Energy 90 (2015) 671–679

Contents lists available at ScienceDirect

Energy

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

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

Cyril Voyant [a,*], Ted Soubdhan [b], Philippe Lauret [c], Mathieu David [c], Marc Muselli [a]

[a] University of Corsica, CNRS UMR SPE 6134, 20250, Corte, France
[b] Laboratory LARGE, University of Antilles and Guiana, Guadeloupe, France
[c] Laboratory PIMENT, University of Reunion, France

ARTICLE INFO

Article history:
Received 10 April 2015
Accepted 22 July 2015
Available online 12 August 2015

Keywords:
Solar forecasting
Time series
Clear sky models
Fractal dimension
Mutual information
Log-return

ABSTRACT

In this paper we propose to determinate and to test a set of 20 statistical parameters in order to estimate the short term predictability of the global horizontal irradiation time series and thereby to propose a new prospective tool indicating the expected error regardless the forecasting methods used. The mean absolute log return, which is a tool usually used in econometrics but never in global radiation prediction, 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. This study gives a judgment for engineers and researchers on the installation or management of solar plants and could help in minimizing the energy crisis allowing to improve the renewable energy part of the energy mix.

1. Introduction

Solar radiation is one of the main energy sources for physical, biological and chemical processes, occupying the most important role in many engineering applications [1–3]. The process of converting sunlight to electricity without combustion allows creating power without direct pollution. Thereby it is necessary to propose some prediction models [4] to use ideally this technology and in order to integrate solar energy *PV* production systems in the energetic mix [5,6]. Thus, solar energy forecasting is used to predict the amount of *PV* energy available in near terms [7]. Several methods have been developed by experts around the world and the mathematical formalism of TS (times series) has often been used [8–10]. *TS* is a set of numbers that measures the status of some activities over time. It is the historical record of measurements taken at equally spaced intervals with a consistency in the activity and the measuring method [11]. Some of the best predictors found in literature are ARMA (autoregressive and moving average) [8,12,13], Bayesian inferences [14–16], Markov chains [17], k-Nearest-Neighbors predictors [18] or artificial intelligence techniques as the ANN (artificial neural networks) [6,9,19,20]. Although these methodologies are potentially good in many areas, we observed in our previous studies on global radiation prediction [14,16,21,22] that the simple model based on the persistence of the clear sky index gives often very good results with acceptable errors [23] for short term forecasting time horizon (less than 1 h). It is quite a standard in the solar forecasting community [24]. Indeed, instead of using directly the GHI (global horizontal irradiation), up to date forecast models predict the clear sky index at different forecast time horizons [25]. The corresponding forecast is obtained through the use of a clear sky model. Also, the clear sky index is very important when one wants to characterize the variability of a site [26,27], therefore a key component in solar forecasting is the clear sky index or equally the clear sky model. Several statistical parameters [26] aim at assessing the variability and consequently the difficulty to forecast the *GHI* [7,27,28]. The goal of this paper is to find a metric characterizing the time series irradiance that is correlated to the forecasting accuracy (*nRMSE*; *nMAE*):

- Based on sound numerical experiments, we study the aforementioned metrics [26,27] 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.).

* Corresponding author. 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 (C. Voyant).

http://dx.doi.org/10.1016/j.energy.2015.07.089
0360-5442/

This 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 results are 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 [29], is also applicable for other estimators [30,31]. This condition usually implies a stable process [32]. This notion directly depends on the fact that certain intrinsic features of the TS (such as mean or variance) change over time or remain constant. To make the time series stationary, we used the CSI (clear sky index) methodology from a CS (clear sky) estimation done with a numerical model (Solis) [33]. The ratio between the *GHI* (x) and the CS (clear sky model) defines the clear sky index *CSI* $\left(CSI(t) = \frac{x(t)}{CS(t)}\right)$.

2.1. Methodology

In order to assess the correlations between the proposed statistical parameters and the forecasting accuracy, we use two years of *GHI* for each site 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 to estimate 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) [34]. 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 [34].

2.2. Data

To validate this study, we choose 8 cities throughout 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. Missing values are observed, this represents less than 2% of the data. A classical cleaning approach is operated in order to identify and remove these data. 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, 75 m 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°26 N, 61°24 W, 11 m asl). The daily average of the solar energy on a horizontal surface is around 5 kWh/m^2. 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–26 °C. Relative humidity ranges from 70% to 80% and the trade winds are relatively constant throughout 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 stations of Ajaccio (41°55′N, 8°44′E, 4 m asl) and Bastia (42°42′N, 9°27′E, 10 m asl). They are located near the Mediterranean Sea and nearby mountains (more than 1000 m altitude at 40 km from the two 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 have been measured on an hourly basis from 1998 to 1999 (exactly two years).

The four continental stations are:

- Northern continental cities; the 3 studied cities are Marseille (43°17′N, 5°22′E, 10 m asl), Nice (42°42′N, 9°27′E, 10 m asl) and Montpellier (43°36′N, 3°52′E, 27 m 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 1000 m asl) the third is located in a flat area. The measures have been recorded during the years 2007 and 2008;
- Southern city; Melbourne is located in the south-eastern part of mainland Australia (37°48′S, 144°57′E, 60 m asl). It has a moderate oceanic climate and is well known for its changeable weather conditions. This is mainly due to Melbourne's location, 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 have been recorded during the years 2008 and 2009.

2.3. Clear sky modeling

Among the clear sky models that can be found in literature, for this study, we have chosen the simplified Solis clear sky model based on radiative transfer calculations and Lambert–Beer relations [35]. 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).

$$CS(t) = H_0.\exp\left(-\tau \Big/ \sin^b(h(t))\right).\sin(h(t)) \quad (1)$$

Where τ is the global total atmospheric optical depth (depends on the aerosol optical depth, pressure, water vapor column and type of place), h is the solar elevation angle, b is a fitting parameter (depends on aerosol optical depth, water vapor column and type