero. The number of points is related to the number of components considered in simplex centroid design and is generalized by formula $(2^c - 1)$, where 'c' is the total number of components.

The ensemble proposed uses weights defined by the mixture analysis, as applied to the results selected by the cluster analysis. Thus, the following Equation (5) models mathematically this combination:

$$\hat{y}_i = \sum_{i=0}^{n} w_i . y_i \tag{5}$$

Here 'w' are the weights, and 'y' are the forecasted values chosen by the cluster analysis.

3.6. Confidence Region

A confidence region allows one to evaluate whether points within the ellipse follow the same characteristics as most of the data in the original set. In this case, a confidence region is built based on insolation data and generation power. Suppose that there is a variance-covariance matrix sigma (S). The following equation defines an ellipse (when the number of variables $p = 2$) centered on the mean ($\overline{x}$) with a constant distance c:

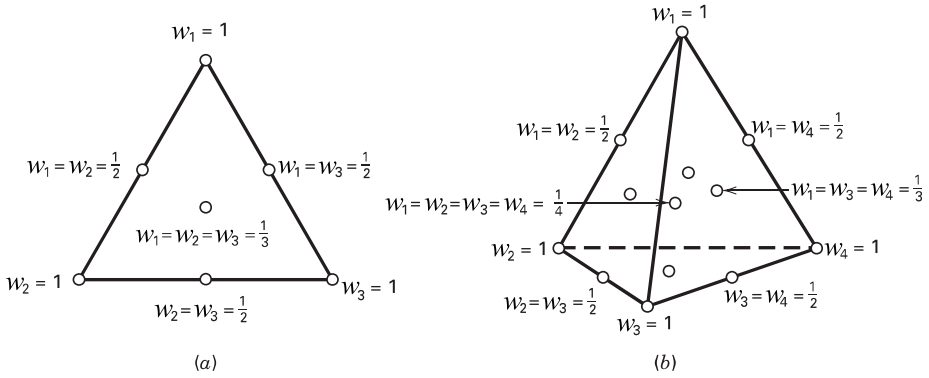


Fig. 5. Simplex design using three components (a) and using four components (b). Adapted from Ref. [16].

$$(x-\overline{x})^{'} S^{-1}(x-\overline{x}) = c^2 \tag{6}$$

A spectral decomposition of the rotated ellipse can be written as:

$$(x-\overline{x})^{'} P\Lambda^{-1} P^{'}(x-\overline{x}) = c^2 \tag{7}$$

In the above, Λ matrix is represented by the eigenvalues of S, denoted by:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{8}$$

The P matrix is defined by the eigenvectors of S:

$$P = \begin{bmatrix} cos\theta & -sen\theta \\ sen\theta & cos\theta \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \tag{9}$$

Thus, in these terms:

$$[P^{'}(x-\overline{x})]^{'} \Lambda^{-1/2}\Lambda^{-1/2}[P^{'}(x-\overline{x})] = c^2 \tag{10}$$

$$\sqrt{[P^{'}(x-\overline{x})]^{'} \Lambda^{-1/2}\Lambda^{-1/2}[P^{'}(x-\overline{x})]} = \sqrt{c^2} \tag{11}$$

$$\Lambda^{-1/2}[P^{'}(x-\overline{x})] = \sqrt{\chi^2} \tag{12}$$

$$[P^{'}(x-\overline{x})] = \sqrt{\chi^2}\Lambda^{1/2} \tag{13}$$

Assuming that $P$is orthonormal, $P^{-1} = P^{'}$:

$$\left[P^{-1}(x-\overline{x})\right] = \sqrt{\chi^2}\Lambda^{1/2} \tag{14}$$

$$(x-\overline{x}) = P\left[\sqrt{\chi^2}\Lambda^{1/2}\right] \tag{15}$$

```
1.  BEGIN
2.      doeMatrix ← load DOE Fractional Factorial Design;
3.      Create an array for each DOE factor;
4.      Fill each array with respective 2-level values;
5.      kwh ← load generation data;
6.      insolation ← load hours of insolation data;
7.      FOR run ← 1 TO numberOfRows( doeMatrix )
8.          Consider if data structure is unstacked or not according to doeMatrix;
9.          Normalize data according to doeMatrix;
10.         IF correlated variables should be applied DO
11.             trainingData ← insolation;
12.         ELSE
13.             trainingData ← kwh;
14.         END IF
15.         FOR day ← 1 TO 7
16.             prepareAnnNeurons( hiddenLayers, inputsNumber, unitsPerLayer );
17.             Create FeedFowardNet;
18.             Set learningRate, numberOfEpochs and trasnferFunction;
19.             trainANN( trainingData );
20.             Add forecastedValue to trainingData;
21.             Denormalize forecastedValue according to normalization method used;
22.             Calculate MAPE;
23.         END FOR
24.     END FOR
25.     Group results based on Ward's linkage method and Euclidean measurement;
26.     Create dendogram;
27.     Separate group with the smallest MAPEs and that have similarity;
28.     Create Mixture DOE using Simplex Centroid Design;
29.     Select the best combination weights that reduces total MAPE;
30.     Draw confidence ellipsoid;
31.     Check if the combined results belong to the confidence ellipsoid;
32. END
```

Fig. 6. Pseudocode of the proposed methodology in each instance presented before.

Table 3
Pearson's correlation coefficient analysis on exogenous variables. All the results are statistically significant with a p-value < 5%.

CORRELATIONS	*kWh*	*Cloudiness*	*Insolation*	*Temperature*	*Precipitation*
Cloudiness	−0.529				
Insolation	0.811	−0.759			
Temperature	0.723	−0.225	0.509		
Precipitation	−0.22	0.409	−0.372	−0.135	
Humidity	−0.683	0.639	−0.75	−0.424	0.331

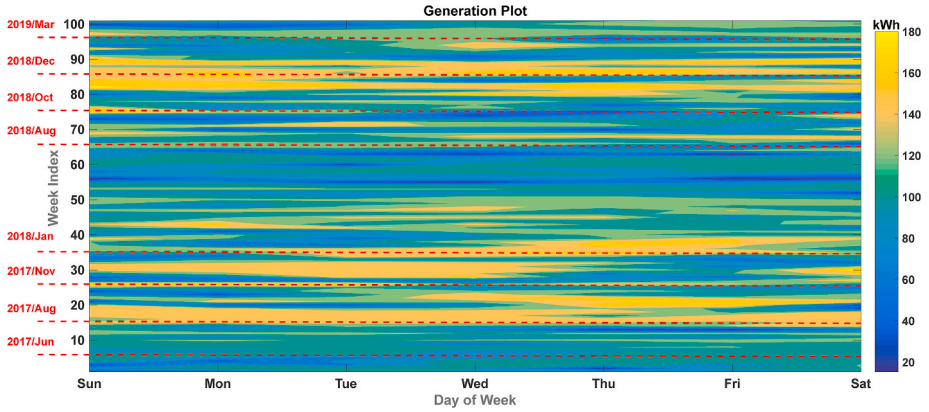


Fig. 7. Contour plot for weekly PV energy analysis of a generation site located in Minas Gerais, Brazil.

Table 4
ANN DOE factors with two levels each.

	(A) Hidden layers	(B) Epochs	(C) Inputs number	(D) Units per layer	(E) Learning rate	(F) Train function options	(G) Transfer function
Level (1)	1	100	3	1.5	0.1	{scg} Scaled Conjugate Gradient	{htg} Hyperbolic Tangent Sigmoid
Level (2)	2	400	5	2	0.9	{lm} Levenberg-Marquardt	{rbf} Radial Basis Function

$$x = \bar{x} + P\left[\sqrt{\chi^2}\Lambda^{1/2}\right] \tag{16}$$

The general equation can be written as:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} + \begin{bmatrix} c\sqrt{\lambda_1}h_{11}cos\propto - c\sqrt{\lambda_2}h_{12}sin\propto \\ c\sqrt{\lambda_1}h_{21}cos\propto + c\sqrt{\lambda_2}h_{22}sin\propto \end{bmatrix} \quad c = \sqrt{\chi^2_{(p,\propto/2)}};\ 0 < \propto < 2\pi \tag{17}$$

Finally, the pseudocode that summarizes the proposed methodology can be seen in Fig. 6. Forecasts using DOE were implemented using Matlab software. In this case, the DOE matrix is read from a text file, and its values work as vector indices representing the factors of the experiment. Each row of the DOE matrix represents an experimental run.

At this point in the paper, it is worth highlighting two important limitations: the first one is that, since DOE leads to a reduction in computational effort, there is no guarantee of obtaining the best forecasting solution (as well as no other forecasting method, presented in the literature, is able to do it [53]). The second one is that there are uncontrollable factors (such as dust on the panels, damaged meteorological sensors, unavailability of data in the time series, etc.) that were not considered in this study and that can cause variation in results from one plant to another.

However, the key to good results is to choose the levels of DOE factors in a systematic way. Even if there is a change from one place to another, the DOE allows an analysis of the interaction between the factors and how the relationship between the variables [54] interferes in the accuracy of the result. From this perspective, a forecasting system can be implemented to automatically recognize the factors that are negatively influencing the results. The factors discussed above are processed to predict one week ahead, and the MAPE is calculated. Therefore, using Minitab software, the cluster analysis and mixture DOE (MDE) are performed. The confidence ellipsoid is created using a formulation in Excel software.

4. Case study for a Brazilian site

The purpose of this study is to predict the PV generation for the horizon of one week ahead, using daily discretized data. There are 686 observations, covering the period from May 21, 2017 to April 6, 2019. The desired forecast week is the range from March 31 to April 6 (2019).

These data were collected from a solar plant with an installed generating capacity of approximately 35 kWh, located at the Federal Institute of Southern Minas Gerais State, city of Carmo de Minas, Brazil. The exogenous time series data available for this analysis included cloudiness, hours of insolation, temperature, precipitation, and humidity, and were obtained from National Meteorological Institute - Brazil (INMET). The choice of an exogenous variable for the preparation of the experiment was selected using Pearson's correlation coefficient.

Table 3 summarizes the correlations found, and points to the hours of insolation as the variable with the highest index, at 0.811. This variable, hours of insolation, will be used in the composition of an experimental factor, whose two levels indicate the presence of a correlated variable or not (only the generation data is used in this last case).

For this specific problem, the contour plot shown in Fig. 7, allows for a different way of analysing generation data, and for seeing trends more clearly than in the time series graph. This graph works as a frequency map, where the color denotes the level of electricity generation. The extremes are blue and yellow, where the first represents the minimum

Table 5
Time series DOE factors with two levels each.

	(H) Unstacked by day of week	(I) Normalization method	(J) Using correlated variable	(K) Number of points
Level (1)	{t} True	{fs} Feature Scaling	{n} No	98
Level (2)	{f} False	{mv} Max Value	{y} Yes	679

Table 6
Fractional factorial DOE and associated mean absolute percentage errors (MAPEs) for PV generation forecast considering one week ahead. It uses a 2^{11-6} design.

Run	ANN factors							Time series factors				MAPEs								Std
	A	B	C	D	E	F	G	H	I	J	K	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Mean	
1	1	100	3	1.5	0.1	scg	hts	t	fs	n	98	35.43	35.88	6.96	13.89	35.64	22.51	46.41	**28.10**	14.06
2	2	100	3	1.5	0.1	lm	hts	t	mv	y	98	26.04	26.52	22.85	27.95	16.42	67.13	42.83	**32.82**	17.10
3	1	400	3	1.5	0.1	lm	rbf	t	fs	n	679	10.52	32.90	15.73	20.21	24.44	20.48	4.51	**18.40**	9.28
4	2	400	3	1.5	0.1	scg	rbf	t	mv	y	679	21.09	44.03	13.57	8.28	17.43	28.85	47.19	**25.78**	14.99
5	1	100	5	1.5	0.1	lm	rbf	f	mv	n	98	11.94	12.38	17.76	11.04	18.72	2.81	3.62	**11.18**	6.18
6	2	100	5	1.5	0.1	scg	rbf	f	fs	y	98	13.22	19.40	12.67	21.79	11.08	40.82	41.96	**22.99**	13.14
7	1	400	5	1.5	0.1	scg	hts	f	mv	n	679	9.40	9.48	18.41	7.96	16.69	5.16	4.70	**10.26**	5.35
8	2	400	5	1.5	0.1	lm	hts	f	fs	y	679	3.96	46.86	38.57	49.79	36.61	73.20	74.60	**46.23**	24.09
9	1	100	3	2	0.1	scg	rbf	f	mv	y	679	4.00	5.81	0.08	8.18	1.34	25.09	26.10	**10.08**	10.93
10	2	100	3	2	0.1	lm	rbf	f	fs	n	679	0.50	0.78	8.10	4.55	17.70	1.18	2.09	**4.99**	6.22
11	1	400	3	2	0.1	lm	hts	f	mv	y	98	20.10	43.31	35.23	46.15	33.32	69.02	70.39	**45.36**	18.61
12	2	400	3	2	0.1	scg	hts	f	fs	n	98	4.23	0.87	7.81	1.53	12.10	10.42	10.14	**6.73**	4.53
13	1	100	5	2	0.1	lm	hts	t	fs	y	679	21.08	24.36	17.10	24.86	6.13	31.87	41.64	**23.86**	11.17
14	2	100	5	2	0.1	scg	hts	t	mv	n	679	1.47	16.27	21.31	42.11	29.68	24.04	22.10	**22.42**	12.39
15	1	400	5	2	0.1	scg	rbf	t	fs	y	98	26.50	18.87	10.58	17.90	4.66	37.76	44.31	**22.94**	14.25
16	2	400	5	2	0.1	lm	rbf	t	mv	n	98	16.65	1.45	5.12	21.58	16.72	3.79	29.81	**13.59**	10.50
17	1	100	3	1.5	0.9	scg	hts	f	fs	y	679	3.53	8.45	35.65	46.85	33.94	69.65	39.52	**33.94**	22.55
18	2	100	3	1.5	0.9	lm	hts	f	mv	n	679	3.60	1.57	12.33	11.29	23.88	7.34	8.86	**9.84**	7.30
19	1	400	3	1.5	0.9	lm	rbf	f	fs	y	98	21.07	19.43	12.69	21.81	11.10	40.85	41.99	**24.14**	12.49
20	2	400	3	1.5	0.9	scg	rbf	f	mv	n	98	10.79	8.29	7.72	8.07	1.91	30.40	5.24	**10.34**	9.27
21	1	100	5	1.5	0.9	lm	rbf	t	mv	y	679	15.61	26.62	19.12	29.46	3.50	60.79	55.17	**30.04**	20.90
22	2	100	5	1.5	0.9	scg	rbf	t	fs	n	679	21.06	16.15	24.55	20.20	25.12	36.45	3.22	**20.97**	10.07
23	1	400	5	1.5	0.9	scg	hts	t	mv	y	98	26.55	24.59	26.94	35.87	14.82	61.94	64.27	**36.43**	19.24
24	2	400	5	1.5	0.9	lm	hts	t	fs	n	98	2.08	18.23	8.26	51.24	14.78	12.52	1.45	**15.51**	16.95
25	1	100	3	2	0.9	scg	rbf	t	mv	n	98	0.69	1.04	0.07	0.43	22.73	6.75	10.41	**6.02**	8.36
26	2	100	3	2	0.9	lm	rbf	t	fs	y	98	23.57	23.11	15.59	25.07	12.57	40.40	40.82	**25.87**	11.03
27	1	400	3	2	0.9	lm	hts	t	mv	n	679	8.47	12.84	11.88	6.79	14.05	7.80	10.26	**10.30**	2.74
28	2	400	3	2	0.9	scg	hts	t	fs	y	679	3.15	10.35	13.69	51.31	5.32	39.40	35.71	**22.71**	19.08
29	1	100	5	2	0.9	lm	hts	f	fs	n	98	9.72	13.27	13.85	12.62	19.72	7.79	6.25	**11.89**	4.49
30	2	100	5	2	0.9	scg	hts	f	mv	y	98	18.79	43.25	35.17	46.11	33.26	68.95	70.31	**45.12**	18.89
31	1	400	5	2	0.9	scg	rbf	f	fs	n	679	22.17	17.78	22.42	16.13	23.51	3.03	2.24	**15.33**	9.07
32	2	400	5	2	0.9	lm	rbf	f	mv	y	679	17.25	11.01	4.70	13.41	3.52	30.78	29.66	**15.76**	10.96

generation, and the second represents the maximum generation. Along the horizontal axes, the day of the week can be seen, and along the vertical axes, the index of the week.

For instance, by the 20th week there is a yellow zone, which means that it was sunny and the generation was high. By the 60th week, there is low generation, visually identified by the blue color zone. The dotted lines, in red color, indicate the approximate year/month of that week.

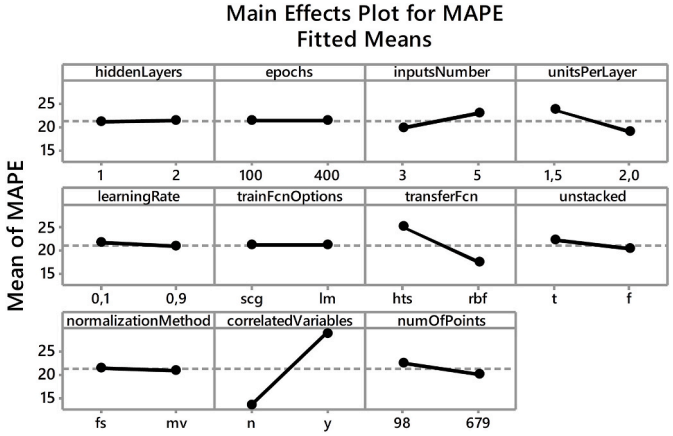


Fig. 8. Main effects for MAPE.

Through these lines, it is possible to identify the times of the year when the generation is higher or lower.

4.1. Pre-experimental planning

As there are 11 factors, divided into 7 neural network factors and 4 time series factors, the experiment was prepared with an effort reduction of 2^{11-6}. This means that only 32 experimental runs are required to solve this fractional factorial DOE, whose resolution is IV.

The ANN was parameterized and implemented using Matlab software. The ANN factors are listed in Table 4, along with their respective levels.

The time series factors were considered as a function of the data structure (unstacked or not), type of normalization, use of correlated variables (hours of insolation), and number of points (few or many). These levels are summarized in Table 5.

4.2. Forecasting data using ANN and DOE

The conduct of the experiments can be observed from Table 6. Here, the alphabetic identifications and respective levels are mapped from the previous two tables (Tables 4 and 5). In each experimental run, it is possible to observe the MAPEs of each day of the week, followed by the weekly average. The minimum average value found for the target week was 4.99%, and the maximum value was 46.23%. The last column

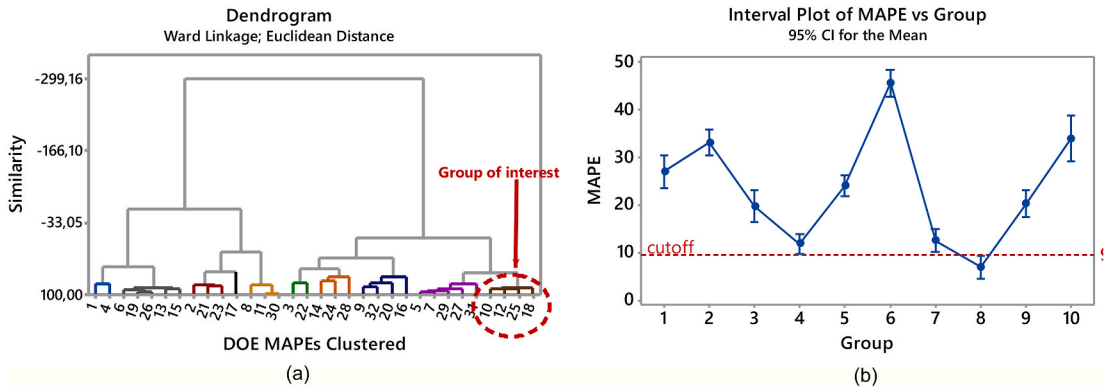


Fig. 9. (a) Dendrogram using Ward's linkage with Euclidean distance (b) One-way analysis of variance using Tukey's comparison procedure.

Table 7
DOE mixture for ensemble weights definition.

Mixture points (for each run number selected)				Individual MAPEs							Mean	Std
10 (*w1*)	12 (*w2*)	18 (*w3*)	25 (*w4*)	Sun	Mon	Tue	Wed	Thu	Fri	Sat		
1.000	0.000	0.000	0.000	0.50	0.78	8.10	4.55	17.70	1.18	2.09	4.99	6.22
0.000	1.000	0.000	0.000	4.23	0.87	7.81	1.53	12.10	10.42	10.14	6.73	4.53
0.000	0.000	1.000	0.000	3.60	1.57	12.33	11.29	23.88	7.34	8.86	9.84	7.30
0.000	0.000	0.000	1.000	0.69	1.04	0.07	0.43	22.73	6.75	10.41	6.02	8.36
0.500	0.500	0.000	0.000	2.36	0.83	7.95	3.04	14.90	4.62	4.02	5.39	4.74
0.500	0.000	0.500	0.000	2.05	0.40	10.21	7.92	20.79	4.26	5.48	7.30	6.81
0.500	0.000	0.000	0.500	0.59	0.91	4.01	2.06	20.22	2.78	4.16	4.96	6.87
0.000	0.500	0.500	0.000	3.91	0.35	10.07	6.41	17.99	1.54	0.64	5.84	6.39
0.000	0.500	0.000	0.500	2.46	0.96	3.87	0.55	17.42	8.58	10.28	6.30	6.15
0.000	0.000	0.500	0.500	2.14	0.27	6.13	5.43	23.30	0.29	0.77	5.48	8.22
0.333	0.333	0.333	0.000	2.78	0.03	9.41	5.79	17.89	0.63	0.27	5.26	6.54
0.333	0.333	0.000	0.333	1.80	0.90	5.28	1.88	17.51	5.33	6.15	5.55	5.67
0.333	0.000	0.333	0.333	1.59	0.08	6.78	5.14	21.44	0.59	0.18	5.11	7.66
0.000	0.333	0.333	0.333	2.84	0.11	6.69	4.13	19.57	3.28	3.90	5.79	6.38
0.250	0.250	0.250	0.250	2.25	0.28	7.04	4.23	19.10	2.16	2.40	5.35	6.42
0.625	0.125	0.125	0.125	1.37	0.53	7.57	4.39	18.40	0.49	0.15	4.70	6.62
0.125	0.625	0.125	0.125	3.24	0.58	7.42	2.88	15.60	6.29	6.27	6.04	4.85
0.125	0.125	0.625	0.125	2.93	0.65	9.68	7.76	21.49	2.59	3.23	6.90	7.17
0.125	0.125	0.125	0.625	1.47	0.66	3.48	1.90	20.92	4.46	6.40	5.61	7.03

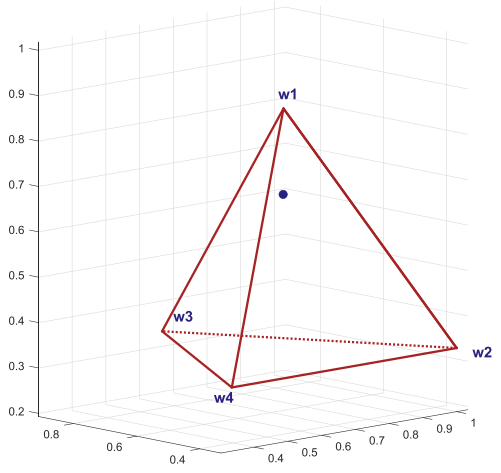


Fig. 10. Tetrahedron representing a configuration of weights that will make up the combination (*w1*, *w2*, *w3*, *w4*) = (0.625, 0.125, 0.125, 0.125).

presents the standard deviation for each run.

As this is a highly volatile data series (owing to the nature of the climate uncertainties associated with the generation process), there may be errors in the forecasting process that vary from week to week, and also from region to region.

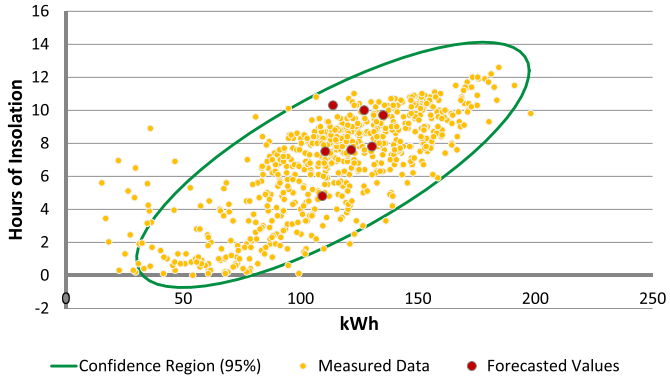


Fig. 11. Confidence ellipsoid for desired forecasting week.

From this numerical perspective, the configuration that contributed to the reduction of MAPE was: two hidden layers; 100 epochs; three neurons in the input layer; six neurons in the intermediate layers (corresponding to the calculation of $2 \times 3 = 6$); a learning rate equal to 0.1; the Levenberg-Marquadt training algorithm; RBF; the sequential data series (disregarding unstacked); a feature scaling normalization method; the use of the PV generation series only (disregarding the exogenous variable insolation); and the use of all available observations (679 measurements).

In contrast and still considering a numerical perspective, the configuration that contributed negatively (i.e., to the increase of MAPE) was as follows: two hidden layers; 400 epochs; five neurons in the input layer; eight neurons in the intermediate layers ($1.5 \times 5 = 7.5$, rounded to 8); a learning rate equal to 0.1; the Levenberg-Marquadt training algorithm; a tangent sigmoid hyperbolic transfer function; a sequential data series (disregarding unstacked); a feature scaling normalization method; the use of hours of sunshine as an exogenous variable in the network inputs; and the use of all observations available (679 measurements).

Only the numerical evaluation can induce one to misinterpret those factors that significantly impact the quality of the forecast. Therefore, the proposed methodology, based on DOE, allows to identify which factors are statistically significant in terms of promoting improvements in MAPE reduction. This inference is important to guide forecasts for different periods.

Through a graph of the main effects, we can see that two factors obtained p-values less than 5%: the transfer function, and the use of correlated variables (in this case, hours of insolation).

From Fig. 8 it is possible to see this happening; interpreting this chart is easy. Each box represents a factor and its variation, from level 1 to level 2. The more horizontal the line is, the stronger the indication that if there is a variation in the factor level, there will be no or little variability in the response (which in this case is the MAPE). A steeper the slope of the line indicates that this factor tends to impact the response when there are variations in its levels.

At this point, the predictions and their respective errors were calculated. The best forecasts aim to promote a combination that reduces the average MAPE for the desired week. Naturally, the choice of components to compose this combination, known as the ensemble, must be conducted systematically. From the dendrogram (Fig. 9a), it is easy to identify a group of interest, composed by four elements of the previous experimental runs: lines 10, 12, 18 and 25.

Statistical separation was performed based on Ward's linkage method, with a Euclidean measurement distance and 10 clusters or groups. A one-way analysis of variance using Tukey's comparison procedure (Fig. 9b) was performed to ensure that the groups were statistically different from each other, and that they had the lowest mean MAPE (numerically). In Fig. 9b, the cut line appears with a mean MAPE of approximately 9.5%. Group 8 of Fig. 9b corresponds to the group of interest from Fig. 9a.

Having chosen the group with the best forecasts, considered individually, it is expected that the combination of these results will lead to a reduction in the weekly mean MAPE. Based on this, the next section discusses the applied MDE-based strategy for ensemble formation.

4.3. Combining results

As discussed previously, the main motivation of this study is to construct a systematic combination of predicted values to decrease the prediction error. The combination technique was based on the definition of the weights that would provide the lowest MAPE, using the MDE. For this specific case, a group containing four elements was selected, representing the predicted values of experimental runs 10, 12, 18 and 25, which are *w1*, *w2*, *w3* and *w4*, respectively. Table 7 shows the mixture data.

Each line assumes a combination with a predetermined weight. In this case, it should be noted that the first four lines of this mixture design consider only the individual forecast, weighing 1, and zero for the others.

From Table 7 it is possible to verify that the weights setting that reduces the average MAPE has a larger effect for the first forecast component, indicated by *w1*. Thus, the geometric representation of this weight distribution is shown in Fig. 10, where there is the formation of a tetrahedron. The positioning coordinates of the ideal ensemble point transcribe as (*w1*, *w2*, *w3*, *w4*) = (0.625, 0.125, 0.125, 0.125). In this case, the mean MAPE dropped to 4.70%.

4.4. Confidence region

The confidence region allows one to verify that the data are statistically similar to each other in the same interval. Points found outside the region delimited by the ellipse do not have the same characteristics, and can be considered statistically different. In this sense, it is aimed to check if the forecast results belong to the ellipsoid. From Fig. 11, the points highlighted in red correspond to the combined forecast, and are within the limits of the ellipsoid.

For ellipsoid construction, the following variance-covariance sigma matrix based on generation and insolation data, and centered on means of 113.8 (kWh) and 6.7 (hours of insolation) was used:

$$S = \begin{bmatrix} 1167.35 & 79.50 \\ 79.50 & 9.21 \end{bmatrix} \tag{18}$$

The eigenvalues were calculated for $\lambda 1 = 1172.78$ and for $\lambda 2 = 3.78$. The associated eigenvectors of *S*were stored in the following matrix:

$$P = \begin{bmatrix} e1 & e2 - 0.068 \\ 0.998 & 0.998 \\ 0.068 & \end{bmatrix} \tag{19}$$

Thus, the research was successful in its goals. The next section lists the final conclusions.

5. Conclusions

In this paper, a methodology for PV generation prediction was proposed. Although the literature reveals a variation in the prediction accuracy and there is no consensus on a generic prediction method that meets all cases, the forecasting area is a constant target of studies.

The originality of the proposed methodology reduces the number of simulated experimental runs through the fractional factorial, and allows the analyst to infer regarding decisions to be made with a high level of confidence. The versatility of the proposed method allows changing the number of factors to be used in the experimental arrangement, the forecast model, and the desired forecast horizon.

The penetration of PV generation in power grids has intensified, and the precision of power forecasting promotes system reliability, in addition to allowing energy management efficiency. The predictive technique chosen, the ANN, is widely used in the literature for this purpose. The ANN had its parametric configuration systematically defined based on DOE, with the experimental design 2^{11-6}.

The implementation of this methodology, when there is interest, should take into account some considerations empirically related to non-controllable factors, which must be carefully observed: deposit of residues on the panels, such as dust; damaged weather sensors (or even when there is no proximity to the generation plant); unavailability of information or data (null values - this compromises learning and the accuracy of forecasting models); intensity and frequency of cloud cover (there are indications that some types of photovoltaic panels produce more energy when the cloud moves away due to the temporary cooling of the cells during the shading period); wind speed (the intensity of wind can dissipate heat, which increases the efficiency of the panels). These factors can interfere with the forecasting result, naturally, since it leads to represent levels of photovoltaic generation that do not match the expected.

Apart from that, DOE allows the analysis of which variables interfere in the forecast result, which can be systematically changed. From this point on, a forecasting system can be implemented to automatically recognize the factors that are negatively influencing the results.

Further works can evaluate different forecasting horizons, different factors, and different forecasting models. Integration with weather variables can be investigated by considering the extraction of common features through principal component analysis, rather than just selecting

the one with the highest correlation.

CRediT author statement

Moreira, M. O.: Conceptualization, Writing - Original Draft, Software Balestrassi, P. P.: Conceptualization, Supervision Paiva, A. P.: Conceptualization, Formal analysis Ribeiro, P. F.: Resources, Bonatto, B. D.: Methodology, Validation.

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.

Acknowledgements

The authors would like to thank the Brazilian Federal Institute of South of Minas Gerais State (IFSULDEMINAS), CAPES, CNPq, and INERGE for supporting this research.

References

[1] Cuk V, et al. Considerations on the modeling of photovoltaic systems for grid impact studies. In: Proceedings of 1st International Workshop on Integration of Solar Power into Power Systems; 2011. p. 35–42. Aarhus, Denmark.
[2] Monteiro RVA, GuimarÃ£es GC, Moura FAM, Albertini MRMC, Albertini MK. Estimating photovoltaic power generation: performance analysis of artificial neural networks, Support Vector Machine and Kalman filter. Elec Power Syst Res 2017;143:643–56. https://doi.org/10.1016/j.epsr.2016.10.050.
[3] Sun S, Wang S, Zhang G, Zheng J. A decomposition-clustering-ensemble learning approach for solar radiation forecasting. Sol Energy 2018;163:189–99. https://doi.org/10.1016/j.solener.2018.02.006.
[4] Agoua XG, Girard R, Kariniotakis G. Short-term spatio-temporal forecasting of photovoltaic power production. IEEE Trans. Sustain. Energy 2018;9(2):538–46. https://doi.org/10.1109/TSTE.2017.2747765.
[5] Michaelson D, Mahmood H, Jiang J. A predictive energy management system using pre-emptive load shedding for islanded photovoltaic microgrids. IEEE Trans Ind Electron 2017;64(7):5440–8. https://doi.org/10.1109/TIE.2017.2677317.
[6] Dall'Anese E, V Dhople S, Johnson BB, Giannakis GB. Optimal dispatch of residential photovoltaic inverters under forecasting uncertainties. IEEE J. Photovoltaics Jan. 2015;5(1):350–9. https://doi.org/10.1109/JPHOTOV.2014.2364125.
[7] Lorenz E, Hurka J, Heinemann D, Beyer HG. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2009;2(1):2–10. https://doi.org/10.1109/JSTARS.2009.2020300.
[8] Zhang Y, Beaudin M, Taheri R, Zareipour H, Wood D. Day-ahead power output forecasting for small-scale solar photovoltaic electricity generators. IEEE Trans. Smart Grid 2015;6(5):2253–62. https://doi.org/10.1109/TSG.2015.2397003.
[9] Huang C, Chen S, Yang S, Kuo C. One-day-ahead hourly forecasting for photovoltaic power generation using an intelligent method with weather-based forecasting models. IET Gener Transm Distrib 2015;9(14):1874–82. https://doi.org/10.1049/iet-gtd.2015.0175.
[10] Yang M, Huang X. Ultra-short-term prediction of photovoltaic power based on periodic extraction of PV energy and LSH algorithm. IEEE Access 2018;6:51200–5. https://doi.org/10.1109/ACCESS.2018.2868478.
[11] Zhang X, Li Y, Lu S, Hamann HF, Hodge B, Lehman B. A solar time based analog ensemble method for regional solar power forecasting. IEEE Trans. Sustain. Energy Jan. 2019;10(1):268–79. https://doi.org/10.1109/TSTE.2018.2832634.
[12] Lee W, Kim K, Park J, Kim J, Kim Y. Forecasting solar power using long-short term memory and convolutional neural networks. IEEE Access 2018;6:73068–80. https://doi.org/10.1109/ACCESS.2018.2883330.
[13] Das UK, et al. Forecasting of photovoltaic power generation and model optimization: a review. Renew Sustain Energy Rev 2018;81:912–28. https://doi.org/10.1016/j.rser.2017.08.017.
[14] Ackermann T, et al. Smart modeling of optimal integration of high penetration of PV-Smooth PV," Final Rep. Smooth PV Proj. under PV ERA NET Call. 2013.
[15] Bracale A, Carpinelli G, De Falco P. A probabilistic competitive ensemble method for short-term photovoltaic power forecasting. IEEE Trans. Sustain. Energy 2017;8(2):551–60. https://doi.org/10.1109/TSTE.2016.2610523.
[16] Montgomery DC. Design and analysis of experiments. John wiley & sons; 2017.
[17] Galicia A, Talavera-Llames R, Troncoso A, Koprinska I, Martínez-Álvarez F. Multi-step forecasting for big data time series based on ensemble learning. Knowl Base Syst 2019;163:830–41. https://doi.org/10.1016/j.knosys.2018.10.009.
[18] Zhen Z, et al. Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image. IEEE Trans Ind Appl 2020;56(4):3385–96.
[19] Theocharides S, Makrides G, Livera A, Theristis M, Kaimakis P, Georghiou GE. Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. Appl Energy 2020;268:115023. https://doi.org/10.1016/j.apenergy.2020.115023.
[20] Sangrody H, Zhou N, Zhang Z. Similarity-based models for day-ahead solar PV generation forecasting. IEEE Access 2020;8:104469–78. https://doi.org/10.1109/ACCESS.2020.2999903.
[21] Pan C, Tan J. Day-ahead hourly forecasting of solar generation based on cluster Analysis and ensemble model. IEEE Access 2019;7:112921–30.
[22] Wen L, Zhou K, Yang S, Lu X. Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting. Energy 2019;171:1053–65. https://doi.org/10.1016/j.energy.2019.01.075.
[23] Ozoegwu CG. Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number. J Clean Prod 2019;216:1–13. https://doi.org/10.1016/j.jclepro.2019.01.096.
[24] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy 2018;148:461–8. https://doi.org/10.1016/j.energy.2018.01.177.
[25] Bugała A, et al. Short-term forecast of generation of electric energy in photovoltaic systems. Renew Sustain Energy Rev 2018;81:306–12. https://doi.org/10.1016/j.rser.2017.07.032.
[26] Deo RC, Şahin M. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renew Sustain Energy Rev 2017;72:828–48. https://doi.org/10.1016/j.rser.2017.01.114.
[27] Sivaneasan B, Yu CY, Goh KP. Solar forecasting using ANN with fuzzy logic pre-processing. Energy Procedia 2017;143:727–32. https://doi.org/10.1016/j.egypro.2017.12.753.
[28] Cervone G, Clemente-Harding L, Alessandrini S, Monache LD. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. Renew Energy 2017;108:274–86. https://doi.org/10.1016/j.renene.2017.02.052.
[29] Lima FJL, Martins FR, Pereira EB, Lorenz E, Heinemann D. Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. Renew Energy 2016;87:807–18. https://doi.org/10.1016/j.renene.2015.11.005.
[30] Amrouche B, Le Pivert X. Artificial neural network based daily local forecasting for global solar radiation. Appl Energy 2014;130:333–41. https://doi.org/10.1016/j.apenergy.2014.05.055.
[31] de Freitas Viscondi G, Alves-Souza SN. A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting. Sustain. Energy Technol. Assessments 2019;31:54–63. https://doi.org/10.1016/j.seta.2018.11.008.
[32] Sobri S, Koohi-Kamali S, Rahim NA. Solar photovoltaic generation forecasting methods: a review. Energy Convers Manag 2018;156:459–97. https://doi.org/10.1016/j.enconman.2017.11.019.
[33] Barbieri F, Rajakaruna S, Ghosh A. Very short-term photovoltaic power forecasting with cloud modeling: a review. Renew Sustain Energy Rev 2017;75:242–63. https://doi.org/10.1016/j.rser.2016.10.068.
[34] Antonanzas J, Osorio N, Escobar R, Urraca R, Martinez-de-Pison FJ, Antonanzas-Torres F. Review of photovoltaic power forecasting. Sol Energy 2016;136:78–111. https://doi.org/10.1016/j.solener.2016.06.069.
[36] van der Meer D, Chandra Mouli GR, Morales-España Mouli G, Elizondo LR, Bauer P. Energy management system with PV power forecast to optimally charge EVs at the workplace. IEEE Trans. Ind. Informatics Jan. 2018;14(1):311–20. https://doi.org/10.1109/TII.2016.2634624.
[37] Wang H, Shen J. An improved model combining evolutionary algorithm and neural networks for PV maximum power point tracking. IEEE Access 2019;7:2823–7. https://doi.org/10.1109/ACCESS.2018.2881888.
[38] López M, Valero S, Rodriguez A, Veiras I, Senabre C. New online load forecasting system for the Spanish Transport System Operator. Elec Power Syst Res 2018;154:401–12. https://doi.org/10.1016/j.epsr.2017.09.003.
[39] Vagropoulos SI, Kardakos EG, Simoglou CK, Bakirtzis AG, Catalão JPS. ANN-based scenario generation methodology for stochastic variables of electric power systems. Elec Power Syst Res 2016;134:9–18. https://doi.org/10.1016/j.epsr.2015.12.020.
[40] Akhter MN, Mekhilef S, Mokhlis H, Mohamed Shah N. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. IET Renew Power Gener 2019;13(7):1009–23.
[41] Hansen LK, Salamon P. Neural network ensembles. Pattern Anal. Mach. Intell. IEEE Trans. 1990;12:993–1001. https://doi.org/10.1109/34.58871.
[42] Pontes FJ, Amorim GF, Balestrassi PP, Paiva AP, Ferreira JR. Design of experiments and focused grid search for neural network parameter optimization. Neurocomputing 2016;186:22–34. https://doi.org/10.1016/j.neucom.2015.12.061.
[43] Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl Soft Comput 2019:105524. https://doi.org/10.1016/j.asoc.2019.105524.
[44] Appiah AY, Zhang X, Ayawli BBK, Kyeremeh F. Long short-term memory networks based automatic feature extraction for photovoltaic array fault diagnosis. IEEE Access 2019;7:30089–101. https://doi.org/10.1109/ACCESS.2019.2902949.
[45] Balestrassi PP, Popova E, Paiva AP, Lima JWM. "Design of experiments on neural network's training for nonlinear time series forecasting. Neurocomputing 2009;72(4):1160–78. https://doi.org/10.1016/j.neucom.2008.02.002.
[46] Xiao W, Nazario G, Wu H, Zhang H, Cheng F. A neural network based computational model to predict the output power of different types of photovoltaic cells. PloS One 2017;12(9):1–8. https://doi.org/10.1371/journal.pone.0184561.
[47] Yeon S, Yu B, Seo B, Yoon Y, Lee KH. ANN based automatic slat angle control of Venetian blind for minimized total load in an office building. Sol Energy 2019;180:133–45. https://doi.org/10.1016/j.solener.2019.01.027.

[48] Kůrková V. "Kolmogorov's theorem and multilayer neural networks. Neural Network 1992;5(3):501–6. https://doi.org/10.1016/0893-6080(92)90012-8.
[49] Mittal M, Bora B, Saxena S, Gaur AM. Performance prediction of PV module using electrical equivalent model and artificial neural network. Sol Energy 2018;176:104–17. https://doi.org/10.1016/j.solener.2018.10.018.
[50] Al-Majidi SD, Abbod MF, Al-Raweshidy HS. A particle swarm optimisation-trained feedforward neural network for predicting the maximum power point of a photovoltaic array. Eng Appl Artif Intell 2020;92:103688. https://doi.org/10.1016/j.engappai.2020.103688.
[51] Amaral RPF, V Ribeiro M, de Aguiar EP. Type-1 and singleton fuzzy logic system trained by a fast scaled conjugate gradient methods for dealing with binary classification problems. Neurocomputing 2019;355:57–70. https://doi.org/10.1016/j.neucom.2019.05.002.
[52] Pertl M, Douglass PJ, Heussen K, Kok K. Validation of a robust neural real-time voltage estimator for active distribution grids on field data. Elec Power Syst Res 2018;154:182–92. https://doi.org/10.1016/j.epsr.2017.08.016.
[53] Abbasimehr H, Shabani M, Yousefi M. An optimized model using LSTM network for demand forecasting. Comput Ind Eng 2020;143:106435. https://doi.org/10.1016/j.cie.2020.106435.
[54] Lira JOB, Riella HG, Padoin N, Soares C. CFD + DoE optimization of a flat plate photocatalytic reactor applied to NOx abatement. Chem. Eng. Process. - Process Intensif. 2020;154:107998. https://doi.org/10.1016/j.cep.2020.107998.