Forecasting of preprocessed daily solar radiation time series using neural networks

Christophe Paoli a, Cyril Voyant a,b, Marc Muselli a,* Marie-Laure Nivet a

a University of Corsica, CNRS UMR SPE 6134, 20250 Corte, France
b Hospital of Castelluccio, Radiotherapy Unit, BP 85, 20177 Ajaccio, France

Received 21 October 2009; received in revised form 17 July 2010; accepted 25 August 2010
Available online 28 September 2010

Abstract

In this paper, we present an application of Artificial Neural Networks (ANNs) in the renewable energy domain. We particularly look at the Multi-Layer Perceptron (MLP) network which has been the most used of ANNs architectures both in the renewable energy domain and in the time series forecasting. We have used a MLP and an ad hoc time series pre-processing to develop a methodology for the daily prediction of global solar radiation on a horizontal surface. First results are promising with nRMSE ~ 21% and RMSE ~ 3.59 MJ/m². The optimized MLP presents predictions similar to or even better than conventional and reference methods such as ARIMA techniques, Bayesian inference, Markov chains and k-Nearest-Neighbors. Moreover we found that the data pre-processing approach proposed can reduce significantly forecasting errors of about 6% compared to conventional prediction methods such as Markov chains or Bayesian inference. The simulator proposed has been obtained using 19 years of available data from the meteorological station of Ajaccio (Corsica Island, France, 41° 55′ N, 8° 44′ E, 4 m above mean sea level). The predicted whole methodology has been validated on a 1.175 kWc mono-Si PV power grid. Six prediction methods (ANN, clear sky model, combination... ) allow to predict the best daily DC PV power production at horizon d + 1. The cumulated DC PV energy on a 6-months period shows a great agreement between simulated and measured data (R² > 0.99 and nRMSE < 2%).

© 2010 Elsevier Ltd. All rights reserved.

Keywords: Time series forecasting; Pre-processing; Artificial Neural Networks; PV Plant Energy Prediction

1. Introduction and review

An optimal use of the renewable energy needs its characterization and prediction in order to size detectors or to estimate the potential of power plants. In terms of prediction, electricity suppliers are interested in various horizons to estimate the fossil fuel saving, to manage and dispatch the power plants installed and to increase the integration limit of renewable energy systems on non-interconnected electrical grids. In this study, we focus on the prediction of global solar irradiation on a horizontal plane for daily horizon, which might interest electricity suppliers. In this way, we have investigated both in time series forecasting, which is a challenge in many fields, and in artificial intelligence techniques, which are becoming more and more popular in the renewable energy domain (Mellit and Kalogirou, 2008; Mellit et al., 2009) and particularly for the prediction of meteorological data such as solar radiation (Mubiru, 2008; Mubiru and Banda, 2008; Kalogirou, 2001; Hocaoglu et al., 2008). Thereby many research works have shown the ability of Artificial Neural Networks (ANNs) to predict time series of meteorological data. Moreover, if we compare to conventional algorithms based on linear models, ANNs offer an attractive alternative by providing nonlinear parametric models. Through the proposed study, we will
particularly look at the Multi-Layer Perceptron (MLP) network which has been the most used of ANN architectures in the renewable energy domain (Mellit and Kalogirou, 2008; Mellit et al., 2009) and time series forecasting (Crone, 2005; Faraway and Chatfield, 1995; Hu and Hwang, 2002; Jain et al., 1996). We compare this tool with other classical predictors like ARIMA, Bayesian inference, Markov chains and k-Nearest-Neighbors predictors. Previous works (Paoli et al., 2009) have shown that a first pre-processing treatment based on an simple extraterrestrial normalization allows to obtain a better correlation between simulated and measured radiation signals. During this study, we use an ad hoc time series pre-processing step before using neural networks. Indeed, a data pre-processing including seasonal adjustment can improve ANN forecasting performances (Zhang and Qi, 2005). Two strategies of pre-treatment are compared: the first based on the clearness index and the second using the clear sky index. Contrary to these previous studies (Paoli et al., 2009), we validate the simulator on a 1.175 kWp mono-Si PV power grid connected on a wall of our laboratory. This validation step presents both additional difficulties: the PV modules are 80° tilted to horizontal and are localized at 10 km from the original station.

The paper is organized as follow: Section 2 describes the context in which this research was done and the data we have used and introduces time series forecasting, presents several conventional prediction and modeling methods of daily irradiation data including ANNs. Using both clearness and clear sky indexes pre-processing treatments for ANNs, all these predictors are then tested and compared with the same 19 years data set in Section 3 (data used for training and validation of the ANN comes from the meteorological station of Ajaccio airport). All referenced and ANN original methods are compared on a seasonal point of view using classical statistical coefficients (MAE, MBE, RMSE, nRMSE, \( R^2 \)). The global approach is validated and discussed on real tilted PV modules (1.175 kWp) for the calculation of the DC energy. At last, the definition of a new parameter \( \Delta \) that reflects the tendency of the prediction (the first derivative linked to the signal predicted slope) is presented and discussed. The Section 4 concludes and suggests perspectives.

2. Methodology

2.1. Context

In this work, measured global daily radiation data taken from a meteorological ground station are used to forecast global solar irradiation for the next day. The global radiation consists of three types of radiation: direct, diffuse and ground-reflected (Liu and Jordan, 1962). The ground-reflected radiation does not concern us because we try to predict the radiation on a horizontal surface. For clear sky, global radiation is relatively easy to model because it is primarily due to the distance from the sun to the sensor (Hay and Davies, 1980; Ineichen et al., 1990; Liu and Jordan, 1962; Perez et al., 1990; Reindl et al., 1990). With cloudy sky, these are mostly stochastic phenomenon, which depend on the local weather.

The study proposes to analyze the radiation time series (MJ m\(^{-2}\)) measured at the meteorological station of Ajaccio (Corsica Island, France, 41°55’N, 8°44’E, 4 m above mean sea level) maintained by the French meteorological organization (Méteo-France). Ajaccio has a ‘Mediterranean’ climate, hot summers with abundant sunshine and mild, dry, clear winters. It is located near the sea but there is also relief nearby (~30 km). This geographical configuration can make nebulosity difficult to forecast.

The data used represent the global horizontal solar radiation and are available on an hourly basis for a period from January 1971 to December 1989. A compact XARIA-C station (DEGREANE HORIZON), certified by Méteo-France, is used to record horizontal radiation (Kipp and Zonen CMP6) with a 3 s time sampling. The sensor works in a range 0–90,000 J/m\(^2\) and requires a
quality control every year. Only average measures over 1 hour are stored in the national database. These last values are integrated to extract a daily time series of the global horizontal radiation.

2.2. Time series forecasting

A time series (Faraway and Chatfield, 1995) is a collection of time ordered observations \( x_t \), each one being recorded at a specific time \( t \) (period). Time series can appear in a wide set of domains such as finance, production or control, just to name a few. In first approximation, a time series model (\( \hat{x}_t \)) assumes that past patterns will occur in the future. In fact a time series model could be used only to provide synthetic time series statistically similar to the original one. The modeling of the series begins with the selection of a suitable mathematical model (or class of models) for the data. Then, it is possible to predict future values of measurements (Brockwell and Davis, 2006).

2.3. Most popular forecasting methods

We have identified in the time series forecasting literature the following methods: ARIMA (classical method), Bayesian inferences, Markov chains and k-Nearest-Neighbors predictors. We present in the next sub-sections each method with the model obtained after a specific optimization.

2.3.1. ARIMA techniques

The ARIMA techniques (Hamilton, 1994) are reference estimators in the prediction of global radiation field. It is a stochastic process coupling autoregressive component (AR) to a moving average component (MA). This kind of model is commonly called SARIMA \((p, d, q)(P, D, Q)\) and is defined as follows \((p, d, q, P, D, Q\) are integers and \(B\) the lag operator):

\[
k_p(B)K_p(B^S)(1 - B)^d(1 - B^S)^DX_t = m_q(B)M_Q(B^S)u_t
\]

where \(X_t\) is a time series, \(k_p\) are the parameters of the autoregressive part, \(m_q\) are the parameters of the moving average part and \(u_t\) is an error term distributed as a Gaussian white noise, \(K_p\) and \(M_Q\) are parameters of seasonal autoregressive and seasonal moving average part. The \(d\) and \(D\) are nonseasonal and seasonal difference and \(S\) is seasonal period. The optimization of these parameters must be made depending on the type of the series studied. To help us in this study, we chose to use Grocer, a very complete toolbox compatible with the Scilab® software (Dubois and Michaux, 2008). The criterion adopted to consider when an ARMA model ‘fits’ to the time series is the normalized root mean square error obtained by: \(\text{nRMSE} = \sqrt{\frac{(x - y)^2}{< (x)^2> / <(y)^2>}}\) where \(x\) represents the measurement and \(y\) the prediction, using averaged values. This parameter is generated by the prediction of 2 years of radiation not used during the ARMA parameters calculation step. After several experiments the model best suited to the study of global radiation is an ARMA \((2,2)\), its representation is:

\[
X_t = u_t + \sum_{i=1}^{2} k_i \cdot X_{t-i} + \sum_{j=1}^{2} m_j \cdot u_{t-j}
\]

where \(k_1 = -1.47, k_2 = 0.47, m_1 = -1.19\) and \(m_2 = 0.22\). A Student’s \(T\)-test (introduced by William Sealy Gosset in 1908) has been used to verify that these coefficients were significantly different from zero and residual auto-correlogram tests have been computed to verify white noise error terms.

Moreover, we also wanted to study a simplest model among the different ARIMA techniques. We have chosen to study the AR type and we have found that the most interesting in this case was the AR(8).

2.3.2. Bayesian inference

The second classical technique we have chosen is called Bayesian inference (Diday et al., 1982; Celeux and Nakache, 1994; Pole et al., 1994). In this method, evidences or observations are used to update or to newly infer the probability that a hypothesis may be true. To estimate the probability that the series is in the state \(y_k\) at time \(t\), it is useful to use the Bayes theorem. This ideology is expressed mathematically by the following formula (where \(N\) represents the initial and known measured, \(k\) is the class and \(y_k\) the value of this class, \(P\) is the conditional probability):

\[
X_t = y_k = \arg \max_k (P(X_t = y_k | N)) = \arg \max_k (P(X_t = y_k) \cdot P(N | X_t = y_k))
\]

A practical way to solve this equation is to make the assumption of conditional independence as follows:

\[
P(N | X_t = y_k) = \prod_{j=1}^{J} P(X_{t-j} | X_t = y_k)
\]

which leads to:

\[
X_t = y_k = \arg \max_k \left( P(X_t = y_k) \cdot \prod_{j=1}^{J} P(X_{t-j} | X_t = y_k) \right)
\]

To use this kind of predictor, we must establish the conditional probability table of the series and so quantify the last term of the equation. For the optimization we have used Matlab® software and we have been able to identify that the best prediction was obtained with 50 classes \((0 < k < 51)\) and an order of \(J = 3\).

2.3.3. Markov chains

In forecasting domain, some authors have tried to use so-called Markov processes (Logofet and Lesnaya, 2000; Muselli et al., 2001), specifically the Markov chains. A Markov process is a stochastic process with the Markov property. Markov property means that, given the present state, future states are independent of the past states. In
other words, the description of the present state fully captures all the information that could influence the future evolution of the process. Future states will be reached through a probabilistic process instead of a deterministic one. The proper use of these processes needs to calculate firstly the matrix of transition states. The transition probability of state $i$ to the state $j$ is defined by $p_{ij}$. The family of these numbers is called the transition matrix of the Markov chain $R$, we have:

$$p_{ij} = P(X_t = j|X_{t-1} = i)$$

(6) the formulation of the prediction can be resumed by:

$$X_{t+n} = X_t \cdot R^n$$

(7)

The choice of the dimension of the transition matrix (number of class), and order of the chain (determination of the lag prediction) was done on Matlab®. We obtained an optimal prediction for 50 classes and an order of 3:

$$X_t = y_k = \arg\max_k ((X_{t-r} \cdot R + X_{t-2} \cdot R^2 + X_{t-3} \cdot R^3) \cdot e_k)$$

(8) where $e_k$ represents the set of basic vectors of the transition matrix.

2.3.4. K-Nearest Neighbors

The k-Nearest Neighbors’ algorithm (k-NN) (Sharif and Burn, 2006; Yakowitz, 1987) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It can also be used for regression. Unlike previous models, this tool does not use a learning base. The method consists in looking into the history of the series for the case the most resembling to the present case. By considering a series of observations $X_t$, to determine the next term $X_{t+1}$, we must find among anterior information, which minimize the quantity defined on Eq. (9) ($d$ is the quadratic error).

$$r_o = \arg\min_r (d(X_t, X_{r-r}) + d(X_{t-1}, X_{r-r-1}) + \cdots + d(X_{t-k}, X_{r-r-k}))$$

(9)

In this study we have chosen a $k$ equal to 10. After this argument of the minimum search, the prediction can be written:

$$X_{t+r} = X_{t+r+1}$$

(10)

2.4. Artificial Neural Networks for prediction

Artificial Neural Networks (ANN) are intelligent systems that have the capacity to learn, memorize and create relationships among data. We present in this section an overview of this methodology in the context of time series prediction. An ANN is made up by simple processing units, the neurons, which are connected in a network by a large number of weighted links where the acquired knowledge is stored and over which signals or information can pass. An input $x_j$ is transmitted through a connection, which multiplies its strength by a weight $w_{ij}$ to give a product $x_j w_{ij}$. This product is an argument to a transfer function $f$, which yields an output represented as: $y_i = f(x_j w_{ij})$ where $i$ is an index of neurons in the hidden layer and $j$ is an index of an input to the neural network (Mubiru, 2008; Cortez et al., 2002). In a Multi-Layer Perceptron (MLP), neurons are grouped in layers and only forward connections exist. This provides a powerful architecture, capable of learning any kind of continuous nonlinear mapping, with successful applications ranging from Computer Vision, Data Analysis or Expert Systems, etc. (Cortez et al., 2001; Faraway and Chatfield, 1995; Jain et al., 1996; Hu and Hwang, 2002; Crone, 2005). One of the most interesting characteristic of ANNs is their ability to learn and to model a phenomenon. Fig. 1 gives the basic architecture for a MLP application to time series forecasting during the learning task. To establish the prediction, a fixed number $p$ of past values are set as inputs of the MLP, the output is the prediction of the future value of the time series. This method called “sliding-window technique” uses a moving time window to select $N$ times, $p$ inputs data for training.

In order to determine the best network configuration, we have tried to study all the parameters available in this network architecture. The principal parameters witch influence the number of local minima and the complexity of the network and its learning are: the number of input (these are the neurons which represent the input layer of the neural net); the number of hidden layer and their number of neurons; the activation (or transfer) function; the learning algorithm and the comparison function used during the learning phase. Additionally we have to think about the normalization of data, the learning sampling size and data distribution between the learning, test and validation phases. For convenience we have chosen to optimize the parameters separately with an intuitive order of preference for all the ANN used in this work. Thus the chosen parameter optimization follows the pattern below:

1. Number of input: 1–15 inputs.
2. Number of hidden layers and neurons: 1–3 hidden layers, and 1–50 neurons by layer.
3. Transfer function of hidden layer: linear (with or without saturating), exponential and sigmoid.
4. Normalization of data: size of the interval for standardization between 0.6 and 1.
5. Learning sample size: 500–6000 samples.
7. Comparison function: mean squared normalized error (MSNE), mean square error (MSE), sum of square error (SSE), mean absolute error (MAE).
Customization: new study on the number of hidden neurons, and the hidden layer to complete the process.

As a result of this iterative process, the selected network has three layers: input, hidden and output layers. There was no significant difference in the use of 1, 2 and 3 hidden layers architectures. Considering this fact one hidden layer was used in order to minimize the complexity of the proposed ANN model. We tried several input layer configurations, best results were obtained with eight inputs which received the endogenous entries $S_{t-1}, \ldots, S_{t-8}$ normalized on $[0, 1]$. We found that three neurons on the hidden layer were sufficient. Finally, as we wanted to predict the solar radiation at horizon 1, we have one neuron on the output layer $b_{St}$. Concerning the activation functions the best results were obtained with the Gaussian (hidden layer) and linear (output layer) functions. Concerning the training algorithm, many experiences enabled us to choose the Levenberg–Marquardt optimization (second-order algorithm) with 5000 epochs and $\mu$ decrease factor (learning rate) of 0.5, all other parameters have default values. The ANN configuration leads to an output signal corresponding to a nonlinear auto-regression to the data entry:

$$\hat{S}_t = \sum_{j=1}^{3} W_{2,j,1} \cdot e^{- \left( \sum_{k} W_{1,k,j} S_{t-k} + b_{1,j} \right)^2} + b_{2,1}$$

where parameters $W_{i,j,k}$ represent interconnection weights between layers and $b_{i,j}$ bias correction coefficients. The reader can find in the Appendix A the full list of these values for the most efficient ANN predictor.

We used the Matlab® software and its neural network toolbox to implement the network. The Matlab® learning, testing and validation were set respectively to 80%, 10% and 10%. These three steps constitute the training of the network, after which the weights and bias values are optimized. The training has concerned the years 1971–1987 and the performance function was the mean square error (MSE). In the learning step, the data are presented to the ANN following a sequential order. The predicted values of global radiation are compared to the years 1988–1989, which were not used during training.

3. Results and discussion

3.1. Results with non-stationary time series

In this section we examine if the proposed network was really interesting in terms of daily irradiation prediction. Indeed, we have compared its performance with the forecasting results obtained with a naive predictor (365 values representing the average over 17 years of the considered day), order three Markov chains, order three Bayesian inferences, an order 10 $k$-NN, an order 8 AutoRegressive...
AR(8). Table 1 and Fig. 2 present results obtained in the case of prediction for a period of 2 years (1988–1989). The main comparison function used is the nRMSE for daily global radiation prediction \((d+1)\). Moreover, classical statistical parameters enable us to evaluate the prediction quality: the absolute error is measured by the root mean square error (RMSE), the Mean Bias Error (MBE) which gives an idea about the error in terms of under or over estimation, determination coefficient \((R^2)\) and mean absolute error (MAE).

As can be seen the predictors other than AR(8) and ANN give the same results, slightly better than those obtained with a naive predictor. Although AR and ANN could be considered as the best predictors, nRMSE are important considering from the point of view of energetic applications such as grid management with decentralized grid-connected systems. In order to improve these results, the possibility to determine a signal pre-treatment as input of AR and ANN is discussed. As proposed in (Diday et al., 1982), it appeared interesting to make stationary the time
series as much as possible. So we try to find a treatment in order to eliminate seasonal components without changing the other information. This treatment is called a seasonal adjustment process. In the present study, we have considered the solar irradiation like a seasonal phenomenon. The choice of the methodology used depends on the nature of the seasonality. In our case the seasonality is very pronounced and repetitive, so very deterministic and not stochastic. In the next sections, we present how we have used the physical modeling of radiation to determine a stationarization methodology for the daily signal.

3.2. Result with stationary time series

3.2.1. Pre-processing proposed

In this section we present all the steps of an ad-hoc time series pre-processing. Fig. 3 summarizes the different steps we want to experiment. The ‘classical methods’ like k-NN, Bayesian inference and also naïve predictor don’t require the use of pre-processing. In the literature, very few references relate ways the data stationary. Concerning the Markov chains, we assumed that the simplest model without pre-processing was sufficient in comparison with others ‘classical methods’. For the pre-processing, a first treatment (step 1) allows us to clean the series of non-typical points related to sensor maintenances or absence of measurement. 4.1% of measurements were missing and replacing by the hourly average over the 19 years for the given day.

Steps 2 and 3 will be described in the next section and lead to a series corrected utilizing clear sky or extraterrestrial normalizations. In case of stationary transformation by clear sky model the term $S_{corr}^{d,y}$ will be replaced by $S_{corr}^{d,y}$. The step 4 consists to use one of the forecasting methods which have been outlined in the previous section with the optimized MLP. Finally, step 5 allows to reverse the pre-processing treatment and to obtain the prediction of global irradiation. The VC parameter represents the signal dispersion. In the remainder of this paper, we choose the following naming convention: $X_t$ designates the time series, and $X_{d,y}$ the modeling of the variables, where $d$ is the day of the year $y$. To remove the seasonal dependencies, we chose to study both the “clearness index” and “clear sky index”. According to Bird and Hulstrom (1981), the global radiation on a horizontal plane under clear sky (defined as Earth’s standard atmosphere without cloud) is connected to the extraterrestrial radiation and solar elevation angle. In daily case, the periodic phenomenon of irradiation is only due to the extraterrestrial solar radiation coefficient for day $d$ ($H_{0}^d$) (Badescu, 2008). Thus we apply on the original series $X_{d,y}$ the ratio to trend method (Bourbonnais and Terraza, 2008). This leads to a new series $S_{d,y}$ known

![Fig. 3. Summarize of the protocol followed to obtain the predicted irradiation. VC is the variation coefficient (ratio to standard deviation to the means).](image-url)
as *clearness index* series (often noted as *K* in scientific press):

\[ S_{d,y} = \frac{X_{d,y}}{H_0^d} \quad (12) \]

In parallel, we wanted to test another method to remove seasonal effects, which is done by using the clear sky model for *clear sky index*. There are a lot of methods to determine this model. In our case, we preferred the simplified “Solis clear sky” model (Ineichen, 2008) based on radiative transfer calculations and the Lambert–Beer relation (Mueller et al., 2004). In this case the clear sky global horizontal irradiance (*H*<sub>gh, clear sky</sub>) reaching the ground is defined by:

\[ H_{gh, clear sky} = H_0 \cdot e^{-(\tau \sin^2(h))} \cdot \sin(h) \quad (13) \]

where \( \tau \) is the global total atmospheric optical depth (–0.37 in our case), \( h \) is the solar elevation angle and \( b \) is a fitting parameter (0.35 for us). The daily integration of the *H*<sub>gh, clear sky</sub> parameter allows us to determine the daily solar ratio for modeling *H*<sub>d</sub><sup>gh, clear sky</sup>. A series of test (not presented in this paper) has allowed to validate the Solis model on horizontal and 80° tilted daily global radiation We obtain a relation “equivalent” to Eq. (12):

\[ S_{d,y}^{corr} = \frac{X_{d,y}}{H_{gh, clear sky}} \quad (14) \]

These treatments aim to create a new distribution without periodicity. Although the previous pre-treatment tends to make the time series stationary, a test of Fisher shows that seasonality was not optimal. According to Bourbonnais and Terraza (2008), after using a ratio to trend method (Eqs (12) and (14)) to correct rigid seasonality, a ratio to moving average can be used. This second ratio can be applied when there is no analytical expression of the trend. In this case, we find that *H*<sub>d</sub> led a new seasonality which is difficult to model. That is the reason why we considered a moving average ratio to overcome this flexible seasonality:

\[ Y_{d,y} = \frac{S_{d,y}}{\sum_{i=-m}^{m} S_{d,y+i}} \quad (15) \]

In the present case, 2\( m + 1 = 365 \) days (period of original time series), we obtain \( m = 182 \). The term \( S_{d,y} \) is equivalent to \( S_{d,y}^{corr} \) they refer only to the chosen stationarization methodology. To complete the process, we use the 365 seasonal factors \( (Y_{d,y}) \). These are coefficients which allow overcoming seasonality by a moving average ratio described above (Bourbonnais and Terraza, 2008). In order not to distort the series, we have considered that the total sum of the components of the series is the same before and after the report (final seasonal factors \( Y_{d,y} \)) of the Eq. (18). The transition coefficients \( (N = 18, \) the number of years of history) and the average coefficients of the regular 365 days are given by Eq. (16). A new series seasonally adjusted that represents only the stochastic component of global radiation is given by Eq. (19).

\[ Y_{d,y} = \frac{1}{N} \sum_{y=1}^{N} Y_{d,y} \quad (16) \]

\[ \bar{Y}_{d,y} = \frac{1}{365} \sum_{d=1}^{365} Y_{d,y} \quad (17) \]

\[ Y_{d,y} = \frac{Y_{d,y}}{\bar{Y}_{d,y}} \quad (18) \]

\[ S_{d,y}^{corr} = S_{d,y} \quad (19) \]

Thereafter, the juxtaposition of different values of \( S_{d,y}^{corr} \) or \( S_{d,y}^{corr} \) will be resumed on stationary time series noted \( S_{d} \) (extraterrestrial stationarization) and \( S_{d}^{corr} \) (clear sky stationarization). After this step we needed to verify the effectiveness of pre-treatment on the time series. To achieve this we used a Fast Fourier Transform (FFT). The spectrum of the original series shows an important peak for the value 365 days that disappears after the pre-treatment.

### 3.2.2. Comparison of ARIMA and ANN

In Table 2, the statistical errors nRMSE obtained in the case of pre-treatment with the two best predictors (AR/ARMA and ANN) are shown. The results are slightly better with pre-processing. The reduction of the error prediction is low, but the difference between the two methods is significant, given the average and the confidence interval. A nRMSE reduction from 21.0% to 20.2% represents a gain of ~1%.

<table>
<thead>
<tr>
<th></th>
<th>Clearness index methodology (%)</th>
<th>Clear sky index methodology (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(8) without pre-processing</td>
<td>21.18 ± 0</td>
<td>21.18 ± 0</td>
</tr>
<tr>
<td>ARMA(2,2) with pre-processing</td>
<td>20.31 ± 0</td>
<td>20.32 ± 0</td>
</tr>
<tr>
<td>ANN[8,3,1] without pre-processing</td>
<td>20.97 ± 0.15</td>
<td>20.97 ± 0.15</td>
</tr>
<tr>
<td>ANN[8,3,1] with pre-processing</td>
<td>20.17 ± 0.1</td>
<td>20.25 ± 0.1</td>
</tr>
</tbody>
</table>

Table 3 details in both ANN pre-processing normalizations (clear sky and clearness indexes) the annual prediction errors obtained for the years 1988 and 1989. In this study (global horizontal radiation) no significant differences between the both stationarization methodologies (“clearness index” and “clear sky index”) are noted.

The confidence interval is calculated after eight training-simulations providing information on the prediction robustness. The weights are initiated before each simulation. With small confidence intervals, it can be said that there are very few local minimums. The ANN learning step error is obtained for RMSE < 10^-4 (fixed before the treatment), leading to the stop of the learning. The monthly average error is the error (nRMSE) generated by monthly average values of the global radiation prediction. It is in fact the error of the monthly sum of global radiation. As can be seen the prediction is different from an average of 4% of the measured data. The negative MBE (~0.37 MJ/
m² and −0.32 MJ/m²) means that the solar potential is underestimated over the year. With non-ordinary low irradiation days, a tendency to overestimate is noted. The conclusion is based on the simulations of global radiation during these special days. We defined non-ordinary low irradiation days as days with thick cloud cover causing a very low sunshine duration. The determination coefficient $R^2$ is greater than 0.8 that is good in relation to the noise present in measurements. There should be a compromise between RMSE and nRMSE. The nRMSE, computed from the mean global radiation obtained on the season, are useful for comparison and optimization (20.17% for clearness index and 20.25% for clear sky index). But for the absolute interpretation of received energy we must look at the RMSE. In order to better understand the predictor performances, Fig. 4 shows the errors of prediction and distinguishes the seasons for the years 1988 and 1989 for a clear sky pre-processing. Best results in term of forecast are obtained in summer. These results can be used for example by energy managers who need to avoid using hydraulic power plants in dry season.

The spring season is the most difficult to predict with accuracy given the climate instabilities of this period. The absolute error is consistent. However, we find that in summer the error does not exceed 3.24 MJ/m² (900 Wh/m²), while the irradiation is important. MBE are found negative, which indicates an underestimation. The MBE is not significantly different from one season to another. Thus we always have the same prediction error, whatever the season. We systematically overestimate the days when the irradiation is minimal (winter). Moreover it is very difficult to predict the days when the irradiation had to be theoretically important. We would undoubtedly have improved the results optimizing an ANN by season, but it would complicate the process, and tend to decrease the procedure robustness.

After presenting the pre-treatment to be performed on the series and after verifying its effectiveness, we propose in the next section to validate this process on a facade PV system.

### 3.3. Frontage PV

In order to validate the approach proposed, we decided to use the previous simulator on a real frontage PV system installed recently in our laboratory (Vignola near Ajaccio-Campo del’Oro). Two additional difficulties have to be overcome: the PV modules are tilted 80° to horizontal and are located at 10 km from the meteorological station of Ajaccio-Campo del’Oro.

#### 3.3.1. System description

The system has a nominal power of 6.525 kWp composed by 1.8 kWp and 4.725 kWp amorphous and monocrystal PV modules respectively built in 6 independent power subsystems (Fig. 5). PV power predictions from ANN methodology described in this paper have been computed from the central part of the PV plant on the front side exposed to the south (azimuth zero) and tilted at 80° (Index “PV” on Fig. 5). The PV system consists of 9 SUNTECH 175S-24Ac with 1.175 kWp nominal power connected to a 1.85 kW SUNNY BOY SMA inverter for PV production on the grid. The irradiance sensor used (Fig. 5) is an INGENIEURBÜRO SI-12TC calibrated by the PTB Braunschweig (German national metrology Institute): scale range between 0 and 1200 W/m² requiring an annual quality control for calibration.

For the daily PV power calculation, we use in first approximation, a classical linear production based on a...
constant PV plant efficiency $\eta_{PV}$ \( \sim 15.3\% \) \((R^2 = 0.9967)\), with

$$E_{PV}(\text{MJ}) = \eta_{PV}I_{\beta}S$$

(20)

where $I_{\beta}$ is the daily global irradiation on the PV system \((\text{MJ} \ m^{-2}, \beta = 80^\circ)\), $S$ is the usable surface of the PV system under consideration \((S = 10.125 \ m^2)\).

We propose to predict the DC electrical energy produced by PV modules, at 1 day horizon between January 15th 2009 and June 15th 2009. Unfortunately, no historical irradiation data (on synoptic Campo del’Oro station and on our laboratory) are available for $\beta = 80^\circ$. So as it is shown in Fig. 6 we planed to use an ANN trained on the site of Campo del’Oro (10 km from the PV modules) with horizontal global radiation data. The training phase described in Section 3 is performed; only the prediction phase is different because we need to transpose the tilted data to horizontal data for use the ANN (and the reverse process as output).

3.3.2. Global radiation on a tilted plane

Before continuing, it is necessary to develop a methodology to convert the horizontal data for the tilt angle of the PV. This approach used the PVSYST software version 4 (www.pvsyst.com) to determine the global horizontal and inclined clear sky radiation. This software is very recognized in the field of renewable energies (Bouhouras et al., 2009; Qoaidar and Steinbrecht, 2009; Spanos and Duckers, 2004). Eq. (21) has been built to convert the horizontal global radiation where $\beta$ is the angle of inclination and $I$ the global radiation. On the basis of a 6-months period, this equation leads to a nRMSE = 14% from measured and simulated daily tilted irradiation between $80^\circ$ and $0^\circ$ \((\text{RMSE} = 2.4 \ \text{MJ}/\text{m}^2)\). To validate this methodology, we have compared the tilt results with a methodology more sophisticated based on the Climed-2 and Klucher models. The first model is used to determine the diffuse and beam fraction, and the second to tilt the diffuse component. In hourly case, the Climed-2–Klucher methodology is very convincing, (Noorian et al., 2008; Notton et al., 2006a, 2006b) but in daily case (integration of hourly data), the results are similar with the methodology proposed in this article. For both methodologies the nRMSE values are almost identical. Because the Climed–Klucher methodology is only applicable on some locations near the Mediterranean Sea, we have chosen to develop in this article only the PVSYST method which is applicable everywhere. In the Eq (21), the diffuse and ground-reflected radiations are implied in the bracketed term:

$$I_{\beta} = I_{\beta=0} \left(\frac{I_{\beta}}{I_{\beta=0}}\right)^{\text{clear sky}}$$

(21)

3.3.3. Predictor models and results

In order to conduct an objective study, several forecasting experiments have to be completed in parallel. Their comparisons allow to generate interpretable results of PV power prediction (Eq. (20)). We have chosen to use models that have shown the best results in the horizontal case presented in previous sections, and also a naïve predictor for quantify the prediction quality. The selected models are:

(A) Prediction omitting cloud cover: clear sky modeling.

The Solis model (Section 3.2.1) is used and coupled with the PVSYST process (Eq. (21)) to determine the tilted clear sky global radiation on Vignola site.

Fig. 5. Frontage PV system \((6.525 \ \text{kWp})\). The PV plant under study is mentioned by the “PV” mark. Tilted \((80^\circ)\) daily solar radiation $I_{\beta}$ is measured on the top of the system facing south.

Fig. 6. PV energy prediction methodology based on a daily horizontal irradiation ANN simulator.
(B) Prediction by average values

Daily solar radiation averages ($\beta = 0^\circ$) are computed from the 19 years of historical data available on the site of Campo del’Oro, then inclined at $80^\circ$ (Eq. (21)).

(C) ANN without pre-processing

ANN is used with time series not made stationary, directly on the raw data.

(D) ANN with the clearness index

This method follows the schematic of Fig. 6. The time series is made stationary with the processing based on extraterrestrial radiation.

(E) ANN with clear sky index

Equivalent to (D) but the processing used is the ratio to clear sky model (Solis).

(F) ARMA(2,2) with clear sky index

Equivalent to (E) but the ANN are replaced by the linear methodology (ARMA).

Identical irradiation data from Campo del’Oro meteorological station are used as inputs for the processes (B)–(F) except model A where a physical model Solis computes clear sky conditions. Table 4 presents errors obtained on the period of prediction (January–June) and details of the bi-monthly errors. As can be seen on the 6-months period (January 15–June 15, 2009), the ANN with clear sky index (E) methodology slightly outperforms the other. Indeed, all results computed are consistent (nRMSE = 34.1%, RMSE = 0.81 MJ, MBE = -0.03 MJ and MAE = 0.62 MJ). But the differences are small among (C)–(F) experiments i.e. between the use of ANN or ARMA predictor, with or without pre-processing. The max value of RMSE is 0.85 MJ ((C), ANN without pre-processing), the min is 0.81 MJ ((E), ANN clear sky) and the difference is only 0.04 MJ/day. Only clear sky (A) and average (B) models are the worst forecasters (respectively nRMSE = 46.6% and 47.6%). Bi-monthly errors show different behavior for prediction processes. The methods with clear sky index are the most relevant for the months for which the signal is very noisy (January–February prediction with (E), ANN, and March–April with (F), ARMA). In the case of May–June prediction, the simple method ANN without pre-processing (C) is the most relevant with nRMSE = 15.5% while the sophisticated methods with pre-processing (D)–(F) generate nRMSE $\geq 16.2\%$. The MBE factor is very weak for (C)–(F) (max = -0.12 MJ in (F) case is very acceptable). In the case of the wall-PV power system with 80° tilt, the maximum radiation value is obtained during the winter months concluding that the “E” process is the most relevant. ANN and ARMA perform almost similar, denoting the stochastic nature of the time series and thus the impossibility to predict the cloud effect on solar radiation.

These two processes present no significant differences; only these two predictors are considered in the following. Fig. 7 shows the cumulative daily predicted PV power vs. measured ones from January 15 to June 15, 2009. This type of chart is very interesting to check the gap between measurement and simulation over the long-term (seasonal vari-

<p>| Table 4 |
| Bi-monthly prediction errors and 6-months prediction errors on the electrical energy from PV plant (1.175 kWp, $\beta = 80^\circ$) for 1 day horizon from 15 January 2009 to 15 June 2009. The mean and variance are the characteristics of the global radiation time series on the period. |</p>
<table>
<thead>
<tr>
<th>A clear sky</th>
<th>B average</th>
<th>C ANN</th>
<th>D ANN clearness</th>
<th>E ANN clear sky</th>
<th>F ARMA clear sky</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>January–February</strong></td>
<td><strong>Mean:</strong> 2.39 MJ/m²</td>
<td><strong>Std. dev.: 0.062</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRMSE (%)</td>
<td>56.4</td>
<td>57.1</td>
<td>41.7</td>
<td>38.7</td>
<td>37.9</td>
</tr>
<tr>
<td>RMSE (MJ)</td>
<td>1.47</td>
<td>1.49</td>
<td>1.09</td>
<td>1.01</td>
<td>0.99</td>
</tr>
<tr>
<td>MBE (MJ)</td>
<td>1.04</td>
<td>1.02</td>
<td>-0.13</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>MAE (MJ)</td>
<td>1.06</td>
<td>1.12</td>
<td>0.98</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>March–April</strong></td>
<td><strong>Mean:</strong> 2.27 MJ/m²</td>
<td><strong>Std. dev.: 0.059</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRMSE (%)</td>
<td>49.3</td>
<td>45.1</td>
<td>38.3</td>
<td>38.3</td>
<td>37.9</td>
</tr>
<tr>
<td>RMSE (MJ)</td>
<td>1.22</td>
<td>1.11</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>MBE (MJ)</td>
<td>0.79</td>
<td>0.58</td>
<td>-0.01</td>
<td>-0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>MAE (MJ)</td>
<td>0.79</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>May–June</strong></td>
<td><strong>Mean:</strong> 1.95 MJ/m²</td>
<td><strong>Std. dev.: 0.031</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRMSE (%)</td>
<td>17.9</td>
<td>18.9</td>
<td>15.5</td>
<td>16.2</td>
<td>16.4</td>
</tr>
<tr>
<td>RMSE (MJ)</td>
<td>0.35</td>
<td>0.37</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>MBE (MJ)</td>
<td>0.20</td>
<td>0.08</td>
<td>-0.03</td>
<td>-0.07</td>
<td>-0.09</td>
</tr>
<tr>
<td>MAE (MJ)</td>
<td>0.66</td>
<td>0.79</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>January–June</strong></td>
<td><strong>Mean:</strong> 2.19 MJ/m²</td>
<td><strong>Std. dev.: 0.055</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRMSE (%)</td>
<td>46.6</td>
<td>47.6</td>
<td>35.6</td>
<td>34.7</td>
<td>34.1</td>
</tr>
<tr>
<td>RMSE (MJ)</td>
<td>1.11</td>
<td>1.13</td>
<td>0.85</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>MBE (MJ)</td>
<td>0.65</td>
<td>0.64</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>MAE (MJ)</td>
<td>0.66</td>
<td>0.79</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
</tr>
</tbody>
</table>
lations and under or overestimations of the PV energy). All figures show a good correlation between measured and predicted data ($R^2 > 0.99$). We understand easily that (C)–(F) methods are similar and although nRMSE are important, the accumulated predictions are almost identical to the measures (2.1% for ANN without pre-processing, 1.5% for ANN with clearly sky and clarity index and also for ARMA with clear sky index).

It is important to understand why the daily errors are significant. The first hypothesis is based on the high frequency noise series. In fact, the sampling frequency is the same as the frequency of the noise. It seems very unlikely that the ANN can predict “extra-ordinary days” at least if the previous day’s cloud cover is ordinary. The second hypothesis is that ANNs (and ARMA) do not take risks and propose irradiation value centered, on a mean value with a small standard deviation. The output of the network is then an “improved average” that fits the precursor trend of previous lags. Table 5 presents the decomposition for the ANN prediction errors (“E” process).

The decomposition of the error is quite significant. We understand easily that the total error committed is important: the combination of pre-processing and modeling data (tilting, normalization for ANN) induces an error of about 20%. Furthermore the year 2009 (January–June) is not representative, there was an escalation of 13% error due to specific meteorological conditions (rainy period). Moreover, the 80° module tilt is very destabilizing, because the winter is a period of high climatic instability, but also a period where irradiation (80°) is the most important.

To try to compare all prediction methods (A–F) in a last step, we tried to present a new factor that reflects the tendency of prediction (the first derivative linked to the signal predicted slope). In the case of the measured signal, we use the $\Delta_m$ coefficient defined as follows ($I$ is the measured values of PV electricity energy): $\Delta_m = I(t+1) - I(t)$, and in the case of prediction using the coefficient $\Delta_p$ ($\hat{I}$ represents the values predicted for the DC PV energy): $\Delta_p = \hat{I}(t+1) - \hat{I}(t)$.

It is interesting to note that a Student’s $T$-test between the coefficient $\Delta_m$ and $\Delta_p$ for ANN and ARMA shows that predictions with the ANN process better represent the reality ($t = 0.364$ and $p = 0.71$ for ANN vs. $t = 0.584$ and $p = 0.56$ for ARMA), recall that $p$-value is the probability

---

**Fig. 7.** Cumulative predictions and measures. (a) ANN without pre-processing, (b) ANN with clearness index, (c) ANN with clear sky index and (d) ARMA with clear sky index.

---

**Table 5**

<table>
<thead>
<tr>
<th></th>
<th>nRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-processing and modeling</td>
<td>20</td>
</tr>
<tr>
<td>Specificity of year 2009</td>
<td>13</td>
</tr>
<tr>
<td>Electrical conversion</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Total observed error</td>
<td>34</td>
</tr>
</tbody>
</table>
that the averages between the forecasting and measured values are identical. So according to this indicator, on average, ANN is more representative than ARMA for the trend estimation. Table 6 presents the coefficient comparing the quality of prediction based on the ratio $\Delta = \Delta_m/\Delta_p$ for all prediction methods (A)–(F). This type of parameter “$\Delta$” allows determining four interpretation intervals:

- $<0$: number of very bad interpretation, for example: $\Delta_m > 0$, $\Delta_p < 0$, and vice versa,
- $[0;1 − ε]$ [and] $1 + ε$: area where one commits an error of prediction; take $ε = 0.2$. In the first interval on a $\Delta_p > \Delta_m$ and the other interval $\Delta_p < \Delta_m$,
- $[1 − ε; 1 + ε]$; good prediction zone, the sign and the absolute value is correct.

Table 6 presents the “$\Delta$” values and its results for the four classes considering for each interval, the number of events obtained on the studied period.

With this methodology we must recognize that ANN and ARMA are a little different. (A) and (B) processes are not efficient for daily DC PV energy predictions (Fig. 8a). It must be said that it is very difficult to distinguish the quality of prediction made with ANN or ARMA (Fig. 8b). Only the Student test used with the coefficient $\Delta_m$ and $\Delta_p$ allows separating the prediction quality of the two models with ANN slightly higher than ARMA.

### 4. Conclusion and perspectives

In this paper, an ANN prediction approach developed to determine global irradiation at daily horizon $(d + 1)$ which can help electrical managers with grid-PV power systems connected is presented. This prediction model has been compared to other prediction methods (AR, ARMA, k-NN, Markov Chains, etc.). We have used an ad hoc time series pre-processing (based on clearness or clear sky indexes) and a time series prediction designed MLP. Although the location was very specific, with the proximity to the sea and the mountain that can greatly affect nebulosity, we have obtained relevant results.

Without pre-processing, AR(8) and ANN models presents better daily RMSE of about 3.6 MJ/m$^2$ (998 Wh/m$^2$) and nRMSE $\sim 21\%$ compared to Markov chain, Bayes, k-NN methods where nRMSE $\sim 25$–$26\%$. However, annual pre-processing ANN methods based on clearness index and clear sky index reduce forecasting errors of about 5–6% (nRMSE $\sim 20\%$) compared to classical predic-

![Fig. 8. Distribution of $\Delta$-values (X-axis) representing the quality of prediction. (a) A–B processes not relevant, (b) C–F processes corresponding to best ANN and ARMA simulators.](image)

tors as Markov chains. The choice of a pre-processing built on clearness index (with $H_{gh}^c$) of clear sky index ($H_{gh, clear sky}^d$) leads to comparable results.

These simulation tools have been successfully validated on the DC energy prediction of a 1.175 kWp mono-Si PV power system connected to the grid. On a seasonal point of view, ANN with clear sky pre-processing, represents for winter months (January, February) an adequate solution (nRMSE $\sim 37\%$). For summer months, ANN without pre-processing gives the best results (nRMSE $\sim 15\%$). Cumulative simulated and measured PV productions are in very good agreement ($R^2 > 0.99$) validating the whole prediction process (nRMSE $< 2\%$ for a 6-months period).

Finally, a new differential variable $\Delta$ has been introduced to study predicted tendency errors based on the first derivative predicted signal. According to a Student test, ANN and ARMA simulated daily irradiation profiles confirm the good accuracy for the predicted tendency at $d + 1$. These two methodologies are similar.

In the future, this tool (ANN) could eventually help the system manager, for implanting new PV systems especially on isolated electrical grid where 30% for renewable energy power system represents an integration limit. With operational prediction tools, this threshold can be increased to 50% allowing to limit the utilization of fossil plants to supply electricity.

Moreover, it seems important to study shorter time horizons (hour). As a matter of fact, electrical managers are also interested to horizons that can range from $\frac{1}{2}$ hour

<table>
<thead>
<tr>
<th>$\Delta$-Values</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;0$</td>
<td>59</td>
<td>45</td>
<td>33</td>
<td>32</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>0–0.8</td>
<td>17</td>
<td>24</td>
<td>54</td>
<td>53</td>
<td>56</td>
<td>53</td>
</tr>
<tr>
<td>0.8–1.2</td>
<td>22</td>
<td>23</td>
<td>11</td>
<td>5</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>$&gt;1.2$</td>
<td>38</td>
<td>44</td>
<td>38</td>
<td>46</td>
<td>43</td>
<td>43</td>
</tr>
</tbody>
</table>

Table 6: Tendency of prediction with $\Delta$-values distributions for all processes.
References

Badescu, V., 2008. Modelling Solar Radiation at the Earth Surface, Collectivity of Corsica. We thank the French National
Meteorological Organization (Météo-France) which has
supervised the data collection from their data bank for Campo del’Oro synoptic station. We thank Jean PANIGHI (University of Corsica) for the supply of irradiance
sensors technical characteristics.

Acknowledgements

This work was partly supported by the Territorial
Collectivity of Corsica. We thank the French National
Meteorological Organization (Météo-France) which has
supervised the data collection from their data bank for Campo del’Oro synoptic station. We thank Jean PANIGHI (University of Corsica) for the supply of irradiance
sensors technical characteristics.

Appendix A

MLP weight values after training with clarity index for the
global radiation (most efficient configuration). W1 cor-
responds to the matrix where each element \( W_{i,j} \) is
the weight related to the number of input neurons (from 1
to 8) and the number j of hidden neurons (from 1 to 3).
W2 corresponds to a vector with three elements represent-
ing the weights between the three hidden nodes and the
output. B1 is the bias vector related to each hidden neuron
(three values) and B2 the output bias value.

\[
W_1 = \begin{pmatrix}
1.26 & 0.02 & 0.26 & -0.001 & -0.08 & 0.34 & 0.12 & 0.13 \\
0.02 & -0.19 & 0.29 & -0.89 & 0.41 & 0.64 & 0.78 & -0.85 \\
0.21 & 0.50 & -0.14 & -0.09 & -0.61 & -0.48 & 0.49 & -0.02
\end{pmatrix}
\]

\[
W_2 = \begin{pmatrix}
0.47 \\
-0.07 \\
-0.07
\end{pmatrix}
\]

\[
B_1 = \begin{pmatrix}
-0.90 \\
-0.53 \\
1.02
\end{pmatrix}
\]

\[
B_2 = (-0.13)
\]

References


