Statistical parameters as a means to a priori assess the accuracy of solar forecasting models

Cyril Voyant[1*], Ted Soubdhan[2], Philippe Lauret[3], Mathieu David[3], Marc Muselli[1]

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

2-Laboratory LARGE, University of Antilles and Guiana, Guadeloupe, France

3-Laboratory PIMENT, University of Reunion, France

*Corresponding author. Address: Université de Corse, UMR SPE 6134, Route des Sanguinaires, 20000 AJACCIO, France. Tel.: +33 4 95 52 41 30; fax: +33 4 95 45 33 28. E-mail address: voyant@univ-corse.fr

Abstract

In this paper we propose to determinate and to test a set of statistical parameters (20) to estimate the predictability of the global horizontal irradiation time series and thereby propose a new prospective tool indicating the expected error regardless the forecasting methods a modeller can possibly implement. The mean absolute log return, which is a tool usually used in econometry, proves to be a very good estimator. Some examples of the use of this tool are exposed, showing the interest of this statistical parameter in concrete cases of predictions or optimizations.

Keywords

Solar forecasting, time series, clear sky models, fractal dimension, mutual information, log-return

1. Introduction

Solar radiation is one of the principal energy sources for physical, biological and chemical processes, occupying the most important role in many engineering applications [1]. The process of converting sunlight to electricity without combustion allows creating power without direct pollution. Thereby it is necessary to propose some prediction models [2] to use ideally this technology and in order to integrate solar energy PV production systems in the energetic mix [3]. Thus, solar energy forecasting is used to predict the amount of solar energy available in near terms [4]. Several methods have been developed by experts around the world and the mathematical formalism of Times Series (TS) has often been used [5]. TS is a set of numbers that measures the status of some activity over time. It is the historical record of some activity, with measurements taken at equally spaced intervals with a consistency in the activity and the method of measurement [6]. Some of the best predictors found in literature are Autoregressive and moving average (ARMA) [5,7,8], Bayesian inferences [9,10], Markov chains [11], k-Nearest-Neighbors predictors [12] or artificial intelligence techniques as the Artificial Neural Network (ANN) [9-11]. Although these methodologies are potentially good in many areas, we observed in our previous studies on global radiation prediction [9,13,14] that the simple model based on the persistence of the clear sky index gives often very good results with acceptable errors [15] for short term forecasting time horizon (less than 1 hour). It quite a standard in the solar forecasting community [16]. Indeed, instead of using directly the global horizontal irradiation (GHI), up to date forecast models predict the clear sky index at different forecast time horizons [17]. The corresponding forecast is obtained through the use of a clear sky model. In addition, forecasts based on persistence on the clear sky index exhibit rather good performance for forecast horizons < 1h. Also, the clear sky index is very important when one want to characterize the variability of a site [18,19], therefore a key component in solar forecasting is the clear sky index or equally the clear sky model. Several statistical parameters [18] aim at assessing the variability and consequently the difficulty to forecast the GHI [4,19,20]. The goal of this paper is to find a metric that is correlated to the forecasting accuracy (nRMSE; nMAE):

-Based on sound numerical experiments, we study the aforementioned metrics [18,19] or even simple metric (variation coefficient, mean, standard deviation, etc.);

-Conversely, other parameters related to financial econometric community are studied (return, absolute log return, etc.).

The paper is organized as follow: Section 2 describes the data and the statistical parameters used. Section 3 exposes the prediction methodology comparing the statistical parameters. In the two following sections, the comparison

result is shown concerning 8 different locations and through 3 illustrations, we show that this new metric enables to a priori assess the accuracy of the forecasting methods based on the clear sky index series.

2- Materials and methods

To estimate a time series prediction, a stationary hypothesis is often necessary. This result, originally shown for ARMA methods [21], can be also applicable for other estimators [22,23].This condition usually implies a stable process [24]. This notion is directly linked to the fact that whether certain feature such as mean or variance change over time or remain constant. To make the time series stationary, we used the clear sky index (CSI) methodology from a clear sky (CS) estimation done with a numerical model (Solis) [25]. The ratio between the GHI (x) and the clear sky model (CS) defines the clear sky index CSI ($CSI(t) = \frac{x(t)}{CS(t)}$).

2.1. Methodology

In order to assess the correlations between the proposed statistical parameters and the forecasting accuracy, we employ for each site two years of GHI and a repeated random sub-sampling validation is done in order to overcome the specificities of some years (data resampling). It is a common technique for estimating the performance of a classifier improving over the holdout method (corresponding to a 2-fold cross validation). This method randomly splits the dataset into training and validation data. For each such split, the model is fit to the training data (80% of data), and the predictive accuracy is assessed using the validation data (20% of data). The results are then averaged over the splits (mean prediction error). The advantage of this method (over k-fold cross validation) is that the proportion of the training/validation split is not dependent on the number of iterations (folds) [26]. The disadvantage of this method is that some observations may never be selected in the validation subsample, in practice, we have chosen 10 resamples. Note that, when the number of random splits goes to infinity, the repeated random sub-sampling validation becomes arbitrary close to the leave-p-out cross-validation.

2.2. Data

To validate this study, we choose 8 cities distributed around the world: 4 insular cities (2 in northern hemisphere, 1 in the northern tropical zone and 1 in the southern tropical zone), 3 continental cities in the north hemisphere and 1 continental city in the southern hemisphere. All these stations are part of a national measurement network and the measurement standards are almost equivalent. The three Islands (4 stations) are:

-Reunion Island; it exhibits a particular meteorological context dominated by a large diversity of microclimates. Two main regimes of cloudiness are superposed: the clouds driven by the synoptic conditions over the Indian Ocean and the orographic cloud layer generated by the local reliefs. The data used to build the models are measured at the meteorological station of St Pierre (21°20'S ; 55°29'E, 75m a.s.l) located in the southern part of Reunion Island. Measurements are available on an hourly basis and two years of data (2011 and 2012).

-Guadeloupe Island, we have used a two years database (2011 and 2012) of GHI measured on an hourly basis at the Meteo France meteorological station of le Raizet (16°26N, 61°24W, 11m asl) The daily average for the solar load on a horizontal surface is around 5 kWh/m². A constant sunshine combined with the thermal inertia of the ocean makes the air temperature variation quite weak, between 17°C and 33°C with an average of 25°C to 26°C. Relative humidity ranges from 70% to 80% and the trade winds are relatively constant all along the year. As for Reunion Island, two main regimes of cloudiness are superposed: the clouds driven by the synoptic conditions over the Atlantic Ocean and the orographic cloud layer generated by the local reliefs.

-Corsica Island, the data used to build the models, are GHI measured at the meteorological station of Ajaccio (41°55'N, 8°44'E, 4m asl) and Bastia (42°42'N, 9°27'E, 10m asl). They are located near the Mediterranean Sea and nearby mountains (1000 m altitude at 40km from the sites). 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. The data representing the global horizontal solar radiation were measured on an hourly basis from 1998 to 1999 (exactly two years). As for all experimental acquisitions, missing values are observed, here, this represents less than 2% of the data. A classical cleaning approach is then operated in order to identify and remove this data.

The four continental stations are:

-Northern continental cities; the 3 studied cities are Marseille (43°17'N, 5°22'E, 10m asl), Nice (42°42'N, 9°27'E, 10m asl) and Montpellier (43°36'N, 3°52'E, 27m asl). These locations (metropolitan France) are characterized by the same climate, namely a Mediterranean climate with mild, humid winters and warm to hot, mostly dry summers. If concerning the two first cities are near mountains (over 1000m asl) the third is located in a flat area. The measures were recorded during the years 2007-2008.

-Southern city; Melbourne is located in the south-eastern part of mainland Australia (37°48'S, 144°57'E, 60m asl). It has a moderate oceanic climate and is well known for its changeable weather conditions. This is mainly due to Melbourne's location situated on the boundary of the very hot inland areas and the cool southern ocean. This temperature difference is more pronounced in the spring and summer months and can cause very strong cold

fronts to form. These cold fronts can be responsible for all sorts of severe weather from gales to severe thunderstorms and hail, large temperature drops, and heavy rain. The measures were recorded during the years 2008-2009.

2.3. Clear sky modelling

Among the clear sky models that can be found in literature, for this study, we have choose the simplified "Solis clear sky" model based on radiative transfer calculations and the Lambert-Beer relation [27]. In previous studies, this model has shown its effectiveness to fit the global radiation of cloudless days. In this case, the clear sky global horizontal irradiance (*CS*) reaching the ground is defined by the equation 1.

$$CS(t) = H_0.\exp(-\tau / \sin^b(h(t))).\sin(h(t)) \quad \text{Equation 1}$$

Where τ is the global total atmospheric optical depth, *h* is the solar elevation angle, *b* is a fitting parameter and H_0 the global radiation on the top of atmosphere. Concerning the global radiation forecasting, it is a common practice to filter out the data in order to remove night hours and to compare objectively the studied predictors. This choice is justified because during these periods there is obviously no significant solar radiation to generate electricity (i.e. low potential overnight). We chose to apply a selection criterion based on the solar zenith angle (SZA=90°-*h*): solar radiation data for which the solar zenith angle is greater than 80° have been removed. This transformation is equivalent to a filtering related to the solar elevation angle (*h*) lower than 10°. In addition, this filtering process allows to discard data with less precision as measurement uncertainties associated to pyranometers are typically much higher than ± 3.0% for SZA > 80°. Note that for the sunrise and sunset, the prediction is also very difficult (mainly for the mountainous area) owing to the geographic shield. All the clear sky models are linked to the atmospheric parameters [20], in the case of Solis model the aerosol optical depth is very important and can dramatically alter the output.

2.4. Statistical parameters

Whether in the field of renewable energy or financial markets, it is now common to speak of "prediction". In fact, the estimated future value of a meteorological variable (such as global irradiation) or of a financial product may, under the aspect of time series analysis, be treated in the same way. Generally, in order to estimate the future value of a variable, it is essential to have information on its past evolution. A time series of a given variable is intuitively defined as an ordered sequence of past values [28]. To use the formalism of the TS, it is necessary to

consider first some definitions. The current value at *t* of the time series is noted x_t (representing in our case the GHI or the CSI) where *t*, the time index, is between 1 and *n*, with *n* is the total number of observations. For the horizon 1 (the simplest case; one hour head in our case), the general formalism of the prediction will be represented by Equation 2 where ϵ represents the error between the prediction and the measurement, f_n the model to estimate and *t* the time index taking the (n-p) following values: *n, n-1,..., p+1, p*. The variable *p* is the number of model parameters (it is assumed that $n \gg p$). [29]

$$x(t+1)=f_n\,(x(t),x(t-1),\ldots.,x(t-p+1)\,)+\epsilon(t+1) \qquad \text{Equation 2}$$

Studies in finance and econometrics have yielded many models more or less sophisticated. These were taken in the context of other subjects, including the prediction of global solar radiation.

In this paper, we want to apply some statistical parameters on different time series and discuss about their impact on the error of prediction generated by different prediction models. In financial modelling or econometrics, a lot parameters were developed (return, volatility, etc.). In the following, we propose to adapt some of these parameters to solar radiation forecasting. The first studied parameter is the simple ratio at the time *t* which is defined in the equation 3.

$$ratio(t) = \frac{CSI(t)}{CSI(t-1)} \qquad \text{Equation 3}$$

This new time series (hourly step in our case) represents the increase (ratio>1) or the decrease (ratio<1) of the global radiation at time *t.* We define also the simple mathematical average of a series of ratio (mean ratio; *r*) generated over a period of time. An average ratio is calculated in the same way, a simple average for any set of numbers. Note that mean absolute ratio is equivalent to the mean ratio regarding the global irradiation (all the values are positive). From the ratio parameter, it is possible for each step to define the simple return or arithmetic return, *r=ratio(t)-1*.

One of the benefit of using returns is normalization: measuring all variables in a comparable metric, thus enabling evaluation of analytic relationships among two or more variables despite originating from CSI series of unequal values (or from different sites). This is a requirement for many multidimensional statistical analysis and machine learning techniques. For example, interpreting an equity covariance matrix is made wise when the variables are both measured in percentage. Usually it is not the return which is used but the log-return (*logr*) defines by the equation 4.

$$logr(t) = log\,\frac{CSI(t)}{CSI(t-1)} = \log\big(CSI(t)\big) - \log\big(CSI(t-1)\big) \qquad \text{Equation 4}$$

The log return has the nice property of log-normality. If we assume that CSI is distributed log normally (which, in practice, may or may not be true for any given series), then *log(ratio)* is conveniently normally distributed. Unfortunately there are a number of points that initially discourage acceptance of the idea of returns.

Two other important statistical propertis of the return or log-return are the skewness and kurtosis of the distribution. Skewness is the normalized third central moment and it describes the symmetry of the random variable with respect to its mean. Kurtosis is the normalized fourth central moment and it describes the behavior of the tail of the distribution. It is independent of scale and location parameters and so can be used as a comparison coefficient between the empirical data and the normal distribution. Together skewness and kurtosis summarize the extent of asymmetry and tail thickness of the distribution. In figure 1 is shown the log-return distribution computed in Ajaccio ($1128.\exp(-((x-0.024)/0.1258)^2$; $R^2=0.9951$) considered as normally distributed.

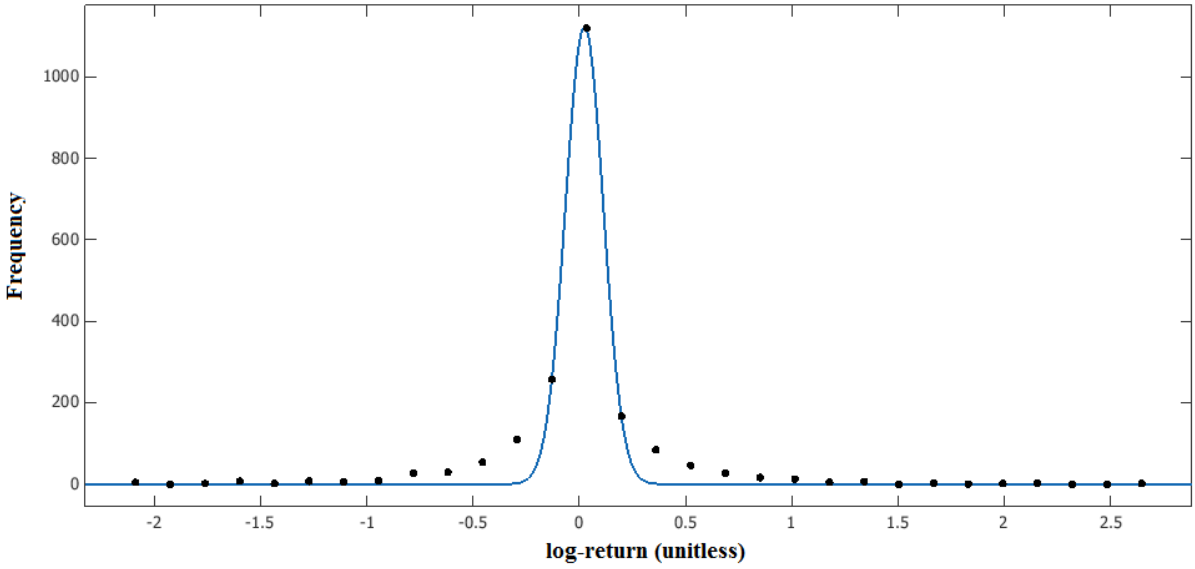

Figure 1. Gaussian fit for log-return distribution (line) and measured values (point).

The equation 5 defines the *ih* moment about the mean of series *logr* (where E is the expectation operator).

$$\mu_i = E[(logr(t) - E[logr(t)])^i] \qquad \text{Equation 5}$$

The second central moment μ_2 is the variance (this square root represents the standard deviation noted *std* in the next). The third and fourth central moments are used to define the standardized moments which are used to define skewness (*skew*) and kurtosis (*kurt*), respectively. In statistics, the Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test statistic *JB* is defined

by the equation 6 in the case of the log-return. The lower is JB and more the series can be described by a normal distribution.

$$JB = \frac{n}{6}\left(skew\left(logr(t)\right)^2 + \frac{1}{4}\left(kurt\left(logr(t)\right) - 3\right)^2\right) \quad \text{Equation 6}$$

In order to better take into account the intermittency in the CSI series (no compensation effect between positive and negative values), it is possible to define the absolute value of the ratio, return or log-return. In the case of the ratio the mean of the absolute log-return is done by operating a temporal mean on $\left|logr\left(CSI(t)\right)\right|$ (see equation 7).

$$|logr(t)| = abs\left(\log\left(CSI(t)\right) - \log\left(CSI(t-1)\right)\right) \quad \text{Equation 7}$$

Note that this parameter doesn't follow a Gaussian distribution from its construction. If the previous parameters allow to consider the noise in the series (high frequency at 1 hour^{-1}), we choose also to study a more classical parameter allowing to estimate the seasonality of the *CSI* series (low frequency at 1 year^{-1}): the coefficient of variation (*CV* defined for the *CSI* in the equation 8).

$$CV = std(CSI(t))/E\left(CSI(t)\right) \quad \text{Equation 8}$$

The next studied parameter (*V*) is based on the Marquez formula [19] and is described in the equation 9. It is almost equivalent to the $|logr(t)|$ with a L^2-norm and without the log transformation.

$$diff = CSI(t) - CSI(t-1) \quad \text{and} \quad V = \sqrt{mean(diff^2)} \quad \text{Equation 9}$$

An extension of this metric (called *P*) proposed by Perez et al. [18] is based on the dispersion of the quantity *diff* as shown in the next equation (equation 10). Note that if $mean(diff) \to 0$ so $V \to P$.

$$P = std(diff) \quad \text{Equation 10}$$

One of the other intrinsic characteristics of each time series is the number of lag statistically dependent. In fact to predict *CSI(t+1)*, it is often not necessary to know all the previous value of the time series. The order of dependency can be computed by two ways: the autocorrelation and the auto-mutual information. In contrast to the linear dependence measured by autocorrelation, auto-mutual information supplies a measure of general dependence. Mutual information answers the following question: given the observation of *CSI(t)*, how accurately can one predict *CSI(t+* τ*)*? Thus, successive delay coordinates are interpreted as relatively independent when the mutual information is small. In practice we, the first minimum of the auto-mutual information (*AMI*) is considered as τ and corresponds (not exactly but this approximation is sufficient in the study context) to information dimension (*ID*). The equation 11 describes the *AMI* computing ($p\left(CSI(t), CSI(t-\tau)\right)$ is the joint probability and $p\left(CSI(t)\right)$ the marginal probabilities.

$$AMI(\tau) = \sum(\log \frac{p(CSI(t),CSI(t-\tau))}{p(CSI(t))\,p(CSI(t-\tau))})p(CSI(t), CSI(t-\tau)) \quad \text{Equation 11}$$

The interested reader can refer to [30] for more information concerning the Shannon entropy and the mutual information. The last studied parameter is the fractal dimension (*FD*) computed with the Box counting method. A fractal dimension is a ratio providing a statistical index of complexity comparing how detail in a pattern (strictly speaking, a fractal pattern) changes with the scale at which it is measured. Box counting is a method of gathering data for analyzing complex patterns by breaking a dataset into smaller and smaller pieces, typically "box"-shaped, and analyzing the pieces at each smaller scale in order to quantify box-counting dimension and fractal scaling. Suppose that *N(ε)* is the number of boxes of side length *ε* required to cover the set, the Fractal dimension is defined by the equation 12 [31].

$$FD = \frac{\log(N(\varepsilon))}{\log(\frac{1}{\varepsilon})} \quad \text{Equation 12}$$

The table 1 summarizes the studied statistical parameters.

*Clear sky index defined in the section 2

	Initial time series*	**Ratio (Eq 3)**	**Log-retrun (Eq 4)**	**Absolute log-return (Eq 7)**
Mean	-	*mean(ratio)*	*mean(logr)*	*Mean(abs_logr)*
Standard deviation	-	*std(ratio)*	*std(logr)*	*std(abs_logr)*
Kurtosis (Eq 5)	-	*kurt(ratio)*	*kurt(logr)*	*kurt(abs_logr)*
Skewness (Eq5)	-	*skew(ratio)*	*skew(logr)*	*skew(abs_logr)*
Jarque-Nera stat (Eq 6)	-	*JB(ratio)*	*JB(logr)*	*JB(abs_logr)*
Marquez parameter (Eq 9)	*V*	-	-	-
Perez parameter (Eq 10)	*P*	-	-	-
Coefficient of variation (Eq 8)	*CV*	-	-	-
Information dimension (Eq 11)	*ID*	-	-	-
Fractal dimension (Eq 12)	*FD*	-	-	-

Table1. List and names of the studied statistical parameters (- means that the parameter is not tested here)

Table1. List and names of the studied statistical parameters (- means that the parameter is not tested here)

3- Forecasting methodologies

In this study, we chose to use 3 forecasting methodologies, if the two first related to the persistence (simple and scaled) are very easy to use, the third one is a more sophisticated artificial intelligence tool: the artificial neural networks.

The first type of forecasting method studied is the persistence model; the simplest way of producing a forecast. The persistence assumes that the conditions at the time of the forecast will not change (Eq 13; x_t are the global radiation time series elements).

$$\hat{x}(t+1) \overset{P}{\rightarrow} x(t) \qquad \text{Equation 13}$$

To take into account the fact that the TS is periodical (due to the solar geometry), it is possible to correct the persistence form with a scale term noted S_t (Eq 14).

$$\hat{x}(t+1) \overset{SP}{\rightarrow} x(t).(S(t)) \qquad \text{Equation 14}$$

The last factor of this product can be computed using a clear sky modeling regardless the time *t*. If the clear sky time series is named *CS(t)*, the scaled persistence become:

$$\hat{x}(t+1) \overset{SP}{\rightarrow} x(t).\left(\frac{CS(t+1)}{CS(t)}\right) \qquad \text{Equation 15}$$

Note that the scaled persistence is in fact a persistence of the ratio $\frac{x(t)}{CS(t)}$ which is also called clear sky index (*CSI*).

In the discussion part, a prediction using multilayer perceptron (MLP; particular artificial neural network) is presented. To forecast the time series, a fixed number *p* of past values are set as inputs of the MLP, the output is the prediction of the future value. Considering the initial time series equation (Equation 2), we can transform this formula to the non-linear case of one hidden layer MPL with *b* related to the biases, *f* and *g* to the activation functions of the output and hidden layer, and ω to the weights (see equation 16). The number of hidden nodes (*H*) and the number of the input nodes (*In*) allow detailing this transformation [32,33]:

$$\widehat{CSI}(t+1) \overset{\text{MLP}}{\longrightarrow} f\left(\sum_{i=1}^{H} o_i\, \omega_i^2 + b^2\right) \text{ with } o_i = g\left(\sum_{j=1}^{In} CSI(t-j+1)\, \omega_{ij}^1 + b_i^1\right) \qquad \text{Equation 16}$$

In the presented study, the MLP has been computed with the Matlab© software and its Neural Network toolbox. The optimization of the number of input nodes is done with the automutual information and the number of hidden neurons is taken equal to the input nodes number. The results shown in the next are related to the best networks among 10 different trainings coupled with a random weight initialization. Interested readers can consult previous papers for more details [15]. The error metrics used during these manipulations are the nRMSE $\left(=\frac{\sqrt{E[(\widehat{CSI}-CSI)^2]}}{E[CSI]}\right)$ and the nMAE ($\left(=\frac{\sqrt{E[|\widehat{CSI}-CSI|]}}{E[CSI]}\right)$. In this section, we have defined the normal and scaled persistence and *MLP* predictions of the *CSI*, in the following the correlations are related to this parameter.

4- Results and parameters validation

In order to estimate the a priori parameters linked to prediction quality, we expose in the table 2 the Spearman and the Pearson correlation factors between nRMSE (and nMAE) versus the 20 statistical parameters mentioned above (a p-value<0.05 meaning a statistical dependence between variables). These correlation factors are computed over the 8 locations: Ajaccio, Bastia, Saint-Pierre, Melbourne, Marseille, Montpellier, Nice and Le Raizet. In this first study, we decide to show only the scaled persistence predictor. In fact, with the simple persistence, there is no significant conclusion and no significant correlation and with the MLP. A known problem in training ANN is that the training process can be trapped in a local minimum generating sometimes a high variance between the 10 runs (see k-fold part in section 2) and so a difficulty of interpretation. Moreover, some authors have shown that the global radiation predictions done with a scaled persistence is often better than the prediction done with more complicated models [12,16]. Concerning the two correlation coefficient, the Spearman coefficient is computed on ranks and so depicts monotonic relationships while the Pearson coefficient is on true values and depicts linear relationships. The interpretation of both allows to test if the correlation is monotonic and/or linear.

	nRMSE				**nMAE**			
	Pearson		Spearman		Pearson		Spearman	
Parameters	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
mean(ratio	0.305	0.462	0.619	0.115	0.142	0.738	0.548	0.171
mean(logr)	-0.012	0.978	0.190	0.665	0.045	0.916	0.048	0.935
mean(abs_logr)	**0.864**	**0.006**	0.619	0.115	**0.848**	**0.008**	0.548	0.171
std(ratio)	0.302	0.468	0.500	0.216	0.138	0.744	0.286	0.501
std(logr)	0.551	0.156	0.595	0.132	0.433	0.284	0.524	0.197
std(abs_logr)	0.436	0.281	0.571	0.151	0.303	0.466	0.500	0.216
kurt(ratio)	0.478	0.231	0.286	0.501	0.338	0.413	0.000	1.000
kurt(logr)	0.245	0.559	-0.286	0.501	0.077	0.856	-0.548	0.171
kurt(abs_logr)	0.277	0.507	0.190	0.665	0.110	0.795	-0.071	0.882
skew(ratio)	0.419	0.302	0.286	0.501	0.282	0.498	0.000	1.000
skew(logr)	-0.244	0.560	-0.048	0.935	-0.321	0.438	-0.143	0.752
skew(abs_logr)	0.245	0.558	-0.143	0.752	0.078	0.854	-0.405	0.327
JB(ratio)	0.440	0.275	0.286	0.501	0.284	0.495	0.000	1.000

JB(logr)	0.285	0.494	-0.333	0.428	0.120	0.777	-0.571	0.151
JB(abs_logr)	0.295	0.478	0.190	0.665	0.130	0.759	-0.071	0.882
V	0.330	0.425	0.238	0.582	0.426	0.293	0.357	0.389
P	0.330	0.425	0.238	0.582	0.426	0.293	0.357	0.389
CV	0.469	0.241	0.357	0.389	0.447	0.267	0.357	0.389
FD	-0.177	0.676	0.183	0.657	-0.284	0.496	-0.052	0.914
ID	-0.329	0.427	-0.176	0.679	-0.473	0.236	-0.454	0.261

Table 2. Correlations between 18 statistical parameters computed a priori and the error of prediction done with the scaled persistence (nRMSE/nMAE over 10 repeated random sub-sampling validations). In bold the correlation significantly different from zero (p-value<0.05; i.e. statistical dependence between variables)

In fact, although parameters as Kurtosis or Fractal dimension seem interesting and directly linked to the predictability of the time series, the result of this study is that only the mean absolute log-return is linked to the error of prediction concerning the two studied metrics. The relationship is monotonic and linear between the two compared elements. The spearman factor generates any evidence of this link. In the figure 2 is represented for the 8 cities the plot between nRMSE/nMAE and the absolute log-return

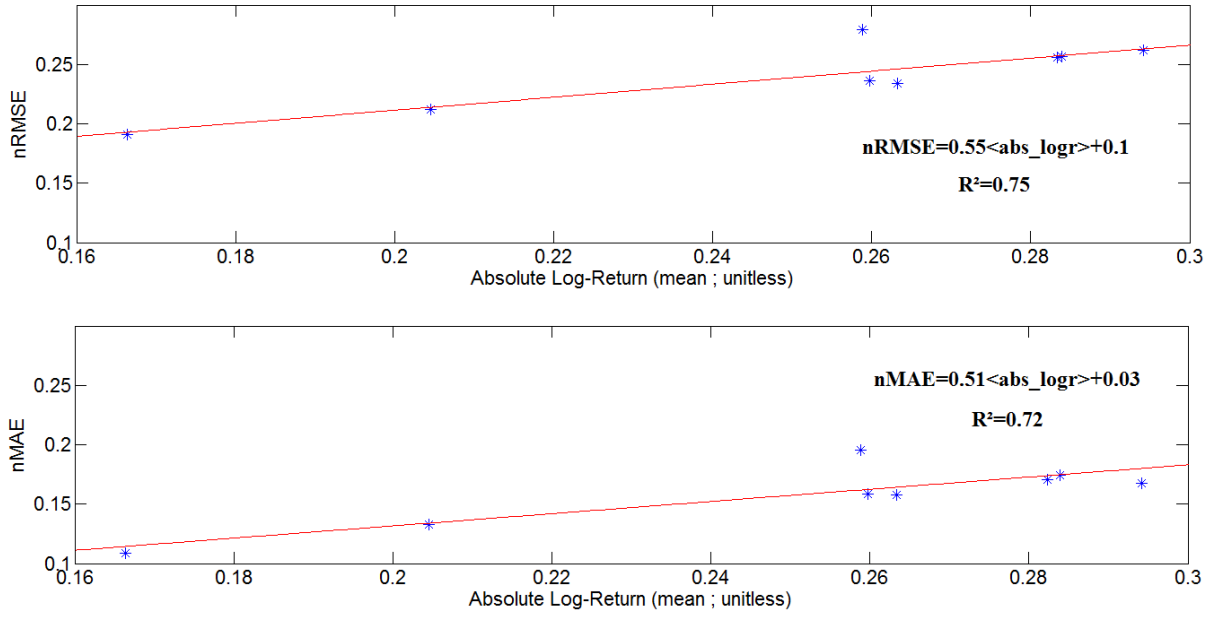

Figure 2. Linear dependence between the prediction error and the mean of the absolute log-return

Although there are a few points in these figures, it is evident that there is a relation (perhaps not linear) between these two kinds of parameters.

In order to increase the number of point we have introduced a phase shift in the clear sky modeling. To perform it, we modify 50 times the solar elevation (see equation 1) generating a delay or an advance of the clear sky model relative to measures. The figure 3 shows the results of this new study for the Ajaccio City.

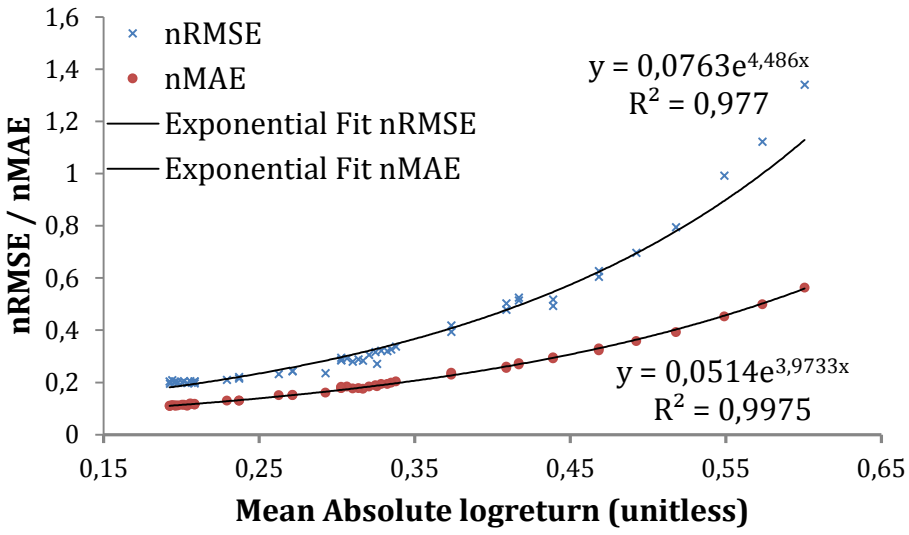

Figure 3. Relationship between mean absolute log-return and prediction error in Ajaccio

Previously we proposed that the link between prediction error and mean absolute log-return was linear here, with the increase of the number of points, we can easily notice that the relation is exponential. Knowing the mean absolute log-return provide an a priori information about the expected error.

These results also shows the importance of a good clear sky modelling in the performance of a forecasting model. A wrong parameterization of the clear sky model raises grossly the error of prediction. Concerning others cities, the trend of the curve linking error metric and mean absolute logreturn is also an exponential growth, the values of the fit parameters are slightly different but the interpretation is equivalent. For these reasons we choose to not show the other cases in this study.

5- Discussion about the interest of an a priori indicator of predictability

We will try through three examples to identify the interest of the mean absolute log-return, the first and second one refers to his relative aspect and the third to the absolute aspect of it's use.

-Example 1: we search to establish the prediction localized in a single location (Guadeloupe; Le Raizet) and we have 3 clear sky models available to make the time series stationary (Solis model, Kasten Model [] and Bird model []). In this case, we can imagine to compare the mean absolute log-return generated by the three clear sky model, we obtain for Solis 0.246, for Kasten 0.276 and for Bird 0.2749. Taking into account what we have exposed previously, we decide before to try to forecast the global radiation that the model Solis is the more interesting for this study because it generates the lower mean log-return which we consider as a "quality index". For this study only the nRMSE was used and two years of measurements (one year for training and one year for the prediction). The result of the MLP prediction are related to 10 runs and the network that induces the lower nRMSE is keep. In the next table (table 3) are shown the result of the prediction for Guadeloupe.

	Solis	**Kasten**	**Bird**
MLP	0.2512	0.2662	0.2627
Scaled Persistence	0.2673	0.2815	0.2791
Persistence	0.3760	0.3760	0.3760

Table 3. nRMSE for the irradiation prediction in Guadeloupe for three clear sky models

In this example, the anticipated order based on the mean log-return interpretation is exactly similar to the ranking obtained after the modeling by MLP or scaled persistence. Concerning the persistence, the a priori parameter is not interesting.

-Example 2: We search to optimize a clear sky index in order to predict the global radiation in Ajaccio via a scaled persistence methodology. The CS chosen is Solis but we have a parameter called aerosol optical depth (ADO) taking value between 0 and 1. The purpose of this example is to use an a priori parameter to estimate the best value of AOD to consider. If we plot the mean absolute logreturn considering the AOD (see figure 4) we observe a minimum value of mean(abs_logr) for an AOD between 0.15 and 0.25 (exactly 0.18).

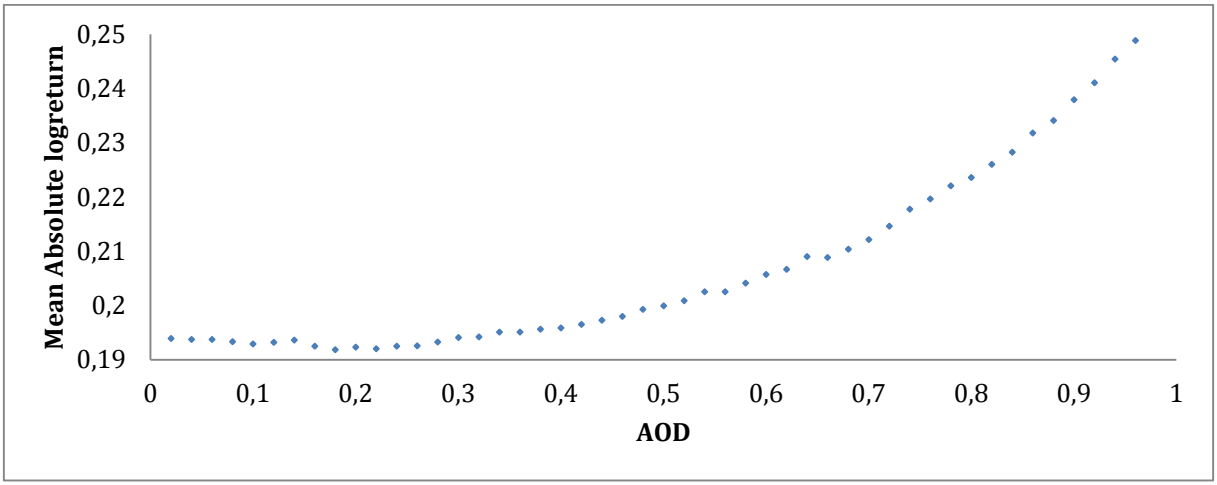

Figure 4. Relationship between mean absolute log-return and aerosol optical depth

According to our study, we imagine that the best value of AOD (giving the best result of prediction) is close to 0.18. Note that the AERONET group (http://aeronet.gsfc.nasa.gov/new_web/aerosols.html) proposes a mean value of AOD for 500 and 350nm (wavelength) respectively equal to 0.17 and 0.19 (computed during year 2002 to 2012 in Ajaccio). In the figure 5, we expose now the nMAE (obtained with a scaled persistence estimation and the repeated random sub-sampling) versus the AOD and we infer that the optimum value is 0.17 (very close to the value determined with the mean absolute logreturn approach). In this case the absolute log return method is efficient and allow to optimize a clear sky model, making thus, stationary a global radiation time series without having to test predictions related to a myriad of configurations.

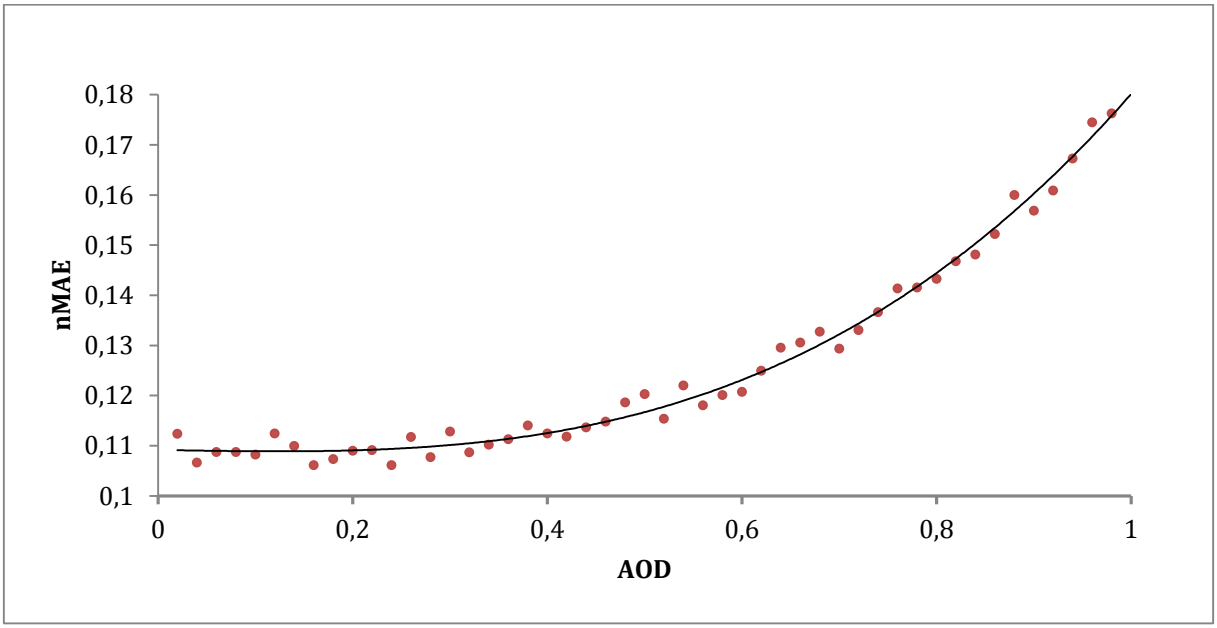

Figure 5. Relationship between error metric (nMAE) and aerosol optical depth

-Example 3: We search the global radiation prediction error in Reunion island (mean absolute log-return=0.2195) generated by scaled and simple persistence and want to know if this model is better than a prediction done with MLP (nRMSE=0.2054 and nMAE=0.1454). There are two methods, first one is to compute the global study and to compute the error of prediction, and second one, we decide to use the curve generated in the previous part (figure 3). Knowing the a priori parameter (mean(abs_logr)=0.2195), we can try to apply it on the fit equation generated for the Ajaccio case (nRMSE=0.0763exp(4.4486mean(abs_logr)) and nMAE=0.0514exp(3.9733mean(abs_logr)) and conclude that nRMSE~0.2042 and nMAE~0.1229 and finally that the scaled persistence is better than MLP in this case. The true error metrics computed from a scaled persistence modeling are nRMSE=0.2123 and nMAE~0.1325. For the two error metrics, there is a difference close to 1 percentage point, modifying the interpretation of the predictor ranking. In this case the mean absolute log-return can't be used and the operator is forced to do all the steps of the prediction whatever the site studied. Note that if we use now the equations of the figure 2 which are generated after the study of the 8 cities, results and interpretations are different (nRMSE=0.2207 and a nMAE= 0.1419) but the difference is also close to 1 percentage point.

6- Conclusion

In this paper, we have shown that the use of well-chosen statistical parameters could help the modeler and so the PV manager to optimize the clear sky index and to establish the ranking related to different clear sky index without do the experiences. We have shown that the methodology works for scaled persistence and MLP modeling but is not efficient for the simple persistence. One way of generalization could be to extend it to others machine learning estimators (SVM, Bayesian neural network, Gaussian process, etc.). Among all the parameter tested, only one has proved efficiency (mean absolute log-return), but with others cities or/and more cities the result could have been different. It is essential to use and test econometrics tools (log-return, volatility, etc.) to estimate simple prediction quality indexes, our predictions would be perhaps better and it would be time saving. Concerning the absolute aspect of the log-return and the use of the fit equation linking error metric and absolute log-return, we didn't show the feasibility of it, for us it is too dependent of the location of the study. For the same location, the methodology

is possible, the determination coefficients are close de 1 for nRMSE and nMAE, but transposed a curve generated in a city into an other cities, it is certainly not realistic.

References

[1] Badescu V. Modeling solar radiation at the earth's surface: recent advances. Springer; 2008.

[2] Diagne M, David M, Boland J, Schmutz N, Lauret P. Post-processing of Solar Irradiance Forecasts from WRF Model at Reunion Island. Energy Procedia 2014;57:1364–73. doi:10.1016/j.egypro.2014.10.127.

[3] Bofinger S, Heilscher G. solar electricity forecast : approach and first results, 2006.

[4] Perez R, Hoff T, Dise J, Chalmers D, Kivalov S. The Cost of Mitigating Short-term PV Output Variability. Energy Procedia 2014;57:755–62. doi:10.1016/j.egypro.2014.10.283.

[5] 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.

[6] Soubdhan T, Abadi M, Emilion R. Time Dependent Classification of Solar Radiation Sequences Using Best Information Criterion. Energy Procedia 2014;57:1309–16. doi:10.1016/j.egypro.2014.10.121.

[7] Brockwell PJ, Davis RA. Time series: theory and methods. 2nd ed. New York: Springer-Verlag; 1991.

[8] Bourbonnais R, Terraza M. Analyse des séries temporelles : application à l'économie et à la gestion. 2e éd. Paris: Dunod; 2008.

[9] Lauret P, Fock E, Randrianarivony RN, Manicom-Ramsamy JF. Bayesian neural network approach to short time load forecasting. Energy Convers Manag 2008;49:1156–66.

[10] Lynch SM. Bayesian Statistics. Encycl. Soc. Meas., New York: Elsevier; 2005, p. 135–44.

[11] Diday E, Lemaire J, Pouget J, Testu F. Éléments d'analyse de données. Dunod; 1982.

[12] Voyant C, Paoli C, Muselli M, Nivet M-L. Multi-horizon solar radiation forecasting for Mediterranean locations using time series models. Renew Sustain Energy Rev 2013;28:44–52. doi:10.1016/j.rser.2013.07.058.

[13] 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.

[14] Paoli C, Voyant C, Muselli M, Nivet M-L. Solar Radiation Forecasting Using Ad-Hoc Time Series Preprocessing and Neural Networks. Emerg. Intell. Comput. Technol. Appl., vol. 5754, Springer Berlin / Heidelberg; 2009, p. 898–907.

[15] Voyant C, Muselli M, Paoli C, Nivet M-L. Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. Energy 2012;39:341–55. doi:10.1016/j.energy.2012.01.006.

[16] Dambreville R, Blanc P, Chanussot J, Boldo D. Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model. Renew Energy 2014;72:291–300. doi:10.1016/j.renene.2014.07.012.

[17] Kühnert J, Lorenz E, Heinemann D. Chapter 11 - Satellite-Based Irradiance and Power Forecasting for the German Energy Market. In: Kleissl J, editor. Sol. Energy Forecast. Resour. Assess., Boston: Academic Press; 2013, p. 267–97.

[18] 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.

[19] Marquez R, Coimbra CFM. Proposed Metric for Evaluation of Solar Forecasting Models. J Sol Energy Eng 2012;135:011016–011016. doi:10.1115/1.4007496.

[20] Gueymard CA. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renew Sustain Energy Rev 2014;39:1024–34. doi:10.1016/j.rser.2014.07.117.

[21] Diazrobles L, Ortega J, Fu J, Reed G, Chow J, Watson J, et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmos Environ 2008;42:8331–40. doi:10.1016/j.atmosenv.2008.07.020.

[22] 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.

[23] Iqdour R, Zeroual A. The MLP Neural Networks for Predicting Wind Speed, Marrakech, Morocco: 2006.

[24] Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Syst 1989;2:303–14. doi:10.1007/BF02551274.

[25] 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.

[26] Wiens TS, Dale BC, Boyce MS, Kershaw GP. Three way k-fold cross-validation of resource selection functions. Ecol Model 2008;212:244–55. doi:10.1016/j.ecolmodel.2007.10.005.

[27] 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.

[28] Kumar U, Jain VK. Time series models (Grey-Markov, Grey Model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India. Energy 2010;35:1709–16. doi:doi: DOI: 10.1016/j.energy.2009.12.021.

[29] Hamilton J. Time series analysis. Princeton N.J.: Princeton University Press; 1994.

[30] Jiang A-H, Huang X-C, Zhang Z-H, Li J, Zhang Z-Y, Hua H-X. Mutual information algorithms. Mech Syst Signal Process 2010;24:2947–60. doi:10.1016/j.ymssp.2010.05.015.

[31] Muzy JF, Bacry E, Baile R, Poggi P. Uncovering latent singularities from multifractal scaling laws in mixed asymptotic regime. Application to turbulence. EPL Europhys Lett 2008;82:60007. doi:10.1209/0295-5075/82/60007.

[32] Lauret P, Diagne M, David M. A Neural Network Post-processing Approach to Improving NWP Solar Radiation Forecasts. Energy Procedia 2014;57:1044–52. doi:10.1016/j.egypro.2014.10.089.

[33] Zhang G. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003;50:159–75. doi:10.1016/S0925-2312(01)00702-0.

List of figures

List of tables