# Prediction intervals for global solar irradiation forecasting using regression trees methods

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

# Prediction intervals for global solar irradiation forecasting using regression trees methods

Cyril Voyant[1,2*], Fabrice Motte[1] , Gilles Notton[1], Alexis Fouilloy[1], Marie-Laure Nivet[1], Jean-Laurent Duchaud[1],

[1] University of Corsica, CNRS UMR SPE 6134, 20250 Corte, France

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

***Corresponding author: Cyril Voyant, phone: +33 4 95 29 36 66, fax:+ 33 4 95 29 37 97, email: cyrilvoyant@gmail.com**

**Abstract**.

A global horizontal irradiation prediction (from 1 hour to 6 hours) is performed using 2 persistence models (simple and "smart" ones) and 4 machine learning tools belonging to the regression trees methods family (normal, pruned, boosted and bagged). A prediction band is associated to each forecast using methodologies based on: bootstrap sampling and k-fold approach, mutual information, stationary time series process with clear sky model, quantiles estimation and cumulative distribution function. New reliability indexes (gamma index and gamma test) are built from the mean interval length (*MIL*) and prediction interval coverage probability (*PCIP*). With such methods and error metrics, good prediction bands are estimated for Ajaccio (France) with a *MIL* close to 113 Wh/m², a *PCIP* reaching 70% and a gamma index lower than 0.9.

**Keywords**: probabilistic forecasts, bagging, boosting, pruning, mean interval length, prediction interval coverage probability

# 1. Prediction intervals instead of single predictions… Why?

Electrical operators have to ensure an exact balance between electricity production and consumption at any time of the year [1,2]. They often have some difficulties to maintain this stability with conventional (heat engine, nuclear power plant, hydroelectricity, etc.) and incontrollable energy production system (PV and wind farm), almost in small or no interconnected electrical grid (as island ones [3,4]). The consistency of the electrical system is in fact dependent on the ability of the system to accommodate expected and unexpected changes in production and consumption in order to maintain quality and continuity of service to the customers [5,6]. Usually, the prediction of the solar and wind energy system production is necessary to achieve this goal. A lot of studies show that the time series formalism gives the best results for horizons between 1 hour and 6 hours [3,5,7,8], while for deeper horizons the use of satellite data and numerical weather predictions becomes the most attractive methods [9–11]. Satellite-based irradiance models are able to estimate the solar radiation levels (historic, recent and future levels) without the need of installing ground sensors at the location of interest but correction based on measurements improves the results [12]. As clouds are the dominant source of small-scale variability in surface solar radiation and uncertainty in its prediction, for very short term global horizontal irradiance forecast, it is possible and recommended to use sky images as described by Schmidt et al. [13] .The present paper focuses on the first problem also called the nowcasting and proposes a machine learning methodology aiming on prediction intervals [14–16] rather than a single forecasted value [7,17]. In statistical inferences, specifically predictive inferences, a prediction interval is an estimation of the interval in which future observations will be with a given probability. This kind of approach is often denoted probabilistic forecasting and is often used in atmospheric science [18,19] in both regression analysis and frequentist statistics and allows the distribution generation of individual future prediction [20]. For the grid manager, this sort of information (prediction interval or reliability index of prediction) is essential and allows important additional information making it possible to master the management of electrical networks and particularly to increase the intermittent renewable energy part.

The structure of this paper will be: data and predictors description, prediction interval generation, results and then conclusions.

## 2. The data used

The data used to build the models are hourly solar global horizontal irradiation (*GHI*) measured from a meteorological station and a usual cleaning approach is then operated in order to identify and remove the desired data. Mistakes often 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) [21] has been applied to the data. The process to estimate the quality of the data [22] and the procedure applied to flag suspicious or erroneous measurements is described in detail in [20].

### 2.1. Measurements

All the experiments and numerical simulations are related to Ajaccio, during a period of 9 years from 2006 to 2014, (Corsica Island, France). This station is equipped with pyranometers (CM 11 Kipp & Zonen) and standard meteorological sensors (pressure, nebulosity, etc.). It is located near the Mediterranean Sea (100 m) and nearby mountains (1000 m altitude at 40 km from the site). This specific geographical configuration makes nebulosity difficult to forecast. Mediterranean climate is characterized by hot summers with abundant sunshine and mild, dry and clear winters. As the computing power was increasing over the past few decades, the field of machine learning has rapidly advanced in both theory and practice. Machine learning methods are usually based on the assumption that the data generation mechanism does not change over time, thus considering that the used process is stationary. In the next subsection, the method used to make the *GHI* time series stationary is exposed [23].

## 2.2. Clear sky modelling

In previous studies [24,25], it has been demonstrated that the clear sky index (*CSI*) calculated with the simplified Solis model of Ineichen [26] is the most reliable for Ajaccio. This model generates a clear sky hourly irradiation (CS) expressed by Eq. (1), this model requires a fitting parameter (g), the extraterrestrial irradiation ($I_0$), the solar elevation (h) and the total measured atmospheric optical depth ($\tau$):

$$CS(t) = I_0\ (t).exp\left(\frac{-\tau}{sin^g(h(t))}\right).sin\ (h(t)) \quad (1)$$

The simplified Solis clear sky model is based on radiative transfer calculations and the Lambert-Beer relation [26] . The expression of the atmospheric transmittance is valid with polychromatic radiations, however when dealing with global radiation, the Lambert-Beer relation is only an approximation because of the backscattered effects. In view to improve the quality of the *CS* modelling, monthly average of water vapor column and aerosol optical depth at 700nm were introduced in the model using the aeronet sources [27]. According to [28] this model remains a good fitting function of the global horizontal radiation. The new computed time series (*CSI*) defined by equation (2) can be directly used with the machine learning forecasting:

$$CSI(t) = GHI(t)/CS(t) \quad (2)$$

# 3. The prediction models

In this paper, the time series approach is used, the common notation specifying a time series *CSI* that is indexed by the natural number is written *CSI* = $\{CSI(t)$: $t \in T\}$ where *T* is the time index set. The modelling of a time series can be defined by a linear or non-linear model denoted $f_n$ (see Eq. 3 where *t* = [*n*, *n*-1,…, *p*+1, *p*] and *n* and *p* are respectively the number of observations and of parameters of the model ; $n \gg p$; h is the horizon of prediction and $\epsilon_{t+h}$ the associated error) [29].

$$CSI(t+h) = f_n(CSI(t), CSI(t-1) \ldots., CSI(t-p+1)) + \epsilon_{t+h} \quad (3)$$

To estimate $f_n$ using a machine learning method, a stationary hypothesis is often required and implies to use a stable process [30,31]. A process is defined stable if its mean and/or variance variations remain constant over time. Previous studies [3,32,33] confirmed that the use of the clear sky index (*CSI*) (Eq. (2)) makes the time series stationary hence it can be introduced in a machine learning tool such as regression tree forecasting. A lot of methods of prediction based on machine learning are available, interested readers can refer to [34] concerning on random forest ensemble of support vector regression models, to [35] about Kalman filter and regressor, to [36] for works related to the Kriging, NWP and gradient boosted regression tree and to [37] for a very interesting evaluation of statistical learning configurations.

## 3.1. Naïve and reference methods: the persistence

The persistence is a naïve forecasting method. It is the most cost-effective and provides a benchmark against which more sophisticated models can be compared. The persistence approach [12] considers that the future *GHI* values are equal to the observed *GHI* at time *t* (Eq. 4). It considers that the atmospheric conditions and the solar irradiation remain unchanged between the current time t and the future time *t+h*.

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

A scaled persistence, also called "smart persistence" is defined by Eq. 5, using the clear sky Solis model $CS$. This model takes into account the daily solar irradiance profile. In numerous papers, this model is used as a reference and allows very good predictions (in stable meteorological conditions) [3,38].

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

## 3.2. Predictions based on regression trees

Regression tree learning is a method based on the use of a decision tree as a predictive model. It is particularly used in data mining and in automatic learning and machine learning. In these tree structures, the leaves represent the values of the target variable and the branch lines correspond to combinations of input variables that lead to these values [39–42]. Decision trees have originally been used for decision analysis. They were used to explicitly represent the decisions made and the processes that lead to them. They have since been introduced in machine learning and data mining. A decision tree describes the data but not the decisions themselves. It is a supervised learning technique: we use a set of data containing the past measurements and the target to build the tree. We then validate the tree performances by extrapolating its results to the test data set.

## 3.3. Classic regression tree (RT)

Hastie and Tibshirani [43] proposed a formalization of the classic regression tree models:

$$\widehat{CSI}(t+h) = \sum_{i=1}^{t-1} k_i \times I(CSI(t-i)) \quad (6)$$

Where $k_i$are constant factors, *I* is a function returning 1 if the input is used and 0 if not. Once the tree structure has been constructed, a regression model is computed for each node. The learning phase is an iterative process where the error (defined as the mean of the absolute difference between the predicted and the actual value) will be minimized.

### 3.3.1.Pruned regression trees (RT_pruned)

Pruned regression aims to reduce the number of nodes to make the regression tree more regularizable. Pruned trees are built by increasing the quadratic error tolerance per node. Splitting nodes stops when the quadratic error per node drops below a given tolerance ($\gamma_m$ the split variable). 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. In a pruned RT, *I* (see Eq. (6)) returns 0 more frequently than in the normal mode.

### 3.3.2.Boosted regression trees (RT_boosted)

There is a lot of interest in "ensemble learning" methods that generate many regression models and aggregate their results. For RT, two well-known methods are boosting and bagging of classification trees [44–46]. In boosting, the trees are built successively. The trees improving the prediction are weighted by an extra coefficient. The prediction is then obtained by the weighted linear combination of the trees [47]. Eq. (7) gives the function for additive models applied to the solar forecasting by boosted regression trees.

$$\widehat{CSI}(t+h) = \sum_m \beta_m \, b(\widehat{CSI}(t+h), \gamma_m) \quad (7)$$

The basis function $b(\widehat{CSI}(t+h), \gamma_m)$ represents the individual trees with $\gamma_m$ the split variable, defined by different values at each node and prediction results. $\beta_m$ is the coefficient taken into account in the global algorithm to weight the results obtained by the different trees.

### 3.3.3.Bagged regression tree (RT_bagged)

In bagging, the trees do not depend on earlier trees. Each one is independently constructed using a bootstrap sample of the data set. At the end, a simple majority vote is taken for prediction. The Bagging method is another version of the prediction with regression trees, it was described by Breiman [48]. Bagging means bootstrap aggregating, the model is an aggregation of regression trees which grow from samples of dataset. The subtrees are employed for the prediction and a vote takes place for the prediction (Eq (8)):

$$\widehat{CSI}(t+h) = av_k \, \varphi_k(\widehat{CSI}(t+h)) \quad (8)$$

Where $\varphi_k$ are the different predictors before the aggregation and $av_k$ is the mean of the different predictors.

## 3.4. Experiment set-up

Various steps are necessary for developing forecasting simulations and to objectively compare methodologies in view to draw reliable conclusions. These guidelines are listed below.

### 3.4.1. Feature selection

One step is common to all data driven or machine learning models: the choice of the number of endogenous inputs to consider, this step is called "feature selection". The same methodology is applied for all numerical experiments by using the mutual information applied to the *CSI* time series: the auto-mutual information (*AMI*). This information measures statistical dependence between the current state, $CSI(t)$ and the previous measures $CSI(t-i)$ ($i = [0, \dots, N]$, for N number of observations). In contrast to the correlation coefficients defined by Spearman and Pearson, the *AMI* measures non-monotonic and other more complicated relationships between variables [49]. It is expressed as a combination of marginal and conditional entropies (respectively $H(CSI(t))$ and $H(CSI(t)|CSI(t-i))$) as described in the Eq (9).

$$AMI(CSI(t), CSI(t-i)) = H(CSI(t)) - H(CSI(t)|CSI(t-i)) \quad (9)$$

This quantity is constructed from the amount of randomness of the random variable $CSI(t)$ given that the value $CSI(t-i)$ is known. The maximum of lagged inputs to consider (i.e. number of inputs of the regression tree) corresponds to index $i_m$ of the first minimum of the auto-mutual information [50]. For example, this study in Ajaccio gives a first $AMI$ minimum at the 8th time lag ($i_m$ = 8), hence the regression tree will be constructed with 8 inputs (in Eq (3) $p = i_m$).

### 3.4.2. Filtering process

Concerning the *GHI* forecasting, it is a usual practice to transform the data in order to remove night hours and to objectively compare the studied predictors. This filtering is possible because during night time there is not significant PV electricity production [7]. We chose to apply a filtering criterion based on the solar elevation angle: solar radiation data for which the solar elevation angle is lower than 10°

have been removed. Moreover, this filtering process allows to consider only data associated with high measurement accuracy. Indeed, the measurement uncertainties associated to pyranometers are typically much higher than ± 3.0% for solar elevation angle of less than 10°[3]. Note that for the sunset and sunrise, the prediction is also very difficult (mainly for the mountainous area) owing to the geographic shield.

### 3.4.3.Validation rules and error metrics

The models parameters (i.e. an approximation to the function $f_n$ in equation (3)) are determined with the help of pairs of input and output examples contained in the training data. Once the model is fitted (or trained), it can be evaluated on a test data set totally independent of the training data. In our context, $\mathcal{D} = \{\mathbf{CSI}_\mathrm{i}, CSI(t_i + h)\}_{\mathrm{i}=1}^{\mathrm{N}}$ represents the training data set. The vector $\mathbf{CSI}_\mathrm{i}$ contains the $p$ past values (defining by the first minimum of the auto-mutual information) of the clear sky index for training sample $i$, $\mathbf{CSI}_i = (CSI(t_i), CSI(t_i - 1), \ldots., CSI(t_i - p - 1))^T$. The column vector inputs for all $N$ training cases can be aggregated in the so-called $N \times p$ design matrix **INPUT** and the corresponding measurements are collected in the vector **Output,** so we can write $\mathcal{D} = \{\mathbf{INPUT}, \mathbf{Output}\}$. Similarly, we have $\mathcal{D}_* = \{\mathbf{INPUT}_*, \mathbf{output}_*\}$ for the test data set. During this study, a k-fold methodology has been used. In a *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal sized subsamples [51]. k-fold cross validation should be employed to estimate the accuracy of the model induced from a regression algorithm, because the accuracy resulting from the training data of the model is generally too optimistic [52]. The *k* subsamples are used as training data. The cross-validation process is then repeated *k* times (k=50 in our case), with each of the *k* subsamples used exactly once as the training data. The *k* results from the folds can then be averaged (or differently combined) to produce a single estimation or used to compute probabilistic forecasts. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once [53]. More than 10-fold cross-validation is commonly used, but generally *k* remains an unfixed parameter (50 in our study). In this study, the accuracy of the forecasting models will be

estimated on the basis of the normalized root mean square error (*nRMSE*) and the skill core (*ss*) which are the two most used error criterion in solar radiation forecasting [3,54].

$$nRMSE = \sqrt{\mathrm{E}[(\hat{\mathrm{x}} - \mathrm{x})^2]}/\langle \mathrm{x} \rangle \text{ and } ss = 100.\left(1 - \frac{RMSE_{method}}{RMSE_{smart\ persistence}}\right) \quad (10)$$

According to the definition of the skill score factor *ss* (Eq (10), the scaled persistence model has a forecast skill *ss*= 0% [3]. A value of *ss*= 100% denotes a perfect forecast. Negative values of s indicate that the forecasting model fails to outperform the smart persistence model while positive values of *ss* means that the forecasting method improves on smart persistence. Furthermore, the higher the skill score is, the better the improvement is.

# 4. Probabilistic forecasts and prediction interval generation

Several methods are available to product a bounded prediction [55], however the methodology used in this paper is based on the bootstrap of the training set [56,57]. In our case, bootstrapping refers to the building of several predictors based on different subsets of the training data. The resampling is done with the k-fold method, with subsets containing only 10% of the training data, randomly chosen [55]. For large subsets, all the bootstrapped estimations are equivalent and the prediction distributions do not allow generation of prediction intervals. For each fold *k* (see section 3.4.3), a new learning subset is built and is used to train a new regression tree. Each tree will return a prediction and the *k* predicted values will be used to construct a cumulative distribution function (*CDF*, described in the next subsection) for each step. In our study, we took $k = 50$, leading to 50 predictions per step.

## 4.1. Percentile bootstrap

All bootstrap methods [15,58] are constructed without making assumptions about the underlying distributions from which our observations could have been sampled. With this kind of methods, the data themselves are used to estimate sampling distributions of predictions from the *k* subsets and *k* associated predictors. These estimated sampling distributions are then used to compute the confidence intervals based on percentiles estimation [59]. In descriptive statistics, a percentile is each of the 99

values that divide the data sorted into 100 equal parts, so that each part represents 1/100 of the population sample.

## 4.2. Prediction distributions

The cumulative distribution function (*CDF*) of prediction is computed from the predicted *GHI* probability distribution function (*PDF*). The prediction interval methodology used during this study requires to determine these two kinds of distributions:

-for the *PDF*, we use the fact that when a sufficiently large sample is available, *PDF* is equivalent to the histogram of the predicted $\widehat{GHI}$ [60],

-for the *CDF*, it is easy to compute because it is the normalized integral of the *PDF* [19]. Evaluated at a particular value (denoted $\widehat{GHI}^*$), *CDF* gives the probability that $\widehat{GHI}$ will take a value less than or equal to $\widehat{GHI}^*$ [16,20], it gives the area under the *PDF* from minus infinity to $\widehat{GHI}^*$.

An example of *CDF* of prediction (regression trees in Ajaccio with hourly data) is available in Figure 1.

Figure 1. Example of prediction cumulative distribution function (*CDF*) used during prediction interval generation

With this tool, all the percentiles can be generated in order to compute the prediction intervals. A percentile (or a centile) is a statistical measure indicating the value below which a given percentage of prediction falls. For example, the 30$^{th}$ percentile Q(0.3) is the value below which 30% of the prediction may be found (110 Wh/m² in Fig. 1, dashed line). The 25$^{th}$ percentile Q(0.25) is called first quartile, the 50th one Q(0.5) is the median, and the 75th percentile the third one Q(0.75). The median value Q(0.5) of the *CDF* can be considered as a $\widehat{GHI}$ prediction and the other quantiles are used to define ad-hoc prediction intervals. The 50 available intervals framing the forecast are given with the triplet of Eq (11) (with $n \in [1,50]$ defining the 50 intervals) and $\overline{\widehat{GHI}_n(t+h)} \geq \widehat{GHI}(t+h) \geq \underline{\widehat{GHI}_n(t+h)}$ (where ($\overline{\widehat{GHI}_n(t+h)}$ is the upper bound and $\underline{\widehat{GHI}_n(t+h)}$ the lower bound of the framing).

$$\begin{cases} \widehat{GHI}(t+h) = Q(0.5)|_{CDF(t+h)} \\ \overline{\widehat{GHI}_n(t+h)} = Q(0.5 + n.0.01)|_{CDF(t+h)} \\ \underline{\widehat{GHI}_n(t+h)} = Q(0.5 - n.0.01)|_{CDF(t+h)} \end{cases} \quad (11)$$

Where $CDF(t+h)$ is the *CDF* related to the 50 bootstrapped estimators (50-fold in section 3.4.3) as described in Eq. (12).

$$CDF(t+h) = CDF\left\{\widehat{GHI_k}(t+h)\right\} \text{ with } k \in [1,50] \quad (12)$$

In the figure 1, concerning $n$ = 20, the prediction would be equal to 116Wh/m², the higher bound to 122/m² and the lower bound to 110Wh/m².

## 5. Prediction interval relevance

The uncertainties induced by the global radiation forecasting can be decomposed into three parts [55]: the first one is related to the measure, the second one to the time series characteristics and the last one to the data driven method. Here, the method is based on a methodology of estimation of the uncertainties due to the data driven method, some authors [16,20,61] proposed probabilistic forecasting from data driven methods and exposed the related uncertainty. Two kinds of approaches, a bit similar, not competing but complementary, are proposed to draw a confidence band around predictions. If the works presented in [16,20] are quite similar to our approach, the tools that we use (prediction interval coverage probability and mean interval length) are probably more intuitive than the reliability diagram, the rank histogram, the continuous ranked probability score and its associated skill score. The grid manager needs to have a simple method for estimating the reliability of the forecasting and he must be able to draw easily conclusions and to react rapidly. The most interesting aspect of the present study is the simplicity of the algorithms. Moreover, our prediction band generation methodology is usable for all the machine learning methods and for different time granularities and horizons. New metrics taking into account the aspects of accuracy (measures between the bounds of the bands) and relevance (intervals length) are proposed in order to compare all the proposed bands.

## 5.1. Mean interval length (*MIL*) and prediction interval coverage probability (*PICP*)

The mean interval length (*MIL*) is defined by the difference between upper and lower bounds of the prediction interval (respectively $\overline{\widehat{GHI}_n(t+h)}$ and $\underline{\widehat{GHI}_n(t+h)}$) as described in Eq (13).

$$MIL_n = \langle \overline{\widehat{GHI}_n(t+h)} - \underline{\widehat{GHI}_n(t+h)} \rangle \quad (13)$$

The prediction interval coverage probability (*PICP*) is defined by the probability that the measure at $t+h$ be between the upper and lower prediction bounds [62]. It is estimated by the rule defined in the Eq. (14) ($N$ the number of available data).

$$PICP_n = (100/N).count(j) \text{ with } j : \overline{\widehat{GHI}_n(t+h)} \leq GHI(t+h) \leq \underline{\widehat{GHI}_n(t+h)} \quad \text{Eq 14}$$

To have a *PICP* close to 100% – that is to say, to be sure that the forecast will be, with 100% probability in the *MIL* range – a very large *MIL* must be chosen. But, for a grid manager, the interest of this approach will be inefficient. The goal of the prediction interval is to elaborate a methodology conducing to a good compromise between high value of *PICP* and a low value of *MIL*. Considering the theory of the global radiation calculation under clear sky, 100% of the predicted values will be included between the upper born corresponding to the global radiation in clear sky conditions and the lower one, diffuse radiation under clear sky, considering uncertainties related to the Solis modeling error (see Eq. 2).

## 5.2. A new test: the gamma test

A methodology called gamma test [63,64] have been developed in order to compare two 2D maps. In this paper this method is adapted to the interval comparison. With the two previous parameters *MIL* and *PICP* a gamma factor is computed using Eq. (15) concerning the 50 prediction intervals ($Tol_{MIL}$ and $Tol_{PCIP}$ are two tolerances depending on the considered problem, and $n \in [1,50]$, see Eq. (11)).

$$\Gamma_n = \sqrt{\left(\frac{MIL_n}{Tol_{MIL}}\right)^2 + \left(\frac{1-PCIP_n}{Tol_{1-PCIP}}\right)^2} \quad (15)$$

The higher the index, the less the prediction interval is efficient. With this index, it is possible to construct a statistical hypothesis test. In the start of the procedure, there are two hypotheses, the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$) defined by:

-$H_0$: "the prediction interval is relevant" if $\Gamma_n$<1,

-$H_1$: "the prediction interval is not relevant" if $\Gamma_n$>1

Computing $\Gamma_n$ for the $n$ intervals and each predictor, we propose a simple rule (or test) allowing to validate the prediction interval. This test allows to boundary the Cartesian coordinate plane defined by the two variables *MIL* and (100-*PICP*) (figure 3). This limit is an ellipse:

-inside the ellipse, the hypothesis $H_0$ is retained, it is the "prediction interval is relevant" area,

-outside the ellipse, $H_1$ is retained, it is the "prediction interval is not relevant" area.

We have taken $Tol_{MIL}$=0.5$\langle GHI(t)\rangle$ and $Tol_{1-PCIP}$=50%, meaning that a good interval proposes a *MIL* inferior than 50% of the mean value of the *GHI* and also allows to obtain a *PICP* higher than 50%. Note that this two values may be modified considering the problem. The gamma index is, for our case defined by the equation 16 with a factor $o$ ($\in [-1,1]$) allowing to bias one of the two variables ($MIL_n$ and $PICP_n$). In the following sections we have chosen $o$ = 0 in order to make the readability of the article easier.

$$\Gamma_n = \sqrt{(1+o)^2\left(\frac{MIL_n}{0.5\langle GHI(t)\rangle}\right)^2 + (1-o)^2\left(\frac{1-PCIP_n}{50\%}\right)^2} \quad (16)$$

# 6. Results

The data used in this application are solar global horizontal irradiations *GHI* measured in Ajaccio with an hourly time granularity for the years 2006-2014. In the next we apply the previous methodologies to the real case of the *GHI* prediction interval measured in-situ.

As previously mentioned, the machine learning methods can be applied only to stationary time series and to satisfy this requirement, the solar irradiations *GHI* are transformed in clear sky index *CSI*. Thus, CSI are predicted and at last, a reverse process is applied to obtain the corresponding *GHI*.

Only endogenous data were used as input and the number of inputs was calculated by auto-mutual information methods (AMI) (paragraph 3.4.1). The application of this AMI showed that the 8 previous data must be used to forecast the future data.

## 6.1. Predictors comparison (h+1)

Using the methodology previously described, some prediction intervals obtained with classical Regression tree, were calculated as shown in Fig. 2.

Figure 2. *GHI* prediction concerning 3 intervals defined from percentiles

This kind of figure is not enough interesting to consider a ranking of intervals or to estimate the best model of forecast to propose to the grid manager (the 1348$^{th}$ and 1366$^{th}$ hours are ill-modeled but is this a bad model?). In this scope, it is essential to compare the different parameters exposed previously (*nRMSE*, $\Gamma_{min}$, *MIL* and *PICP*). In Table 1, the values of these parameters for the 6 tested predictors (4 kinds of regression trees (classic, pruned, bagged and boosted) and 2 models of persistence (classic and smart one). For each regression tree, we show results related to classical forecasts realized using a time series formalism without quantiles estimation (denoted "classic" in the table) and to forecasts based on the Q(0.5) quantile estimation (denoted "median Q(0.5)").

Table 1. Predictors comparison for a time horizon h+1 hour (in bold the best results)

The $\Gamma_{min}$ values in Table 1 are related to the lowest value computed among the 50 prediction intervals descriptions for each regression trees type (see section 3.4.3) as shown in Eq (17).

$$\begin{cases} n_0 = \operatorname{argmin}_n(\Gamma_n) \\ \Gamma_{min} = \Gamma_{n_0} \end{cases} \quad (17)$$

The ellipse related to gamma index and the comparison of the predictors are plotted in Fig. 3. Each mark is related to a prediction interval based on the quantile estimation (50 marks per RT predictor).

Figure 3. Comparison of the predictors and of the predictions interval definition using the *MIL-PICP* plot. ✖ for the RT_pruned, ○ for the RT_bagged, △ for RT and ✚ for RT_boosted.

In Table 1 and Figure 3, only two regression trees pass the gamma test (RT and RT_boosted). The marks the closest to the origin of inner product space defined by *MIL* and 100-*PICP* (Fig. 3) show the best intervals and quantiles to consider. For all the models, the best configurations are obtained for *n* between 35 and 45 (see Eq. (11)) because theses couples of *MIL* and 100-*PICP* are those which minimize *MIL* and maximize 100-*PICP*. For all the tested models, the predictions of the Q(0.5) estimation give very good results compared to the classical approaches. Excepted for RT (*nRMSE*=0.2477), the machine learning methods propose results better than the smart persistence. In the next paragraph, a comparison of the parameters of prediction intervals will be analyzed depending on the horizon of forecasts.

## 6.2. Influence of the prediction horizon

In this part, only the best predictor, RT_boosted, is used for more clarity. Table 2 shows the values of the gamma index, *MIL*, *PICP* and skill Score for RT_boosted. These values are reported according to the time horizons between 1h and 6h by hourly time step.

Table 2. Prediction interval evolution concerning the considered horizon for the RT_boosted

For horizons upper than 2 hours, the best $\Gamma_{min}$ is related to a *MIL* close to 100 Wh/m² while for h+1 horizon it is 134 Wh/m². The *PICP* decreases from the horizon h+2. It is a consequence of the variability of the *CSI* (and *GHI*) and of the decrease of forecastability with the horizon. Note also that the skill score is, for h+2 and h+3 lower than 1, that is meaning that, for these horizons, the smart persistence is better than RT_boosted. But it has to be kept in mind that the smart persistence cannot generate confidence bands. In the next part, the performance of the band estimation will be improved with the Solis model.

### 6.3. Use of model of knowledge to bound the prediction intervals

In order to improve the prediction bounds, it is possible to use the Solis clear sky model [28]. Indeed, the measured horizontal global irradiation is most of the time lower than the global irradiation estimated with a clear sky and higher than the horizontal diffuse irradiation (the only solar component present by cloudy condition and minimum when the sky is clear). In Fig. 4, the scheme related to this correction.

Figure 4. Solis model as improvement of the band generation. In Gray the prediction band, in black the measurements and in blue the upper and lower bounds computed with the Solis model

In theory, this correction is attractive, because it allows, with a same *PCIP* to decrease the *MIL* and so to decrease the gamma index. In practice, the fact to use the Solis model is very interesting but the numerical uncertainties of Solis model modify slightly the *PICP* (less than 5 percentage points). The result of this improvement is shown in Table 3 for the RT_boosted predictor and a horizon h+1.

Table 3. Impact of the clear sky model improvement on the result of the prediction bands

For the same *PICP*, the *MIL* is decreased by 15% using the Solis model as limit of confidence bands. The gamma is strongly modified from 0.88 to 0.74, so it is improved by 16% with this simple modification. If we consider a large band defined only with the global and the diffuse clear sky modelling (Solis model), the *MIL* is equal to 408.87 Wh/m² and the *PICP* is close to 100% (all the measured values are between the global and the diffuse clear sky limits). The gamma index becomes 0.94 instead of 0.74 with the RT_boosted.

## 7. Comments and conclusions

In this paper, some results related to the *GHI* probabilistic forecasting were exposed with 2 persistence models (simple and smart one based on the Solis model) and 4 machine learning tools related to the regression trees (normal, pruned, boosted and bagged). A prediction band methodology was

elaborated, based on the bootstrap sampling and the cumulative distribution function (*CDF*) of prediction for horizons varying between 1 to 6 hours. A new validation tool was built based on the mean interval length (*MIL*) and prediction interval coverage probability (*PICP*) and called the gamma test. With these methods and the error metrics, a reliable prediction band was elaborated for Ajaccio with a *MIL* close to 113 Wh/m², a PCIP reaching 70% and a gamma index lower than 0.9. The proposed graphical tool would allow the grid manager to better assess the risk taking on the forecast. In future, this methodology will be applied in an on-line system based on the Tilos Island through the TILOS ((Technology innovation for the Local Scale, Optimum integration of Battery Energy Storage) H2020 project.

# References


[1] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renew. Sustain. Energy Rev. 27 (2013) 65–76. doi:10.1016/j.rser.2013.06.042.

[2] M.H. Agha, R. Thery, G. Hetreux, A. Hait, J.M. Le Lann, Integrated production and utility system approach for optimizing industrial unit operations, Energy. 35 (2010) 611–627. doi:doi: DOI: 10.1016/j.energy.2009.10.032.

[3] P. Lauret, C. Voyant, T. Soubdhan, M. David, P. Poggi, A benchmarking of machine learning techniques for solar radiation forecasting in an insular context, Sol. Energy. 112 (2015) 446–457. doi:10.1016/j.solener.2014.12.014.

[4] G. Notton, Problematic Integration of Fatal Renewable Energy Systems in Island Grids, in: Renew. Energy Serv. Mank. Vol II, Springer International Publishing, 2016: pp. 245–255. http://link.springer.com/chapter/10.1007/978-3-319-18215-5_22 (accessed July 24, 2017).

[5] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renew. Sustain. Energy Rev. 27 (2013) 65–76. doi:10.1016/j.rser.2013.06.042.

[6] T.M. Lai, W.M. To, W.C. Lo, Y.S. Choy, Modeling of electricity consumption in the Asian gaming and tourism center--Macao SAR, People's Republic of China, Energy. 33 (2008) 679–688. doi:doi: DOI: 10.1016/j.energy.2007.12.007.

[7] C. Voyant, F. Motte, A. Fouilloy, G. Notton, C. Paoli, M.-L. Nivet, Forecasting method for global radiation time series without training phase: comparison with other well-known prediction methodologies, Energy. 120 (2017) 199–208.

[8] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renew. Energy. 105 (2017) 569–582.

[9] R. Bubnová, G. Hello, P. Bénard, J.F. Geleyn, Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system, Mon. Weather Rev. 123 (1995) 515–535.

[10] P. Lauret, M. Diagne, M. David, A Neural Network Post-processing Approach to Improving NWP Solar Radiation Forecasts, Energy Procedia. 57 (2014) 1044–1052. doi:10.1016/j.egypro.2014.10.089.

[11] C. Voyant, M. Muselli, C. Paoli, M.-L. Nivet, Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, Energy. 39 (2012) 341–355. doi:10.1016/j.energy.2012.01.006.

[12] S. Ener Ruşen, A. Hammer, B. Akinoglu, Estimation of daily global solar irradiation by coupling ground measurements of bright sunshine hours to satellite imagery, Energy. 58 (2013). doi:10.1016/j.energy.2013.05.062.

[13] T. Schmidt, J. Kalisch, E. Lorenz, D. Heinemann, Evaluating the spatio-temporal performance of sky-imager-based solar irradiance analysis and forecasts, Atmos Chem Phys. 16 (2016) 3399–3412. doi:10.5194/acp-16-3399-2016.

[14] H.A. Nielsen, H. Madsen, T.S. Nielsen, Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts, Wind Energy. 9 (2006) 95–108. doi:10.1002/we.180.

[15] A.M. Alonso, D. Peña, J. Romo, Forecasting time series with sieve bootstrap, J. Stat. Plan. Inference. 100 (2002) 1–11. doi:10.1016/S0378-3758(01)00092-1.

[16] A. Grantham, Y.R. Gel, J. Boland, Nonparametric short-term probabilistic forecasting for solar radiation, Sol. Energy. 133 (2016) 465–475. doi:10.1016/j.solener.2016.04.011.

[17] M. Benghanem, A. Mellit, S.N. Alamri, ANN-based modelling and estimation of daily global solar radiation data: A case study, Energy Convers. Manag. 50 (2009) 1644–1655. doi:10.1016/j.enconman.2009.03.035.

[18] Daniel S. Wilks, Statistical methods in the atmospheric sciences, 2. ed., [Nachdr.], Elsevier [u.a.], Amsterdam, 2009.
[19] D.S. Wilks, D.S. Wilks, Statistical Methods in the Atmospheric Sciences An Introduction., Elsevier Science, Burlington, 2014. http://public.eblib.com/choice/PublicFullRecord.aspx?p=269991 (accessed February 19, 2016).
[20] M. David, F. Ramahatana, P.-J. Trombe, P. Lauret, Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models, Sol. Energy. 133 (2016) 55–72.
[21] GEOSS, (n.d.). https://www.earthobservations.org/geoss.php (accessed July 26, 2017).
[22] M. Korany, M. Boraiy, Y. Eissa, Y. Aoun, M.M. Abdel Wahab, S.C. Alfaro, P. Blanc, M. El-Metwally, H. Ghedira, K. Hungershoefer, others, A database of multi-year (2004–2010) quality-assured surface solar hourly irradiation measurements for the Egyptian territory, Earth Syst. Sci. Data. 8 (2016) 105–113.
[23] M. Sugiyama, M. Kawanabe, Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT Press, 2012. http://www.jstor.org/stable/j.ctt5hhbtm (accessed July 12, 2017).
[24] C. Paoli, C. Voyant, M. Muselli, M.-L. Nivet, Solar Radiation Forecasting Using Ad-Hoc Time Series Preprocessing and Neural Networks, in: Emerg. Intell. Comput. Technol. Appl., Springer Berlin / Heidelberg, 2009: pp. 898–907. http://dx.doi.org/10.1007/978-3-642-04070-2_95.
[25] C. Paoli, C. Voyant, M. Muselli, M.-L. Nivet, Use of Exogenous Data to Improve Artificial Networks Dedicated to Daily Global Radiation Forecasting, in: Valencia, Spain, 2010. https://acrobat.com/#d=jlEALth2cTsgdH1H7hnGog (accessed June 17, 2010).
[26] P. Ineichen, A broadband simplified version of the Solis clear sky model, Sol. Energy. 82 (2008) 758–762. doi:10.1016/j.solener.2008.02.009.
[27] Aerosol Robotic Network (AERONET) Homepage, (n.d.). https://aeronet.gsfc.nasa.gov/ (accessed July 28, 2017).
[28] R.W. Mueller, K.F. Dagestad, P. Ineichen, M. Schroedter-Homscheidt, S. Cros, D. Dumortier, R. Kuhlemann, J.A. Olseth, G. Piernavieja, C. Reise, L. Wald, D. Heinemann, Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module, Remote Sens. Environ. 91 (2004) 160–174. doi:10.1016/j.rse.2004.02.009.
[29] C. Voyant, M. Muselli, C. Paoli, M.-L. Nivet, Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation, Energy. 36 (2011) 348–359. doi:10.1016/j.energy.2010.10.032.
[30] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366. doi:10.1016/0893-6080(89)90020-8.
[31] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. 2 (1989) 303–314. doi:10.1007/BF02551274.
[32] C. Voyant, M. Muselli, C. Paoli, M.-L. Nivet, P. Poggi, P. Haurant, Predictability of PV power grid performance on insular sites without weather stations : use of artificial neural networks, in: Hambourg, 2009. http://docs.google.com/viewer?a=v&q=cache:PPeVFTGYZHEJ:arxiv.org/pdf/0905.3569+Predictability+of+PV+Power+Grid+Performance+on+Insular+Sites+without+Weather+Stations+:+Use+of+Artificial+Neural+Networks+%3F&hl=fr&gl=fr&pid=bl&srcid=ADGEEShgfR86_Oo5vmgw_vXKEhhBGzh0uh_muG7O7rYl9sUNJkbCCWHVvGMqAOaoenKdBVW_BWsQzIsfM6qct9c7ChYlhIRBwZMNGbmFF-UXC6FnsT8f7cP6aZS1Nt6YFmLSsckPKrPV&sig=AHIEtbTnkvnLST0La4vl58ZXkKXdIpT8nQ (accessed June 30, 2010).
[33] C. Voyant, M. Muselli, C. Paoli, M.-L. Nivet, Hybrid methodology for hourly global radiation forecasting in Mediterranean area, Renew. Energy. 53 (2013) 1–11. doi:10.1016/j.renene.2012.10.049.
[34] M. Abuella, B. Chowdhury, Random Forest Ensemble of Support Vector Regression Models for Solar Power Forecasting, ArXiv170500033 Cs. (2017). http://arxiv.org/abs/1705.00033 (accessed February 21, 2018).
[35] H.-Y. Cheng, Hybrid solar irradiance now-casting by fusing Kalman filter and regressor, Renew. Energy. 91 (2016) 434–441. doi:10.1016/j.renene.2016.01.077.

[36] J. Davis, Gradient Boosted Regression Trees for Forecasting Daily Solar Irradiance from a Numerical Weather Prediction Grid Interpolated with Ordinary Kriging, (n.d.). http://www.academia.edu/6806374/Gradient_Boosted_Regression_Trees_for_Forecasting_Daily_Solar_Irradiance_from_a_Numerical_Weather_Prediction_Grid_Interpolated_with_Ordinary_Kriging (accessed February 21, 2018).
[37] D. Gagne, A. Mcgovern, S. Haupt, J. Williams, Evaluation of statistical learning configurations for gridded solar irradiance forecasting, Sol. Energy. 150 (2017) 383–393. doi:10.1016/j.solener.2017.04.031.
[38] C. Voyant, T. Soubdhan, P. Lauret, M. David, M. Muselli, Statistical parameters as a means to a priori assess the accuracy of solar forecasting models, Energy. 90 (2015) 671–679.
[39] S.K. Aggarwal, L.M. Saini, Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest, Energy. 78 (2014) 247–256. doi:10.1016/j.energy.2014.10.012.
[40] M.P. Almeida, O. Perpiñán, L. Narvarte, PV power forecast using a nonparametric PV model, Sol. Energy. 115 (2015) 354–368. doi:10.1016/j.solener.2015.03.006.
[41] A. Lahouar, J. Ben Hadj Slama, Day-ahead load forecast using random forest and expert input selection, Energy Convers. Manag. 103 (2015) 1040–1051. doi:10.1016/j.enconman.2015.07.041.
[42] G.K.F. Tso, K.K.W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy. 32 (2007) 1761–1768. doi:10.1016/j.energy.2006.11.010.
[43] T. Hastie, R. Tibshirani, Generalized additive models, Stat. Sci. 1 (1986) 297–318.
[44] L. Breiman, Bagging Predictors, Mach. Learn. 24 (1996) 123–140. doi:10.1023/A:1018054314350.
[45] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.
[46] S. Ben Taieb, G. Bontempi, A.F. Atiya, A. Sorjamaa, A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition, Expert Syst. Appl. 39 (2012) 7067–7083. doi:10.1016/j.eswa.2012.01.039.
[47] G. De'ath, Boosted Trees for Ecological Modeling and Prediction, Ecology. 88 (2007) 243–251. doi:10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.
[48] L. Breiman, Bagging predictors., Mach. Learn. 24 (1996) 123–140. doi:10.1023 /A:10180 54 314350.
[49] C. Voyant, G. Notton, C. Paoli, M.L. Nivet, M. Muselli, K. Dahmani, Numerical weather prediction or stochastic modeling: an objective criterion of choice for the global radiation forecasting, Int. J. Energy Technol. Policy. (2014). https://hal.archives-ouvertes.fr/hal-00934872 (accessed August 19, 2015).
[50] these cyril voyant pdf free ebook download, n.d. http://ebookbrowse.com/these-cyril-voyant-pdf-d299673866 (accessed November 16, 2012).
[51] J.D. Rodriguez, A. Perez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 569–575.
[52] T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognit. 48 (2015) 2839–2846. doi:10.1016/j.patcog.2015.03.009.
[53] T.S. Wiens, B.C. Dale, M.S. Boyce, G.P. Kershaw, Three way k-fold cross-validation of resource selection functions, Ecol. Model. 212 (2008) 244–255.
[54] A.H. Murphy, Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, Mon. Weather Rev. 116 (1988) 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.
[55] C. Voyant, G. Notton, C. Paoli, A. Fouilloy, F. Motte, C. Darras, Uncertainties in global radiation time series forecasting using machine learning: The multilayer perceptron case, Energy. 125 (2017) 248–257.
[56] B. Chen, Y.R. Gel, N. Balakrishna, B. Abraham, Computationally efficient bootstrap prediction intervals for returns and volatilities in ARCH and GARCH processes, J. Forecast. 30 (2011) 51–71. doi:10.1002/for.1197.
[57] B. Efron, Bootstrap Methods: Another Look at the Jackknife, Ann. Stat. 7 (1979) 1–26. doi:10.1214/aos/1176344552.

[58] P. Bühlmann, Bootstraps for Time Series, Stat. Sci. 17 (2002) 52–72. doi:10.1214/ss/1023798998.
[59] R.R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, Academic Press, 2016.
[60] P. Givord, X. D'Haultfoeuille, La régression quantile en pratique, (2013).
[61] J.R. Trapero, Calculation of solar irradiation prediction intervals combining volatility and kernel density estimates, Energy. 114 (2016) 266–274.
[62] R. M, K. I, A. Vg, 2D-interval forecasts for solar power production, Sol. Energy. 122 (2015) 191–203.
[63] C. Voyant, P. Haurant, M. Muselli, C. Paoli, M.-L. Nivet, Time series modeling and large scale global solar radiation forecasting from geostationary satellites data, Sol. Energy. 102 (2014) 131–142.
[64] P. Haurant, C. Voyant, M. Muselli, M.L. Nivet, C. Paoli, Hourly global radiation prediction from geostationary satellite data, in: 2013.