Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability

Alexis Fouilloy, Cyril Voyant, Gilles Notton, Fabrice Motte, Christophe Paoli, Marie-Laure Nivet, Emmanuel Guillot, Jean-Laurent Duchaud

Solar irradiation prediction with machine learning: forecasting models selection method depending on weather variability

Alexis Fouilloy[1]*, Cyril Voyant[2,3], Gilles Notton[1], Fabrice Motte[1], Christophe Paoli[1], Marie-Laure Nivet[1] , Emmanuel Guillot[4], Duchaud Jean-Laurent[1]

[1]University of Corsica, CNRS UMR SPE 6134, Centre Georges Péri, Route des Sanguinaires, 20000 Ajaccio, France

[2]Castelluccio Hospital, Radiotherapy Unit, BP 85, 20177 Ajaccio, France

[3]University of La Réunion - PIMENT Saint-Pierre, Réunion

[4]PROMES-CNRS Laboratory, 7 Rue du Four solaire, 66120 Font-Romeu-Odeillo-Via, France

fouilloy_a@univ-corse.fr; cyrilvoyant@gmail.com; notton@univ-corse.fr;motte_f@univ-corse.fr; cpaoli@univ-corse.fr; nivet@univ-corse.fr; duchaud_jl@univ-corse.fr; emmanuel.guillot@promes.cnrs.fr

Abstract: Eleven statistical and machine learning tools are analyzed and applied to hourly solar irradiation forecasting for time horizon from 1 to 6 hours. A methodology is presented to select the best and most reliable forecasting model according to the meteorological variability of the site. To make the conclusions more universal, solar data collected in three sites with low, medium and high meteorological variabilities are used: Ajaccio, Tilos and Odeillo. The datasets variability is evaluated using the mean absolute log return value. The models were compared in term of normalized root mean square error, mean absolute error and skill score. The most efficient models are selected for each variability and temporal horizon: for the weak variability, auto-regressive moving average and multi-layer perceptron are the most efficient, for a medium variability, auto-regressive moving average and bagged regression tree are the best predictors and for a high one, only more complex methods can be used efficiently, bagged regression tree and the random forest approach.

Keywords: Time Series forecasting, machine learning, variability, ARMA, ANN, Regression tree, Gaussian process, SVR

* Corresponding author: Alexis FOUILLOY, email: fouilloy_a@univ-corse.fr

1. Introduction

The management of an electrical network is a very complex task, particularly on small grids like islands which are generally not interconnected to the mainland grid (Diagne, David, Lauret, Boland, & Schmutz, 2013). The equality between production and consumption is the most difficult challenge; moreover, the increased rate of intermittent and stochastic renewable energy sources into the energy mix adds a layer of difficulty. For an efficient management of the energy mix and a safest electricity supply, production and consumption must be planned before. There are many forecasting methods for solar radiation, all the methods have advantages and inconvenient and their accuracy depends on the forecasting horizon and on the geographical situation. The three main categories of models are sky imaging, physical models (or numerical weather prediction, e.g. NWP) and machine learning models. The sky imaging models are mainly used for short term forecasting (1 to 60 minutes), the physical models concern numerical weather prediction models often used for long term forecasting (several days). The medium-term forecasting (from 1 to few hours) is often made by machine learning methods, a large state of the art was made concerning the solar power forecasting (Voyant, Notton, et al., 2017). This study is related to the comparison of eleven machine learning models in order to forecast solar global irradiation and to compare their performances in three different sites to make our results more universal. The main goal is to propose a prioritization methodology to highlight the best forecasting method between the tested ones according to the level of irradiation variability considering the complexity of the method. To distinguish the different weather conditions and to characterize the global behavior of the solar time series in the three sites, a property of time series named variability is calculated, the characterization of this time series property can be made by various statistical parameters (Voyant, Soubdhan, Lauret, David, & Muselli, 2015). The solar forecasting is realized with an hourly time granularity for 1 hour to 6 hours' time

horizons. The comparison between the models is evaluated by statistical indexes and classical error metrics, the eleven tested models are:

- Persistence (P) and smart persistence (SP) can be considered as naïve models, they are often used in intraday solar radiation forecasting and are usually taken as reference to compare the performances with other models (Lauret, Voyant, Soubdhan, David, & Poggi, 2015).
- Auto regressive moving average (ARMA) and artificial neural network (ANN) are two types of models allowing to realize respectively linear and non-linear regressions. ARMA is certainly the most studied tool especially in econometry, but also in global irradiation forecasting (Boland, David, & Lauret, 2016). Artificial neural network (ANN) is taken as well-known global radiation forecasting method (Voyant, Motte, et al., 2017).
- Regression tree based models family is composed by five models with different levels of complexity and optimization. Classical regression tree (RT) and pruned regression tree (RT-pruned) are sometimes used in very short term forecasting and allow good results compared with other machine learning models (Troncoso, Salcedo-Sanz, Casanova-Mateo, Riquelme, & Prieto, 2015). Boosted regression tree (RT-boosted) consists in combine the responses of a collection of weak classifiers which once averaged form a strong classifier to make prediction (De'ath, 2007). Then the aggregated versions of regression trees denoted bagged regression tree (RT-bagged) (Breiman, 1996) are developed to build an aggregated predictor in order to improve accuracy of the forecast. Finally, we propose also the random forest method (RF) which is another improvement of regression tree methods with more robustness (Breiman, 2001).

- Models based on Kernel estimation like Gaussian processes (GP) and support vector regression (SVR). The first one is a nonparametric kernel-based probabilistic model giving very good results in time series prediction (Rasmussen, 2004), The second one (SVR) can be considered as a generalization of linear classifiers and can be used to forecast solar irradiation and can be improved by several methods of optimization (Jiang & Dong, 2017).

The study is developed as follow: section 2 describes the materials and methods, the data and the models. Section 3 presents the results and the discussion about the forecasting reliability, the variability of solar data and the link between variability and efficiency of the forecasting models.

2. Materials and methods

This section presents information related to the preliminary stage before the forecasting phase. Fig 1 shows the different steps (detailed afterwards) of this methodology.

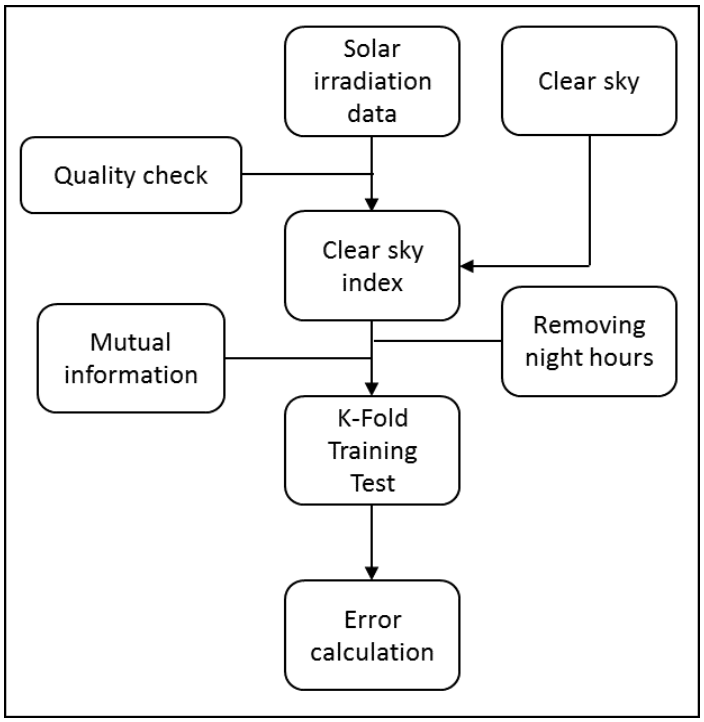


Figure 1. Flowchart of the model development before forecasting

It consists in:

1. A quality control of the solar data: often, mistakes appear in the temporal series of solar data due to problems with the acquisition system; an automatic quality check used in the frame of GEOSS project (Group on Earth Observation System of System) (http://www.earthobservations.org/geoss.php) has been applied to the data;
2. A data preprocessing: the night hours are removed, due to problems occurring at the sunset and the sunrise (mask effects and bad response of pyranometers) (Badescu, 2008; Iqbal, 1983), a filter is applied on the datasets which remove all the data that correspond to a solar elevation angle up to 10°.
3. Stationarity: the solar irradiation time series contains some properties like seasonality and periodicity. Most of the machine learning forecasting methods are applicable only to stationary time series (Hornik, Stinchcombe, & White, 1989) describe the necessary steps in order to use data in forecasting with machine learning. Thus, the solar data time series must be made stationary using a solar radiation model by clear sky and computing the clear sky index (paragraph 2.1), this method is described in (Paoli, Voyant, Muselli, & Nivet, 2010).
4. At last, the choice of the number of input data for each forecasting tool is realized by auto mutual information method (Luo, Shi, Zheng, Gang, & Cai, 2017) and described in paragraph 2.2.

2.1 The clear sky model (CS) and clear sky index (CSI)

The solar radiation by clear sky computing is the third step of the preprocessing method, it consists in calculate the maximum global horizontal irradiation (GHI) at the ground level in cloudiness conditions (taking into account scattering and absorption by atmosphere). Almost

all the machine learning approaches are based on a stationary hypothesis of the time series. they are usually based on the assumption that the data generation mechanism does not change over time. For the global irradiation forecasting in dividing each measured solar irradiation by its corresponding clear sky value, the periodicity of the solar irradiation time series is then removed; thus, a new time series is created with values between 0 and 1 and the forecasting method is then applied to predict only the cloud occurrences and not the fact that GHI is more important at 12h than at 6h, or in summer than in winter. The CS model chosen is the SOLIS model developed by Mueller (Mueller et al., 2004) and based on the studies of Ineichen et al. (Ineichen, 2006, 2008); it gives excellent results when it was compared with solar measurements in Europe. It is based on radiative transfer models, Beer Lambert function, and their integration on the solar spectra uses the information's of news satellites ERS-2/ENVISAT. The calculation of the clear sky irradiation CS(t) is given by:

$$CS(t) = H_0 . e^{\frac{-\tau}{sin^b(h(t))}} . sin(h(t)) \quad (1)$$

h(t) is the solar height in degrees and H_0 is the extraterrestrial irradiation (irradiation on the top of the atmosphere before being scattered and absorbed by the atmosphere (Iqbal, 1983)); τ is the global total atmospheric depth and the parameter b is a fitting parameter, both of them depending on the meteorological characteristics of the site, Mueller and Ineichen study the different parameters used in modelling of clear sky (Ineichen, 2008; Mueller et al., 2004). τ and b depend on the site and their values can be found in (https://aeronet.gsfc.nasa.gov/).

The clear sky index (CSI) is calculated by:

$$CSI(t) = \frac{GHI(t)}{CS(t)} \quad (2)$$

All the machine learning tools described in the next section can be only applied to stationary time series. CSI is forecasted from previous CSI values and GHI is then obtained using Eq. 2; the errors metrics are given in term of GHI in absolute and relative values.

2.2. Input matrix dimension of the forecasting tool and k-fold sampling

The next step of the preprocessing is to choose the dimension of the input matrix of the forecasting tools. The data are presented under a time series (TS) formalism, a time series is a series of data points indexed in function of the time ($CSI(t)$). The approach consists in predicting the future clear sky index (at different time horizons) from past observed data. Mathematically, the formulation is:

$$CSI(t+h) = f(CSI(t), CSI(t-1), \dots, CSI(t-n)) + \epsilon(t+h) \qquad (3)$$

$\epsilon(t+h)$ is a random white noise and *f* depending on the model.

The future time step (t+h) is forecasted from a given number of observed data at previous times (t, t-1…, t-n) which must be optimally determined. In other words, the objective is to calculate the value of n (number of previous values) and to obtain ϵ as low as possible. The choice of n, i.e. the dimension of the input matrix, is made by an auto mutual information method, in (Luo et al., 2017; Parviz, Nasser, & Motlagh, 2008) some details concerning this method. This auto mutual information is a property of a time series and depends on each dataset. It determinates the degree of statistical dependence of the variables. The lag corresponding to the first minimum of this parameter corresponds to the best n to consider.

In machine learning method, the times series is divided into two groups: a training and a testing group. The first one is used to train the model i.e. inputs and output are given to the model and a training algorithm determines the parameter values of the model; then, once these parameters known, the model is tested on a second set of data called testing set (never used for the training) and the error metrics are calculated only on this testing set.

Sometimes, the model is trained on a given percentage of first data and tested on the complementary percentage of last data. But when this data repartition, there exist the risk that the model was trained on a period for which some specific meteorological phenomena occur,

conducing to a non-generalizable model. In view to avoid this problem and to make the training less dependent of particular meteorological periods, the k-fold method is used during the validation phase; it consists in dividing the dataset in k samples (here k=10), each sample is used at least one time for the training and one time for the test, and this process is repeated as many time as necessary. Thus, the results are independent of the set of data used for the training because using only one data set (with its own statistical particularities) can reduce the robustness of the conclusions.

Note that the k-fold cross-validation is only used during the retrospective model's comparison but never in operational mode or prediction in real installation. In a retrospective context, the concept of present, future and past data does not make sense…all the data are part of the past. What is essential is that the data used for training are never used during the test. The k-fold methodology induces that the training set is not defined "in the past" with respect to the testing set. Testing and training data are interleaved.

2.3. Meteorological stations and data

The datasets are time series of horizontal global solar irradiation measurements (GHI) in three different sites (Fig 2).

Figure 2. Geographical situation of experimental sites

The first dataset was provided by PROMES laboratory (CNRS UPR 8521) located in south of France at Odeillo (Pyrénées Orientales, France, 42°29 N, 2°01 E, 1650 m asl) (red circle in Fig. 2), the station is located in the mountains, at about 100 km from the Mediterranean Sea and presents often a high nebulosity. The second dataset comes from measurements realized in the laboratory at Ajaccio (Corsica, France, 41°55 N, 8°44 E, 4m asl) (blue circle in Fig. 2) at about 100 m from the Mediterranean Sea. The third dataset is constituted by measurements realized in Tilos island (Greece, 36°24 N, 27°22 E, 96 m asl) (green circle in Fig. 2), Tilos is a small Greek island in the Dodecanese archipelago, the tallest mountain is about 650 m high.

The meteorological conditions (variability of solar radiation) in these three stations are very different, the solar radiation is outlined and will be confirmed in paragraph 5.

The periods of the solar data collect and the number of validated measures (after quality check and night data extraction) are given in Table 1.

Table 1. Period of availability and number of validated data for the three sites.

Location	Period of collected data	Number of validated data

Odeillo	01/01/2000 to 12/31/2002	26280
Ajaccio	01/01/1998 to 12/31/2000	26280
Tilos	01/01/2016 to 12/31/2016	8617

3. The forecasting models

Eleven forecasting models are briefly described and more information on each of them are available in literature. These models are classified in three categories: naïve models, classical machine learning models and regression trees-based models. The symbol $\widehat{\ }$ indicates that the value is predicted, without this symbol the value is measured.

3.1. Naive models

The two naive models are generally used as a reference in view to compare it with more sophisticated models. Indeed, if a complex model is not more efficient than these two models which are easy to implement and not requiring historical data, their use is not justified. The first model is the simplest one, the persistence, it is the repetition of the measure at the instant t to the predicted value at the instant t+h (h being the forecasting horizon) (Diagne et al., 2013):

$$\widehat{GHI}(t+h) = GHI(t) \quad (4)$$

$\widehat{GHI}(t)$ and $GHI(t)$ are respectively the predicted and measured hourly global horizontal solar irradiation at time t. This model is very simple to implement but it gives generally results with low accuracy.

The daily profile of the solar radiation can be added in using the Solis clear sky model to give the smart persistence, a simple improvement of the previous model developed for solar forecasting and used as reference on comparison with other models (Voyant et al., 2015) .

$$\widehat{GHI}(t+h) = GHI(t).\frac{CS(t+h)}{CS(t)} \quad (5)$$

The accuracy of such models decreases rapidly with the time horizon and is generally not adequate for a horizon higher than one hour.

3.2. Classical machine learning models

3.2.1. Auto Regressive Mobile Average (ARMA)

The ARMA model, used in energy consumption forecasting (de Oliveira & Cyrino Oliveira, 2018) includes two parts, an auto regressive one and a mobile average one. This model predicts the future values, as described in Ref (De Gooijer & Hyndman, 2006; Faraday & Chatfield, 1998), from a linear combination of past values and a past residue:

$$\widehat{CSI}(t+h) = \varepsilon(t) + \sum_{i=0}^{p} \varphi_i \,.\, CSI(t-i) + \sum_{i=0}^{q} \theta_i \,.\, \varepsilon(t-i) \quad (6)$$

CSI(t+h) being the clear sky index at time t+h, φ and θ the ARMA parameters deduced by a least square method, p and q are the model orders and ε(t) is the error or a noise related to a normal distribution.

3.2.2. Artificial neural network (ANN): MultiLayer Perceptron (MLP)

A feed forward MLP, a type of ANN, with one hidden layer and one output layer is used. The utilization of this technique for application on energy systems and in forecasting and modelling solar radiation (Kalogirou, 2000; Mellit, 2008). The equation for a MLP with one hidden layer of m neurons, one output neuron and n input variables is given by:

$$\widehat{CSI}(t+h) = \sum_{j=1}^{m} \omega_j \,.\, \left(g\left(\sum_{i=0}^{n-1} \omega_{i,j}.\, CSI(t-j) + b_j\right)\right. \quad (7)$$

with CSI the input vector of n clear sky indexes, $\widehat{CSI}(t+h)$ the predicted value, b_j the biases of the hidden neuron j and $\omega_{i,j}$ the weights between the input i and the hidden node j, g is the transfer function,ω_j the weight between the output and the hidden neuron j

3.2.3. Gaussian Process (GP)

The Gaussian process, nonlinear model, is a Gaussian distribution with an infinity of variables, (Rasmussen, 2004) details the work and implementation of GP. Every predicted CSI is represented by the sum of a function $f(CSI(\boldsymbol{\tau}))$ and an independent Gaussian noise $\mathcal{N}(O, \sigma_n^2)$ with a variance σ_n^2 and $\mathrm{CSI}(\tau) = (\mathrm{CSI(t)}, \mathrm{CSI(t-1)}, \cdots, \mathrm{CSI(t-n)})$:

$$\widehat{CSI}(t+h) = f\big(CSI(\boldsymbol{\tau})\big) + \mathcal{N}(O, \sigma_n^2) \quad (8)$$

GP is defined as a mean function $m(CSI(\boldsymbol{\tau}))$ and a covariance function *k*. *k* is an exponential squared function relying $\widehat{CSI}(t_p + h)$ with $\widehat{CSI}(t_q + h)$, t_p and t_q being two successive instants:

$$k\big(\widehat{CSI}(t_p + h), \widehat{CSI}(t_q + h)\big) = \sigma_f^2 exp\left[\frac{-\big(CSI(t_p) - CSI(t_q)\big)^2}{2l^2}\right] + \delta_{pq}\sigma_n^2 \quad (9)$$

δ_{pq} is Kronecker delta, σ_f^2 and σ_n^2 are the hyper parameters of the covariance function and are responsible of the complexity of the model, l is a length parameter.

3.2.4. Support Vector Regression (SVR)

SVR is a Kernel based model, firstly developed for regression problems and applied now for forecasting purposes (Lauret et al., 2015; Vapnik, 2013). With a training dataset $D = \{CSI(\tau), CSI(t+h)\}$; the predicted CSI can be expressed by:

$$\widehat{CSI}(t+h) = \sum_{\tau=1}^{t-1} \alpha_\tau \,.\, k_{rbf}(CSI(t+h), CSI(t-\tau)\ \ ) + b \quad (10)$$

and the Kernel radial basis function is:

$$k_{rbf}\big(CSI(t_p), CSI(t_q)\ \ \big) = exp\left[\frac{-\big(CSI(t_p) - CSI(t_q)\big)^2}{2\sigma_f^2}\right] \quad (11)$$

α_i is the Lagrange multipliers, solutions of a quadratic problem, b is the bias determined by specific conditions.

3.2.5. Regression Trees based methods

Decision trees based on "If-Then" rules are used for classification and produces understandable models due to their graphical representation, they are adapted to work on forecasting problems, like solar radiation forecasting (Aggarwal & Saini, 2014; Burrows, 1997). They have been then extended for predicting numerical values of attributes and led to regression trees which are a decision tree in which the leaf nodes have been set as regression models, and therefore, continuous numeric values can be predicted. Some results are given on forecasting of photovoltaic production system using regression trees (Persson, Bacher, Shiga, & Madsen, 2017).

A. Standard and pruned regression trees (RT and RT-pruned)

Hastie and Tibshirani (Hastie & Tibshirani, 1986) proposed a formalization of the RT models:

$$\widehat{CSI}(t+h) = \sum_{i=1}^{t-1} k_i . I(CSI(t-i)) \quad (12)$$

with k_i constant factors, I a function which return 1 if input is used and 0 elsewhere; The trees are built by splitting the data based on the values of predictive attributes. A regression model is then computed for each node. The pruning aspect of RT is operated with an elevation of the quadratic error tolerance per node. Splitting nodes stops when the quadratic error per node drops below tolerance. For normal RT, the tolerance is close to zero, while for the pruned RT, a higher value is chosen using a heuristic method based on the minimizing of the global error of prediction (Pedro, Coimbra, David, & Lauret, 2018) .

B. Boosted and bagged regression trees (RT-boosted and RT-bagged)

It is interesting in "ensemble learning" methods to generate many regressors and aggregate their results. Two usual methods are boosting and bagging of RT, these two methods are improvement of the classical regression trees.

The boosting method consists in assembly weak RT classifiers and take the average of predictions in order to improve the efficiency (Huang & Perry, 2016). A weak predictor is a simple single split RT and the next trees give more weights to the data badly predicted at the previous point, De'Ath explains in details the method to use boosted RT in modelling and forecasting applications (De'ath, 2007). The function for additive models applied to solar forecasting by boosted RT is:

$$\widehat{CSI}(t+h) = \sum_m \beta_m \, b(\widehat{CSI}(t+h), \gamma_m) \quad (13)$$

The function b represents the individual trees with γ_m the split variable and β_m the weight at each node. The Bagging method (for bootstrap aggregating) is another improvement of the RT (Breiman, 1996); the model is an aggregation of RT which grows from samples of dataset :

$$\widehat{CSI}(t+h) = av_k \, \varphi_k(\widehat{CSI}(t+h)) \quad (14)$$

φ_k are the different predictors before the aggregation and av_k the mean of the different predictors.

C. Random forests

Random forests [9] add an additional layer of randomness to bagging. In a random forest, the dataset is equally divided in samples but each regression tree grows differently, each node is split using the best among a subset of predictors randomly chosen at that node (Ibrahim & Khatib, 2017). This improvement by randomness gives robustness to the model and decreases the over-training risks.

4. Evaluation of models performance

To evaluate the accuracy of the models, three different statistics are used. The mean absolute error (MAE), defines the absolute value of the gap between the observed and the predicted value.

$$MAE = \frac{\sum_{i=1}^{N} |\widehat{GHI}(i) - GHI(i)|}{N} \quad (15)$$

with $\widehat{GHI}(i)$ the predicted variable, $GHI\ (i)$ the observed variable and N the number of data.

The normalized root mean squared error (nRMSE) is more sensitive to large forecast errors, and hence is suitable for applications where small errors are more tolerable and larger errors cause disproportionately high costs, as for example in the case of utility applications. It is probably the reliability factor that is most appreciated and used: it's a good statistical index to evaluate the accuracy of a models, the aim of an operator is to minimize it in order to improve model performances:

$$nRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\widehat{GHI}(i) - GHI(i))^2}}{\overline{GHI}} \quad (16)$$

where $\overline{GHI}$ is the algebraic average of the observed values.

The skill score is an index calculated in order to compare the performance of a given model with a reference model, here the reference model is the smart persistence model (SP) described in 3.1.

$$Skill\ Score = \frac{Metric_{forecast} - Metric_{reference}}{Metric_{perfect\_forecast} - Metric_{reference}} = 1 - \frac{nRMSE_{forecast}}{nRMSE_{reference}} \quad (17)$$

The skill score is always inferior at 1, negative if the forecaster is less performant than the reference, 0 if the performances are similar, and positive if it is better.

5. Evaluation of the variability

The evaluation of the variability of different datasets is made thanks to a statistical parameter, Voyant et al. (Voyant et al., 2015) described several parameters to quantify the variability of a solar irradiation time series and the mean absolute log return appeared to be the most efficient:

$$meanabs\ (logr) = \frac{\sum_{i=1}^{N} |\log(CSI(i)) - \log(CSI(i-1))|}{N} \quad (18)$$

Other variability metrics are available like the P parameter described by Perez *et al*. (Perez, Kivalov, Schlemmer, Hemker Jr., & Hoff, 2012) and computed from the standard deviation (*std*) by $P = std(CSI(t) - CSI(t-1))$. In order to not overburden the manuscript, the variability is only estimated with the mean absolute log return. The variability of the three solar dataset were computed using Eq (18) and the Mean absolute log return values for each site are shown in Table 2.

Table 2. Variability for three CSI datasets according to the mean absolute log return.

Site	**Odeillo**	**Tilos**	**Ajaccio**
Mean absolute log return	0.5028	0.3732	0.1961
Type of variability	Strong	Medium	Weak

A large difference between the three datasets is noted, the variability of the solar irradiance measurements differs from one site to another; thus, the performance of the forecasting methods presented here will have a more universal character in nature because they are the result of calculations realized from various weather conditions. Moreover, it will be possible to draw a correlation between the site variability and the ranking of the models in term of performance.

6. Solar irradiance forecasting and performances of forecasting models

This section presents the results of solar irradiance forecasting for the three sites. The preprocessing is the same for every dataset:

- Calculation of the clear-sky irradiation by the Solis model;
- Calculation of clear sky indexes;

- Remove of night hours and hour with the sun elevation up to 10°.

The auto-mutual information which is a part of information theory, is a way to select the number of input given in the models (Fu et al., 2017; Parviz et al., 2008). For each time series shows that the optimal number of input data n (first minimum of the auto mutual information criteria, see paragraph 2.2) is n=8 for Ajaccio and Odeillo and n=6 for Tilos. Thus, for estimating the future solar irradiation, the n previous measured solar irradiations are used. The training set is about 80% of the dataset and the testing set about 20% considering a k-fold sampling equal to 10.

Tables 3 and 4 show the results of the nRMSE and MAE calculations for the eleven models and a forecast horizon from 1 to 6 hours with an hourly resolution.

Table 3. nRMSE vs forecast horizon, the two best models are highlighted.

	Horizon	**h+1**	**h+2**	**h+3**	**h+4**	**h+5**	**h+6**
Odeillo (variability: 0.5028)	Persistence	40.91%	61.81%	76.18%	84.53%	88.35%	86.25%
	Smart persistence	37.00%	54.59%	66.68%	74.64%	77.77%	77.76%
	ARMA	34.42%	47.12%	49.26%	49.76%	49.16%	48.56%
	MLP	29.88%	43.51%	45.86%	47.87%	48.44%	48.72%
	Regression tree (RT)	36.73%	51.94%	54.95%	56.76%	57.86%	57.45%
	Boosted RT	30.20%	43.24%	45.59%	47.71%	48.57%	48.78%
	Bagged RT	28.80%	42.00%	44.66%	46.63%	47.83%	47.52%
	Pruned RT	29.90%	43.97%	46.47%	48.47%	50.01%	49.84%
	Random forest	28.76%	42.75%	44.89%	46.56%	47.78%	48.34%
	Gaussian process	28.65%	42.15%	45.37%	46.67%	48.54%	48.42%
	Support vector regression	34.18%	57.71%	57.25%	57.76%	58.27%	57.25%

	Horizon	h+1	h+2	h+3	h+4	h+5	h+6
Tilos (variability: 0.3732)	Persistence	27.33%	44.81%	57.12%	64.59%	67.50%	65.13%
	Smart persistence	18.51%	26.20%	31.73%	34.57%	36.73%	36.65%
	ARMA	17.71%	27.42%	29.91%	30.64%	30.85%	31.44%
	MLP	18.97%	30.76%	31.39%	32.40%	32.96%	33.74%
	Regression tree (RT)	25.62%	35.09%	37.45%	37.04%	34.45%	36.37%
	Boosted RT	19.37%	29.63%	32.91%	32.89%	36.17%	33.93%
	Bagged RT	20.11%	30.55%	30.64%	31.09%	33.22%	33.33%
	Pruned RT	20.66%	30.57%	34.05%	33.33%	34.04%	33.49%
	Random forest	19.19%	29.42%	32.28%	32.90%	33.54%	32.30%
	Gaussian process	18.48%	28.87%	32.28%	32.46%	33.53%	33.52%
	Support vector regression	18.73%	39.27%	39.67%	39.78%	40.03%	39.62%
	Horizon	**h+1**	**h+2**	**h+3**	**h+4**	**h+5**	**h+6**
Ajaccio (variability: 0.1961)	Persistence	26.60%	42.62%	54.10%	61.39%	64.51%	63.86%
	Smart persistence	19.26%	26.46%	31.18%	34.15%	36.92%	38.93%
	ARMA	18.35%	29.27%	31.38%	32.25%	33.18%	33.69%
	MLP	18.26%	29.26%	31.31%	32.47%	32.98%	33.84%
	Regression tree (RT)	24.64%	36.88%	38.47%	39.74%	39.95%	41.24%
	Boosted RT	18.75%	29.55%	31.89%	32.51%	33.55%	33.98%
	Bagged RT	18.76%	29.80%	31.10%	32.17%	33.35%	34.02%
	Pruned RT	18.72%	30.88%	32.27%	33.76%	34.01%	35.00%
	Random forest	18.97%	29.63%	31.62%	32.38%	33.37%	33.91%
	Gaussian process	18.97%	30.08%	31.96%	33.29%	33.55%	34.44%
	Support vector regression	18.55%	38.78%	41.03%	41.56%	41.66%	41.60%

Table 4. MAE values (in Wh/m²) vs forecast horizon, the two best models are highlighted

	Horizon	h+1	h+2	h+3	h+4	h+5	h+6
Odeillo (variability: 0.5028)	Persistence	147.07	238.13	297.75	328.68	341.65	331.23
	Smart persistence	124.62	199.81	249.00	277.70	288.13	285.29
	ARMA	130.73	179.45	189.20	190.79	188.41	187.33
	MLP	105.87	162.23	172.95	181.43	185.81	186.29
	Regression tree (RT)	119.94	186.50	201.09	208.70	212.29	210.24
	Boosted RT	107.77	161.64	171.97	180.56	185.47	185.90
	Bagged RT	98.94	156.32	166.26	175.28	179.78	179.82
	Pruned RT	105.68	163.08	173.14	181.84	188.12	186.80
	Random forest	97.48	156.73	167.45	174.15	180.13	181.94

	Gaussian process	99.50	157.09	169.66	175.93	184.45	185.04
	Support vector regression	110.49	214.72	213.05	214.57	215.82	213.61
Tilos (variability: 0.3732)	**Horizon**	**h+1**	**h+2**	**h+3**	**h+4**	**h+5**	**h+6**
	Persistence	143.12	246.30	306.64	353.09	358.97	343.51
	Smart persistence	73.93	113.29	140.99	156.14	165.36	162.97
	ARMA	74.74	128.93	140.58	145.30	145.64	146.64
	MLP	79.09	137.93	144.65	147.69	150.01	149.41
	Regression tree (RT)	94.51	145.97	157.23	155.33	149.38	152.16
	Boosted RT	82.81	132.26	143.15	145.50	149.75	145.07
	Bagged RT	76.13	130.23	133.61	136.03	140.34	140.42
	Pruned RT	89.69	131.89	145.60	144.57	144.19	140.30
	Random forest	73.71	127.95	136.51	140.44	141.65	138.56
	Gaussian process	71.40	127.41	138.11	138.47	142.03	141.78
	Support vector regression	71.27	166.29	168.16	168.84	170.02	168.69
Ajaccio (variability: 0.1961)	**Horizon**	**h+1**	**h+2**	**h+3**	**h+4**	**h+5**	**h+6**
	Persistence	104.61	176.01	220.18	252.23	259.18	255.98
	Smart persistence	55.47	79.46	95.99	107.08	116.37	123.95
	ARMA	60.68	88.84	94.98	97.40	100.59	102.31
	MLP	60.63	89.42	95.19	98.86	100.66	103.31
	Regression tree (RT)	73.28	117.66	125.00	129.83	132.00	135.58
	Boosted RT	61.36	91.76	98.07	100.10	103.32	104.62
	Bagged RT	61.38	93.33	97.27	101.34	104.63	105.96
	Pruned RT	60.80	95.42	99.97	104.98	106.14	108.86
	Random forest	61.18	92.94	99.08	102.03	105.03	106.48
	Gaussian process	61.91	96.06	101.10	104.80	106.27	108.30
	Support vector regression	54.58	126.64	133.29	135.50	135.08	134.62

Conclusions are difficult to drawn for these three locations. More meteorological sites would have been welcome but it is difficult to find reliable and long hourly solar irradiation data set and to propose a synthetic paper with so many forecasting models applied to a larger number of locations. The results generalization is, as often, very complicated and depends also on the error metrics used.

However, some first conclusions can be drawn, the first one the naïve models (persistence and smart persistence) give always bad performances for all the sites and are the worst models. For a high variability (Odeillo), the best performances are obtained with ensemble learning models (bagged regression trees and random forest), the machine learning models are less performant. These results show that ensemble learning models have the capacity to apprehend complicated

phenomena. The high dataset variability explains the bad performances of “classical” models as MLP, ARMA or simple regression trees.

For a medium variability dataset (Tilos), the best results come from the ARMA model, but occasionally smart persistence and random forest gave good results. The poorer performances of machine learning models, for Tilos, is probably due to the small length of the dataset (smaller than for Ajaccio and Odeillo).

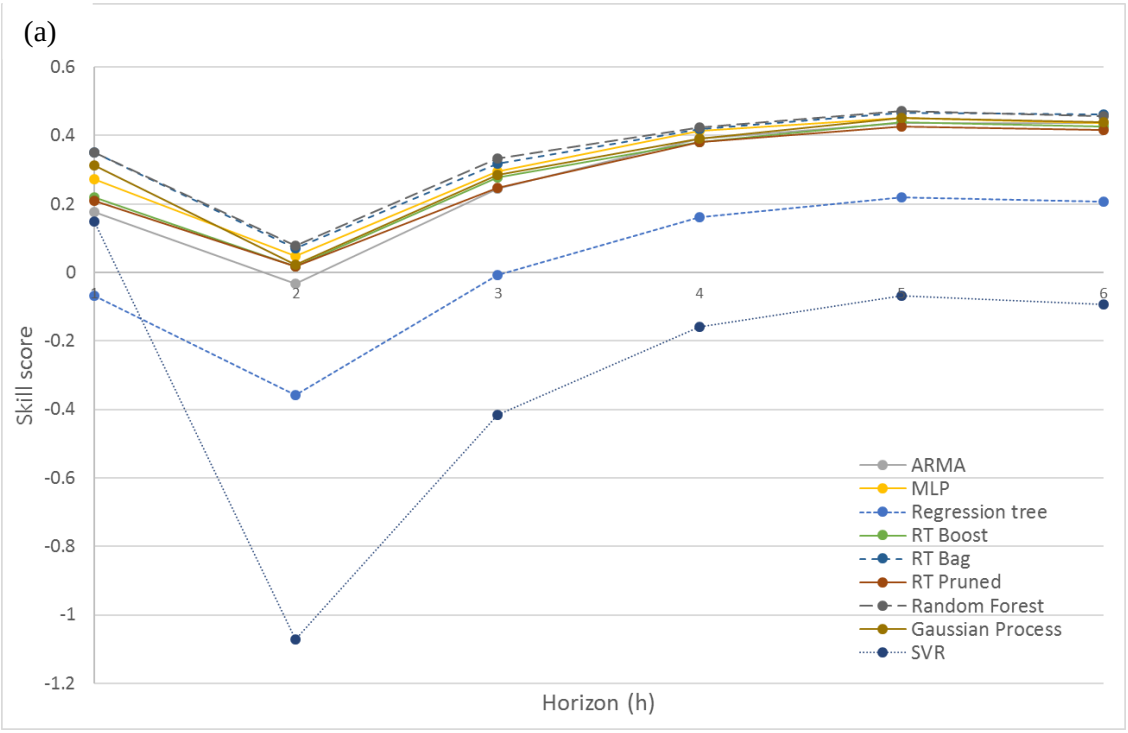
For a low variability dataset (Ajaccio) the best models are ARMA and MLP, followed by smart persistence and bagged regression trees; however, the errors are of the same order of magnitude for all the models excepted for persistence and simple regression tree. With such a variability, machine learning models are strong predictors.

The results from the MAE point of view are similar:

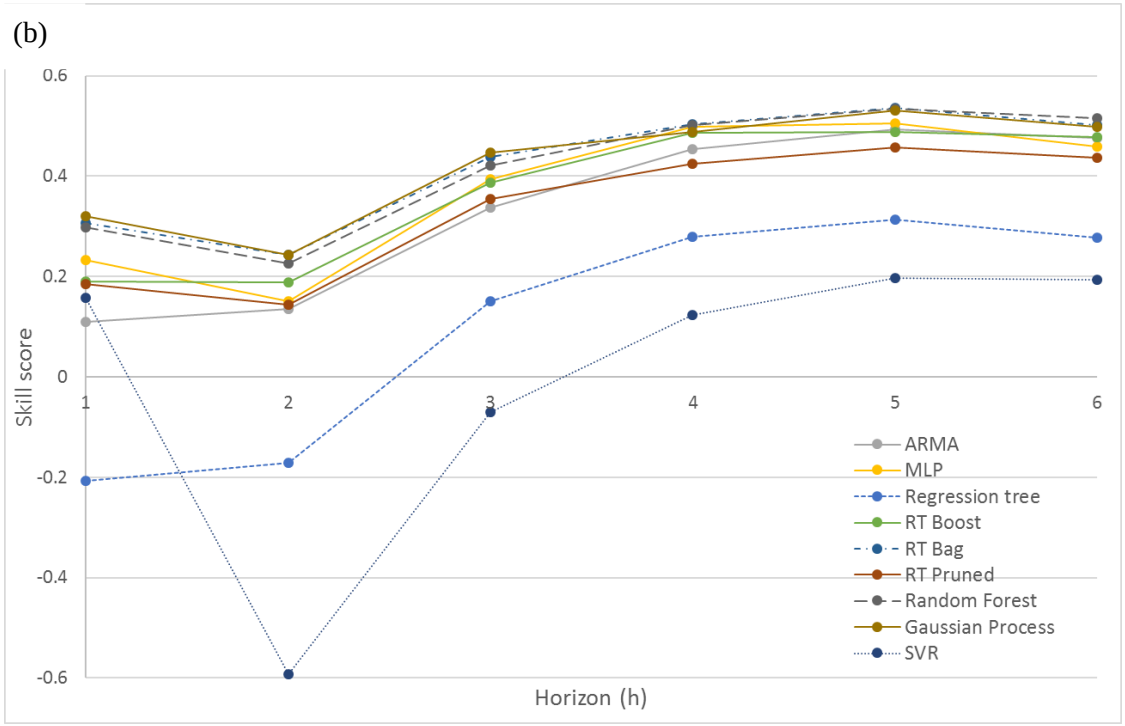
- For a high variability dataset, the ensemble learning models present the smallest absolute error.
- For a medium variability and the two first hours of forecasting horizon, the statistical models give the best performances and for deeper horizons, the ensemble learning models are the best predictors
- For a low variability dataset, excepted for the naïve model, all the models are similar in term of absolute error, the results are of the same order of magnitude.

The skill score (related to the smart persistence reference model) is computed for each model, each horizon and site and is plotted in Fig 3.

(a)


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
Skill score
2
3
4
5
6
Horizon (h)
ARMA
MLP
Regression tree
RT Boost
RT Bag
RT Pruned
Random Forest
Gaussian Process
SVR

(b)


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Skill score
1
2
3
4
5
6
Horizon (h)
ARMA
MLP
Regression tree
RT Boost
RT Bag
RT Pruned
Random Forest
Gaussian Process
SVR

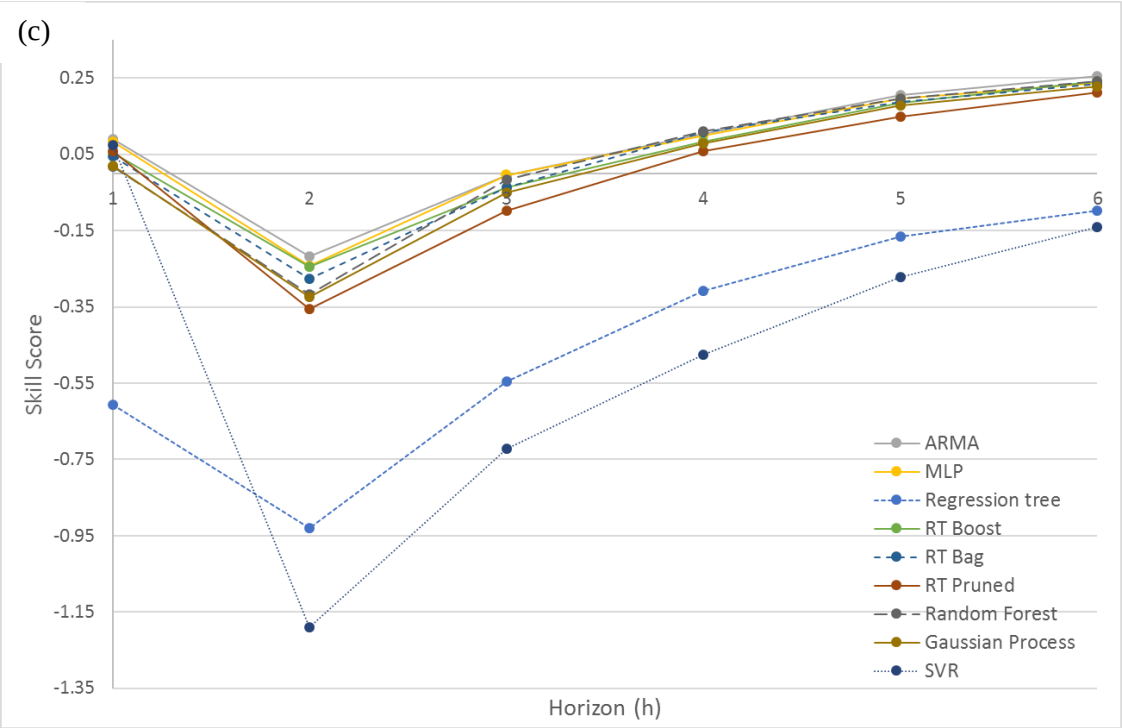


Figure 3. Skill score vs forecast horizon for a) Odeillo, b) Tilos and c) Ajaccio

The skill score compares the performances of the model with the smart persistence. If the performances of the models are generally better than the smart persistence for Odeillo, this fact is not checked for the two other sites. Thus, as expected, the improvement on the forecasting due to the use of machine learning methods compared with naïve models increases with the variability of the dataset to predict.

7. Comments concerning prediction error and variability

Depending on the variability of the site, the models inducing the lowest prediction error are different. In Table 5, the main conclusions concerning the benchmarking study of forecasting model are drawn for the three sites: Ajaccio, Odeillo and Tilos.

For a low variability, (Ajaccio), a linear model ARMA and a classical MLP give very good results. With a lower reliability, the smart persistence or a RT with the bagged mode for the horizons inferior to 4 hours can be also used. For a high variability site (here Odeillo), the

choice is more complicated, ensemble machine learning methods like random forest or bagged RT are more recommended and linear models or smart persistence must be rejected . For an intermediate variability (here Tilos), the two best models are ARMA and Bagged RT, but all the other machine learning models have about the same performances and the model ranking is very difficult.

It is obvious than these results must be confirmed on other meteorological sites with a wide range of variability.

Table 5. Ranking of the prediction methodologies according to the variability

	Weak variability $\boldsymbol{meanabs\,(logr) < 0.2}$	**Medium Variability** $\boldsymbol{0.2 < meanabs\,(logr) < 0.4}$	**Strong variability** $\boldsymbol{0.4 < meanabs\,(logr)}$
Recommended models	ARMA, MLP	ARMA, Bagged RT	Bagged RT, Random forest
Usable models	SP (horizon < 4h), Bagged RT	Random forest, Gaussian process, MLP, SP (horizon <3h)	Gaussian process

8. Conclusion

Eleven statistical and machine learning tools for global solar irradiation forecasting were analyzed and compared in term off performances on three sites with different meteorological characteristics. To characterize the solar data time series measured in each location, an evaluation of the variability was realized. A ranking of the forecasting methods according to the prediction horizon and depending on the site was realized; the main conclusions can be drawn:

- For Ajaccio, with a weak variability, ARMA and MLP are the most efficient tools;
- For Tilos, with a medium variability, ARMA and bagged regression tree are recommended;

- For Odeillo with a higher variability, the forecasting reliability is lower, but the best results were obtained for the bagged regression tree and the random forest approach.

Even as the results are not totally conclusive, it appears that that higher is the variability, more complex is the forecasting tool to be used. In general, when the cloud occurrences are low, the utilization of an ARMA model is sufficient else the regression tree based on the bagging mode is the most reliable. The conclusions drawn here concern only three sites and to confirm them, other simulations must be carried out on well-chosen sites with particular weather characteristics (desert, monsoon, extreme longitude, etc.).

Acknowledgment

This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 646529 through the TILOS project (Technology innovation for the Local Scale, Optimum integration of Battery Energy Storage). The data sets were providing by the CNRS UPR PROMES 8521 laboratory in Odeillo (Pyrénées Orientales, France), the CNRS UMR SPE 6134 Laboratory in Ajaccio (Corsica, France) and the local weather station on TILOS Island.

Bibliography

[1] Diagne M, David M, Lauret P, Boland J, Schmutz N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew Sustain Energy Rev 2013;27:65–76. doi:10.1016/j.rser.2013.06.042.
[2] Voyant C, Notton G, Kalogirou S, Nivet M-L, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: A review. Renew Energy 2017;105:569–82. doi:10.1016/j.renene.2016.12.095.
[3] Voyant C, Soubdhan T, Lauret P, David M, Muselli M. Statistical parameters as a means to a priori assess the accuracy of solar forecasting models. Energy 2015;90, Part 1:671–9. doi:10.1016/j.energy.2015.07.089.
[4] Lauret P, Voyant C, Soubdhan T, David M, Poggi P. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol Energy 2015;112:446–57. doi:10.1016/j.solener.2014.12.014.
[5] Boland J, David M, Lauret P. Short term solar radiation forecasting: Island versus continental sites. Energy 2016;113:186–92. doi:10.1016/j.energy.2016.06.139.
[6] Voyant C, Motte F, Fouilloy A, Notton G, Paoli C, Nivet M-L. Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies. Energy 2017;120:199–208. doi:10.1016/j.energy.2016.12.118.

[7] Troncoso A, Salcedo-Sanz S, Casanova-Mateo C, Riquelme JC, Prieto L. Local models-based regression trees for very short-term wind speed prediction. Renew Energy 2015;81:589–98. doi:10.1016/j.renene.2015.03.071.
[8] De'ath G. Boosted Trees for Ecological Modeling and Prediction. Ecology 2007;88:243–51. doi:10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.
[9] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40. doi:10.1023 /A:10180 54 314350.
[10] Breiman L. Random forests. Mach Learn 2001;45:5–32. doi:10. 1023 /A: 10109 33404 324.
[11] Rasmussen CE. Gaussian Processes in Machine Learning. In: Bousquet O, Luxburg U von, Rätsch G, editors. Adv. Lect. Mach. Learn., Springer Berlin Heidelberg; 2004, p. 63–71. doi:10.1007/978-3-540-28650-9_4.
[12] Jiang H, Dong Y. Global horizontal radiation forecast using forward regression on a quadratic kernel support vector machine: Case study of the Tibet Autonomous Region in China. Energy 2017;133:270–83. doi:10.1016/j.energy.2017.05.124.
[13] Iqbal M. An introduction to solar radiation. 1983.
[14] Badescu V. Modeling solar radiation at the earth's surface: recent advances. Springer; 2008.
[15] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw 1989;2:359–66. doi:10.1016/0893-6080(89)90020-8.
[16] Paoli C, Voyant C, Muselli M, Nivet M-L. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol Energy 2010;84:2146–60. doi:10.1016/j.solener.2010.08.011.
[17] Luo Y, Shi Y, Zheng Y, Gang Z, Cai N. Mutual information for evaluating renewable power penetration impacts in a distributed generation system. Energy 2017;141:290–303. doi:10.1016/j.energy.2017.09.033.
[18] Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M, Cros S, Dumortier D, et al. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens Environ 2004;91:160–74. doi:10.1016/j.rse.2004.02.009.
[19] Ineichen P. Comparison of eight clear sky broadband models against 16 independent data banks. Sol Energy 2006;80:468–78. doi:10.1016/j.solener.2005.04.018.
[20] Ineichen P. A broadband simplified version of the Solis clear sky model. Sol Energy 2008;82:758–62. doi:10.1016/j.solener.2008.02.009.
[21] Parviz RK, Nasser M, Motlagh MRJ. Mutual Information Based Input Variable Selection Algorithm and Wavelet Neural Network for Time Series Prediction. In: Kůrková V, Neruda R, Koutník J, editors. Artif. Neural Netw. - ICANN 2008, Springer Berlin Heidelberg; 2008, p. 798–807.
[22] de Oliveira EM, Cyrino Oliveira FL. Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. Energy 2018;144:776–88. doi:10.1016/j.energy.2017.12.049.
[23] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecast 2006;22:443–73. doi:10.1016/j.ijforecast.2006.01.001.
[24] Faraday J, Chatfield C. Times Series Forecasting with Neural Networks: A Case Study 1998.
[25] Kalogirou SA. Applications of artificial neural-networks for energy systems. Appl Energy 2000;67:17–35. doi:10.1016/S0306-2619(00)00005-2.
[26] Mellit A. Artificial intelligence techniques for modelling and forecasting of solar radiation data: A review. Int J Artif Intell Soft Comput 2008:52–76.
[27] Vapnik V. The Nature of Statistical Learning Theory. Springer Science & Business Media; 2013.

[28] Burrows WR. CART Regression Models for Predicting UV Radiation at the Ground in the Presence of Cloud and Other Environmental Factors. J Appl Meteorol 1997;36:531–44. doi:10.1175/1520-0450(1997)036<0531:CRMFPU>2.0.CO;2.
[29] Aggarwal SK, Saini LM. Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest. Energy 2014;78:247–56. doi:10.1016/j.energy.2014.10.012.
[30] Persson C, Bacher P, Shiga T, Madsen H. Multi-site solar power forecasting using gradient boosted regression trees. Sol Energy 2017;150:423–36. doi:10.1016/j.solener.2017.04.066.
[31] Hastie T, Tibshirani R. Generalized additive models. Stat Sci 1986;1:297–318.
[32] Pedro HTC, Coimbra CFM, David M, Lauret P. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renew Energy 2018;123:191–203. doi:10.1016/j.renene.2018.02.006.
[33] Huang J, Perry M. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. Int J Forecast 2016;32:1081–6. doi:10.1016/j.ijforecast.2015.11.002.
[34] Ibrahim IA, Khatib T. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers Manag 2017;138:413–25. doi:10.1016/j.enconman.2017.02.006.
[35] Perez R, Kivalov S, Schlemmer J, Hemker Jr. K, Hoff TE. Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance. Sol Energy 2012;86:2170–6. doi:10.1016/j.solener.2012.02.027.
[36] Fu X, Sun H, Guo Q, Pan Z, Xiong W, Wang L. Uncertainty analysis of an integrated energy system based on information theory. Energy 2017;122:649–62. doi:10.1016/j.energy.2017.01.111.