Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation

Cyril Voyant\textsuperscript{a,b,*}, Marc Muselli\textsuperscript{a}, Christophe Paoli\textsuperscript{a}, Marie-Laure Nivet\textsuperscript{a}

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

\begin{abstract}
We propose in this paper an original technique to predict global radiation using a hybrid ARMA/ANN model and data issued from a numerical weather prediction model (NWP). We particularly look at the multi-layer perceptron (MLP). After optimizing our architecture with NWP and endogenous data previously made stationary and using an innovative pre-input layer selection method, we combined it to an ARMA model from a rule based on the analysis of hourly data series. This model has been used to forecast the hourly global radiation for five places in Mediterranean area. Our technique outperforms classical models for all the places. The nRMSE for our hybrid model MLP/ARMA is 14.9% compared to 26.2% for the naïve persistence predictor. Note that in the standalone ANN case the nRMSE is 18.4%. Finally, in order to discuss the reliability of the forecaster outputs, a complementary study concerning the confidence interval of each prediction is proposed.
\end{abstract}

1. Introduction

Solar radiation is one of the principal energy sources, occupying a very important role in some engineering applications as production of electricity, heat and cold [1–4]. The process of converting sunlight to electricity without combustion creates power without pollution. It is certainly one of the most interesting themes in solar energy area. To use ideally this technology, it is necessary to understand and create efficient prediction models like done for example in Mueller et al., 2004, Mellit et al., 2005 and Mubiru et al., 2008 [5–7]. Insolation is defined as the solar radiation striking a surface at a certain time and place and is typically expressed in kilowatt hours per meter square (kWh/m\textsuperscript{2}) [8,9]. Many factors determine how much sunlight is available at a given location. We can mention the atmospheric conditions, the Earth's position in relation to the sun, and the site obstructions [4,9]. Atmospheric conditions that can also affect the amount of radiation received on the Earth's surface are the quantity of air molecules, water vapor, dust, ozone and carbon dioxide, the cloud cover, the air pollution, the dust storms, the volcanic eruptions, etc [5,10]. There is an interest to control the solar radiation prediction, as for example to identify the most optimal locations for developing solar power project or to maintain the grid stability in solar and conventional power management. Note that, in Europe, the white paper on Energy (established in 1997) set a target (not yet achieved) of 12% of electricity production from renewable energies by 2010. The new European guidelines set a new target to 20% by 2020. The issues of the solar energy prediction are very important and mobilize a lot of research teams around the world and particularly in the Mediterranean area [3,7,11–13]. In practice, the global radiation (or insolation) forecasting is the name given to the process used to predict the amount of solar energy available in the current and near terms. A lot of methods have been developed by experts around the world [6,14,15]. Often the times series (TS) mathematical formalism is necessary. It is described by sets of numbers that measures the status of some activity over time [16]. In primary studies [1,13,17] we have demonstrated that an optimized multi-layer perceptron (MLP) with endogenous input made stationary and exogenous inputs (meteorological data) can forecast the global solar radiation time series with acceptable errors. This prediction model has been compared to other prediction methods (ARMA, k-NN, Markov Chains, etc.) and we have concluded that MLP and ARMA were similar. Following these studies and in order to see if we can significantly improve our results, we decided to add weather forecast (instead of exogenous data previously used) as new inputs of our model. We assumed that numerical weather prediction (NWP) simulations tools compute data patterns essential for determining solar radiation. These
weather forecast data present two advantages: firstly they are becoming more and more available through the Internet and secondly the models provide spatially distributed data which are very relevant to the regional scale studies. Thanks to an agreement with Météo-France, which is the French meteorological organization (http://france.meteofrance.com), we had the opportunity to freely access to some of the forecasts of the French operational limited area model ALADIN (acronym of Aire Limitée et Adaptation Dynamique) [18–20]. ALADIN (http://www.cnrm.meteo.fr/aladin) is a hydrostatic model developed by Météo-France in collaboration with the European Centre for Medium Range Weather Forecasts (ECMWF). It is the result of a project launched in 1990 by Météo-France with the aim of developing a limited area model and today fifteen countries are participating in the common work. We propose in this paper an original technique to predict hourly global radiation time series using meteorological forecasts from the ALADIN NWP model. After optimizing our MLP with ALADIN forecast data and endogenous data previously made stationary with an ad-hoc method, we combine it to an auto-regressive and moving average (ARMA) model from rules based on the analysis of hourly data series. Finally we present all forecasting results with confidence intervals in order to give more complete information to a final user, such as a power manager.

The paper is organized as follows. Section 2 describes the data and forecast data computed by the ALADIN numerical weather model. After recalling the principles of time series forecasting and the need to make stationary a time series, we present in Section 3 the forecasting models (ARMA and ANN) that have allowed us to build our original hybrid method, and the variables selection approach used. Section 4 includes final results and experiences conducted during this study and showing that forecasting results can significantly be improved by selecting ANN or ARMA models according to their performances. The reliability of the predictions is also considered by computing the confidence intervals. Section 5 concludes and suggests perspectives.

2. Radiation time series and numerical weather prediction forecast data

In this study we have used two types of data: radiation time series and meteorological forecasts from the ALADIN NWP model. In order to verify the robustness of our approach we chose to apply our methodology on five distinct stations located in Mediterranean coastal area. Fig. 1 shows the location in Mediterranean area of the five weather stations studied.

Concerning endogenous data, the radiation time series (Wh.m$^{-2}$) are measured at coastal meteorological stations maintained by the French meteorological organization: Météo-France. We selected Ajaccio (41°5’N and 8°5’E, seaside, 4 m asl), Bastia (42°3’N, 9°3’E, 10 m asl), Montpellier (43.6’N and 3.9’E, 2 m asl), Marseille (43.4’N and 5.2’E, 5 m asl) and Nice (43.6’N and 7.2’E, 2 m asl). These stations are equipped with pyranometers (CM 11 from Kipp & Zonen) and standard meteorological sensors (pressure, nebulosity, etc.). The choice of these particular places is explained by their closed geographical and orographical configurations. All stations are located near the Mediterranean Sea with mountains nearby. This specific geographical configuration makes nebulosity difficult to forecast. Mediterranean climate is characterized by hot summers with abundant sunshine and mild, dry and clear winters. The data representing the global horizontal solar radiation were measured on an hourly basis from October 2002 to December 2008 (more than 6 years). The first four years have been used to setup our models and the last two years to test them. Note that in the presented study, only the hours between 8:00AM and 04:00PM (true solar time) are considered. The others hours are not interesting from energetically point of view and their predictions are complicated because it is very difficult to make stationary the measures of sunrise and sunset (stationarity scheme detailed in the next section). A first treatment allows us to clean the series of non-typical points related to sensor maintenances or absence of

![Fig. 1. The five studied stations marked in the Mediterranean sea: Ajaccio, Bastia, Montpellier, Marseille and Nice.](image)
measurement. Less than 4% of measurements were missing and replaced by the hourly average for the given hour.

The second types of data we decided to use are the meteorological forecasts from the ALADIN NWP model. Météo-France proposed us a free access to some of the forecasts issued from their numerical weather prediction model called ALADIN-France. This model is a bi-spectral limited area, based on the assimilation of daily measurements, and driven using, for boundary data, the outputs of the ARPEGE (acronym of “Action de Recherche Petite Echelle Grande Echelle”) global model provided also by French meteorological services. The model evolves in average every six months because the ALADIN code follows the ARPEGE one in its permanent evolution and we are actually on the 37th cycle. For a better description of the model and its parameterization, the interested reader may refer for e.g. to refs. [18–20]. The French NWP system is organized around the production of analyses at 00, 06, 12 and 18UTC, and the range of the forecast is 54 h. The actual horizontal resolution of ALADIN-France is approximately 9.5 Km, with 60 levels vertically. The ALADIN model has more than twenty outputs available at a temporal resolution of 1 h; note that the global radiation is recently available. These values are computed in all points of the computing grid with mesh size of 9.5 Km. Considering these facts we had three major choices to do. First we had to select the forecast parameters to add as an input of our model. In a second time we had to choose the grid points for our five locations, and finally we add to select the analyses (between 00, 06, 12 and 18UTC) and ranges (1–54 h) of the forecasts to take into account. For the ALADIN output, we based our choice on preceding works [17] in which we analyzed the benefit of taken into account exogenous variables. In these studies we made some computations about the correlation between the global radiation and a lot of exogenous meteorological parameters. Among the 23 ALADIN possible outputs we chose those which seem to have a straight link to solar radiation [17]. These data are pressure (P, Pa), nebulosity (N, Octas), rain precipitations (RP, mm) and temperature (T, °C). The second choice was on the selection of the grid points. We decided to simply use a proximity criterion and we chose the points of the ALADIN computational grid nearer to our five selected stations. We decided to take into account the analyses closer to the sunrise: we select the 06AM analyze with a 12 h head forecast horizon. This couple of analyze and range allows us to cover, with a unique ALADIN extraction per day, the central hours we consider, that is to say 8:00AM to 4:00PM. However, we are aware that some of our choices can be discussed: first the periodic changes in the ALADIN model (more than ten cycles) may affect the training capability of the ANN, secondly an explicit help of a professional forecaster may comfort us in the choice of NWP model parameters, we will return to these restrictions in our conclusion.

3. Forecasting models

This section details the prerequisites that have enabled us to build our prediction model. After recalling the principles of time series forecasting and the need to make stationary a time series and how to validate it, we present the ARMA and ANN models. To conclude this section we present our final prediction model as a combination of ARMA and ANN models.

3.1. Time series analysis, need of stationarity and validation

There are different approaches to model time series (TS) [16]. We studied ARMA [15,21] and ANN [22,23], considering they are the most effective according to our previous works. However the common base of all these models seems to be that a TS $x_t$ can be defined by a linear or non-linear model called $f_n$ (see Equation (1))

$$x_{t+1} = f_n(x_t, x_{t-1}, \ldots, x_{t-p+1}) + \varepsilon_{t+1}$$

To estimate the $f_n$ model, a stationarity hypothesis is often necessary. This condition usually implies a stable process [23–26]. This notion is directly linked to the fact that whether certain feature such as mean or variance change over time or remain constant. In fact, the time series is called weak-sense or weakly stationary if the first two moments stay also correct for neural network studies [29]. In fact, all artificial networks are considered like functions approximation tools on a compact subset of $\mathbb{R}$. Moreover standard MLP (with at least 1 hidden layer) are asymptotically stationary, it converges to its stationary distribution, (i.e. lim$_{h \to 0} = f$). Moreover, this kind of network can approximate any continuous and multivariate function. They cannot show “explosive” behavior or growing variance with time [21,30]. In practice a varying process may be considered to be close to stationary if it varies slowly and it is the modeling condition to use the MLP. Note that, the network can be trained to mimic a non-stationary process on a finite time interval. But the out-of-sample or prediction performance will be poor. Indeed, the network inherently cannot capture some important features of the process. Without pre-process, ANN and ARMA can be unappealing for many of the non-stationary problems encountered in practice [31]. One way to overcome this problem is to transform a non-stationary series into a stationary (weakly or stronger if possible) one and then model the remainder by a stationary process.

In our case, we have developed a sophisticated method to make the global radiation stationary ($X_t$, modeling of cycles). The original series has two periodicities (angular frequency $\omega_1$ and $\omega_2$) very difficult to overcome (Equation (4)) because the amplitude $a(t)$, $b(t)$ and $R_t$ the cloud occurrence, have not simple expressions. Without the last term, this equation represents a clear sky estimation.

$$X_t = (a(t)\cos(\omega_1 t) + b(t)\cos(\omega_2 t)) \cdot R_t$$

The first periodicity is a classic yearly seasonality which can be erased with a ratio to trend: multiplicative scheme induced by the nature of the global radiation series in Equation (5). In previous studies [13], we have demonstrated that the clear sky index obtained with Solis model [5,32] is the more reliable for our locations. As the second daily seasonality is often not completely erased after this operation, then we use a method of seasonal correction (corrected for seasonal variance) based on the moving average [27,28].
The chosen method is essentially interesting for the case of a deterministic nature of the series seasonality (true for the global radiation series) but not for the stochastic seasonality [33]. The steps we follow to make the series stationary are as following:

1. Use of Solis model to establish the clear sky model of the considered location:

\[ H_{gh}(t) = H_0(t) \cdot \exp \left( \frac{-t}{\sin^2(h(t))} \right) \cdot \sin(h(t)) \]  (5)

2. Calculate the ratio to trend to overcome the periodicity, the result is the Clear Sky Index CSI

\[ CSI_t = X_t / H_{gh}(t) \]  (6)

3. Calculate the moving average (MM(t)) considering that 2.\Delta T + 1 is equal to the periodicity of the series. In the case of a 9 h periodicity, \( \Delta T \) corresponds to 4 h

\[ MM_t = \langle CSI_t \rangle_{t-[2-\Delta T, t+\Delta T]} \]  (7)

4. Operate a new ratio to trend with the moving average. The new coefficient are called the periodic coefficients \( C(t) \)

\[ C_t = CSI_t / MM_t \]  (8)

5. Compute of the average over one year \( E[C_t] \) (done on 365 \times 9 = 3285 values because only 9 h per day have been considered)

6. Compute of the new stationary series CSI*\(_{t}(t)\)

\[ CSI_t^* = CSI_t / E[C_t] \pmod{3285 \text{ hours}} \]  (9)

The Fig. 2 shows the stationarity methodology concerning the global radiation time series prediction.

To validate the stationary processing, we have chosen two tests. The first is the variation coefficient of the series \( (VC_X = \sqrt{\text{Var}(X) / \text{E}(X)^2}) \). Although this tool is easy to use, it is only dedicated to the cross comparison. It is not an absolute criterion allowing to distinguish the stationary and non-stationary process. To overcome this problem, there are a lot of available non-stationary tests: variance test of Fisher, unit root, Dickey–Fuller, KPSS, just to name a few. For more details, the reader can refer to Hamilton [29], Pollock [31], and Bourbounais [27]. The quality of each method seems equivalent and some tests are not adapted to the seasonal case but to the extra-seasonality trends or difference stationary. We have chosen a very classic statistic test based on variance analysis (variance ratio to period by residue): the Fisher test detailed by Bourbounais in [27]. This test involves a TS without trend as it is the case for the insolation. As a matter of fact its annual average value is relatively constant at long term. In case of the periodic effect is significant, we can consider that the TS has a seasonal component. This test is proposed for the seasonality of the TS: daily and yearly phenomena.

Before to perform this test it is necessary to construct a new scheme of the series, so CSI*\(_{ij}(t) = m_{ij} + e_{ij}, \) with \( m_{ij} = a_{ij} + b_{ij} (a \text{ and } b \text{ are respectively related to the daily and yearly seasonality}), \) where \( e_{ij} \) is the associated residue described be a normal law \( N(0, \sigma^2) \).

The test is constructed from the null hypothesis \( H_0 \) meaning no periodic influence in the TS, and the alternative hypothesis \( H_1 \) meaning TS with periodic influence. While details can be found in [27], principle of the Fisher test is described by the following equations:

1. Compute of the “empirical” Fisher statistic \( F_c = V_p / V_b \) (\( V_p \) the variance on period in Equation (10) and \( V_b \) the residue variance in Equation (11)), \( p \) is the number of measures per period and \( N \) the number of periods.

\[ V_p = \frac{1}{p-1} \sum_{j=1}^{p} N \cdot \left( \langle CSI^*_t \rangle_i - \langle CSI^*_t \rangle_j \right)^2 \]  (10)

\[ V_b = \frac{1}{(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{p} \left( \langle CSI^*_t \rangle_i - \langle CSI^*_t \rangle_j \right)^2 \]  (11)

\[ \langle CSI^*_t \rangle_i = \left( \frac{1}{p} \right) \sum_{j=1}^{p} CSI^*_t \]  (12)

For the a time series (daily effect) the parameter \( N = 6 \) and \( p = 3285 \), for the b series (yearly effect), \( N = 6 \times 365 = 2190 \) and \( p = 9 \) (not 24 h because only the hour between 8:00AM and 04:00PM are considered).

2. Determination of the critical Fisher value for the degree of freedom \( v_1 = (p-1) \) et \( v_2 = (N-1)(p-1) \) and the \( \alpha = 0.05 \), so the value corresponding to:

\[ F_{v_1;v_2}^{0.05} = F_{\text{limit}} \]  (13)

3. If \( F_c > F_{\text{limit}} \) the \( H_0 \) hypothesis is rejected and \( H_1 \) is accepted. If the \( F_c \) calculated from the data is greater than the critical value of \( F \) distribution for desired false-rejection probability \( \alpha = 0.05 \), the TS is described as seasonal. Furthermore, we can estimate than, more the \( F_c \) coefficient is important, and more the seasonality component is important.

In section dedicated to the experiments we will present the results obtained for the variation coefficient and the Fisher test on our CSI stationary processing with and without periodic coefficients for the five places studied.

3.2. ARMA

The ARMA model is certainly the most used with the prediction problems [27,29,34]. The ARIMA techniques are especially 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 ARMA \((p, q)\) and is defined with \( p \) and \( q \) parameters (Equation (14)).

\[ \left( 1 - \sum_{t=1}^{p} \phi_t \cdot L^t \right) \cdot x_t = \left( 1 + \sum_{t=1}^{q} \theta_t \cdot L^t \right) \cdot e(t) \]  (14)

Where, \( x_t \) is a time series, \( \phi \) and \( \theta \) are the parameters of the autoregressive and moving average part, \( L \) is the lags operator and \( e \) is the error term. Please cite this article in press as: Voyant C, et al., Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, Energy (2012), doi:10.1016/j.energy.2012.01.006
is an error term distributed as a Gaussian white noise. The optimization of these parameters must be made depending on the type of the series studied. In the presented study, we chose to use Matlab© software and the Yule–Walker fitting method [29]. The criterion adopted to consider when an ARMA model ‘fits’ to the global radiation time series is the normalized root mean square error described by the Equation (15) [35].

\[ nRMSE = \sqrt{\frac{E[(X - \hat{X})^2]}{E[X^2]}} \]  

The prediction error is generated by the prediction of two years of radiation not used during the ARMA parameters calculation step. Several experiments are needed to obtain the best model. Residual auto-correlogram tests have been computed to verify that the error term is a white noise. Before using this method of forecasting, the global radiation time series is made stationary with the method described in the Section 3.1 (clear sky index with seasonal adjustment) and then, centered and reduced. The models after optimization are very simple: ARMA(1,0) for Ajaccio, Bastia, Montpellier and Nice, and ARMA(2,0) for Marseille. In the case of the ARMA(2,0) the prediction can be expressed by the Equation (16).

\[ \hat{X}(t + 1) = \sum_{i = 0}^{1} \phi_i X(t - i) + \epsilon(t + 1) \]  

3.3. Neural network and time series forecasting

Although a large range of different architecture of ANN is available [36,37], MLP remains the most popular estimator in times series forecasting [3,12,22,24,30,33,34,36,38–41,43,45] and especially in global radiation prediction [6,7,11,13]. In particular, feed-forward MLP networks with two layers (one hidden layer and one output layer) are often used for modeling and forecasting time series. Several studies [22,33,40,41] have validated this approach based on ANN for the non-linear modeling of time series against other class of ANN such as radial basis or recurrent networks. According these readings, radial basis network and Elman network are not really appropriate for hourly forecasting of global radiation. In addition Dreyfus et al. explain in the book: Apprentissage Statistique (Eyrolles 2008, 450p), that the radial basis network have local non-linearity, their influence zone is limited in the space which is not the case of MLP with sigmoid activation function. Concerning the recurrent network, they are used for dynamic models and variables linked by differential equations or discrete-time by difference equations and it is not really the case of the global radiation. For all these reasons, we have considered that the MLP is the neural network the most relevant concerning the hourly global radiation series.

To forecast the time series, a fixed number \( p \) of past values are set as inputs of the MLP, the output is the prediction of the future value [42,43]. Considering the initial time series equation (Equation
(1), we can transform this formula to the non-linear case of one hidden layer MLP with \( b \) related to the biases, \( f \) and \( g \) to the activation function of the output and hidden layer, and \( \omega \) to the weights. The number of hidden nodes \( H \) and the number inputs node \( (m) \) allow to detail this transformation (Equation (17));

\[
\text{CSI}_{t+1} = f \left( \sum_{i=1}^{H} y_i^* \alpha_i^2 + b^2 \right)
\]

\[
y_i = g \left( \sum_{j=1}^{m} \text{CSI}_{t-j+1} \alpha_i^j + b_i^j \right)
\]

(17)

In the presented study, the MLP has been computed with the Matlab\textsuperscript{\textregistered} software and its Neural Network toolbox. The characteristics chosen and related to previous work are the following: one hidden layer, the activation functions are the continuously and differentiable 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\textsuperscript{[40]} and is represented by the Equation (18) \((J(\omega)\) is the Jacobian matrix, \(J^T(\omega)\) this transposed and \(\epsilon(\omega)\) the error between the \(N\) simulations and the \(N\) measures).

\[
\Delta \omega = (\omega)^{k+1} - (\omega)^k = \pm J^T(\omega) \epsilon(\omega) + \eta \theta^k \frac{1}{N} J^T(\omega) \epsilon(\omega)^k \]

\[
J(\omega) = \begin{bmatrix}
\frac{\partial e_1(\omega)}{\partial \omega_1} & \frac{\partial e_1(\omega)}{\partial \omega_2} & \cdots & \frac{\partial e_1(\omega)}{\partial \omega_M} \\
\frac{\partial e_2(\omega)}{\partial \omega_1} & \frac{\partial e_2(\omega)}{\partial \omega_2} & \cdots & \frac{\partial e_2(\omega)}{\partial \omega_M} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial e_N(\omega)}{\partial \omega_1} & \frac{\partial e_N(\omega)}{\partial \omega_2} & \cdots & \frac{\partial e_N(\omega)}{\partial \omega_M}
\end{bmatrix} \quad (18)
\]

\[
\epsilon(\omega) = \begin{bmatrix}
e_1(\omega) \\
e_2(\omega) \\
\vdots \\
e_N(\omega)
\end{bmatrix} = \begin{bmatrix}
\text{CSI}_1 - \text{CSI}_1 \\
\text{CSI}_2 - \text{CSI}_1 \\
\vdots \\
\text{CSI}_N - \text{CSI}_1
\end{bmatrix} \quad (18)
\]

In our case the parameter \( m \) takes the value 0.1 and 0.001 when the error, respectively, decreases or increases. Inputs are normalized on \([-0.9,0.9]\) because the truncated interval gives better results than the full interval for the conditions of this study. Training, validation and testing data sets were respectively set to 80%, 20% and 0% (Matlab parameters). These phases concern the four first years and the global solar radiation test forecasting the two last years. The prediction methodology used in this paper is presented on the Fig. 3. The pre-input layer consists on endogenous data at time \( t \), \( t-1 \), etc. and the clear sky index at time \( t+1 \). We encounter problems related to the difficulty to find a significant limit to this coefficient. To work around the apparent classical student \( T \)-test permissiveness and to respect the parsimony principle, we fixed empirically a limit not really justified theoretically. In the presented paper, we proposed a pre-input layer selection method in order to choose from a pool of available data. According to the parsimony principle we have limited the potential inputs to 10 lags for the endogenous case \((L = 10)\) and 2 lags for the each ALADIN output parameters. The total number of exogenous nodes is 8. These 18 data in total are grouped in the pre-input layer and are shown on the scheme of the Fig. 3. This method also uses the statistical Student \( T \)-test to guide the choice but this time it is applied on the coefficients of a multiple linear regression. Note that concerning the ALADIN data, we consider that the physical signification of the lags upper than one are not very relevant because they are predictions of anterior time steps and not measures. In first we generate a regression model for the 18 pre-inputs nodes. The Equation (19) shows the simplest case of 10 endogenous nodes and \( Me \) nodes of exogenous parameter \( E \)\textsuperscript{[25]}:

\[
\text{CSI}_t^* (t+1) = \text{CSI}_t^* (t+1) + \epsilon(t+1)
\]

\[
= \sum_{j=1}^{10} \alpha_j^R X_{t-j+1} + \sum_{p=1}^{Me} \alpha_p^R E_{t-p+1} + b^R L + \epsilon_{t+1}
\]

(19)

This equation can be expressed by matrix form like shown in Equation (20) or in the developed form in Equation (21). \( Y \) is a vector with the elements \( \text{CSI}_t^* (t+1) \), \( S' \) a matrix with concatenated endogenous and exogenous data and \( W^R \) a vector with all the adjustable parameters of the regression. For the estimation of this model there is \( N_T \) data in the measurement history.

\[
\begin{bmatrix}
\text{CSI}_{t+1}^* \\
\vdots \\
\text{CSI}_{t+N_T}^*
\end{bmatrix} = \begin{bmatrix}
1 & \text{CSI}_{t-1}^* & \cdots & \text{CSI}_{t-N+1}^* & E_t & \cdots & E_{t-Me+2} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & \text{CSI}_{t-N+1}^* & \cdots & \text{CSI}_{t-N+2}^* & E_{t-N+2} & \cdots & E_{t-Me+N+2}
\end{bmatrix} \begin{bmatrix}
b^R L \\
\alpha_1^R \\
\vdots \\
\alpha_{Me}^R \\
\epsilon_{t+1} \\
\epsilon_{t+2-N+1}
\end{bmatrix} + \begin{bmatrix}
\epsilon_{t+1} \\
\epsilon_{t+2-N+1}
\end{bmatrix}
\]

(20)

\[
Y = S' W^R + \epsilon
\]

(21)

The optimization corresponds to solve this equation to find correct values of linear regression weights. The least square method is used. The classical estimator is defined by \( W^R = (S'^T S)^{-1} S'^T Y \) where \( S'^T \) is the transpose of the matrix \( S' \). The dimension of the vector \( W^R \) is 19 (10 variables for the insolation, 8 for exogenous data, and 1 for the regression constant). The next step is to verify if some weights \( \text{related to each variable CSI}_t^* \), \( N, RP, P \) and \( T \) called \( X_j \) for \( 1 < j < 12 \) are not significantly different from zero. We use a Student \( T \)-test for all coefficients. The statistic used \( t_j \) corresponds to \( \frac{\alpha_j^R}{\sigma_j} \) \( \sigma_j \) is the standard deviation of the parameter \( j \). It is often easier to interpret it with the notation \( t_j = \frac{\alpha_j^R}{\sigma_j} \) \( \epsilon \) is the error of prediction).

The weight is not statistically equal to zero if \( t_j \) is higher than the value 1.96 (large sample with 5% alpha level). Note that this test can be replaced by an equivalent one, based on the confidence interval of the weight considered \( \text{[} \alpha_j^R \pm t_j \cdot \sigma_j, \alpha_j^R + t_j \cdot \sigma_j \text{]} \). If the sign of the two limits are different, then we can assimilate the weight \( \alpha_j \) to zero. This is equivalent to search the sign of the produce of the two limits.
defined by $\text{sign}((\omega_j^R + t_j \cdot \sigma_j) - (\omega_j^R - t_j \cdot \sigma_j)) = \text{sign}((\omega_j^R)^2 - t_j^2 \sigma_j^2)$. The value must be strictly positive in order to consider $\omega_j$ different to zero. The variables with an associate weight equal to zero are the least correlated with the insolation at the lag $t + 1$. It is the criterion chosen to select (or not) the variable in the input layer of the neural network.

In fact our pre-input layer selection method can be resumed by the following rule: if $\omega_j^R = 0$ then $\omega_j = 0 \forall t$. The application of this rule allows reducing the dimension of the input layer.

After this step, it is necessary to optimize the number of hidden nodes. The technique used is relatively standard; it consists to try several configurations by varying the number of nodes. In previous experiments [13], we have seen that often the number of hidden nodes must be comparable to the number of input nodes. After optimization (choice of input and hidden nodes, activation function, etc.) the output of the network can be expressed by the following expression (Equation (22)) with $\hat{p}$, $\hat{n}$, $\hat{r}$ and $\hat{t}$ and by the number of prediction of each meteorological data use (respectively $p$, $n$, $rp$ and $t$). These predictions done with the model ALADIN depend on latitude, longitude, orography, temperature, humidity, etc [18].

$$
\text{CSI}_t^* (t + 1) = b^2 + \sum_{i=1}^{H_c} \omega_i^2 \cdot \tanh \left( b_1 \sum_{j=0}^{g} \omega_i^j \cdot \text{CSI}_t^* (t - j) \right) + \sum_{p=1}^{P_e \cdot [1 - 2]} \omega_{l-p}^1 \cdot \hat{P}(t - p + 2) + \sum_{u=1}^{N_e \cdot [1 - 2]} \omega_{l-u+ps}^1 \cdot \hat{N}(t - u + 2) + \sum_{v=1}^{R_e \cdot [1 - 2]} \omega_{l-v+ps+n}^1 \cdot \hat{RP}(t - v + 2) + \sum_{w=1}^{R_e \cdot [1 - 2]} \omega_{l-w+ps+np+rp+w}^1 \cdot \hat{T}(t - w + 2) \right) (22)
$$

Note that some elements of the weighting matrix ($\omega_j^R$) are equal to zero after the methodology related to the linear regression presented previously. The computing time is about 5 min with a Core i5 3.20 GHz and 8Gb of RAM. This computing time includes the clear sky generation, the seasonal adjustment, the statistical test, the ANN training and the average prediction computed for five ANN with different initializing.

### 3.4. Hybrid methodology

The idea of a hybrid methodology has started with the observation that standalone predictors are inefficient in some cases and very efficient in other specific cases. One of the solution is to combine predictors to understand and model the dynamic of the signal in a hybrid method: it is the base of the data mining techniques. Among the ANN based predictors hybridization, we can mention the ARMA-MLP [44,45], the fuzzy inferences-MLP [22], the wavelet-MLP [46], the Markov-MLP [7] or Bayes-MLP [47]. We oriented our choice to methods for which we obtained the best results: ARMA and MLP (multivariate case) with time series made stationary and propose a new hybrid model formed by these two sub-predictors. As there is a multitude of possible arrangements to construct it, we have opted to a model selection based on the linear/non-linear transition of the global radiation process. Linearity test exists (Lagrange multipliers), but are not really efficient when the two phenomena are concomitant in the same series. Our hypothesis is very simple: when the nebulosity is low, we consider the series with linear behavior and when the nebulosity is high, we consider the series with non-linear behavior. Thus, for the linear approach we use ARMA and for the non-linear approach, we select the PMC predictor. Zhang [45] confirms that a hybrid model having both linear and non-linear modeling abilities could be a good alternative for predicting time series data. By combining different models, different aspects of the underlying pattern may be captured. In order to construct our model selection we could only consider the day position in the year: during the winter months PMC would be...
selected and ARMA during the summer months. In this case, a selection by rigid seasonality would be considered. Unfortunately this approach will not take into consideration sunny days in winter or cloudy days in summer, even if they are few in number. Consequently, we decided to design a methodology based on a non-rigid, non-repetitive and non-well-marked seasonality. Our hybrid model is constructed as a stochastic model, which depends solely on the mistake made the previous hour. If the ARMA method was better at time \( t \) (equivalent to sunny period, and indirectly to a linear phenomenon) then it will be still ARMA at \( t + 1 \) else it will be the MLP. We can summarize this method with the following rule (Equation (23)) where \( e \) is the residue of the prediction:

\[
\text{if} \quad |\epsilon^{AR}(t)| \leq |\epsilon^{ANN}(t)| \quad \text{then} \quad \hat{x}(t+1) = \hat{x}^{AR}(t+1) \quad \text{else} \quad \hat{x}(t+1) = \hat{x}^{ANN}(t+1)
\]

The Fig. 4 details the process followed by the hybrid methodology.

The next section presents, results obtained using this hybrid model on the five places located in Mediterranean area presented in Section 2.

4. Experiments, results and discussion

This section includes all the experiments and results conducted during this study. The first part of this section is dedicated to the stationary efficiency of our preprocessing methodology. The second part shows intermediate results on MLP trained with endogenous and ALADIN forecast data. Finally we present results obtained with the hybrid model presented in the previous section.

4.1. Experiences about series made stationary

As we have seen previously in Section 3.1, the different methods dedicated to the time series forecasting often requires stationary time series. It should be emphasized that if the available data are non-stationary, they must be made stationary before applied to ARMA or ANN model. Based on this observation, it is necessary to study stationary methods and to see if it should be interesting in the case of the global radiation time series forecasting. The two tested processes, the clear sky index (CSI) and the CSI with use of seasonal correction and periodic coefficients (CSI\textsuperscript{*}) are compared to the original series without treatment. To know if the preprocessing makes the series stationary, we use criterions like the variation coefficient (VC\textsubscript{c}) considered as a dispersion rate and the Fisher T-test related to the daily and yearly stationarity. The values given on the Table 1 are established after a normalization process between \([-0.9,0.9]\). According to the Fisher Table, we found that threshold of the F-Test is \( F_{0.05} = 1 \).

<table>
<thead>
<tr>
<th></th>
<th>Ajaccio</th>
<th>Bastia</th>
<th>Montpellier</th>
<th>Marseille</th>
<th>Nice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original series</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC</td>
<td>-3.45</td>
<td>-2.48</td>
<td>-3.36</td>
<td>-3.62</td>
<td>-3.21</td>
</tr>
<tr>
<td>F\textsubscript{c}\text{(yearly)}</td>
<td>12.99</td>
<td>12.44</td>
<td>11.12</td>
<td>15.07</td>
<td>13.72</td>
</tr>
<tr>
<td>CSI</td>
<td>6.81</td>
<td>1.01</td>
<td>7.02</td>
<td>6.53</td>
<td>7.27</td>
</tr>
<tr>
<td>F\textsubscript{c}\text{(daily)}</td>
<td>3.25</td>
<td>3.05</td>
<td>3.39</td>
<td>4.08</td>
<td>3.44</td>
</tr>
<tr>
<td>CSI\textsuperscript{*}</td>
<td>2.06</td>
<td>2.07</td>
<td>2.07</td>
<td>2.07</td>
<td>2.08</td>
</tr>
<tr>
<td>F\textsubscript{c}\text{(daily)}</td>
<td>0.83</td>
<td>0.87</td>
<td>0.72</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>F\textsubscript{c}\text{(daily)}</td>
<td>2.05</td>
<td>2.04</td>
<td>1.41</td>
<td>1.80</td>
<td>1.52</td>
</tr>
</tbody>
</table>

The VC\textsubscript{c} parameter is not the more interesting as a stationarity criterion but it is very simple to use. The two other parameters show that the CSI\textsuperscript{*} processing is the most efficient. There is only one case where the \( F \) is not minimal with the CSI\textsuperscript{*} (Bastia, \( F\textsubscript{c}\text{(daily)} = 2.04 \)). Indeed, with this method, 8 of 10 parameters (two parameters by city) are lower than the Fisher thresholds. Note that with the others cases, only one empirical Fisher parameter indicates a quasi-stationarity in Bastia (\( F\textsubscript{c}\text{(daily)} = 1.01 \)). It is certainly because of this location the weather is very unstable, generating a very noisy global radiation time series. On the Fig. 5, we can see the impact of the CSI and CSI\textsuperscript{*} preprocessing on a global radiation time series for the difficult Bastia case study. We obtained similar results for the four other locations.

While the CSI\textsuperscript{*} processing seems the most interesting, the CSI seems graphically give the first moment constant (average fixed around the year \( \sim -0.4 \)), but let the second central moment about the mean of the series with periodicity (variance more important during winter than summer). We can see that the curve related to the CSI\textsuperscript{*} process seems the most non-seasonal, contrary to the CSI process. In winter the standard deviation has not increased. The results presented in the next section will argue if the stationarity of the series increases the prediction quality.

4.2. Results with MLP and ALADIN forecast data

This subpart proposes to analyze the impact of adding data from the ALADIN numerical model to a MLP. We have chosen to compare it with four other models (see Table 2). The first model which is considered as the reference is based on ARMA, the second is an endogenous MLP with our clear sky index preprocessing (CSI), the third is an endogenous MLP with CSI and seasonal correction (CSI\textsuperscript{*}), and the last is a MLP with ALADIN forecast data. In the following, we adopt a canonical form to present the MLP architecture obtained after optimization steps: \( \text{end}^{0}, R^{PP N^{p}}, P^{i}, L^{i}) \times H \times S \), where \( e, r, n, p, t \) and \( s \) are the numbers of endogenous data inputs, precipitation, nebulosity, pressure and temperature, \( H \) and \( S \) are the number of hidden and output nodes. The exact description of the compared methods is:
I. ARMA + PC: The best ARMA model with the CSI* preprocessing;  
II. ANN: The best endogenous MLP with the CSI preprocessing. Optimizations have been done with a standard method based on the interpretation of Partial Autocorrelation Factor (PACF) [17];  
III. ANN + PC: The best MLP with CSI* preprocessing. Optimizations have been done with a standard method based on the autocorrelation factor and cross-correlation;  
IV. ANN + ALADIN: An optimized MLP with the CSI preprocessing and pre-input layer selection method using ALADIN forecast data in input;  

Fig. 5. Effects of the stationarity processing on the Bastia time series during 365 days (Oct 2002–Oct 2003). The CSI* and the CSI are normalized.
This methodology is based on the prediction of models I and V. If we analyze the annual error, the methodology V, based on the utilization of an optimized MLP with ALADIN forecasting data and seasonal adjustment with periodic coefficients (CSI preprocessing) is the most relevant. If we analyze seasonally the errors, we find only four cases (Ajaccio, Bastia and Nice during spring and Montpellier during summer) which have a better forecaster: ARMA method. For the other cases, the ARMA method seems equivalent to the ANN method without pretreatment (method II) but is worse than the ANN when the stationary of the series is the same (method III). While the distinct adding of CSI’ preprocessing (method III; error is 13.0%) and ALADIN forecast data (method IV error is 13.2%) to the ANN model improves the prediction, their combined adding (method V) potentiates this effect. Because the sunlight is abundant and there is low occurrence of clouds, the summer season is the most favorable to prediction.

### 4.3. Results with the hybrid methodology

This subpart proposes to present results of the hybrid method (see the Table 3) defined in section 3.d. The ARMA and MLP models used are those presented in the previous section: models I and V. This methodology is based on the prediction of “ARMA” and “ANN” with ALADIN data, clear sky index and seasonal adjustments. The ratio ARMA/ANN represents the numbers of hourly ARMA simulations versus ANN simulations. We can observe that the distribution of predictors used is 1/3 for ARMA and 2/3 for ANN. This observation confirms the results previously presented. In summer, the error is for the first time lower than 10% (Montpellier). All the seasons and cities benefit of the use of this method coupling ARMA and ANN. The summer is the season where the gain is the less significant, certainly because it is the season where the standalone methodologies (ANN and ARMA) are sufficient.

Ultimately, We can see that all the methodologies proposed in this paper have decreased the prediction error. In average of the five cities, the total nRMSE decrease of 11.3% against a naive persistence predictor which has an average nRMSE equal to 26.2%. A step very interesting is the mixing between the ANN and ARMA, but do not forget that this result is only true with the type of ANN and ARMA considered. In the Fig. 6, we can see the matching between measure and simulation (hybrid method presented previously) for all the cities. We can presume that there is a strong correlation between the two quantities. Except to Bastia, the cloudy periods (like the 71st-81st hour interval) seem correctly predicted. Concerning the case of Bastia, the model is not able to anticipate the nebulosity, and improvements are certainly necessary.

### 4.4. Evaluation of the prediction relevance

Considering that the prediction methodology proposed here is mainly designed for a power manager, it is necessary to couple the forecasted value to a confidence interval. The presented study proposes to compute it during the training step of the MLP, and then to use it during the prediction. The simulator gives for each

### Table 3

Comparison between five forecasting methods for the five cities studied. The error is the nRMSE and best results are in bold.

<table>
<thead>
<tr>
<th>Models</th>
<th>Annual</th>
<th>Winter</th>
<th>Spring</th>
<th>Summer</th>
<th>Autumn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>25.4</td>
<td>32.9</td>
<td>25.3</td>
<td>20.0</td>
<td>32.3</td>
</tr>
<tr>
<td>I. ARMA(1,0)</td>
<td>18.9</td>
<td>23.9</td>
<td>19.0</td>
<td>11.8</td>
<td>21.4</td>
</tr>
<tr>
<td>II. ANN (Endo1–10) × 15 × 1</td>
<td>20.3</td>
<td>27.2</td>
<td>20.4</td>
<td>13.7</td>
<td>24.1</td>
</tr>
<tr>
<td>III. ANN + PC (Endo1–2,4,6) × 15 × 1</td>
<td>18.6</td>
<td>25.3</td>
<td>18.4</td>
<td>12.2</td>
<td>24.2</td>
</tr>
<tr>
<td>IV. ANN + ALADIN (Endo1–5,6,8,10) × 2(1–2)2 × 15 × 1</td>
<td>19.0</td>
<td>26.8</td>
<td>19.1</td>
<td>12.3</td>
<td>23.0</td>
</tr>
<tr>
<td>V. ANN + ALADIN + PC (Endo1–5,6,8,10) × 2(1–2)2 × 15 × 1</td>
<td><strong>17.8</strong></td>
<td><strong>24.9</strong></td>
<td><strong>18.1</strong></td>
<td><strong>11.7</strong></td>
<td><strong>22.0</strong></td>
</tr>
<tr>
<td>Persistence</td>
<td>27.1</td>
<td>35.0</td>
<td>27.1</td>
<td>22.6</td>
<td>34.4</td>
</tr>
<tr>
<td>I. ARMA(1,0)</td>
<td>21.1</td>
<td>26.7</td>
<td><strong>20.3</strong></td>
<td>15.8</td>
<td>26.9</td>
</tr>
<tr>
<td>II. ANN (Endo1–10) × 10 × 1</td>
<td>22.8</td>
<td>27.3</td>
<td>23.4</td>
<td>16.1</td>
<td>25.7</td>
</tr>
<tr>
<td>III. ANN + PC (Endo1–2,4,6) × 15 × 1</td>
<td>20.8</td>
<td>24.9</td>
<td>21.4</td>
<td>14.9</td>
<td>24.9</td>
</tr>
<tr>
<td>IV. ANN + ALADIN (Endo1–3,8,10) × 2(1–2)2</td>
<td>21.3</td>
<td>25.8</td>
<td>21.7</td>
<td>15.1</td>
<td>24.0</td>
</tr>
<tr>
<td>V. ANN + ALADIN + PC (Endo1–3,8,10) × 2(1–2)2 × 15 × 1</td>
<td><strong>19.9</strong></td>
<td><strong>24.4</strong></td>
<td><strong>20.5</strong></td>
<td><strong>14.2</strong></td>
<td><strong>23.3</strong></td>
</tr>
<tr>
<td>Persistence</td>
<td>26.9</td>
<td>32.6</td>
<td>25.9</td>
<td>24.6</td>
<td>33.2</td>
</tr>
<tr>
<td>I. ARMA(1,0)</td>
<td>20.1</td>
<td>23.5</td>
<td>18.7</td>
<td>15.5</td>
<td>21.9</td>
</tr>
<tr>
<td>II. ANN (Endo1–10) × 10 × 1</td>
<td>20.8</td>
<td>22.4</td>
<td>20.2</td>
<td>17.9</td>
<td>19.3</td>
</tr>
<tr>
<td>III. ANN + PC (Endo1–2,4,6) × 15 × 1</td>
<td>19.3</td>
<td>20.4</td>
<td>18.8</td>
<td>16.0</td>
<td>19.8</td>
</tr>
<tr>
<td>IV. ANN + ALADIN (Endo1–3,8,10) × 2(1–2)2</td>
<td>19.3</td>
<td>20.3</td>
<td>18.6</td>
<td>16.8</td>
<td><strong>18.1</strong></td>
</tr>
<tr>
<td>V. ANN + ALADIN + PC (Endo1–3,8,10) × 2(1–2)2 × 15 × 1</td>
<td><strong>18.6</strong></td>
<td><strong>20.1</strong></td>
<td><strong>17.9</strong></td>
<td><strong>15.5</strong></td>
<td><strong>19.2</strong></td>
</tr>
<tr>
<td>Persistence</td>
<td>25.3</td>
<td>32.9</td>
<td>25.3</td>
<td>20.0</td>
<td>32.3</td>
</tr>
<tr>
<td>I. ARMA(2/0)</td>
<td>18.9</td>
<td>23.9</td>
<td>19.0</td>
<td>11.8</td>
<td>21.4</td>
</tr>
<tr>
<td>II. ANN (Endo1–10) × 10 × 1</td>
<td>19.0</td>
<td>22.5</td>
<td>20.7</td>
<td>11.3</td>
<td>18.8</td>
</tr>
<tr>
<td>III. ANN + PC (Endo1–2,4,6) × 15 × 1</td>
<td>16.9</td>
<td>20.6</td>
<td>17.8</td>
<td>10.5</td>
<td>17.1</td>
</tr>
<tr>
<td>IV. ANN + ALADIN (Endo1–3,8,10) × 2(1–2)2</td>
<td>17.4</td>
<td>20.4</td>
<td>18.5</td>
<td>10.4</td>
<td><strong>16.4</strong></td>
</tr>
<tr>
<td>V. ANN + ALADIN + PC (Endo1–3,8,10) × 2(1–2)2 × 15 × 1</td>
<td><strong>16.3</strong></td>
<td><strong>19.6</strong></td>
<td><strong>16.6</strong></td>
<td><strong>10.3</strong></td>
<td><strong>16.4</strong></td>
</tr>
<tr>
<td>Persistence</td>
<td>26.4</td>
<td>32.1</td>
<td>24.5</td>
<td>21.1</td>
<td>37.1</td>
</tr>
<tr>
<td>I. ARMA(1,0)</td>
<td>20.7</td>
<td>23.5</td>
<td><strong>17.6</strong></td>
<td>12.4</td>
<td>37.5</td>
</tr>
<tr>
<td>II. ANN (Endo1–10) × 10 × 1</td>
<td>20.9</td>
<td>21.7</td>
<td>18.8</td>
<td>11.7</td>
<td>32.3</td>
</tr>
<tr>
<td>III. ANN + PC (Endo1–2,4,6) × 15 × 1</td>
<td>20.1</td>
<td>20.5</td>
<td>19.1</td>
<td>11.4</td>
<td>30.9</td>
</tr>
<tr>
<td>IV. ANN + ALADIN (Endo1–3,8,10) × 2(1–2)2</td>
<td>20.1</td>
<td>20.4</td>
<td>18.1</td>
<td>11.4</td>
<td><strong>32.3</strong></td>
</tr>
<tr>
<td>V. ANN + ALADIN + PC (Endo1–3,8,10) × 2(1–2)2 × 15 × 1</td>
<td><strong>19.4</strong></td>
<td><strong>19.9</strong></td>
<td><strong>18.6</strong></td>
<td><strong>11.0</strong></td>
<td><strong>30.1</strong></td>
</tr>
</tbody>
</table>

### Table 2

Comparison between five forecasting methods for the five cities studied. The error is the nRMSE and best results are in bold.
Fig. 6. Comparison between measured and simulated global radiation done with the hybrid methodology “ARMA” and “ANN + ALADIN + PC + CSI”. (The 250 h shown corresponding to half-February to half-March 2008).
Fig. 7. Prediction marked with confidence interval of the global radiation in Ajaccio during the 50th first hours of January 2007. The line is related to the measure.
hour two parameters: the $h + 1$ horizon global radiation and a parameter representing the confidence we can give to this value. Before to compute this parameter it is necessary to explain the parameter $CI(t)$ represented in the equation (24). In fact it is the absolute residue error of the prediction during the training sample. The training set includes 4 years, and each year includes $365 \times 9 = 3285$ h, so $4 \times 3285$ elements. Then an hourly average is done allowing to transform this $CI(t)$ series to a new series $CI'(t)$ like described in the equation (25).

$$CI(t) = |e(t)| \forall t \in [1; 4 \times 3285]$$

$$CI'(t) = \left(\frac{1}{4}\right) \sum_{i=1}^{4} CI(t + (i - 1) \cdot 3285) \forall t \in [1; 3285]$$

This index is necessary to judge the relevance of the prediction. For example, the Fig. 7 shows for Ajaccio the prediction and the confidence interval (average $\pm CI$). Three models are evaluated: the ANN with only endogenous data, the ANN with ALADIN and the periodic coefficients and the hybrid model “ARMA” with “ANN + ALADIN + PC”. The period considered is one of the most complicated to forecast. It corresponds to the winter with a lot of cloudy days. It is interesting to see that the simple model endogenous ANN describes well the radiation, however, there are some outliers points (15th or 41st hour), with a confidence interval not significant. In the second curve related to the ANN + ALADIN + PC, the error is most regular, the atypical points seems non-existent. The third curve (both of ANN and ARMA) is visually the most interesting, despite that the confidence interval is yet incorrect for some points. In fact, with the hybrid method (ARMA + ANN), the ANN forecast is preponderant during the cloudy days; the ARMA predictions are used only during the sunny days. So the two last methods presented are relatively similar for the month considered.

5. Conclusion

We proposed in this paper an original technique to predict hourly global radiation time series using meteorological prediction model. We optimized a multilayer perceptron (MLP) with ALADIN forecast data and endogenous data previously made stationary. We have used an innovative pre-input layer selection method and we have combined our optimized MLP to an auto-regressive and moving average (ARMA) model from a rule based on the analysis of hourly data series. This model has been successfully used to forecast the hourly global horizontal radiation for five places in Mediterranean area. In addition, this paper has allowed determining a stationarity process (method and control) for the global radiation time series. This step is crucial to correctly forecast the future values of insolent (annual average nRMSE gain over the five locations equals to 1.7%). The use of ALADIN forecasting data as an input of a MLP has shown a really great interest to improve the prediction (average nRMSE gain of 0.7%). These results could certainly be improved by a better comprehension of the complexity of the model and collaboration with a professional forecaster of Meteo-France who could help us in selection of data especially concerning the runs to consider. In last, the use of a hybrid method coupling ANN and ARMA predictors decrease much the prediction error (average nRMSE gain to 3.5%). If we compare the results with a standard model like persistence method (a typically naïve predictor) the nRMSE error is reduce by about 11.3%. The detail of the prediction error decrease following the different steps of the forecasting methodology is done on the Fig. 8. The last important point treated in this paper is the proposition of using confidence intervals in order to estimate the reliability of the prediction. The perspectives of this work are related to the generalization of our model. We would show that the superior performance of this model is not likely to be a consequence of data mining (or data snooping). In fact, it should be sure that the model constructed in this way is not of limited practical value.

Concerning the practical and policy implication of the results shown here, the conclusions developed here, are certainly compatible with the renewable energies deployment. Because of their intermittent nature, some energy sources (like PV or wind energy) have to be included in a limited way in power systems to ensure the electrical grid stability. There are two methodologies to overcome this problem: the storage of the overflow or the prediction of the energy sources. The work presented in this paper concerns the second problematic and deals with the 1 h forecasting horizon. It is certainly the most important horizon for a non-connected power system manager, especially in the insular case, where the energy autonomy must be considered at medium or long-term. As a matter of fact it corresponds to the starting delay of conventional energy sources like diesel engines or gas turbines. The idea is to use at the right time other energy productions not dependent on weather, and to avoid cuts of the electrical power or maybe a blackout. Actually the energy providers use often methods based on persistence to predict the global radiation. Even if this type of predictor is effective in sunny days, the present study shows that more efficient forecasters exist. Now that our approach have been validated it would be interesting to see if it could be apply to other ANN architectures (TDNN: Time Delay Neural

<table>
<thead>
<tr>
<th>Methodology</th>
<th>nRMSE (%)</th>
<th>RMSE Gain %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>26.2 (±0.8)</td>
<td>-5.4</td>
</tr>
<tr>
<td>ANN without stationary process</td>
<td>20.8 (±1.2)</td>
<td>-1.7</td>
</tr>
<tr>
<td>ANN with CSI and seasonal adjustments</td>
<td>19.1 (±1.3)</td>
<td>-0.7</td>
</tr>
<tr>
<td>ANN with CSI, seasonal adjustments and ALADIN modeling</td>
<td>18.4 (±1.2)</td>
<td>-3.5</td>
</tr>
<tr>
<td>Both of ANN (precedent case) and ARMA</td>
<td>14.9 (±0.9)</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 8. Prediction error decrease following the different steps of the forecasting methodology.

Please cite this article in press as: Voyant C, et al., Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, Energy (2012), doi:10.1016/j.energy.2012.01.006
Network: RBFN: Radial Basis Function Neural Network; etc.). Furthermore it would be relevant to study an approach for solar irradiance forecasting 24 h ahead using several ANN connected, to decrease the time step and to test the methodology on a real PV module. In addition, an important work will be to simplify the model while keeping an acceptable prediction. Indeed, even if our method is attractive, it could be complex and costly to implement for an electric power manager in grid stability context of power supply mixing renewable and conventional energy.

Acknowledgment

This work was partly supported by the Territorial Collectivity of Corsica. We thank the French National Meteorological Organization (Météo-France) and the CNRM (Centre National de Recherches Météorologiques) and particularly Ms. Delmas who has supervised the data collection from their data bank for the five synoptic stations and from the ALADIN model.

References

[10] Mellit A, Eleuch H, Benghanem M, Elaoun C, Massi Pavan A. An adaptive wavelet-network model while keeping an acceptable prediction. Indeed, even if our module. In addition, an important work will be to simplify the decrease the time step and to test the methodology on a real PV irradiance forecasting 24 h ahead using several ANN connected, to Furthermore it would be relevant to study an approach for solar.

Nomenclature

Time series generalities

\( x_t \): Time series measure and prediction at time \( t \)
\( \hat{x}_t \): Global time series measure and prediction at time \( t \) \( [\text{Wh/m}^2] \)
\( \phi_n^{\text{har}} \): Angular frequencies of a signal \( s(t) \)
\( R_C \): Time series calculated from global radiation and representing the cloud cover occurrence

\( a/b(t) \): Continuous functions representing the amplitude of the global radiation \( [\text{Wh/m}^2] \)
\( f_C \): Regression model group

\( c_s \): Residue of regression at time \( t \) 1 r
\( E(t) \): Expected value of \( X(t) \); the first moment of \( X(t) \) is the average, and the second the variance
\( \gamma(h) \): Covariance between two variables without time dependence

Time series preprocessing

\( H_C(t) \): Corrected extraterrestrial solar radiation coefficient at time \( t \) \( [\text{Wh/m}^2] \)
\( H_C \): Clear sky global horizontal radiation \( [\text{Wh/m}^2] \)
\( c \): Global total atmospheric optical depth
\( a \): Fitting parameter of the Solis clear sky model

Please cite this article in press as: Voyant C, et al., Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, Energy (2012), doi:10.1016/j.energy.2012.01.006
h: Solar elevation angle at time $t$ (rad)

CSI: Clear sky index at time $t$

CSI*: CSI with seasonal adjustments at time $t$ and related to $i$ (period number; number of year, or days) and $j$ (number of measures per period).

MM: Moving average of CSI at time $t$

C: Periodic coefficients at time $t$

Statistical tests

VCX: Variation coefficient (%)

F: Fisher statistic

$V_P, P$: Variance of the period and $p$ number of measures per period

$V_R, R$: Variance of the residue and $N$ the number of period

Nh: Number of hours of prediction contained during one year ($365 \times 9 = 3285$)

CI(t), CI* (t): Confidence interval for the time $t$ of prediction. The star is related to the yearly average

ANN nomenclature

$b_1$: $b_i$ set, values of the biases of the hidden nodes

$b_2$: $b_i^*$ set, value of the output node bias

H: Number of hidden nodes

In: Number of endogenous input nodes

f: Activation function of the output and hidden layer

$\alpha_{ij}$: values of the weight between input nodes $j$ and hidden nodes $i$

$\alpha_{ij}^*$: weight values between hidden node $i$ and output node

$\Delta$: weights optimization for the Levenberg–Marquard algorithm

J: Jacobian matrix

I: Identity matrix

$e(t)$: Error prediction vector

$\mu$: Specific algorithm parameter

endo$: e$ lags endogenous for ANN

N(t), n: ALADIN forecast of the nebulosity at time $t$ and $n$, the number of nebulosity lag in ANN input [octas]

$\bar{P}(t), p$: ALADIN forecast of the pressure at time $t$ and $p$, the number of pressure lag in ANN input [Pa]

$\bar{T}(t), t$: ALADIN forecast of the temperature at time $t$ and $t$, the number of temperature lag in ANN input [°C]

$\bar{R}P(t), r$: ALADIN forecast of the rain precipitation at time $t$ and $r$, the number of precipitation lag in ANN input [mm]

H: Number of hidden nodes in the MLP

In: Number of input nodes in the MLP

Linear regression

$W^R$: $w_i^R$ set representing a vector of the linear regression coefficients

Y: Vector with the CSI*($t + 1$) values

$S$: Matrix with concatenated endogenous and exogenous data

ARMA model

$p, q$: Order of ARMA($p, q$)

L: Lag operator

$\varphi$, $\theta$: $\varphi_i$ are the parameters of the AR model and $\theta_i$ the parameters of the MA model