Forecasting of Three Components of Solar irradiation for Building Applications

Gilles Notton[1,*], *Cyril* Voyant[2,3], *Alexis* Fouilloy[1], *Jean Laurent* Duchaud[1] and *Marie Laure* Nivet[1]

[1]Research centre Georges Peri, University of Corsica Pasquale Paoli, 20000 Ajaccio, France
[2]Castelluccio Hospital, Radiotherapy Unit, BP 85, 20177 Ajaccio, France
[3]University of Reunion Island - PIMENT Laboratory, 15, Avenue René Cassin, BP 97715 Saint-Denis Cedex, France

Abstract. Solar energy and the concept of passive architecture and Net Zero Energy buildings are being increased. For an optimal management of the building energy, a Model Predictive Control is generally used but requires an accurate building model and weather forecast. For a more reliable modelling, the knowledge of the global solar irradiation is not sufficient; three methods, smart persistence, artificial neural network and random forest, are compared to forecast the three components of solar irradiation measured on the site with a high meteorological variability. Hourly solar irradiations are forecasted for time horizons from h+1 to h+6. The random forest method (RF) is the most efficient and the accuracy of forecasts are in term of nRMSE, from 19.65% for h+1 to 27.78% for h+6 for global horizontal irradiation, from 34.11% for h+1 to 49.08% for h+6 for beam normal irradiation, from 35.08% for h+1 to 49.14% for h+6 for diffuse horizontal irradiation. The improvement brought by the use of RF compared to the two other methods increases with the forecasting horizon. A seasonal study is realized and shows that the forecasting during spring and autumn is less reliable than during winter and summer due to a higher meteorological variability.

1 Introduction

There has been a rapid increase in solar passive architectural buildings in the last few years to achieve the net-zero energy concept. Due to the dynamic change in solar radiation, reliable energy generation forecasting is necessary for grid operation in the case of solar energy generation and also for passive solar architectural building design for the optimal thermal performance of buildings [1].

Net Zero Energy Buildings are at the frontier of the energy efficiency and renewable energy sources integration in buildings. A proper design of these buildings, taking into account their connectivity, is a key stone to reach them. Once designed their operation requires of an optimal control system called Model Predictive Control (MPC) [2].

MPC requires a model of the system, real-time controllers and weather forecasts. A lot of attention has been paid to the two first: obtaining good models for buildings and checking the robustness of the control scheme adopted. However, less attention has been paid to the weather forecasting requirements for such applications [2].

MPC is a control that employs an explicit model of the system to be controlled which is used to predict the future output behaviour. This prediction capability allows solving optimal control problems on line, where tracking error, namely the difference between the predicted output and the desired reference, is minimized over a future horizon, possibly subject to constraints on the manipulated inputs and outputs. The result of the optimization is applied according to a receding horizon philosophy: At time t only the first input of the optimal command sequence is actually applied to the system. The remaining optimal inputs are discarded, and a new optimal control problem is solved at time t+ 1. There is extensive literature covering this field [3-4].

Different solar systems require different solar forecasts. For solar concentrating systems the normal beam incident irradiance must be forecast, whereas for non-concentrating systems primarily the global irradiance on a tilted surface is required. Efficient and renewable thermal comfort management inside buildings can be considered as a non-concentrating solar system.. For a more reliable modelling, the knowledge of the horizontal global solar irradiation is not sufficient; it is necessary to know the main solar components (normal beam and horizontal diffuse) mainly for two reasons:

1. To know the solar irradiation incident on each surface whatever the orientation and inclination are; the global irradiation cannot be measured in each building area and the received solar irradiation can be computed from beam, diffuse and global components;

[*] Corresponding author: gilles.notton@univ-corse.fr

2. To calculate more accurately the heat exchanges (diffuse and beam solar irradiance being different influences).

Don’t forget that the forecasting tool can be used for estimating the future thermal but also electrical production (by photovoltaic systems).

The choice of the forecasting methods depends on the forecasting time horizon and the temporal resolution as seen in Fig. 1.

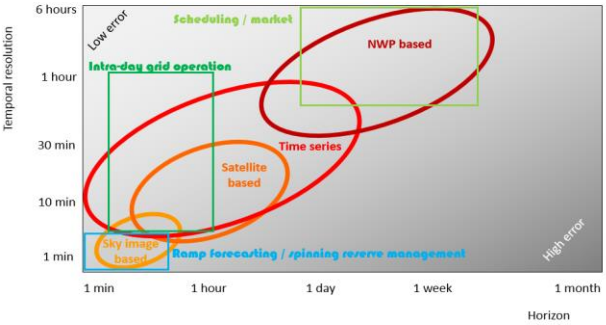

Fig. 1. Forecasting time horizon versus temporal resolution [5].

The existing methods can be classified in four different sets [5-6]:

- Time series based methods: this set holds approaches based on statistical models solely ground on past measurements;
- NWP (Numerical Weather Forecast) based methods: this set holds approaches based on weather forecasts provided by a specialized provider;
- Satellite imagery based methods: this set holds approaches based on images of the earth taken by satellite;
- Sky image based methods: this set holds approaches based on observations of cloud cover from the ground with an in situ camera.

The objective of this paper being to forecast hourly solar irradiation from 1 to 6 hours ahead, we concentrated our attention on methods developed for now-casting. The most efficient existing methods for such a forecast for these time horizons are time series analysis, artificial intelligence methods and deep learning methods [7].

If forecasting methods are largely developed for global solar irradiation (GHI), those developed for Beam Normal Irradiation (BNI) are few in number and those for Diffuse Horizontal Irradiation component (DHI) are, to the best of our knowledge, practically non-existent.

2 Pre-processing

2.1 Cleaning and filtering

An automatic quality control used in the frame of GEOSS project (Group on Earth Observation System of System) [8] is applied to the solar data. Before introducing the solar data into the machine learning process, the data must be cleaned and filtered. The data are filter out in order to remove night hours. The data near sunset and sunrise are generally not reliable (instrumental errors especially due to the cosine response and mask effect of the surrounding mountains), a pre-processing operation is applied based on the solar elevation: solar radiation data with a solar elevation is lower than 10° are removed [9]. 3 years of hourly data have been used in this study. After cleaning and filtering the total number of hourly data for each solar component (global, beam and diffuse) is 10559 (about 60% of the data were not used (2% for outliers data and 58% for sun height less than 10°).

2.2 Stationarization

Machine learning methods are efficient tools for forecasting time series with a stationary behavior. To make solar irradiation data stationary and to separate the climatic effect and the seasonal effects, the solar data are generally transformed in non-dimensional variables called "clearness index", and denoted *kt*, given by the ratio of the solar radiation on the earth to that outside the atmosphere and defined by equation (1) [10]:

$$kt = \frac{GHI}{G_0} \tag{1}$$

with GHI the global irradiation at the earth's horizontal surface for a given location and G_0 the global solar radiation on the top of atmosphere. It is the clearness index series kt that induces randomness, caused by the diversity of atmospheric components (dust, aerosols, clouds motion, and humidity) on the solar irradiation measured at earth ‘surface.

The extraterrestrial irradiation G_0 can be efficiently replaced by the clear sky solar irradiation G_{CS} taking into account the climatic conditions of the meteorological site; thus the clearness index is replaced by the clear sky index $k_{g,cs}$ defined by:

$$k_{g,cs} = \frac{GHI}{GHI_{CS}} \tag{2}$$

with GHI_{CS} the Global Horizontal solar Irradiation in clear sky conditions.

For the other components of solar radiation (direct and diffuse), similar index can be defined such as k_b [11] and k_d [12]:

$$k_{b,cs} = \frac{BNI}{BNI_{CS}} \text{ and } k_{d,cs} = \frac{DHI}{DHI_{CS}} \tag{3}$$

With BNI, the Beam Normal Irradiation and DHI, the Diffuse Horizontal Irradiation. BNI is often called DNI (Direct Normal Irradiation) but in this paper Direct is replaced by Beam for avoiding confusion between Direct and Diffuse.

Various models of clear sky solar irradiations are available in the literature which differ from each other mainly in the inputs needed by each model. The most widely used clear sky models are the Solis model developed by Mueller *et al* [13] and simplified by Ineichen [14], the European Solar Radiation Atlas (ESRA) model [15] and the Reference Evaluation on Solar Transmittance 2 (REST2) model [16].

Thus, we decided to use the simplified Solis clear sky model [14], it allows to calculate GHICS, DNICS and to deduce DHICS by Eq. (4).

$$GHI_{CS}=DHI_{CS}+BNI_{CS} \text{ x } \cos \theta_z \tag{4}$$

θ_Z is the zenithal angle calculated at the middle of the hour [10]. For a better accuracy, the monthly mean values of aerosol optical depth and water vapour column are used as inputs in the Solis model; these averaged values were calculated from data measured at the Pic du Midi site (180 km from Odeillo, our study site presented below and same altitude) available in the data basis AERONET (AErosol Robotic NETwork) for four years (2001-2004) [17].

This clear sky model was validated for each month by comparison with experimental solar radiation data (GHI, BNI, DHI) measured in clear sky conditions. For illustration purpose, experimental and modelled solar irradiances by clear sky are plotted in Fig. 2 for 1 day in April and in September.

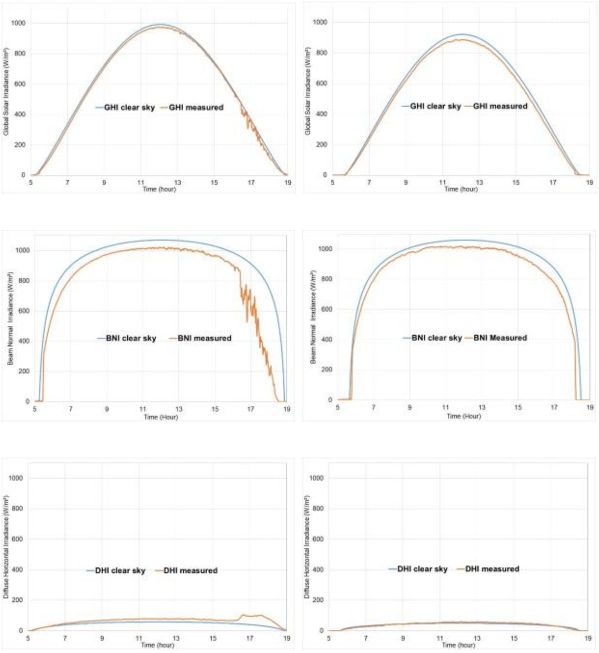

Fig. 2. Experimental and modelled solar irradiance curves in clear sky conditions for April (left) and September (right), GHI (top), BNI (middle) and DHI components (below) (hour in true solar time).

A good concordance is noted between modelled and experimental curves; the diffuse solar irradiance by clear sky is always lower to the irradiance by partially or cloudy skies because this component is minimum when the sky is clear and maximum in cloudy conditions.

2.3 Choice of input data

The purpose of this paper is to predict the future solar irradiation (at different time horizons) based on the past observed data i.e. mathematically:

$$\hat{X}_{t+h} = f(X_t, X_{t-1}, X_{t-2}, X_{t-3}, \dots, X_{t-n}) \tag{5}$$

A variable X with the symbol $\widehat{\ }$ represents a forecasted data, without this symbol, X is a measured data. The solar data at future time step (t+h) X^_(t+h) is forecasted from the observed data X measured at the times (t, t-1…, t-n); thus, the objective is to determine the value of n i.e. the dimension of the input matrix; to do it, an auto mutual information method [18] is used. The auto mutual information is a property of the time series, depends on each dataset and is characteristic to the degree of statistical dependence between X_{t+h} and X_{t-i} with 0≤i≤n.

Another preprocess called k-fold sampling is used with the dataset [19]: it consists in dividing randomly the data set into a training data set (80%) and a test data set (20%); this process is repeated k times and the value of the reliability metrics given in this paper are the average value on the k-fold. Here k is taken equal to 10. Thus, the results are independent of the set of data used for the training because using only one data set (with its own statistical particularities) can reduce the robustness of the conclusions.

3 The meteorological site

The meteorological data GHI, BNI and DHI were provided by the PROMES laboratory (CNRS UPR 8521), located in south of France at Odeillo (Pyrénées Orientales, France, 42°29 N, 2°01 E, 1550 m asl), the station is located in the mountains, at about 100 km from the Mediterranean sea and presents often a high nebulosity (Fig. 3). The solar data are measured and stored with a 1 minute time interval of measurement.

This meteorological station being situated in altitude, the climate is very perturbed, the rainfall continues to be present during the driest months and this station is classified according to the Köppen-Geiger classification in Cfb (i.e. hot temperate climate without dry period and temperate summer). Consequently, the variability of the solar radiation is high and its forecasting is all the more difficult to realize. This variability of the solar irradiation can be quantified thanks to various parameters, Voyant *et al* [20] tested some of them and deduced that the more significant was the mean absolute log return; the mean absolute log return was calculated for Odeillo and are, for the three components:

- 0.6109 for GHI
- 0.9945 for BNI
- 0.4732 for DHI

It appears that the variability of BNI is higher than for GHI and DHI and should be more difficult to predict.

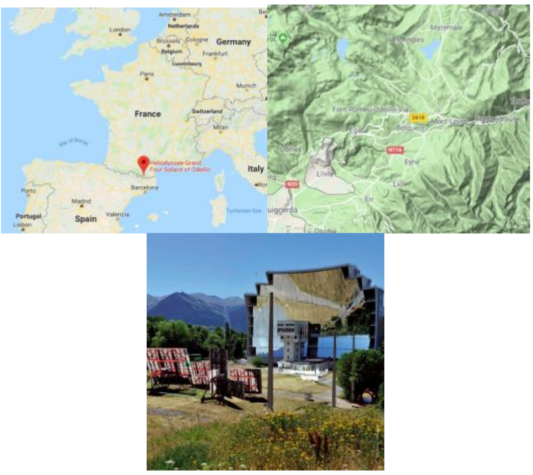

Fig. 3. Meteorological site (Odeillo)

4 Brief description of the forecasting methods

Each forecasting methodology is described shortly in this paragraph: Scaled (or Smart) persistence, Artificial Neural Network and Random Forest. The first method uses a naïve model, easy to implement and requiring no training step i.e. no historical data set; it is generally used as a reference model in view to compare it with more sophisticated models in terms of accuracy. The second and third models belong to the family of machine learning methods, more complex to implement but generally more reliable too.

4.1 Smart-Persistence

The persistence model, the simplest forecasting model, assumes that the future value is same as the previous one (Eq. (6)). Persistence forecast accuracy decreases significantly with forecasting horizon [21].

$$\hat{X}_{t+h} = X_t \text{ with X=GHI, BNI or DHI} \qquad (6)$$

The smart persistence is an improved version of the persistence one taking into account the diurnal solar cycle: the clear sky solar radiation profile over the day is used [20]:

$$\hat{X}_{t+h} = X_t.\frac{X_{CS_{t+h}}}{X_{CS_t}} \text{ with X=GHI, BNI or DHI} \qquad (7)$$

This smart persistence model is applied in this paper to the three solar components and used mainly as a reference model.

4.2 Artificial Neural Network (ANN): Multi-layer Perceptron

It is the more known and used machine learning method for forecasting purposes. ANN is a nonlinear approximator implementing a simple pattern of elements interconnected one another. The ANN used here is the Multilayer Perceptron (MLP) with feed-forward back propagation often used in solar forecasting estimation and prediction [22-23]. The hidden layer receives input data and send an output signal to the output layer. A neuron receives signals from other previous neurons or input data unidirectionally for a feed-forward MLP configuration (Fig 4).

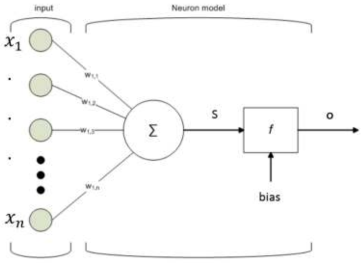

Fig. 4. Neuron model [24].

For the k-th neuron of the hidden layer, a weight $w_{k,j}$ taken various values determined during the training phase, is linked to each input x_j; An activation function f is applied to the weighted sum $(S = \sum_{j=1}^{n} w_{k,j} x_j)$ for calculating an output if this sum exceeds a given bias $f(\sum_{j=1}^{n} w_{k,j} x_j + bias)$. This output is then distributed to other next neurons. A sigmoid function for hidden layers and a linear one for the output layer were taken as activation functions. For the regression of the time series X_t, the mathematical expression for a MLP with one hidden layer of m neurons, one output neuron and n input variables is a function described by:

$$\hat{X}_{t+h} = \sum_{k=1}^{m} \omega_{*k}\cdot(f\left(\sum_{j=1}^{n} \omega_{k,j}.X_{t-j+1} + b_k\right) + b_o \qquad (8)$$

With X the input vector (n x 1) of clear sky index $k_{g,cs}$,$k_{b,cs}$ or $k_{d,cs}$, $\hat{X}_{t+h}$ the output value corresponding to the predicted values of the model at horizon t+h, b_k and b_o the bias related to the hidden neuron k and to the output, and $\omega_{k,j}$ the weights between the j-th measured input and the k-th hidden neuron. f is the transfer function of the hidden neurons, ω_{*k} the weight between the output and k-th hidden neurons. The optimization of the MLP is made by the Levenberg–Marquardt learning algorithm: several configurations with a different number of hidden nodes in the hidden layer are tested (the number of hidden nodes varying between 3 and n+2) and the most efficient is selected.

Once the clear sky index is forecasted $\hat{k}_{t+h}$, the value of the forecasted solar irradiation, $(\widehat{GHI}_{t+h}, \widehat{BNI}_{t+h}$ or $\widehat{DHI}_{t+h})$ is obtained in multiplying $\hat{k}_{t+h}$ by the calculated clear sky irradiation ($(GHI_{CS_{t+h}}, BNI_{CS_{t+h}}$ or $DHI_{CS_{t+h}})$.

4.3 Random Forest

The random forest method belongs to the regression tree (RT) family, it is an improved model of bagging regression tree [7-25].

The binary RT method consists in an iterative split of the data into two groups according to some thresholds and rules [26]; it constructs a set of decision rules on the predictor variables [27] in view to partition the data into smaller groups with binary splits based on a single predictor variable [25]. For RT, the predictor and the threshold or grouping are chosen for maximizing the homogeneity of the corresponding values in the resulting groups. The homogeneity is calculated as the sum of variance of data within each groups, this variance being minimized [26]. Each group is then divided in two subgroups and so on. For each final group (called a leaf), the predicted value is the mean of the values belonging to the leaf. The procedure grows maximal trees and then techniques such as cross validation are used to prune the overfitted tree to an optimal size [28].

It appeared that the output error obtained by a single RT is due to the specific choice of the training data set [25]. Thus, for solving this problem, Breiman [29] proposed to grow several trees and to average their predicted values to yield a more stable final prediction. To avoid having to use too much data for creating several independent trees, samples of data are chosen randomly in the original data set. This method is called bagging (contraction of bootstrap aggregating). The complexity

of the model is tuned with the number of bagged trees, and each individual tree is not pruned [25].
The bagged trees are not statistically independent and the variance of their mean cannot be indefinitely decreased [25] because they are built from the same data set. To reduce this problem, Breiman [30] added a randomization step to bagging, each split of each bagged RT is built in a random subset of the predictors [26]. Numerous trees are growing creating a random forest (RF).
Each subset of data used to grow the tree is replaced in the dataset before growing the next tree. This randomization gives more robustness to the model and decreases the risk of overfitting. At the end, the responses are aggregated to make the forecast. For more precisions about the regression trees based models, the reader can refer to the references given in this section.
A comparison of these three methods (RT, Bagged RT and RF) is given in Prasad et al [25] in term of strengths and limitations.
RF is recognized as one of the most effective machine learning models for forecasting and will be used in this paper.

5 Results

5.1 Statistical index for accuracy evaluation

In this paper, we use four error metrics:

- The mean absolute error (MAE) is appropriate for applications with linear cost functions, i.e. situations where the costs resulting from a poor forecast are proportional to the forecast error:

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} \left| \hat{X}_{t_i} - X_{t_i} \right| \tag{9}$$

- The root mean square error (RMSE) is more sensitive to important forecast errors, and hence is suitable for applications where small errors are more tolerable and larger errors lead to costs that are disproportionate, as in the case of utility applications, for example. It is probably the reliability factor that is the most widely used:

$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} \left(\hat{X}_{t_i} - X_{t_i} \right)^2} \tag{10}$$

These errors are then normalized, the mean value of irradiation is generally used as reference:

$$nRMSE = \frac{RMSE}{\bar{X}} \tag{11}$$

$$nMAE = \frac{MAE}{\bar{X}} \tag{12}$$

5.2 Auto-mutual analysis results

The choice of the number of endogenous inputs was realized by the auto-mutual information method which determines the number of previous solar irradiation data used to predict the future solar irradiation at time h+1 to h+6. The auto-mutual method showed that the number of inputs (n in equation 5) for predicting GHI is 6, for BNI 7 and for DHI 10.

5.3 Annual performances

5.3.1 Global Horizontal Solar Irradiation: GHI

Table 1 gives the values of the performance metrics calculated on the test data set (RMSE and MAE are given in Wh.m^{-2}) for GHI.

Table 1. Performance metrics for GHI (in bold the best predictor for each horizon and each of error metric)

	Model	h+1	h+2	h+3	h+4	h+5	h+6
SP	RMSE	97.7	132.4	157.1	176.5	193.1	202.7
	nRMSE	0.217	0.294	0.348	0.391	0.428	0.449
	MAE	**57.0**	80.8	98.7	112.8	124.6	130.8
	nMAE	**0.126**	0.179	0.219	0.250	0.276	0.290
ANN	RMSE	101.8	126.6	141.9	150.3	154.8	157.3
	nRMSE	0.226	0.281	0.315	0.333	0.343	0.348
	MAE	72.9	91.0	106.8	112.6	117.6	118.6
	nMAE	0.162	0.202	0.237	0.250	0.260	0.263
RF	RMSE	**88.6**	**104.6**	**116.2**	**119.9**	**124.6**	**125.4**
	nRMSE	**0.196**	**0.232**	**0.257**	**0.266**	**0.276**	**0.278**
	MAE	61.5	**73.6**	**84.1**	**86.5**	**90.2**	**91.1**
	nMAE	0.136	**0.163**	**0.186**	**0.192**	**0.200**	**0.202**

As the ranking of the model is almost always identical from a RMSE point of view or MAE point of view, (excepted for h+1) we only present in Fig. 5 the results in term of RMSE and nRMSE expressed in percentage.

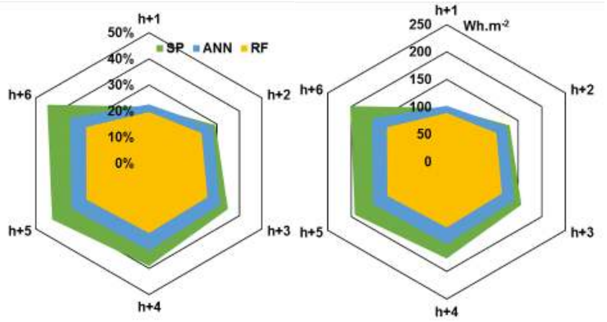

Fig. 5. Comparison of forecasting models for various horizon in term of nRMSE (left) and RMSE (right) for hourly GHI.

The smart persistence, a naive model, was used as a reference, this model has a good RMSE and MAE for a time horizon h+1 but its performances decrease rapidly with the time horizon. We note that the gap in term of performances between ANN and RF increases with the time horizon, from 2.92% at h+1 to 7.07 % at h+6 in nRMSE value.

5.3.2 Beam Normal Solar Irradiation: BNI

Table 2 gives the values of the performance metrics computed on the test data set (RMSE and MAE are given in Wh.m^{-2}) for BNI. The results in term of RMSE and nRMSE are presented in Fig. 6 for BNI.
The forecasting of BNI is more difficult and the performances of the models are less satisfying than with GHI, this is because BNI is more sensitive to meteorological conditions and because the beam radiation intensity is more rapid and of a greater

https://doi.org/10.1051/e3sconf/201911105012

magnitude as suggested by the higher value of the mean absolute log return characterizing the intermittency degree. One more time, RF is the most performant model whatever the time horizon is and the gap in term of nRMSE between ANN and RF passes from 4.11% at h+1 to 12.8% at h+6 justifying even more the use of RF for BNI forecasting than for GHI one.

Table 2. Performance metrics for BNI (in bold the best predictor for each horizon and each of error metric)

	Model	h+1	h+2	h+3	h+4	h+5	h+6
SP	RMSE	207.9	287.6	338.9	378.2	412.7	434.5
	nRMSE	0.374	0.518	0.610	0.680	0.742	0.782
	MAE	**125.2**	187.1	230.3	266.0	298.2	317.7
	nMAE	**0.226**	0.337	0.414	0.478	0.536	0.571
ANN	RMSE	212.3	270.1	297.7	321.6	337.0	344.0
	nRMSE	0.382	0.486	0.536	0.578	0.606	0.619
	MAE	168.1	223.3	244.9	274.6	283.7	299.2
	nMAE	0.303	0.402	0.441	0.494	0.510	0.538
RF	RMSE	**189.5**	**223.7**	**242.5**	**254.1**	**265.4**	**272.8**
	nRMSE	**0.341**	**0.403**	**0.436**	**0.457**	**0.477**	**0.491**
	MAE	141.6	**175.2**	**194.2**	**207.2**	**216.7**	**226.2**
	nMAE	0.255	**0.315**	**0.349**	**0.373**	**0.390**	**0.407**

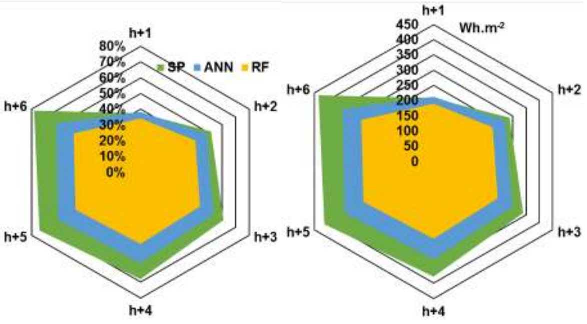

Fig. 6. Comparison of forecasting models for various horizon in term of nRMSE (left) and RMSE (right) for hourly BNI.

5.3.3 Diffuse Horizontal Solar Irradiation: DHI

Table 3 gives the values of the performance metrics calculated on the test data set (RMSE and MAE are given in Wh.m^{-2}) for DHI. The results in term of RMSE and nRMSE are presented in Fig. 7 for DHI.

The metrics values are of the same order of magnitude as for BNI excepted for the smart persistence. The smart persistence presents bad results because the daily profile of the DHI by clear sky (taken into account in this model) is not as well defined as for BNI or GHI. As seen for the two other components, the gap in term of performances between RF and ANN increases with the forecasting time horizon from 5.91% for h+1 to 14.74% for h+6. The use of a forecaster using random forest method for the DHI component gives very correct results.

Table 3. Performance metrics for DHI (in bold the best predictor for each horizon and each of error metric)

Models	metrics	h+1	h+2	h+3	h+4	h+5	h+6
SP	RMSE	87.3	96.5	104.0	110.3	114.2	115.3
	nRMSE	0.636	0.697	0.751	0.796	0.824	0.833
	MAE	76.3	80.5	84.1	86.9	88.9	89.6
	nMAE	0.551	0.582	0.607	0.628	0.641	0.647
ANN	RMSE	56.8	70.3	79.7	84.6	86.2	88.5
	nRMSE	0.410	0.507	0.575	0.611	0.622	0.639
	MAE	40.6	52.3	60.5	64.3	65.1	66.4
	nMAE	0.293	0.377	0.437	0.464	0.470	0.479
RF	RMSE	**48.5**	**58.0**	**63.1**	**66.0**	**67.6**	**68.1**
	nRMSE	**0.351**	**0.419**	**0.456**	**0.477**	**0.488**	**0.491**
	MAE	**33.6**	**41.0**	**44.7**	**47.7**	**48.9**	**50.1**
	nMAE	**0.243**	**0.296**	**0.323**	**0.344**	**0.353**	**0.357**

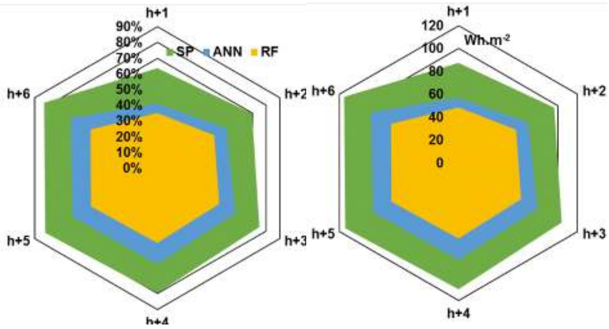

Fig. 7. Comparison of forecasting models for various horizon in term of nRMSE (left) and RMSE (right) for hourly DHI.

5.3.4 Comparison

It is impossible to compare the performances of the models according to the solar component in term of absolute value of RMSE (or MAE) because the daily DHI, BNI and GHI are very different. Thus, we plotted in Fig. 8, a comparison of the performances in term of relative RMSE for the three models. For each forecasting horizon, the three first histograms correspond to the smart persistence, then ANN, then RF.

As previously underlined, GHI is forecasted with a better accuracy compared with BNI and DHI. It is probably due to the fact that in GHI, the two components, DHI and BNI, have compensating effects (when diffuse increases, beam decreases) and the speed of variation of GHI is less important than for DNI. Concerning the performances of the forecasters for GHI, the relative low reliability is probably due to clear sky index which is higher than 1 and can reach high values.

The accuracy obtained for BNI and DHI using ANN and RF are of the same order of magnitude; in contrast, the smart persistence is not adapted at all for forecasting DHI for the reason previously explained.

With SP and ANN methods, DHI and BNI are predicted with a nRMSE nearly twice as high than for GHI, with random forest method this difference is reduced when the forecasted horizon increases and for (h+6) the accuracy obtained for DHI and BNI prediction is the same than for GHI prediction.

It seems that random forest have for all these reasons is the best predictor.

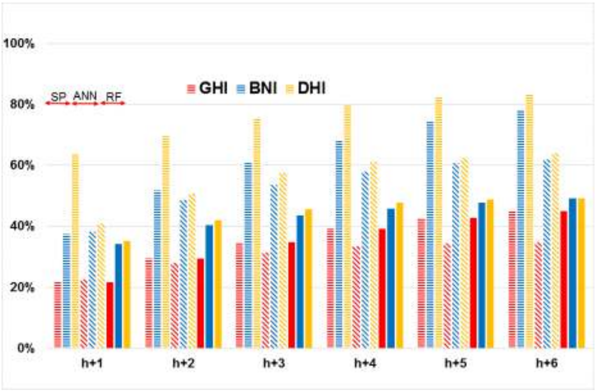

Fig. 8. Comparison of forecasting models for the three components.

5.4 Seasonal performances

It is interesting to observe the performances of the model season by season and thus to observe the impact of the meteorological conditions (and of its variability) on the accuracy of each model. As it is difficult to compare the performances of the models according to the season in term of absolute value of RMSE (or MAE) because according to the season the average hourly irradiation are different, only nRMSE values are given in Tables and Figures.

5.4.1 Global Horizontal Solar Irradiation: GHI

The performance metrics for GHI are presented in Table 4. and in Fig 9.

Table 4. nRMSE for GHI by season for the three forecasting models: Smart Persistence, ANN MLP, Random Forest.

		h+1	h+2	h+3	h+4	h+5	h+6
SP	Winter	0.194	0.266	0.316	0.355	0.385	0.409
	Spring	0.292	0.375	0.440	0.479	0.522	0.541
	Summer	0.174	0.233	0.275	0.313	0.328	0.347
	Automn	0.238	0.321	0.378	0.427	0.477	0.507
ANN	Winter	0.204	0.267	0.303	0.325	0.355	0.360
	Spring	0.261	0.328	0.366	0.380	0.392	0.395
	Summer	0.163	0.214	0.242	0.255	0.273	0.265
	Autumn	0.240	0.307	0.350	0.379	0.404	0.411
RF	Winter	0.181	0.211	0.238	0.253	0.258	0.271
	Spring	0.232	0.276	.0300	0.304	0.314	0.314
	Summer	0.156	0.177	0.194	0.204	0.213	0.207
	Autumn	0.211	0.250	0.286	0.300	0.320	0.315

Whatever the forecasting horizon and the model are, the best results are obtained for summer then for winter; in summer, the occurrence of clear sky irradiations is higher and in winter, the occurrence of cloudy ones too; the solar irradiation during intermediate seasons are more difficult to forecast. In spring and secondly in autumn, the difference in performance term between the three models is the highest, thus when the variability of the GHI is high, the utilization of a more complex forecasting tool is necessary.

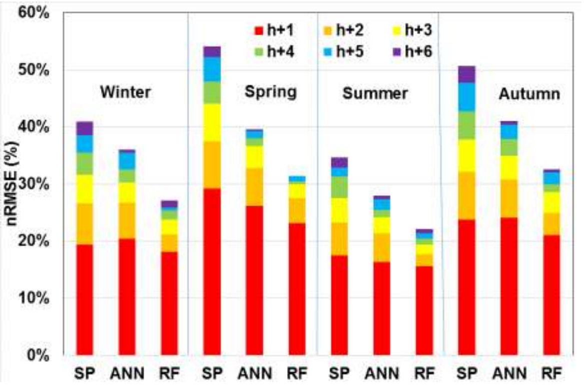

Fig. 9. Comparison of forecasting models performances for various seasons in term of nRMSE for hourly GHI.

5.4.2 Beam Normal Solar Irradiation: BNI

The nRMSE for BNI are presented in Table 5 and Fig. 10.

Table 5. nRMSE for BNI by season for the three forecasting models: Smart Persistence, ANN MLP, Random Forest.

		h+1	h+2	h+3	h+4	h+5	h+6
SP	Winter	0.444	0.515	0.557	0.591	0.617	0.648
	Spring	0.548	0.714	0.820	0.900	0.978	1.00
	Summer	0.365	0.471	0.537	0.587	0.616	0.626
	Automn	0.533	0.681	0.755	0.811	0.871	0.913
ANN	Winter	0.318	0.407	0.457	0.499	0.528	0.543
	Spring	0.483	0.619	0.685	0.735	0.777	0.800
	Summer	0.298	0.377	0.423	0.442	0.460	0.470
	Autumn	0.474	0.603	0.651	0.704	0.737	0.755
RF	Winter	0.255	0.317	0.351	0.389	0.394	0.405
	Spring	0.441	0.526	0.562	0.598	0.625	0.632
	Summer	0.284	0.324	0.354	0.368	0.382	0.383
	Autumn	0.421	0.488	0.530	0.556	0.586	0.601

The same remarks can be made than for GHI concerning the more important difficulty to forecast BNI in spring and autumn. The gap between the worst and the best model is higher than in GHI case and it appears clearly that SP is not adapted to a BNI forecasting.

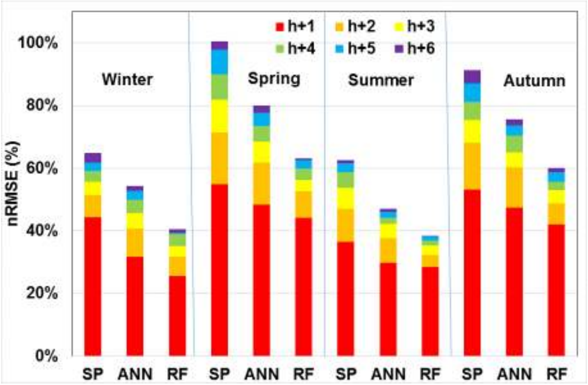

Fig. 10. Comparison of forecasting models performances for various seasons in term of nRMSE for hourly BNI.

5.4.3 Diffuse Horizontal Solar Irradiation: DHI

The nRMSE for DHI are presented in Table 6 and Fig. 11.

Table 6. nRMSE for DHI by season for the three forecasting models: Smart Persistence, ANN MLP, Random Forest.

		h+1	h+2	h+3	h+4	h+5	h+6
SP	Winter	0.682	0.755	0.813	0.862	0.887	0.901
	Spring	0.550	0.616	0.681	0.737	0.770	0.778
	Summer	0.653	0.707	0.741	0.766	0.789	0.793
	Automn	0.657	0.732	0.777	0.816	0.841	0.848
ANN	Winter	0.418	0.544	0.628	0.670	0.674	0.690
	Spring	0.402	0.489	0.554	0.588	0.603	0.619
	Summer	0.363	0.445	0.498	0.524	0.532	0.551
	Autumn	0.431	0.563	0.608	0.646	0.659	0.676
RF	Winter	0.362	0.433	0.497	0.531	0.544	0.550
	Spring	0.344	0.407	0.448	0.468	0.484	0.486
	Summer	0.312	0.377	0.396	0.401	0.408	0.431
	Autumn	0.365	0.436	0.459	0.485	0.507	0.510

As noted previously, the smart persistence is really a bad predictor for DHI, it is probably due to the fact that DHI by clear sky, DHI_{CS}, is lower than DHI by cloudy sky and thus $k_{d,CS}$ is higher than 1 and can vary in a large range, perturbing this method application. For the other methods, the accuracy of the prediction stays in a correct range.

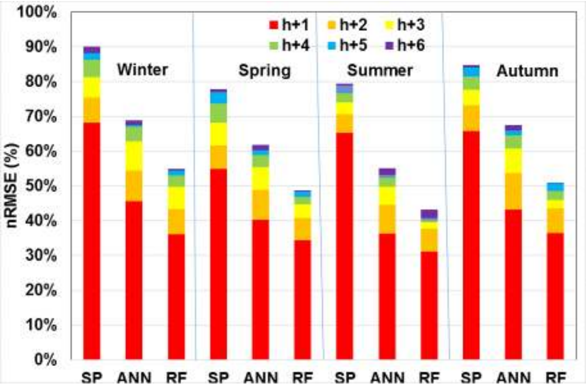

Fig. 11. Comparison of forecasting models performances for various seasons in term of nRMSE for hourly DHI.

6 Conclusion

Three forecasting methods, smart persistence, artificial neural network (multilayer Perceptron) and random forest, were compared and tested on solar data measured in a meteorological site presenting a high variability. The objective was to predict the hourly solar irradiation for a time horizon from h+1 to h+6; these methods were applied on the three solar components: horizontal global, normal beam and horizontal diffuse.

It appears that random forest method allows to predict these three components with a good accuracy:

- nRMSE from 19.65% for h+1 to 27.78% for h+6 for GHI;
- nRMSE from 34.11% for h+1 to 49.08% to h+6 for BNI;
- nRMSE from 35.08% for h+1 to 49.14% for h+6 for DHI.

The random forest method gives the best results and the improvement due to the utilization of RF in comparison of ANN is even more important that the forecasting horizon increases; the improvement in term of nRMSE ($nRMSE_{RF}$-$nRMSE_{SC}$) due to a RF use compared to a SP use is:

- For GHI forecasting, +2.02% for h+1 to +17.13% for h+6;
- For BNI, +3.3% for h+1 to 28.36 for h+6;
- For DHI, +28.56% for h+1 to 34.13% to h+6.

A seasonal study was realized and showed that the forecasting during spring and autumn is more difficult to realize than during winter and summer due to a higher variability of the climate on these periods.

BNI and DHI are more complicated to predict than GHI: the BNI and DHI components are more sensitive to meteorological conditions than GHI one (for GHI, the two components, DHI and BNI, have compensating and smoothing effects (when diffuse increases, beam decreases) and the variability of BNI is more important with a higher speed of variation and a higher amplitude. For DHI, the fact that the clearness index is higher than 1 and can reach high values (contrary to BNI and GHI with a clearness index between 0 and 1) perturbs the forecasting process.

Acknowledgement

The authors thank the CNRS UPR PROMES 8521 laboratory in Odeillo (Pyrénées Orientales, France) for the solar data.

References

1. A.Naveen Chakkaravarthy, M.S.P. Subathra, P. Jerin Pradeep, N. Manoj Kumar. J. Renewable. Sustainable Energy. **10**, 035103 (2018)
2. R. Enriquez, M.J. Jimenez, M. del Rosaio Heras, Energy Procedia **91**, 1024-1032 (2016)
3. E. Camacho, C. Bordons. *Advanced Textbooks and Signal Processing*, Springler-Verlag (2007)
4. A. Bemporad, M. Morari. *Robutness in Identification and Control,* Springer-Verlag (1999).
5. M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz. Rene. Sustain. Energy Rev **27**, 65–76 (2013)
6. G. Notton, C. Voyant. Advances in Renewable Energies and Power Technologies, Elsevier Science. ISBN 978-012-8131855
7. C. Voyant, G. Notton, S. Kalogirou, M.L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Renew. Energy **105,** 569-582 (2017)
8. Global Earth Observation System of Systems (GEOSS). www.earthobservations.org/geoss.php, accessed May 2017.

9. P. Lauret, C. Voyant, T. Soubdhan, M. David, P. Poggi, Sol Energy **112**, 446–457 (2015).
10. M. Iqbal, An introduction to solar radiation, Academic Press, Canada, 1983.
11. A. Kaur, L. Nonnenmacher, H.T.C. Pedro, F.M. Coimbra, Ren. Energy **86**, 819-830 (2016).
12. J. Hofierka, M. Súri, Proc. Open source GIS - GRASS users conference 2002, Trento, Italy, 11-13 September 2002, 1-19 (2002).
13. R. Mueller, K. Dagestad, P. Ineichen, M. Schroedter-Homscheidt, S. Cros, D. Dumortier, Remote Sensing Environ. **91**, 160–174 (2004).
14. P. Ineichen, Sol. Energy **82**, 758-762 (2008).
15. C. Rigollier, O. Bauer, L. Wald, Sol. Energy **68**, 33–48 (2000).
16. C.A. Gueymard, Sol. Energy **82**, 272–285 (2008).
17. NASA, Goddard Space Flight Center. AERONET database. http://aeronet.gsfc.nasa.gov, accessed April 2017.
18. D. Huang, T.W.S. Chow, Neurocomputing **63,** 325–343 (2005).
19. T.S. Wiens, B.C. Dale, M.S. Boyce, G.P. Kershaw, Ecol Model **212**, 244–255 (2008).
20. C. Voyant, T. Soubdhan, P. Lauret, M. David, M. Muselli, Energy **90**, 671–679 (2015).
21. R. Huang, T. Huang, R. Gadh, N. Li, Proc. Smart Grid 2012 IEEE Third International Conference Communications, Tainan, Taiwan, 5-8 November 2012.
22. S. Kalogirou, Renew. Sustain. Energy Rev. **5**, 373-401 (2001).
23. A. Mellit, A.M. Pavan, Sol. Energy **84**, 807–821 (2010).
24. K. Dahmani, R. Dizene, G. Notton, C. Paoli, C. Voyant, M.L. Nivet, Energy **70**, 374-381 (2014).
25. A.M. Prasad, L.R. Iverson, A. Liaw, Ecosystems **9**, 181–199 (2006).
26. M. Zamo, O. Mestre, P. Arbogast, O. Pannekoucke, Sol. Energy **105**, 792–803 (2014).
27. L. Breiman, J. Freidman, R. Olshen, C. Stone, Classification and regression trees, Belmont, Canada, Wadsworth, 1984.
28. T.M. Therneau, E.J. Atkinson, Technical report **61**, Rochester (MM): Mayo Clinic, (1997).
29. L. Breiman, Machine Learn. **26**, 123–140 (1996).
30. L. Breiman, IMS Wald Lecture 2, (2017).