Use of exogenous data to improve an artificial neural networks dedicated to daily global radiation forecasting

Christophe Paoli, Cyril Voyant,
Marc Muselli, Marie-Laure Nivet
University of Corsica
CNRS UMR SPE 6134
(Ajaccio | Corte), France
{christophe.paoli, cyril.voyant,
marc.muselli, marie-laure.nivet}@univ-corse.fr

Cyril Voyant
Hospital of Castelluccio
Radiotherapy Unit
Ajaccio, France

Abstract—**This paper present an application of Artificial Neural Networks (ANNs) in the renewable energy domain and, more particularly, to predict solar energy. We 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 previous studies, we have demonstrated that an optimized ANN with endogenous inputs can forecast the solar radiation on a horizontal surface with acceptable errors. Thus we propose to study the contribution of exogenous meteorological data to our optimized PMC and compare with different forecasting methods used previously: a naïve forecaster like persistence, an ANN with preprocessing using only endogenous inputs and an ANN with pre-processing using endogenous. Although intuitively the use of meteorological data may increase the quality of prediction, the obtained results are relatively mixed. The use of exogenous data generates a decrease of nRMSE between 0.5% and 1% for the two studied locations. The absolute error (RMSE) is decreased by 52 Wh/m²/day in the simple endogenous case and 335 Wh/m²/day for the persistence forecast.**

Renewable energy; solar energy; time series forecasting; artificial neural networks; pre-processing; multi-layer perceptron; prediction

I. INTRODUCTION

We present the results of the prediction of global radiation using Artificial Neural Networks (ANN) which are a popular artificial intelligence technique in the forecasting domain [1][2][3][4]. In previous studies [5][6], we have demonstrated that an optimized ANN with endogenous inputs can forecast the solar radiation with acceptable errors. In this present paper, our aim was to answer to the following question: does the use of exogenous meteorological variables (like temperature, humidity, wind speed and direction, pressure gradient etc.) may increase the quality of prediction? We tried to answer to this question with two sites (Ajaccio and Bastia) located on the Island of Corsica (France). The island is characterized by a Mediterranean climate and a hilly terrain. The paper will be organized as follow: in section 2, the ad-hoc time series pre-processing is described and in section 3, the results of the endogenous and exogenous optimization are presented. The results of the prediction are shown in the section 4 for horizontal radiation. The conclusions and perspectives of this work are presented in the last part.

II. METHODOLOGY

An ANN is made up by simple processing units, the neurons, which are connected in a network by synaptic strengths, where the acquired knowledge is stored. In a feed forward ANN also known as a Multi Layer Perceptron (MLP), neurons are grouped in layers and only forward connections exist. A typical feed forward neural network consists of an input, a hidden and an output layers. Each component includes a neuron, 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:

$$y_i = f(\sum_{j=1}^{n} x_j w_{ij}) \tag{1}$$

where i is a neuron index in the hidden layer and j is an index of an input to the neural network. Training is known as the process of modifying the connection weights in some orderly fashion using a suitable learning method.

Figure 1 gives the basic architecture for a MLP application to time series forecasting [7][8][9][10]. A fixed number p of past values is fed to the input layer of the MLP and the output is required to predict a future value of the time series. This method is often called the sliding window technique as the N-tuple input slides over the full training set.

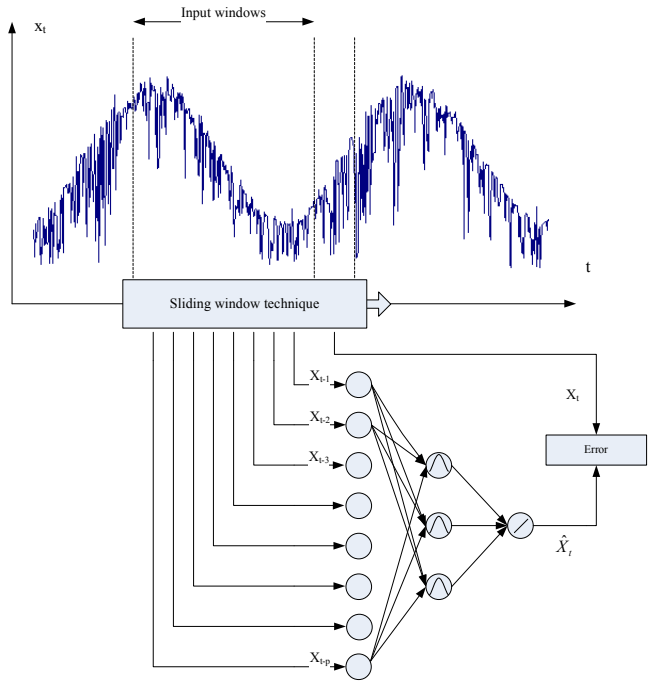

Figure 1. MLP application to time series forecasting (bias nodes are not displayed).

According to previous experimentations [5], we utilized a stationarization method to increase the prediction quality witch consists to use the clear sky model for clear sky index. There are a lot of methods to determine this model. In our case, we have preferred to use the simplified "Solis clear sky" model [11] based on radiative transfer calculations and the Lambert-Beer relation. In this case, the clear sky global horizontal irradiance (H_{gh}) reaching the ground is defined by:

$$H_{gh,clearsky} = H_0 . e^{-(\tau / \sin^b(h))} . \sin(h) \quad (2)$$

where τ is the global total atmospheric optical depth (-0.37 in our case), h is the solar elevation angle and b is a fitting parameter (0.35 for us). The daily integration of the $H_{gh,clearsky}$ parameter allows to determine the daily solar radiation modelling $H^d_{gh,clearsky}$. A series of tests (not presented in this paper) has led to the validation of the Solis model on horizontal global radiation. We have obtained a relation of stationarization equivalent to (2) where X is the measure and S the new time series, (d is the day of the year y):

$$S_{d,y} = \frac{X_{d,y}}{H^d_{gh,clearsky}} \quad (3)$$

This treatment aims to create a new distribution without periodicity. Further the new series generated is equivalent to a nebulosity series. Ideally, the values are fixed to 1 and decrease with the occurrence of a cloud.

Concerning the problem of ANN optimization, 3 independent subparts have to be studied:

- Choice of the hidden layer number and activation function.
- Choice of the endogenous lag number.
- Choice of the exogenous lag numbers for each parameter (daily pressure variation; wind direction; humidity; insulation; nebulosity; precipitation; mean pressure; min-max-mean temperatures, night temperature; wind speed).

Up to now, we have studied the ANN optimization in daily horizon [6]. This experience allows us to used previous results like the model of ANN for this kind of problem and the learning algorithm. The ANN used is so a PMC and we have simulated it with the Matlab software and its NN toolbox. These characteristics are: 1 hidden layer, the activation function are hyperbolic tangent (hidden) and linear (output), the learning algorithm is the Levenberg-Marquardt model (with max fail parameter equal to 5, μ decreases and increases respectively to 0.1 and 0.001, and goals equal to zero), the normalization is done between 0 and 1; the ratio of train, validation and test periods represent respectively 80%, 10% and 10%. We have learned the ANN during the 8 first years and we have computed the global solar radiation during the 2 last years.

III. ENDOGENOUS AND EXOGENOUS PARAMETERS OPTIMIZATION

The partial autocorrelation function (PACF) plays an important role in time series analysis and allows to identify the extent of the lag in an autoregressive model. Thus we have used PACF in the case of PMC to determine the best time lag of the endogenous net-input. In practice, the last correlation coefficient (R_k) different to zero (T-test methodology) induces k endogenous clear sky index like network inputs. The methodology chosen for the exogenous variable selection is similar to the previous one, ie the use of a correlation criterion (Pearson correlation). In our experimental sample size, the significant threshold for the T-test indicates a very low limit. Indeed, the limit is below 0.1 for the sample above 1000 (for the 0.05 critical alpha level). This methodology is not realistic in our case, the threshold for the coefficient R should be more important to select a limited number of exogenous inputs. Thereafter we have chosen a threshold R = 20%, only the higher correlations have been chosen. If we use the previous notation for the clear sky index (S_t) and for an exogenous variable (y_t representing nebulosity, temperature…), the correlation between the variables is:

$$R_k^y = \frac{\sum_{t=k+1}^{N} (S_t - \bar{S})(y_{t-k} - \bar{y})}{\sqrt{\sum_{t=k+1}^{N} (S_t - \bar{S})^2 \sum_{t=k+1}^{N} (y_{t=k+1} - \bar{y})^2}} \quad (4)$$

IV. RESULTS FOR HORIZONTAL RADIATION

In this section we present the results we obtained for the meteorological station of Ajaccio and Bastia located on the Island of Corsica (France).

A. The station of Bastia

To determinate the input number of our PMC, first it is necessary to use the PACF described in the last section. PACF allows quantifying the endogenous number of inputs. On figure 2, we can see that this methodology implies the use of S_t, S_{t-1}, S_{t-2} and S_{t-3} on our ANN to predict S_{t+1}.

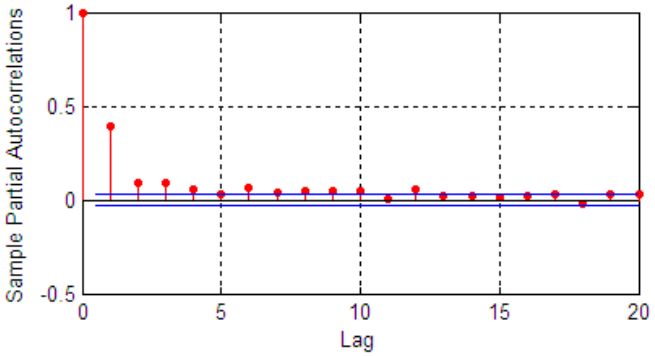

Figure 2. Partial autocorrelation of clear sky index in the station of Bastia. The lines around the zero represent the confidence interval with the PACF.

The second step for optimize the input layer of the PMC is to choose atmospheric variables and their time lag. The correlation study has been done according to the previous section. The determination of the each correlation, specific to one variable and one time lag, allow to determinate the input to add. The figure 3 presents only the variables which are correlated with the clear sky index. We can see that the cut off condition fixed on previous section (threshold R = 20%) implies that the lag 1 is sufficient.

The purpose of this study is not to demonstrate the interest of use an ANN but to search a mechanism for its optimisation. The first step was to stationnarize the global radiation (transition to clear sky index), then to find the inputs number (4 endogenous, 1 for humidity, 1 for insulation and 1 for nebulosity of previous days). We have verified that the use of each exogenous input improves the quality of forecast. In details, the results show that the use of the insulation lag 1, or humidity, or nebulosity, increases not only the mean of the error prediction but also the robustness of the methodology. The variance of the results is lower with the use of exogenous inputs. The other interesting element is the equivalence of the insulation, humidity and nebulosity when they are used separately. For Bastia, the gap of the exogenous inputs utilization is minimal, but is real. The nRMSE is 25.43% against 25.85%, the RMSE is reduce 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). A simplest forecaster as the persistence leads to an error nRMSE = 31.17% (MAE = 1081 Wh/m²/day and RMSE = 1569 Wh/m²/day). This site is well known to be very difficult to predict, this result was already checked in previous studies [12]. The second site mentioned in this article (Ajaccio) is reputed to be easier to predict.

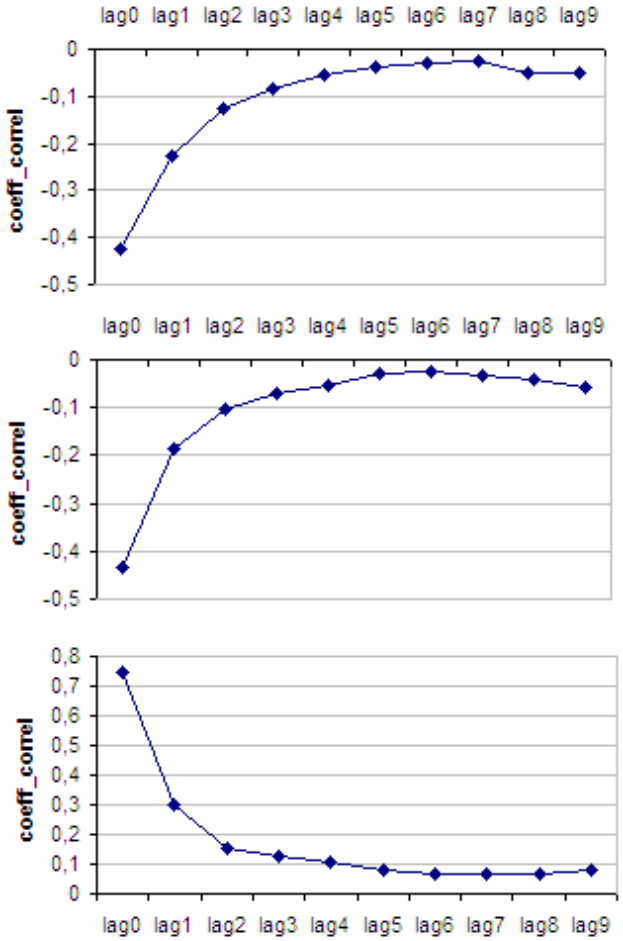

Figure 3. Pearson correlation between the clear sky index, and exogenous variables for Bastia station: top is humidity, middle the nebulosity and bottom the insolation

B. The station of Ajaccio

The optimization methodology is the same that used in the previous case. Applying the autocorrelation function to determine the interesting time lags, we obtain only two parameters (k=2) correlated with the clear sky index (S_{t+1} correlated with S_t and S_{t-1}). The experiment shows that we have to consider only the exogenous inputs such as insulation and nebulosity at a time lag 1. The other variables are equivalent and do not contribute to add information for increase the quality of the prediction. Like Bastia case, only 3 neurons on hidden layer are necessary. Add more neurons will complicate the system without obvious improvements.

The figure 4 shows the comparison between the used of exogenous input and endogenous input (R=0.887 on top) and only endogenous input (R=0.878 on bottom). For Ajaccio, the exogenous methodology generates good results, the nRMSE is 21.54% against 22.50%, the RMSE is reduce by 52 Wh/m² (1087 Wh/m²/day vs 1139 Wh/m²/day), and the mean absolute

error by 73 Wh/m² (839 Wh/m²/day vs 912 Wh/m²/day). The simplest forecaster is the persistence and the error generated is a nRMSE = 27.07% (MAE=971 Wh/m²/day and RMSE=1422 Wh/m²/day).

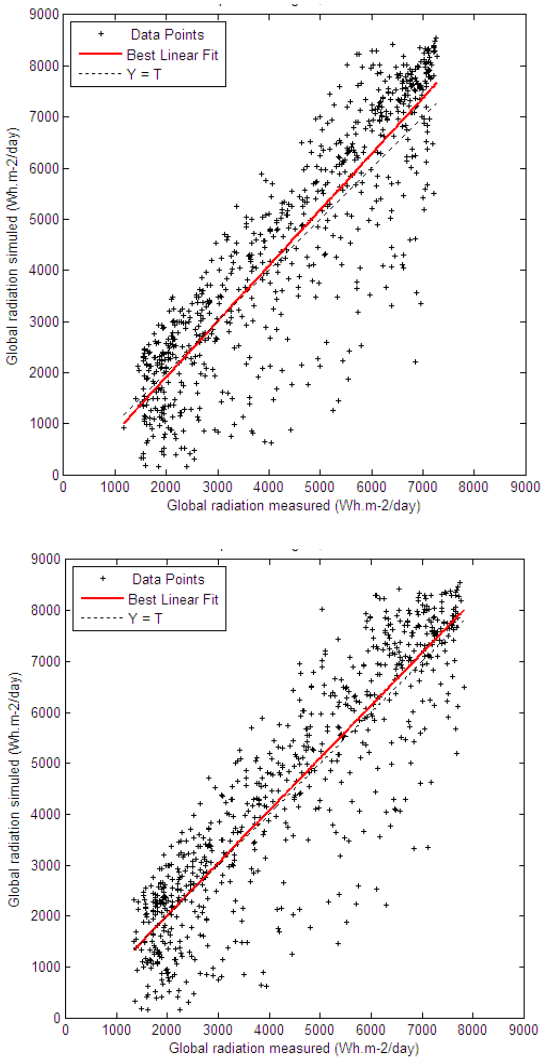

Figure 4. Comparison between the used of exogenous input and endogenous input (R=0.887 on top) and only endogenous input (R=0.878 on bottom) for the site of Ajaccio

V. CONCLUSION

We have proposed in this paper to study the contribution of exogenous meteorological data to improve a MLP network, optimized in previous studies, and to compare with different forecasting methods: a naïve forecaster like persistence, an ANN with preprocessing using only endogenous inputs and an ANN with pre-processing using endogenous. The endogenous case has been easily computed with the use of PACF which has allowed to optimize the number of lag time to consider. For the exogenous variables, we have applied a Pearson correlation coefficient method to optimize the number of considered input neurons.

Although intuitively the use of meteorological data in the input layer of the MLP can only increase the quality of prediction, the obtained results are relatively mixed. The use of exogenous data generates a decrease of nRMSE between 0.5% and 1% for the both studied locations. The absolute error (RMSE) is decreased by 52 Wh/m²/day in the simple endogenous case and 335 Wh/m²/day for the persistence forecast. Thereby we can consider that the results shown in this paper are interesting. On the site of Bastia, the use of the exogenous data on ANN inputs increases a little the prediction quality (only 0.5%). At Ajaccio, the nRMSE is improved by 1%. The last configuration based on the coupling of endogenous and exogenous data begin to be interesting for a power manager.

The next step of our work will be to study the hourly time step. Indeed it should be interesting to verify the idea that the significance of exogenous data increase when the time step of time series decreases.

ACKNOWLEDGMENT

The authors want to thank the Territorial Collectivity of Corsica for its financial support and Météo France for providing the data.

REFERENCES

[1] Kaligirou, S.A.: Artificial neural networks in renewable energy systems applications: a review. Renewable and sustainable energy review, 5, 373-401 (2001)

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

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

[4] Crone, S.F.: 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)

[5] C. Paoli, C. Voyant, M. Muselli, et M. Nivet, "Solar radiation forecasting using ad-hoc time series preprocessing and neural networks," 0906.0311, 2009.

[6] C. Voyant, M. Muselli, C. Paoli, M.L. Nivet, et P. Poggi, "Predictability of PV power grid performance on insular sites without weather stations: use of artificial neural networks," 0905.3569, Mai. 2009.

[7] 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)

[8] Jain, K., Jianchang, M., Mohiuddin, K.M.: Artificial neural networks: A tutorial, IEEE Computer, 29(3), 31–44 (1996)

[9] Hu, Y., Hwang, J.: Handbook of neural network signal processing. ISBN 0-8493-2359-2 (2002)

[10] Crone, S.F.: 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)

[11] P. Ineichen, "A broadband simplified version of the Solis clear sky model," Solar Energy, vol. 82, 2008, pp. 758-762.

[12] G. Notton, P. Poggi, et C. Cristofari, "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, vol. 47, 2006, pp. 1816-1829.