# Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation

Cyril Voyant, Marc Muselli, Christophe Paoli, Marie Laure Nivet

# Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation

Cyril Voyant[1,2], Marc Muselli[1*], Christophe Paoli[1], Marie-Laure Nivet[1]

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

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

**Abstract.**

This paper presents an application of Artificial Neural Networks (ANNs) to predict daily solar radiation. We look at the Multi-Layer Perceptron (MLP) network which is the most used of ANNs architectures. In previous studies, we have developed an ad-hoc time series preprocessing and optimized a MLP with endogenous inputs in order to forecast the solar radiation on a horizontal surface. We propose in this paper to study the contribution of exogenous meteorological data (multivariate method) as time series to our optimized MLP and compare with different forecasting methods: a naïve forecaster (persistence), ARIMA reference predictor, an ANN with preprocessing using only endogenous inputs (univariate method) and an ANN with preprocessing using endogenous and exogenous inputs. The use of exogenous data generates a nRMSE decrease between 0.5% and 1% for two stations during 2006 and 2007 (Corsica Island, France). The prediction results are also relevant for the concrete case of a tilted PV wall (1.175 kWp). The addition of endogenous and exogenous data allows a 1% decrease of the nRMSE over a 6 months-cloudy period for the power production. While the use of exogenous data shows an interest in winter, endogenous data as inputs on a preprocessed ANN seem sufficient in summer.

**Keywords:** Time Series Forecasting**,** Processing, Artificial Neural Networks, PV Plant, Energy Prediction, Stationarity

*Corresponding author: Marc,MUSELLI, tél: +33 4 95 52 41 30, fax (33 4 95 52 53 28) email: marc.muselli@univ-corse.fr

| | |
|---|---|
| ***Time series nomenclature*** | |
| $\hat{x}_t$, $\hat{x}_{d,y}$ | Radiation time series model at time *t* and at day *d* and year *y* |
| $\mathbf{x_t}$, $\mathbf{x_{d,y}}$ | Radiation time series data at time t and at day *d* and year *y* |
| $y_t$ | Exogenous time series data at time t |
| $S_t$ , $S_{d,y}$ | Stationary time series (clear sky index) for the day *d* and the year *y* or for the time *t* |
| ***Clear sky model*** | |
| $H^d_{gh,clearsky}$ | Clear sky global horizontal irradiance [MJ/m²] integrate on the day *d* |
| $H_0$ | Extraterrestrial solar radiation coefficient [MJ/m²] |
| $\tau$ | global total atmospheric optical depth |
| **H** | Solar elevation angle |
| ***B*** | Fitting parameter of Solis clear sky model |
| ***Correlation analysis*** | |
| $\mathbf{r_k}$ | Autocorrelation factor estimation for time lag *k* |
| $\rho_{kk}$ | Estimation of partial autocorrelations for time lag *k* |
| ***R*** | Cross-correlation estimation |
| ***Levenberg-Marquard algorithm*** | |
| $\Delta x$ | ***Parameter to optimize*** |
| ***J*** | ***Jacobian matrix*** |

| | |
|---|---|
| ***Meteorological parameters*** | |
| ***P*** | Pressure average [Pa] |
| ***DGP*** | Daily gradient pressure [Pa] |
| ***N*** | Nebulosity, cloudy height class [Octa] |
| ***T*** | Ambient temperature maximum (TM), minimum (Tm), average (Ta) and night (Tn) at 3:00 AM [°C] |
| ***Ws*** | Wind speed, Average at 10 meters [$m.s^{-1}$] |
| ***PKW*** | Peak wind speed, Maximum speed of 10 meters [$m.s^{-1}$] |
| ***Wd*** | Wind direction measured at 10 meters [deg] |
| ***Su*** | Sunshine duration, direct irradiance from the Sun of at least 120 watts per square meter [h] |
| ***RH*** | Relative humidity, Water vapor in the air [%] |
| ***RP*** | Rain precipitations measured in standard rain gauge [mm] |
| ***Frontage PV parameters*** | |
| $\mathbf{E_{pv,ac/dc}}$ | Photovoltaic wall power (MJ) |
| $\eta_{PV}$ | Plant efficiency (%) |
| $\mathbf{I_\beta}$ | Daily global radiation (tilt of β) [MJ/m²] |
| **S** | Surface of PV wall [m²] |
| **PR** | Performance ratio of the PV plant |

| | | | | |
|---|---|---|---|---|
| $I$ | *Identity matrix* | | | |
| $e(x)$ | *Error term* | | | |
| $\mu$ | *specific algorithm parameter* | | | |

**Nomenclature.**

## 1 Introduction

We present the results of the prediction of global radiation time series using Artificial Neural Networks (ANNs) which are a popular artificial intelligence technique in the forecasting domain [1-6]. Inspired by biological neural networks, researchers in a number of scientific disciplines are designing ANNs to solve a variety of problems in decision making, optimization, control and obviously prediction [7-11], and more particularly time series prediction. A Time Series (TS) [12,13] is a collection of time ordered observations $x_t$, each one being recorded at a specific time t (period). TS are used in a wide set of domains such as finance, production or control, just to name a few. A TS model ($\hat{x}_t$) assumes that past patterns will occur in the future. TS prediction or TS forecasting takes an existing series of data $x_{t-k}$, .. , $x_{t-2}$, $x_{t-1}$ and forecasts the $x_t$ data values. The goal is to observe or model the existing data series to enable future unknown data values to be forecasted accurately. Thus a prediction $\hat{x}_t$ can be expressed as a function of the recent history of the time series, $\hat{x}_t = f(x_{t-1}, x_{t-2}, \ldots x_{t-k})$ [14]. There are a lot of methods to construct this model, some of the best predictors found in literature are ARIMA [13,14], Bayesian inference [15,16], Markov chains [17,18], k-Nearest-Neighbors predictors [19,20] or ANN [21]. In previous studies [22,23], we have demonstrated that an optimized ANN with endogenous inputs can forecast the global solar radiation with acceptable errors. An ANN is made up by simple processing units, the neurons, which are connected in a network by synaptic strengths (weights), where the acquired knowledge is stored. There are a lot of different ANNs that diverge on several features, such as the learning paradigm or the internal architecture

[3]. 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. In a MLP, neurons are grouped in layers and only forward connections exist. This provides a powerful architecture enable to learn any kind of continuous nonlinear mapping. A typical MLP consists of an input, hidden and output layers (see Figure 1).

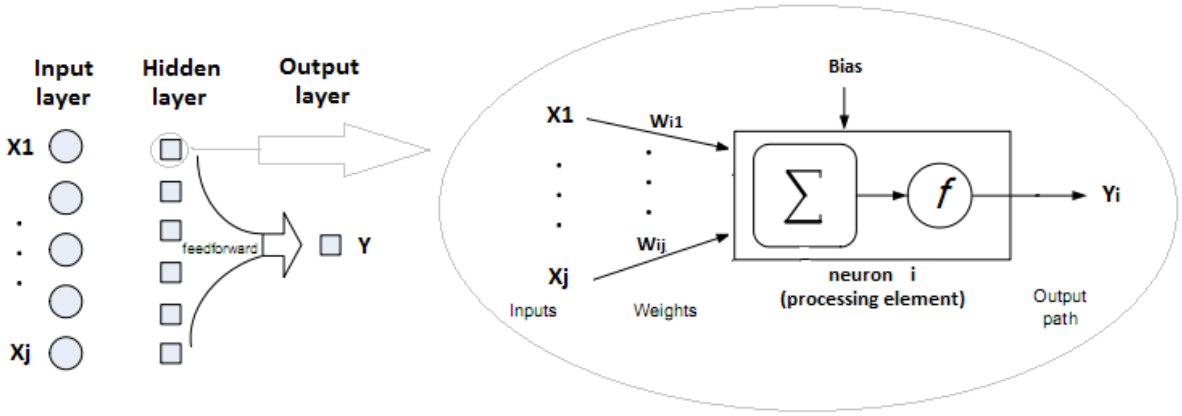



**Figure 1: Example of an MLP (left) and details of a neuron from the hidden layer (right)**

Other components include neurons, weights and a transfer function. 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 $y_i$ represented by $y_i = f(\sum_{j=1}^{n} x_j w_{ij})$ where $i$ is a neuron index in the hidden layer and $j$ is an input index to the neural network. Training is known as the process of modifying the connection weights in some orderly fashion using a suitable learning method or training algorithm. In this process an input is presented to the network along with the desired output which is a real observation and the weights are adjusted so that the neural network attempts to produce the desired output. Another issue involved in designing and training a MLP network is to find a globally optimal solution that avoids local minima. Several methods have been tried to avoid these local minima and the simplest is to try a number of random starting weights and use the one with the best value.

In this present paper, our aim is to answer to two questions: how the use of exogenous variables with ANN (multivariate method) increases the quality of prediction and so, how to select the appropriate data? We try to answer to these questions in three points. First, we propose a useful methodology to optimize the ANN. Among

other things we explain how to choose the network inputs, meaning optimize the number of endogenous and exogenous variables and lags (delays in information between network nodes). Secondly, we test this methodology comparing the performances of the optimized ANN obtained with an ANN taking into account only endogenous inputs. Finally, we validate all approaches using the ANN proposed and other predictors (persistence, ARIMA, ANN with only endogenous input) to forecast the AC production of a PV plant. All the data used come from meteorological stations located on the island of Corsica (France) which is characterized by a Mediterranean climate and a hilly terrain. The paper is organized as follow: section 2 describes the context in which this research was done and the data we used. Section 3 describes the methodology used: the time series preprocessing, the ANN configuration and how we added endogenous and exogenous parameters at different time lags. The results for the prediction of global horizontal radiation are shown and commented in section 4 where the global approach is also validated on a 80° tilted PV plant.

## 2 Context and presentation of data

In this work, measured global daily radiation data from meteorological ground stations are used to forecast global solar irradiation for the next day [24]. The global radiation consists of three components: direct, diffuse and ground-reflected radiations [21,25-28]. The ground-reflected radiation does not concern the first part of this work 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 sensor [29,30]. With cloudy sky, we are in front of a mostly stochastic phenomenon, which depends on the local weather. In Corsica, the official meteorological network (from the French Meteorological Organization called Météo-France) is very poor: only three sites being about 50 km apart are equipped with pyranometers and are enable to measure semi-hourly global horizontal radiation (MIRIA Stations of Degreane industry). Only two of these are equipped with standard meteorological sensors (pressure, nebulosity, etc.): Ajaccio (41°55'N and 8°48'E, seaside, 4 m) and Bastia (42°33'N, 9°29'E, seaside, 10 m). Both locations have a 'Mediterranean' climate, hot summers with abundant sunshine and mild, dry, clear winters. The stations are located near the sea but there is also relief nearby (40 km from Ajaccio and 15 km from Bastia) making nebulosity difficult to forecast. The data representing the global

horizontal solar radiation were measured on an half-hourly basis from January 1998 to December 2007. Half-hourly measured data are transposed in hourly data, which are integrated to produce a new daily data set. The daily time step has been chosen considering the needs of the suppliers which are interested in the estimation of the fossil fuel saving for its electrical thermal plants (191 MW in Corsica, ~ 30%). A first treatment allows us to clean the series of non-typical points related to sensor maintenances or absence of measurement. Less than 4 % of measurements were missing and replaced by the hourly average for the given day. The other meteorological parameters available and studied from January 1998 to December 2007 are pressure (*P*, Pa; average and daily gradient*, measured by numerical barometer during 1 hour), nebulosity (*N*, Octas), ambient temperature (*T*, °C; maximum, minimum, average and night†, measured done during an half hour), wind speed (*Ws*, m/s; average at 10 meters, measured during the 10 last minutes of the half hourly step), peak wind speed (*PKW*, m/s; maximum speed of wind at 10 meters, measured during 30 minutes), wind direction (*Wd*, deg at 10 meters measured during an half hour), sunshine duration (*Su*, h, computed with the global radiation series and the power threshold 120 W.m²), relative humidity (*RH*, % instantaneous measure at the end of the half-hour) and rain precipitations (*RP*, mm, 5 cumulative measures of 6 minutes during the half-hour). The data are transposed into hourly measure by Météo-France service and are available on the official site[x].

## 3 Methodology

In this section we introduce the methodology we follow to forecast global solar irradiation for the next day. First, we present the main steps of an ad-hoc time series processing used to determine a stationarization methodology for the daily signal. The next sub-section details the ANN architecture and how we have constructed it. In the last sub-section we present how we have added endogenous and exogenous parameters at different time lags.

[x] http://climatheque.meteo.fr
* Difference between the mean pressure of day *j* and day *j-1*
† Measured at 3:00 AM

## 3.1 Time series processing

The prediction of the solar energy time series on the earth's surface can be perturbed by the non-stationarity of the signal and the periodicity of the phenomena [31-33,13] (Figure 2.a). We have used physical phenomena in an attempt to overcome the seasonality of the resource (determinist component). In daily case, seasonality is observed on the annual period. According to previous experimentations on horizontal global radiation [31,13] and specially in Corsica [22,23], we have developed a method in order to make the series stationary in an attempt to increase the prediction quality. Our method is based on the clear sky model (with the *clear sky index).* Several methods allow to determine this model. In our case, we have used the simplified "Solis clear sky" model [34,35] based on radiative transfer calculations and the Lambert-Beer relation. In this case, the clear sky global horizontal irradiance ($H_{gh,clearsky}$) reaching the ground is defined by:

$$H_{gh,clearsky} = H_0 . e^{-(\tau / \sin^b(h))} . \sin(h) \qquad \text{Eq 1}$$

where $\tau$ is the global total atmospheric optical depth, $h$ is the solar elevation angle and $b$ is a fitting parameter. According to [34,35] and after several experiences, we have chosen a $\tau = -0.37$ and a $b = 0.35$. The daily integration of the $H_{gh,clearsky}$ parameter allows to determine the daily solar radiation $H^d_{gh,clearsky}$. We have validated the Solis model on a horizontal global radiation with a couple of tests considering one year of daily solar radiation data that are not presented in this paper. We obtain a relation of stationarization (Eq 2) where $X$ is the measure and $S$ the new time series, ($d$ is the day for year $y$):

$$S_{d,y} = X_{d,y} \Big/ H^d_{gh,clearsky} \qquad \text{Eq 2}$$

This treatment aims to create a new distribution without periodicity (Figure 2.b). Moreover the new series generated is equivalent to a nebulosity signal. Ideally, the values are fixed to 1 and decrease with cloudy occurrences. The values superior to 1 are generated by errors measurement, or by the stationarization method which are occasionally not adapted. Obviously, the effect of the proximity to the sea and the mountains are difficult to take into account in a single clear sky model. The values greater than 1 are not corrected during the estimations.

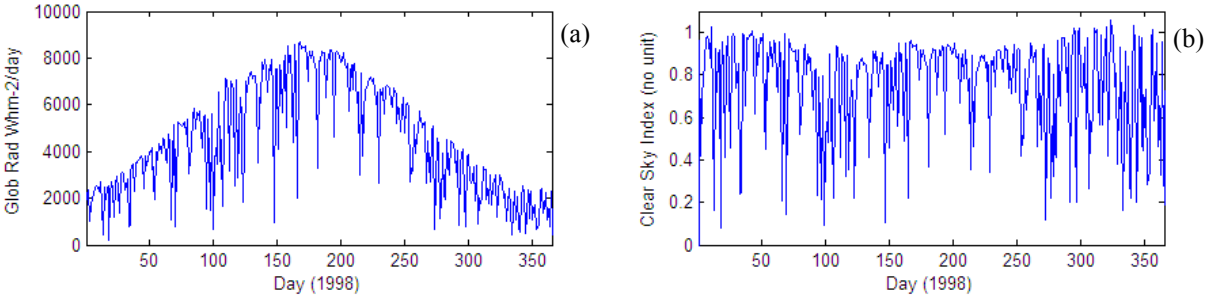



**Figure 2: (a) Daily measured horizontal irradiation (Ajaccio, 1998) ; (b) Impact of the stationarization on the original time series.**

Once the stationarization has been done we have to follow the process in order to find the best network configuration.

## 3.2 ANN configuration

The next step of the methodology has been to find the best ANN configuration. The optimization used is composed by four independent and chronological subparts:

- Choice of the ANN architecture
- Choice of the endogenous lags number
- Choice of the exogenous lags numbers for each parameters: pressure, pressure variation; wind direction; peak wind, humidity; sunshine duration; nebulosity; precipitation; min-max-mean ambient temperature, night ambient temperature and wind speed.
- Optimization of the ANN: data normalization, hidden neurons number, etc.

We have tried to study all the parameters available in this network architecture. The principal parameters which influence the number of local minima, the network complexity and the difficulty of the learning phase are: the inputs number i.e., the number of endogenous time lags and the number of exogenous inputs and time lags for each of them; the hidden layers number and their neurons number; the activation (or transfer) function; the learning algorithm and the comparison function used during the learning phase. Additionally, the normalization of data, the learning sampling size and the data distribution between learning, test and validation phases must be taken into account. For convenience we have chosen to optimize the parameters separately with an intuitive order of preference. So we have optimized parameters by considering each other constant. The optimization of $P_i$

parameter (where $i \leq M$, $M$ is the total number of optimization parameter) consists on finding the parameter value which minimizes the forecasting error ($dE = 0$). Applying this naive assumption, the optimization is therefore to consider separately the parameters:

$$dE = \sum_{i=1}^{M} \frac{\partial E}{\partial P_i}.dP_i \quad \text{With} \quad \forall i, j \;\; P_i \perp P_j \quad \Rightarrow \quad \frac{\partial E}{\partial P_i} = 0 \qquad \text{Eq 3}$$

To exploit the optimization assumption, we developed a chronological process which consists of a sequence of optimization parameters. At each stage we have used the best configuration obtained in the previous steps. We particularly look at the MLP architecture because it has been the most used both in the renewable energy domain and in the time series forecasting. The basic architecture for a MLP application to time series forecasting, fixes number of past values in the input layer and the output is required to predict a future value of the time series [36]. The MLP has been computed with the Matlab software and the Neural Network toolbox. The obtained characteristics are: one hidden layer, the activation functions are hyperbolic tangent (hidden) and linear (output), the Levenberg-Marquardt learning algorithm (with a max fail parameter before stopping training equal to 5). This algorithm is an approximation to the Newton's method [37] and is represented by the equation 4:

$$\Delta x = \left[ J^T(x).J(x) + \mu I \right]^{-1} J^T(x).e(x) \qquad \text{Eq 4}$$

In our case the parameter μ takes the value 0.1 and 0.001 when the error, respectively, decreases or increases. Inputs are normalized on {-1,1}. Training, validation and testing data sets were respectively set to 80%, 10% and 10% (Matlab parameters). These phases concern the 8 first years and the global solar radiation forecasting the 2 last years. The next part explains the methodology used to find out the number of endogenous lags which will be put together with the exogenous ones in the input layer of the MLP (see details on Figure 3).

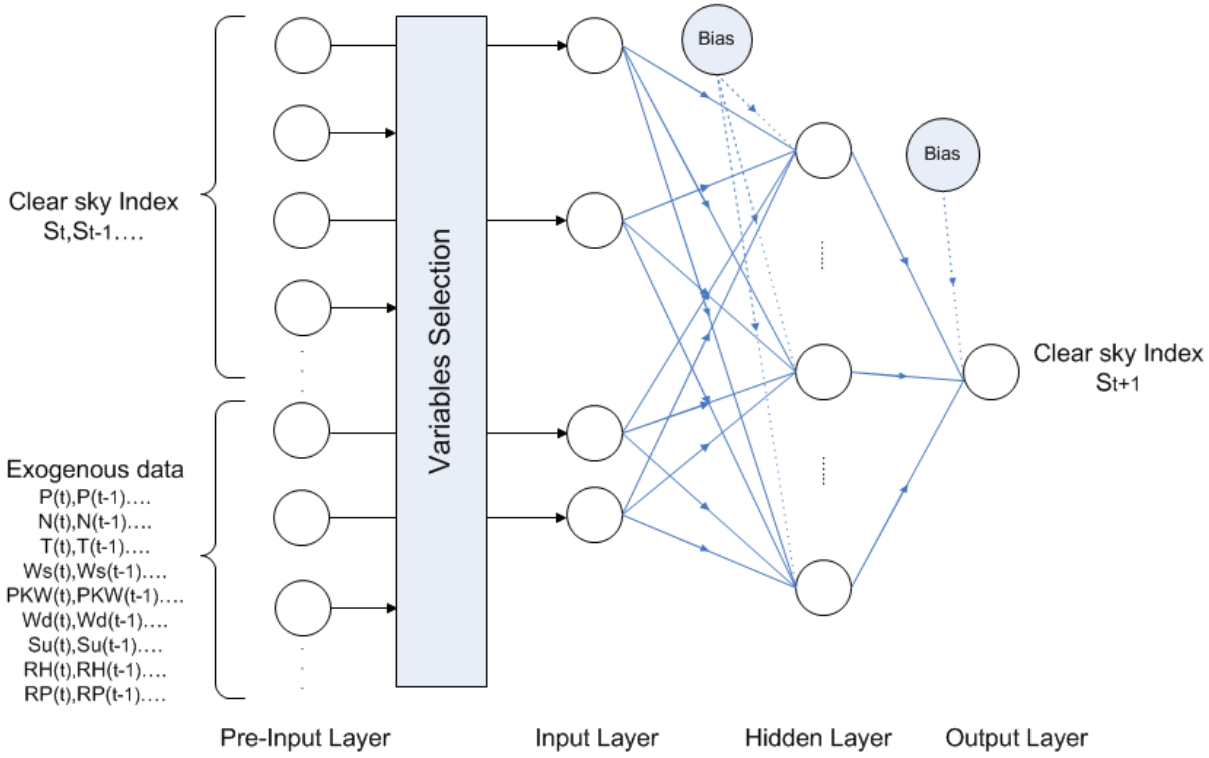



**Figure 3: Detail of the overall real system prediction.**

## 3.3 Methodology to select the time lags for endogenous and exogenous parameters

On the key tasks in time series forecasting is the selection of inputs variables [38,39]. Indeed we studied how to add efficiently endogenous and exogenous parameters at different time lags using correlation criteria and validation test based on student T-test. Since ANN are non linear, their calculations gives an indication rather than a standard tool for finding useful variables and lags [38].

### 3.3.1 The endogenous case

In this sub-section, we present how we have determined the number of endogenous time lags to take into account as inputs of the MLP. We have chosen to follow some of the principles of the Box and Jenkins [40] autoregressive integrated moving average (ARIMA) methodology. The Box and Jenkins approach has been one of the most widely used linear models in times series forecasting. As proposed in this methodology, Partial autocorrelation function (PACF; $\rho_{kk}$) which is an extension of the AutoCorrelation Function (ACF; $\rho_k$) aims at

identifying the extent of the lag in an autoregressive model. We decided to use it in order to select the best endogenous time lag for the neural network input. Using PACF allows to select only the time lag correlated with the future step. It can be considering as follow, it gives the autocorrelation between $x_t$ and $x_{t-k}$, when the linear dependences of $x_{t-1}$ through to $x_{t-k+1}$ have been removed. In other words the PAC is similar to autocorrelation, except that when calculating it, the (auto) correlations with all the elements within the lag are partially out. In practice, the last $\rho_{kk}$ different to zero leads to take into account k endogenous clear sky index in the network input layer.

To compute the PACF we have first to consider the ACF. The ACF estimation ( $r_k$ ) is computed from the chronological series (Eq 5) [16].

$$r_k = \sum_{t=k+1}^{N}(x_t-\bar{x})(x_{t-k}-\bar{x}) \Big/ \sqrt{\sum_{t=k+1}^{N}(x_t-\bar{x})^2 \sum_{t=k+1}^{N}(x_{t-k}-\bar{x})^2} \qquad \text{Eq 5}$$

The empirical estimation of partial autocorrelations, noted $r_{kk}$, are obtained from theoretical partial autocorrelation $\rho_{kk}$ by replacing the $\rho_i$ by their estimations $r_i$. One of the best representations of the PACF is a succession of determinant ratio defined by $\rho_{kk} = \left|P_k^*\right| / \left|P_k\right|$

Where

$$P_k = \begin{bmatrix} 1 & \rho_1 & \dots & \dots & \rho_{k-1} \\ : & 1 & & & : \\ : & & . & & : \\ : & & & . & : \\ \rho_{k-1} & & \dots & \dots & 1 \end{bmatrix} \qquad \text{Eq 6}$$

$\left|P_k^*\right|$ is the determinant of the matrix $P_k$ where the last row is replaced by the vector [ $\rho_1 .. \rho_k$ ]. From k>3, it's generally easier to use the recursive formula:

$$\rho_{ii} = \begin{cases} \rho_1 \quad if\ i=1 \\ \dfrac{\rho_i - \sum_{j=1}^{i-1}\rho_{i-1}\rho_{i-j}}{1-\sum_{1}^{i-1}\rho_{i-1}\rho_j} \quad i=2...k \end{cases} \qquad \text{Eq 7}$$

Finally, in order to quantify the time lags which are significantly correlated, we have to consider the 95% Confidence Interval of the PACF. This interval is the same as the 95% CI of the ACF which is given by $CI = \pm 1.96\sqrt{1/N}$. This formula is obtained considering that the distribution of estimator follows asymptotically a normal distribution. The Student's T-test (introduced by William Sealy Gosset in 1908) has been used to verify that the correlation coefficients were significantly different from zero.

This process of endogenous input lags selection is easy to perform graphically and is presented latter in this paper (Section 4).

### 3.3.2 The exogenous case

A correlation measure is computed in order to determine which of the exogenous parameters are to consider. The correlation between two variables reflects the degree to which the variables are linked (considering the limitation that correlation criteria can only detect linear dependencies between variables). The most common correlation measure is the Pearson's correlation. A correlation of +1 (or -1) means that there is a perfect positive (or negative) linear relationship between variables and a value of 0 implies that there is no linear correlation between the variables. The Pearson correlation coefficient between two variables is defined as $R = \mathrm{cov}(X,Y)/\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}$, where *cov* designates the covariance and *var* the variance. For a TS, the estimation of R is given by *(*Eq: 8):

$$R = \sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) \Big/ \sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2 \sum_{k=1}^{N}(y_k - \bar{y})^2} \qquad \text{Eq 8}$$

where the bar notation stands for an average over the index *k*.

Generally, a Pearson correlation between -0.5 and 0.5 indicates a weak, little or no association between two variables. The link to the Student test (T-test) shows that the score *R* may be used as a statistic test to assess the significance of a variable. In our experimental sample size, the limit of significance according to the T-test indicates a threshold very low. Indeed, the limit is below 0.1 for the sample above 1000 elements and for a 0.05 critical value for alpha level. This methodology is not appropriate in our case because the threshold for the coefficient *R* should be more important to select a limited number of exogenous inputs. We have chosen

intuitively an *R* threshold equal to 20%, only the higher correlation will be chosen. Indeed we have noted that the increase of this value is so restrictive that no exogenous data would be eligible, on the other hand, the decrease of this value is responsible of the augmentation of the number of input nodes, making the PMC architecture too complicated. Moreover, after the correlation analysis, each exogenous parameter is tested separately to exclude those which do not improve the prediction. In fact, a lower value of threshold selects more exogenous data, but after the prediction test, the number of exogenous data chosen is the same that in the 20% threshold. If the bound is fixed to 30%, there is no exogenous data to consider (for Ajaccio and Bastia), so the experience led us to consider the 20 % R threshold as an adequate solution.

If we use the previous notation for the clear sky index ($S_t$) and for an $y_t$ exogenous variable (representing nebulosity, temperature, etc.), the cross-correlation between the variables is:

$$R_k^y = \sum_{t=k+1}^{N}(S_t - \bar{S})(y_{t-k} - \bar{y}) \Big/ \sqrt{\sum_{t=k+1}^{N}(S_t - \bar{S})^2 \sum_{t=k+1}^{N}(y_{t=k+1} - \bar{y})^2} \qquad \text{Eq 9}$$

## 4 Results and discussion

In this section we present the main results obtained with data from meteorological stations of Bastia and Ajaccio. Finally we validate the methodology proposed, using the obtained simulator on a real frontage PV. Although there are several error indicators for measuring performance of time series forecasting, there is none that is uniformly accepted [41]. In this work, the traditional normalized Root Mean Square Error (nRMSE) has been used. The nRMSE is obtained by the formula: nRMSE = $\sqrt{<(x-y)^2>/<(x)^2>}$ where *x* represents the measurement and *y* the prediction.

### 4.1 Selection of endogenous inputs

To determine the MLP input number, the PACF is used as described previously (section 3.3.1). PACF allows to quantify the relevance of the endogenous inputs. Considering the auto-correlogram, the first $\rho_{kk}$ not significantly different of zero will induce the number of endogenous inputs ($|\rho_{kk}| > 1.96\sqrt{1/N}$). Figure 4

presents the PACF of the clear sky index for our two sites Bastia and Ajaccio. As we can see, for the first location (on the left), in order to predict $S_{t+1}$ we have to consider $S_t$, $S_{t-1}$, $S_{t-2}$, $S_{t-3}$ and $S_{t-4}$ as input of the ANN. In fact, the lag $S_{t-4}$ has not been considering because its value is very close to the upper bound of the PACF Confidence Interval. Moreover the experience has shown that no gain was added with this parameter and we have concluded that it was not significant. In fact, using these non relevant time lags will not decrease the prediction error, and will increase the number of local minima during the ANN learning phase. This principle called parsimony means eliminate higher time lags in order to simplify the predictor system. For the second station, Ajaccio (on the right on Figure 4), two parameters (k=2) are correlated with the clear sky index: $S_{t+1}$ correlated with $S_t$ and $S_{t-1}$.

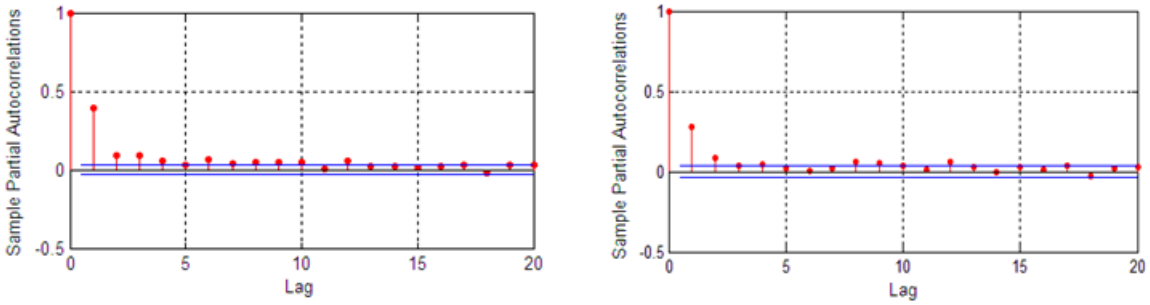



**Figure 3: Partial autocorrelation of the clear sky index (Bastia case on left and Ajaccio case on right). The lines around zero represent the upper and lower bounds of the PACF 95% confidence interval.**

## 4.2 Selection of the exogenous inputs

The second important step in the MLP input layer optimization is the choice of exogenous meteorological variables and their time lags. The cross-correlation study is done according to the description given in the previous section (3.3.2). Figure 5 represents the Pearson cross-correlation between the clear sky index and all the exogenous meteorological variables available. Note that the value of cross-correlation for lag 0 is given as information; it could not be used in a forecast approach. One can distinguish between simulations, where the output (radiation) and inputs (measures) are considered at the same time, and forecasts, where the inputs are measured ahead of time [42]. According to the 20% bound fixed in the previous section (represented on the curves of Figure 5 by the two lines centered around zero) only three exogenous data have to be taken into

account for Bastia. These are relative humidity (RH), sunshine duration (Su) and nebulosity (N). For each one only the first lag 1 is taking into account, considering the data of the previous day. In the Ajaccio case, the time lag 1 is interesting for the variables sunshine duration (Su), pressure (P), differential pressure (DGP), and nebulosity (N).

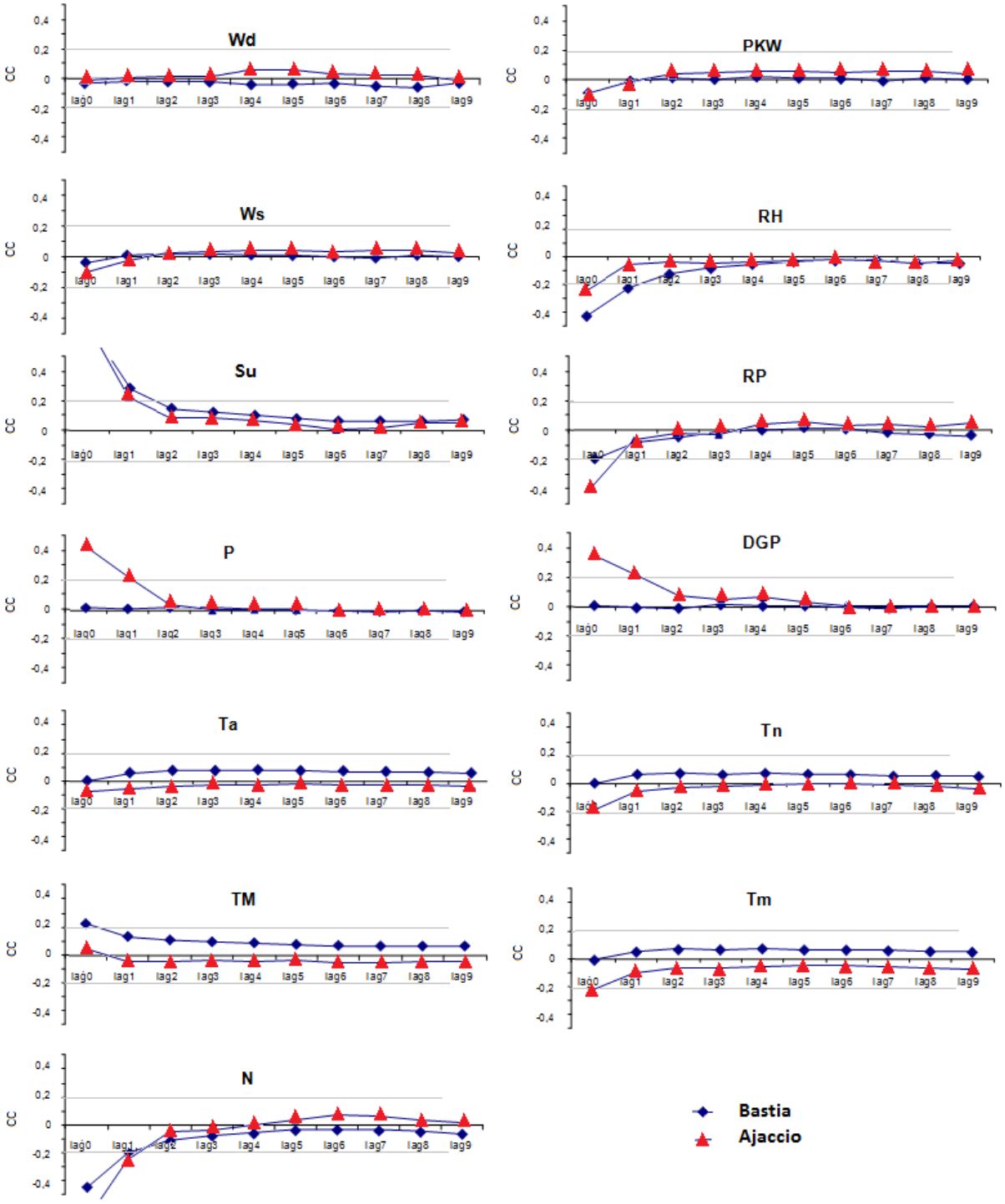



**Figure 4: Pearson cross-correlation between the clear sky index, and exogenous variables for Bastia and Ajaccio stations (wind direction Wd, peak of wind speed PKW, wind speed Ws, relative humidity RH, sunshine duration Su, precipitation RP, pressure P, differential pressure DGP, ambient temperature average Ta, night temperature Tn , max TM and min Tn temperatures and nebulosity N). The lines represent the confidence band chosen.**

The purpose of this study is not only to demonstrate the interest of using ANN technique, but also to search a mechanism for its optimization. Once the most significant variables are determined, another technique is applied to test the real impact of the exogenous variables in order to hold only the most relevant. We have evaluated the influence of all the exogenous parameters on the output taking them one by one. Table 1 (Bastia) and 2 (Ajaccio) show improved forecast results by the use of one exogenous inputs and an arbitrary number of hidden nodes equal to five.

| Lag Endo | Lag Su | Lag RH | Lag N | Hidden Nodes | nRMSE | IC95% |
|---|---|---|---|---|---|---|
| 4 | - | - | - | 5 | 25.85% | 0.46% |
| **4** | **1** | **1** | **1** | **5** | **25.60%** | **0.18%** |
| 4 | 1 | - | - | 5 | 25.65% | 0.17% |
| 4 | - | 1 | - | 5 | 25.66% | 0.09% |
| **4** | **-** | **-** | **1** | **5** | **25.58%** | **0.13%** |

**Table 1: Prediction error with only one exogenous input among those highly correlated for the site of Bastia. Bold letters represents the best configurations.**

For the Bastia case on Table 1, the use of the time lag 1 for relative humidity, sunshine duration or nebulosity decreases not only the mean of the error prediction but also increase the robustness of the methodology. The variance (represented by the CI95%) of the results becomes lower, meaning there are fewer local minima and so a greater accuracy. The other interesting element is the equivalence of the sunshine duration, humidity and nebulosity when they are used separately showing that there must all be taken into account.

| Lag Endo | Lag Su | Lag P | Lag DGP | Lag N | Hidden Nodes | nRMSE | IC95% |
|---|---|---|---|---|---|---|---|
| 2 | - | - | - | - | 5 | 22.50% | 0.16% |
| **2** | **1** | **1** | **1** | **1** | **5** | **21.62%** | **0.16%** |
| 2 | 1 | - | - | - | 5 | 22.04% | 0.08% |
| 2 | - | 1 | - | - | 5 | 22.43% | 0.14% |
| 2 | - | - | 1 | - | 5 | 22.47% | 0.14% |
| **2** | **-** | **-** | **-** | **1** | **5** | **21.79%** | **0.11%** |

**Table 2: Prediction error with only one exogenous input among those highly correlated for the site of Ajaccio.**

On Table 2 for the station of Ajaccio, the most interesting parameters are the nebulosity (N) followed by the sunshine duration (Su). The two other parameters, pressure (P) and daily gradient pressure (DGP) are not very interesting. They are not improving the error toward the endogenous case so their influence could be neglected (the nRMSE is 22.5% for the endogenous case and 22.43% and 22.47% for the pressure and the gradient

pressure). Later in the paper, for Ajaccio, we consider only the lag 1 exogenous inputs for sunshine duration and nebulosity. The other variables are equivalent and do not contribute to improve the prediction quality.

| Lag Endo | Lag Su | Lag RH | Lag N | Hidden Nodes | nRMSE | IC95% |
|---|---|---|---|---|---|---|
| 4 | 1 | 1 | 1 | 1 | 25.52% | 0.10% |
| **4** | **1** | **1** | **1** | **2** | **25.43%** | **0.11%** |
| **4** | **1** | **1** | **1** | **3** | **25.43%** | **0.16%** |
| 4 | 1 | 1 | 1 | 4 | 25.61% | 0.05% |
| 4 | 1 | 1 | 1 | 5 | 25.60% | 0.18% |
| 4 | 1 | 1 | 1 | 6 | 25.59% | 0.42% |
| 4 | 1 | 1 | 1 | 7 | 25.60% | 0.27% |
| 4 | 1 | 1 | 1 | 10 | 25.70% | 0.30% |

**Table 3: Optimization of the hidden layer for the site of Bastia.**

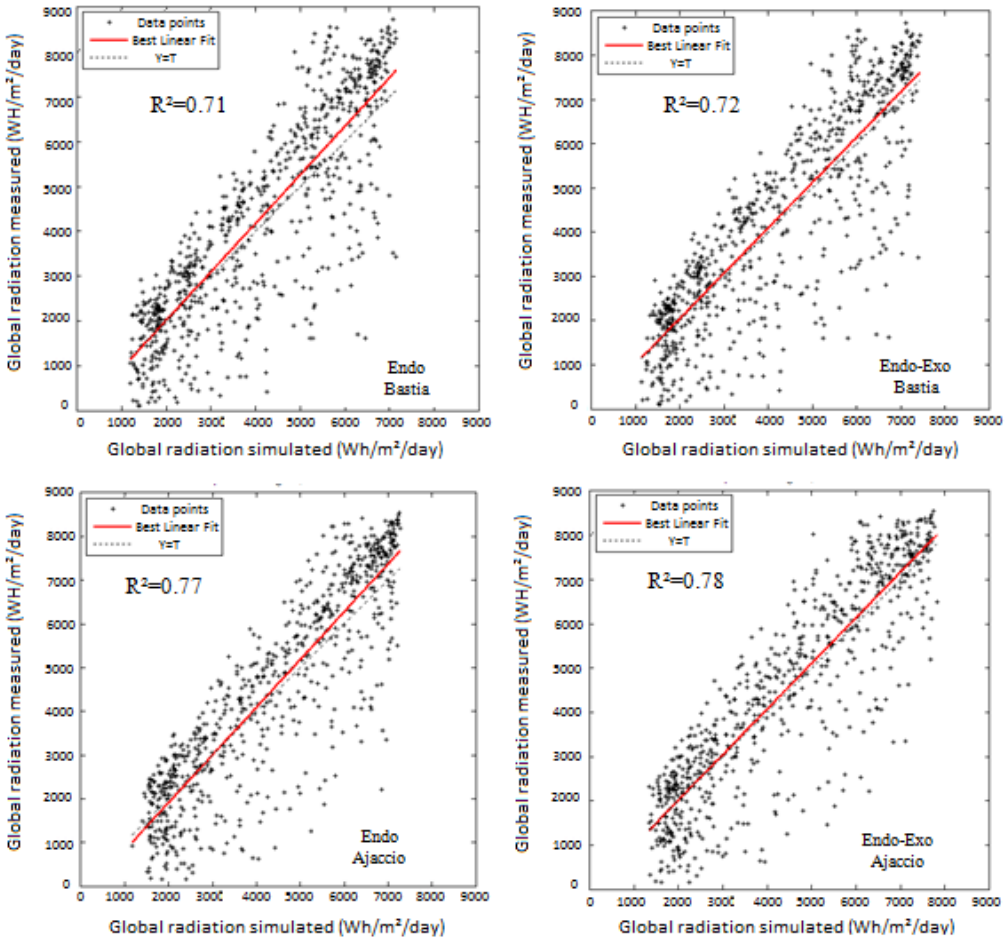
The last step of the optimization is the determination of the optimal number for hidden nodes. Table 3 and 4 are represented the nRMSE error versus the number of hidden nodes between one and ten for the two stations. In Bastia case on Table 3, at the beginning an increase of the number of hidden nodes results in a small decrease of the nRMSE unless trend reversed. In fact all results are almost equivalent even if a small increase of the local minima is observed with an increase of the hidden nodes. Considering this fact we should have to take two nodes for the hidden layer. Actually we have chosen three nodes for the hidden layer because the third number after the decimal point for the nRMSE was smaller than the nRMSE for the two nodes case (25.432% versus 25.438%).

| Lag Endo | Lag Su | Lag N | Hidden Nodes | nRMSE | IC95% |
|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 21.61% | 0.01% |
| 2 | 1 | 1 | 2 | 21.75% | 0.18% |
| **2** | **1** | **1** | **3** | **21.54%** | **0.05%** |
| 2 | 1 | 1 | 4 | 21.67% | 0.11% |
| 2 | 1 | 1 | 5 | 21.61% | 0.12% |
| 2 | 1 | 1 | 6 | 21.83% | 0.53% |
| 2 | 1 | 1 | 7 | 21.87% | 0.20% |
| 2 | 1 | 1 | 10 | 21.78% | 0.19% |

**Table 4: Optimization of the hidden layer for the site of Ajaccio.**

The optimization of the number of nodes on hidden layer for the second station is represented on the Table 4. Like in Bastia case, only three neurons on the hidden layer are necessary.

## 4.3 Forecasting errors




5

6 **Figure 5 : Comparison between the use of exogenous and endogenous inputs (on right) and only endogenous input (on left). The top is related to Bastia and the bottom to Ajaccio.**

For Bastia location, the improvements related to exogenous/endogenous inputs (multivariate data) are minimal but real. On the Figure 6, the global radiation computed by the two methodology are equivalent, although the determination coefficient shows that the use of exogenous data improve the prediction ($R^2$=0.72 versus $R^2$=0.71). For the prediction error, the nRMSE is 25.43% for multivariate method (see Table 3) against 25.85% for univariate method (Table 1). The RMSE is reduced by 20 Wh/m² (1233 Wh/(m².day) vs 1253 Wh/(m².day)), and the mean absolute error by 51 Wh/m² (957 Wh/(m².day) vs 1008 Wh/(m².day)). The naive forecaster based on the persistence leads to a nRMSE = 31.17% (MAE = 1081 Wh/(m².day) and RMSE = 1569 Wh/(m².day)). For Ajaccio, the gain of the exogenous inputs utilization is more interesting. The nRMSE is

21.54% for multivariate method (see Table 4) against 22.50% for univariate method (Table 3). The RMSE is reduce by 52 Wh/m² (1087 Wh/(m².day) versus 1139 Wh/(m².day)), and the mean absolute error by 73 Wh/m² (839 Wh/(m².day) versus 912 Wh/(m².day)). The naive persistence leads to nRMSE = 27.07% (MAE = 971 Wh/(m².day) and RMSE = 1422 Wh/(m².day)). On Figure 7 the comparison between the two methodologies of prediction is shown. Like in Bastia case the determination coefficient indicates that the exogenous methodology improves the prediction ($R^2$=0.78 Versus $R^2$=0.77).

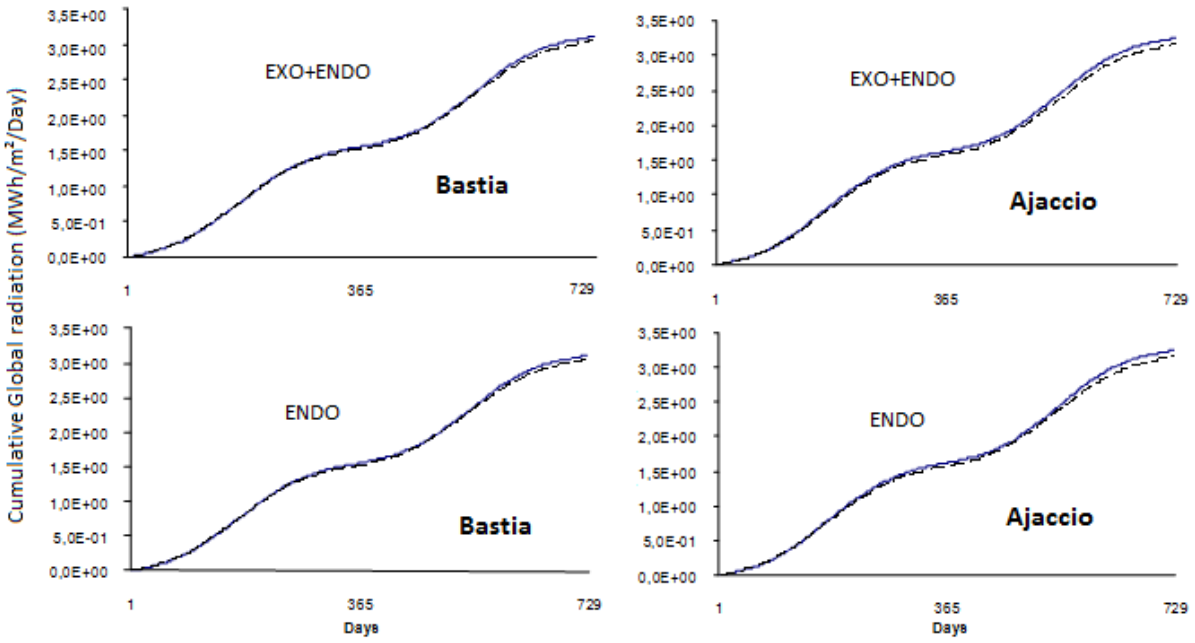



**Figure 6: Comparison of cumulative global radiation between the use of exogenous input and only endogenous input. The dotted line is the prediction and the continuous line is the measure on Bastia and Ajaccio ($R^2$=0.99 for all figures).**

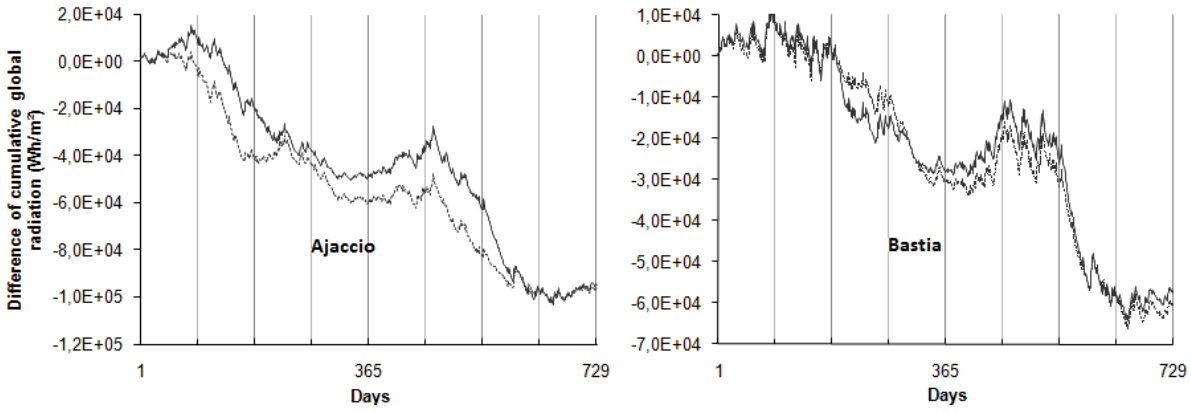



**Figure 8: Differences between measures and ANN simulations in Ajaccio and Bastia (continuous line is ANN with only endogenous data, dotted line the endogenous and exogenous data). A negative value means that the simulation underestimates the global radiation.**

On the Figure 7 and the Figure 8, the cumulative global radiation for both configurations is compared. Figure 8 helps to compare the predictors when the errors generated are low and allows to make a zoom on the error prediction. The trend of the difference of cumulative global radiation is decreasing and is maximal after two years. For Bastia, the error generated (bi-annual integration) by the both simulations are very low (60 kWh/m² for exogenous data use and 57 kWh/m² for only endogenous data, the cumulative measure being equal to 3.1 MWh/m²). There is an inversion of the trend after the 200$^{th}$ day (Figure 8), the difference of cumulative radiation related to the exogenous methodology become more important than the only endogenous method. The cumulative curve is a comparison tool, but the interpretation is not trivial. A predictor can generate a high nRMSE and a low cumulative error like in the persistence case where the cumulative modeling is equal to the original series but the daily error generated is high. It is surprising, but it seems that the sunny period increases the error while the winter period compensates it (rebounds curves on Figure 8 at 100$^{th}$ and 500$^{th}$ day). The two prediction methods with univariate and multivariate data coincide with the cumulative measurements ($R^2$ = 0.99). This site (Bastia) is well known to be very difficult to predict and this result was already found in previous studies [31].

For Ajaccio, Figure 7 shows that the error generated by the two simulations is very low, the determination coefficient does not allow to separate the two predictors (univariate versus multivariate) on the cumulative study

($R^2$ = 0.99 for both methods). In fact, the exact cumulative error value after two years (where there is compensation between negative and positive values of daily prediction) is 96 kWh/m² for multivariate forecasting and 94 kWh/m² for univariate forecasting. The cumulative measure is equal to 3.26 MWh/m². Like for the Bastia case, it seems that the sunny period increase the error while the winter period compensate this. It is certainly caused by the high value of the global radiation on summer.

In our two studies (Ajaccio and Bastia), there is overestimation on cloudy months and underestimation on sunny months. The difference of cumulative global radiation (Figure 8) decreases when the predicted days advance. During the first 20 months, the exogenous case underestimates often the global radiation. But during the four last months, the two cumulative predictions are equal. The two curves related to the multivariate estimation represented on the figure 8 can be fitted with a linear tendency (Equation 10 for Ajaccio and Equation 11 for Bastia).

$$Cumulative\_Error[Wh/m^2] = -146.8\ days\_number\ (R^2 = 0.94) \quad \text{Eq 10}$$

$$Cumulative\_Error[Wh/m^2] = -73.7\ days\_number\ (R^2 = 0.81) \quad \text{Eq 11}$$

We can estimate that the cumulated error is -146.8 Wh/m² (underestimation trend) by day in Ajaccio and -73.7 Wh/m² by day in Bastia. In order to better understand and quantify gains induced by the use of exogenous data, Figure 9 presents monthly results.

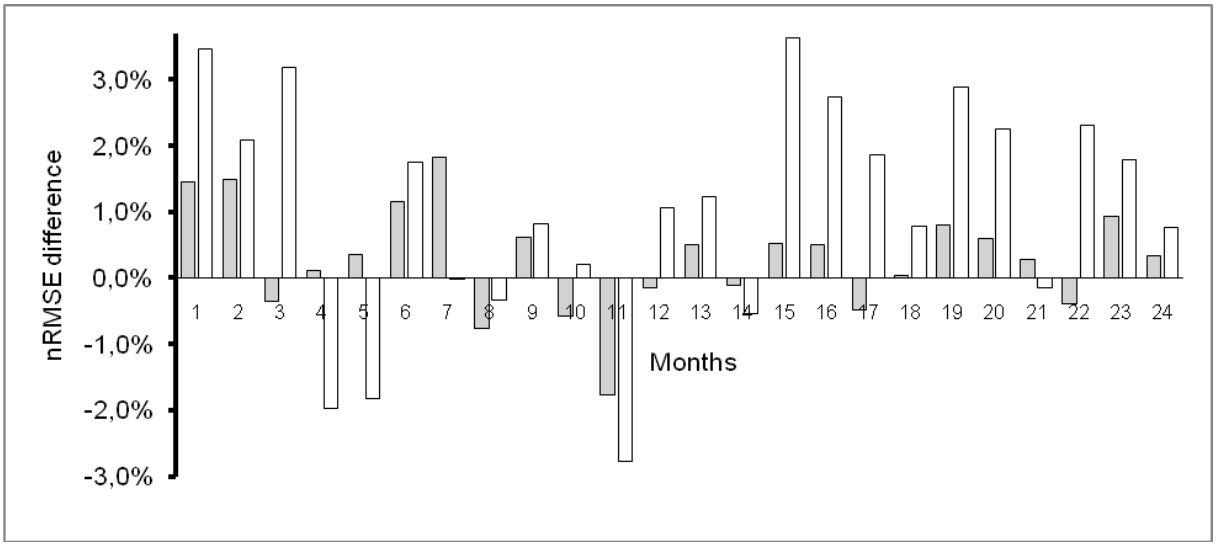


**Figure 7: Differences between nRMSE obtained with endogenous simulation and endo-exogenous simulation for Bastia (grey box) and Ajaccio (white box). A positive value means a diminution of the nRMSE derived from endo-exogenous simulation.**

For Ajaccio, The multivariate methodology improves the forecast, around 1% with a max in March (+3.5%) and a min in November (-2.8%). There are more months where the nRMSE difference is positive (17 months to 24 predicted months) meaning that the use of exogenous parameters is significant better than single use of endogenous inputs. For Bastia, globally, the exogenous methodology improves the quality of the forecast ; the average gain is around 0.5% with a max in July (nRMSE =+1.7%) and a min in November (-1.7%).

## 6.1 The frontage PV system

In the previous section, we have identified a methodology to predict with exogenous variables, the horizontal global radiation. In this new section, we apply the forecast methodology to predict the specific case of the DC production for a tilted PV wall.

**Figure 8: The frontage PV system on the Vignola laboratory walls.**

A frontage PV system has been installed recently in our laboratory (located at Vignola, Ajaccio, on Figure 10). It has a nominal power of 6.525 kW composed by respectively 1.8 kW and 4.725 kW amorphous and mono-crystal PV modules built in 6 independent power subsystems. PV power predictions from ANN methodology described in this paper have been computed from one of this whole PV plant on a frontage side exposed to the

south (azimuth null) and tilted at 80°. The PV system is composed by 9 SUNTECH 175S-24Ac for a 1.175 kW nominal power connected to a 1.85 kW SUNNY BOY SMA inverter for PV production on the grid. The irradiance sensor used 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. The measures are done every 5 minutes and stocked on a dedicated PC. For the PV power calculation ($E_{PV/DC}$), we use in first approximation, a linear production based on a constant PV plant efficiency $\eta_{PV} \sim 15\%$ ($R^2 = 0.997$), with:

$$E_{PV/DC}\,(Wh) = \eta_{PV} \,.I_{\beta}\,.S, \qquad \text{Eq 12}$$

where $I_{\beta}$ is the daily global irradiation on the PV system ($\beta = 80°$), S is the usable surface of the PV system under consideration ($S = 10.125$ m²). The prediction results for PV wall must be corrected (second approximation) to take into account the PR (performance ratio) of the plant which is very much affected by the choice of system component like inverter, temperature, energy loss, etc. and can provide appreciable difference in electricity production in long term. The PR is calculated during the period of measurement by the ratio between the measured efficiency and the theoretical efficiency, we obtain a minimum value equal to 0.71 in winter, a maximum value equal to 0.80 in summer and an annual mean value equal to 0.76 (>0.70 so PV system with a high efficiency). The parameter which allows to determinate the alternating current (AC) available on the output of the PV plant is given by the expression:

$$E_{PV/AC}\,(Wh) = PR.E_{PV/DC}\,(Wh) = PR.\eta_{PV.}\,I_{\beta}\,.S \qquad \text{Eq 13}$$

To predict this energy, we used as a learning set 10 years of global horizontal irradiation available on the site of Ajaccio. The distance of this frontage PV system from the meteorological station used to obtain the training series is about 10 km. Classical models are used to compute tilted irradiation for an 80° angle. We used the Climed-2 method [43] to determine the diffuse fraction then the classical transformation to tilt the beam component, and the Klutcher equations [44] to tilt the diffuse part. The couple of exogenous data (lag 1 for sunshine duration and nebulosity) and the ANN estimation designed previously have been used to predict the DC and AC energy produced by the PV plant (see Figure 11).

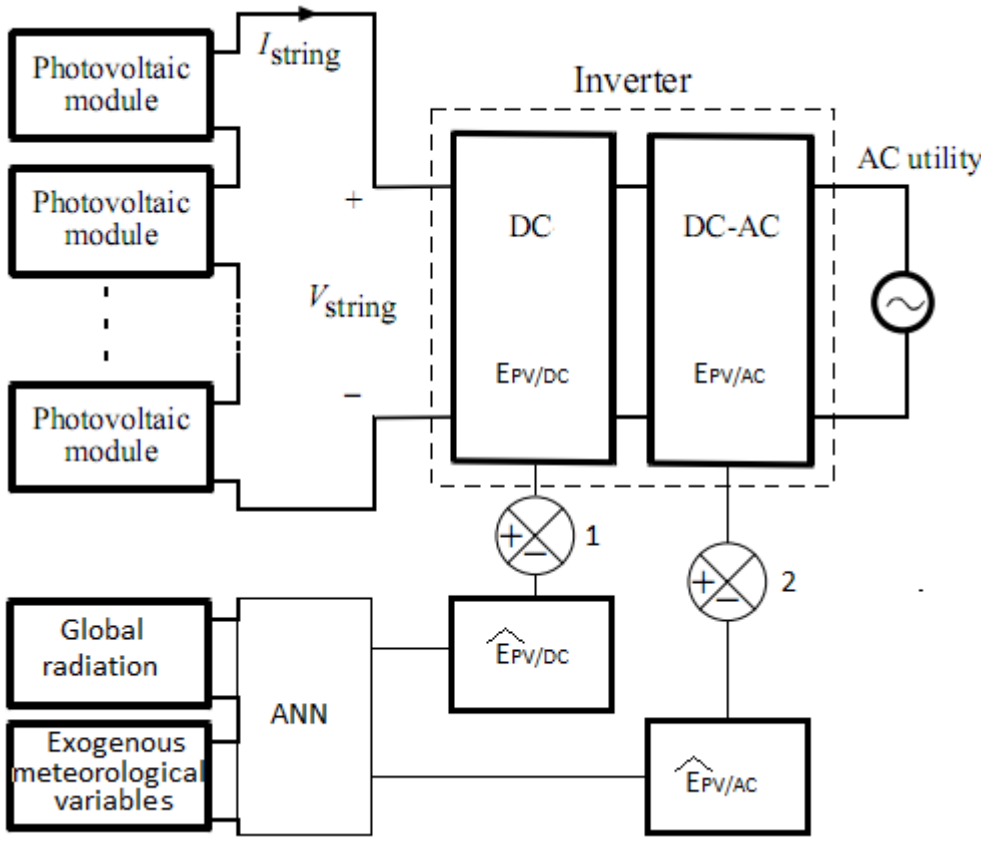



**Figure 11: Block Scheme of the PV wall experience. Comparator 1 and 2 are related to the DC and AC PV energy.**

We have chosen to compare the following solar radiation forecasting approach: the ANN methodology with only endogenous data, ANN with endogenous and exogenous data, persistence and the referenced autoregressive with moving average (ARMA(p,q)) model [45]. The optimization of this model is done separately with an exhaustive study of parameter $p$ and $q$. We have found that the best configuration was the simple ARMA(2,2) with Solis stationarization and centered values. The test of prediction is done with extreme condition, because the period chosen for the prediction is the 6 months between January and June 2009 where the weather was unusually very cloudy. These extreme conditions are illustrated in Table 5 that represents the means and standard deviation of the PV output electrical AC power. For the first months the standard deviation of the DC and AC electrical power measured are respectively 299 Wh and 227Wh, while the means is equal to 666 Wh and 506 Wh. It means a variation coefficient (or dispersion rate) closed to 50% showing the high variability of the PV plant production and the difficulty to offer a reliable forecast of the global horizontal irradiation. Table 6 shows

the results of the AC energy prediction tests for the four used methodologies (ANN with endogenous and exogenous inputs, ANN with only endogenous inputs, ARMA and persistence). Concerning the DC power, the results are equivalent for the purpose of the study which is to compare 3 methodologies of prediction and not to found the best model of the PV plant. The transition between the global radiation forecast and the PV energy forecast is the same for all the 3 methods. For the nRMSE the AC and DC results are identical, and for the other error parameters they must be multiply by 0.76 (=PR). Endogenous and exogenous data as inputs of the ANN allow to decrease the nRMSE by 1% on a 6-months cloudy period for the DC power production (January-June). Moreover, the use of exogenous data shows an interest only in cloudy period (winter season). In summer, endogenous data as inputs on a preprocessed ANN seems to be sufficient. By comparison to a naïve forecaster like persistence or the referenced forecaster (ARMA), an ANN with endogenous and exogenous data improves the DC (and AC) electrical power energy prediction by respectively 9% and 1%. The ARMA process is equivalent to an ANN with only endogenous inputs, both of which are less relevant that the use of ANN with selected meteorological inputs.

Figure 12a shows the cumulative prediction versus the combination of the electric power measurement. The prediction overestimates the energy production. But, we can see that the use of ANN with exogenous data can properly quantify the energy resource (like in the ARMA case the determination coefficient for the cumulative prediction is $R^2$=0.99 and for the persistence $R^2$=1).

| | Jan-June | Jan-feb | March-April | May-June |
|---|---|---|---|---|
| Means (Wh) | *464.2* | *506.1* | *480.3* | *412.6* |
| Std Dev (Wh) | *178.9* | *227.2* | *202.7* | *58.17* |

**Table 5: Means and standard deviation of the AC PV plant electrical power.**

| | predictors | Jan-June | Jan-feb | March-April | May-June |
|---|---|---|---|---|---|
| nRMSE (%) | *ANN endo-exo* | ***33.1*** | ***37.5*** | ***36.1*** | *16.4* |
| | *ANN endo* | *34.2* | *37.9* | *37.9* | *16.4* |
| | *ARMA* | *34.3* | *38.4* | *37.7* | ***16.3*** |
| | *Persistance* | *42.3* | *47.0* | *47.1* | *18.8* |
| RMSE (Wh) | *ANN endo-exo* | ***164.43*** | ***207.22*** | ***188.25*** | *68.48* |
| | *ANN endo* | *169.97* | *209.81* | *197.66* | *68.32* |

| | | | | | |
|---|---|---|---|---|---|
| | *ARMA* | *170.48* | *212.89* | *197.02* | ***68.16*** |
| | *Persistance* | *210.21* | *260.19* | *245.63* | *78.14* |
| MBE (Wh) | *ANN endo-exo* | *7.31* | *27.66* | *9.89* | *-10.21* |
| | *ANN endo* | *-7.02* | *4.35* | *-3.56* | *-19.55* |
| | *ARMA* | *-2.30* | *-1.90* | ***0,84*** | *-24.75* |
| | *Persistance* | ***1.34*** | ***-1.35*** | *1.91* | ***2.40*** |
| MAE (Wh) | *ANN endo-exo* | ***122.43*** | ***177.72*** | ***148.50*** | ***122.43*** |
| | *ANN endo* | *130.77* | *184.04* | *162.29* | *130.77* |
| | *ARMA* | *131.16* | *189.18* | *158.46* | *131.16* |
| | *Persistance* | *140.19* | *195.13* | *179.05* | *140.19* |

**Table 5: AC PV power prediction (80° tilted PV plant). ANN with endogenous and exogenous inputs, ANN with only endogenous data, ARMA process and persistence. In bold the best results for each measurement error.**

This cumulative plot is interesting because it shows the global error generated. A good estimator must have a low value of daily nRMSE and also a low error on the predicted cumulative energy. For 6 months, the absolute global error of the cumulative prediction is less than 4 kWh for the tilted PV wall, while the value of cumulative produced measures is 60 kWh (error ~ 7%). On the Figure 12b, we can see that the measures are very noisy and that the inclination of 80° is very penalizing; the high values of energy are observed in winter when the weather is often cloudy. Moreover, the relative error is much larger in summer in this case than in horizontal case. However, after 90 days (3 months, so early April) the prediction is very consistent to the measures.

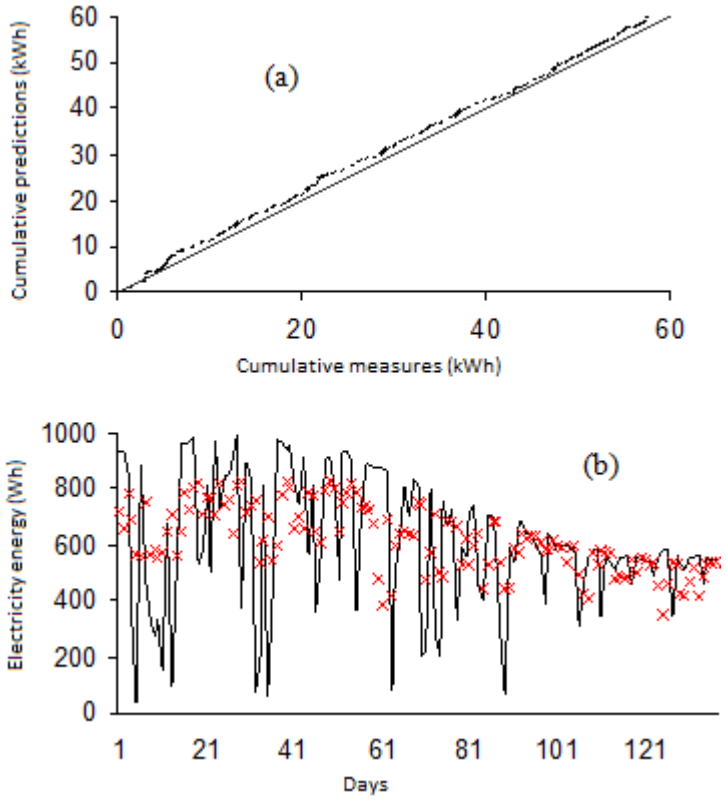



**Figure 9: On (a), dashed points represent the relation between cumulative measure and forecast of the PV electricity, with ANN and endo-exogenous input, ($R^2$=0.99) on the period January-June 2009. The continuous line represents the case of perfect predictor (Y=X). On (b) the electric power measured (line) and the electric power forecasted (marks) are represented for Ajaccio on 80° tilted PV wall.**

## 7 Conclusions

This paper proposes to study the contribution of exogenous meteorological data to an optimized MLP in order to predict solar energy (multivariate forecasting). We have compared this technique with different forecasting methods on two sites located in Corsica Island and the results seem to be quite interesting. On Bastia, the first site studied, the use of the exogenous data on ANN inputs increases a little the prediction quality (only 0.5%), the maximum is in July (nRMSE =+1.7%) and the minimum in November (-1.7%). In Ajaccio, the second site studied, the multivariate forecasting improves the nRMSE by 1%, and this result begins to be interesting for a power manager. The maximum gain is in March (+3.5%) and the minimum in November (-2.8%). If we consider

the cumulative prediction, the results are of course better, but the two predictors appear equivalent in this case ($R^2$=0.99 for both). The results are similar to the concrete case of a tilted PV wall (1.175 kWp): endogenous and exogenous data ANN inputs allow decreasing the nRMSE by 1% on a 6 months-cloudy period for the AC power production (January-June). Moreover, the use of exogenous data shows an interest only in cloudy period (winter season). In summer, endogenous data as inputs on a preprocessed ANN appear to be sufficient. By comparison to a naïve and reference forecaster as respectively persistence and ARMA, an ANN with endogenous and exogenous data improves the AC electrical power energy prediction by respectively 9% and 1%. The ARMA process is equivalent to the "only endogenous input" architecture (univariate method): both are less relevant than the use of ANN with meteorological inputs. All theses results encourage us to study in the future how to adapt this methodology to shorter horizons in order to target the problem of scheduling in a power system. It seems obvious that in this case the meteorological data should have a greater impact on the prediction accuracy.

## References


[1] Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science 2008; 1-1:52-76

[2] Mubiru J. Predicting total solar irradiation values using artificial neural networks. Renewable Energy 2008; 33-10:2329-2332

[3] Mubiru J, Banda E. Estimation of monthly average daily global solar irradiation using artificial neural networks. Solar Energy 2008; 82-2: 181-187

[4] Kalogirou SA. Artificial neural networks in renewable energy systems applications: a review. Renewable and sustainable energy reviews 2001;5: 373-401

[5] Hocaoglu FO, Gerek ON, Kurban M. Hourly solar forecasting using optimal coefficient 2-D linear filter and feed-forward neural networks. Solar energy 2008; 82-8:714-726

[6] Zarzalejo LF, Ramirez L, Polo J. Artificial intelligence techniques applied to hourly global irradiance estimation from satellite-derived cloud index. Energy. 2005 Jul;30(9):1685-1697.

[7] Jain K, Jianchang M, Mohiuddin KM. Artificial neural networks: A tutorial, IEEE Computer 1996;29-3: 31–44

[8] Hu Y, Hwang J. Handbook of neural network signal processing. ISBN 0-8493-2359-2; 2002

[9] Crone SF. Stepwise Selection of Artificial Neural Networks Models for Time Series Prediction Journal of Intelligent Systems, Department of Management Science Lancaster University Management School Lancaster, United Kingdom; 2005

[10] Jiang Y. Computation of monthly mean daily global solar radiation in China using artificial neural networks and comparison with other empirical models. Energy. 2009 Sep;34(9):1276-1283.

[11] Benghanem M, Mellit A. Radial Basis Function Network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy. 2010 Sep;35(9):3751-3762.

[12] Faraway J, Chatfield C. Times series forecasting with neural networks: a case study, Research report 95-06 of the statistics group, University of Bath; 1995

[13] Hamilton JD. Times series analysis. ISBN 0-691-04289-6 ; 1994

[14] Bourbonnais R, Terraza M. Analyse des séries temporelles. ISBN 9782100517077, 318p., Dunod Ed., Paris; 2008

[15] Celeux G, Nakache JP. Analyse discriminante sur variables qualitatives ISBN 2840540274, Polytechnica, 270 p., Paris ; 1994

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

[17] Muselli M, Poggi P, Notton G, Louche A. First Order Markov Chain Model for Generating Synthetic 'Typical Days' Series of Global Irradiation in Order to Design PV Stand Alone Systems. Energy Conversion and Management 2001;42-6 :675-687

[18] Logofet DO, Lesnaya EV. The mathematics of Markov models: what Markov chains can really predict in forest successions. Ecological Modelling 2000;126: 285-298

[19] Sharif M, Burn DH. Simulating climate change scenarios using an improved K-nearest neighbor model. Journal of Hydrology 2006; 325 1-4,179-196

[20] Yakowitz, S. Nearest neighbors method for time series analysis. Journal of Time Series Analysis 1987;8: 235-247.

[21] Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews 2009;13-2:406-419

[22] Paoli C, Voyant C, Muselli M, Nivet ML. Solar radiation forecasting using ad-hoc time series preprocessing and neural networks," Emerging Intelligent Computing technology and Applications 2009:898-907

[23] Voyant C, Muselli M, Paoli C, Nivet ML, Poggi P and Haurant P. Predictability of PV power grid performance on insular sites without weather stations: use of artificial neural networks," EU PVSEC proceeding 24th European Photovoltaic Solar Energy Conference, Hambourg (Germany), DOI : 10.4229/24thEUPVSEC2009-5BV.2.35, 4141-4144

[24] Liu BHY, Jordan RC. Daily sunshine duration on surfaces tilted towards the equator. Trans SHRAE 1962; 67:526–541

[25] Badescu V. Modelling Solar radiation at the earth surface, recent advances. ISBN: 978-3-540-77454-9, Viorel Ed.;2008

[26] Reindl DT, Beckman WA, Duffie JA. Evaluation of hourly tilted surface radiation models. Solar Energy 1990;45-1:9-17

[27] Perez R, Ineichen P, Seals R. Modeling daylight availability and irradiance components from direct and global irradiance. Solar Energy 1990;44-5:271–289

[28] Elminir HK, Azzam YA, Younes FI. Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. Energy. 2007 Aout;32(8):1513-1523.

[29] Hay JE, Davies JA. Calculation of the solar radiation incident on an inclined surface. In: Proc. First Canadian Solar radiation workshop 1980:59–72

[30] Ineichen P, Guisan O, Perez R. Ground-reflected radiation and albedo. Solar Energy 1990;44-4:207–214

[31] Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series, European Journal of Operational Research 2005;160:501-514

[32] Bird Richard E, Hulstrom L. A Simplified Clear Sky Model for Direct and Diffuse Sunshine duration on Horizontal Surfaces, SERI/TR-642-761, Solar Energy Research Institute, Golden, USA, Colorado; 1981

[33] Cao J, Cao S. Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis. Energy. 2006 Dec;31(15):3435-3445.

[34] Ineichen P. A broadband simplified version of the Solis clear sky model. Solar Energy 2008; 82-8:758-762

[35] Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M, Cros S, Dumortier D, Kuhlemann R, Olseth JA, Piernavieja G, Reise C, Wald L, et Heinemann D. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sensing of Environment 2004;91:160-174

[36] Paoli C, Voyant C, Muselli M, Nivet ML. Forecasting of preprocessed daily solar radiation time series using neural network. Solar Energy, In Press

[37] Costa MA, Braga ADP, Menezes BRD. Improving generalization of MLPs with sliding mode control and the Levenberg-Marquardt algorithm. Neurocomputing. 2007 Mar;70(7-9):1342-1347.

[38] Sfetsos A, Coonick AH. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy. 2000;68-2:169-178

[39] Muttil N, Chau K. Machine-learning paradigms for selecting ecologically significant input variables. Engineering Applications of Artificial Intelligence 2007;20-6:735-744

[40] T Box GE.P, Jenkins GM. Time Series Analysis: Forecasting and Control (revised edition), Holden Day, San Francisco;1976

[41] Ahlburg DA. Error measures and the choice of a forecast method. International Journal of Forecasting. 1992;8-1:99-100

[42] Toukourou MS, Johannet A, Dreyfus G. Flash Flood Forecasting by Statistical Learning in the Absence of Rainfall Forecast: A Case Study. Engineering Applications of Neural Networks 2009;43:98–107

[43] Notton G, Poggi P and Cristofari C. Predicting hourly solar irradiations on inclined surfaces based on the horizontal measurements: Performances of the association of well-known mathematical models. Energy Conversion and Management 2006;47:1816-1829

[44] Noorian AM, Moradi I, Kamali GA. Evaluation of 12 models to estimate hourly diffuse irradiation on inclined surfaces. Renewable Energy 2008;33-6:1406-1412

[45] Dubois E, Michaux E. "Grocer: an econometric toolbox for Scilab", 2008, available at http://dubois.ensae.net/grocer.html

## Acknowledgements

This work was partly supported by the Territorial Collectivity of Corsica. We thank the French National Meteorological Organization (Météo-France) for providing the data used in this study.

**List of captions:**

Figure 10: Example of an MLP (left) and details of an neuron from the hidden layer (right)

Figure 2: (a) Daily measured horizontal irradiation (Ajaccio, 1998) ; (b) Impact of the stationarization on the original time series.

Figure 3: Detail of the overall real system prediction.

Figure 4: Partial autocorrelation of the clear sky index (Bastia case on left and Ajaccio case on right). The lines around zero represent the upper and lower bounds of the PACF 95% confidence interval.Figure 5: Pearson cross-correlation between the clear sky index, and exogenous variables for Bastia and Ajaccio stations (wind direction Wd, peak of wind speed PKW, wind speed Ws, relative humidity RH, sunshine duration Su, precipitation RP, pressure P, differential pressure DGP, ambient temperature average Ta, night temperature Tn , max (TM) and min (Tn) temperature and nebulosity N). The lines represent the confidence band chosen.

Figure 6: Comparison between the use of exogenous and endogenous inputs (on right) and only endogenous input (on left). The top is related to Bastia and the bottom to Ajaccio.

Figure7: Comparison of cumulative global radiation between the use of exogenous input and only endogenous input. The dotted line is the prediction and the continuous line is the measure on Bastia and Ajaccio ($R^2$=0.99 for all figures).

Figure 8: Differences between measures and ANN simulations in Ajaccio and Bastia (continuous line is ANN with only endogenous data, dotted line the endogenous and exogenous data). A negative value means that the simulation underestimates the global radiation.

Figure 9: Differences between nRMSE obtained with endogenous simulation and endo-exogenous simulation for Bastia (grey box) and Ajaccio (white box). A positive value means a diminution of the nRMSE derived from endo-exogenous simulation.

Figure 10: The frontage PV system on the Vignola laboratory walls.

Figure 11: Block Scheme of the PV wall experience. Comparator 1 and 2 are related to the DC and AC PV energy.

Figure 12: On (a), dashed points represent the relation between cumulative measure and forecast of the PV electricity, with ANN and endo-exogenous input, ($R^2$=0.99) on the period January-June 2009. The continuous line represents the case of perfect predictor (Y=X). On (b) the electrics power measured (line) and the electric power forecasted (marks) are represented for Ajaccio on 80° tilted PV wall.

Table 1: Prediction error with only one exogenous input among those highly correlated for the site of Bastia. Bold letters represents the best configurations.

Table 2: Prediction error with only one exogenous input among those highly correlated for the site of Ajaccio.

Table 3: Optimization of the hidden layer for the site of Bastia.

Table 4: Optimization of the hidden layer for the site of Ajaccio.

Table 5: Means and standard deviation of the AC PV plant electrical power

Table 6: AC PV power prediction (80° tilted PV plant). ANN with endogenous and exogenous inputs, ANN with only endogenous data, ARMA process and persistence. In bold the best results for each measurement error.