SASEC2015
Third Southern African Solar Energy Conference
11 – 13 May 2015
Kruger National Park, South Africa

INFLUENCE OF GLOBAL SOLAR RADIATION TYPICAL DAYS ON FORECASTING MODELS ERROR

Soubdhan T.[*], Voyant C., Lauret P.
*Author for correspondence
Laboratory LARGE,
University of Antilles and Guiana,
Guadeloupe, France
E-mail: ted.soubdhan@univ-ag.fr

ABSTRACT

In this work, we have led an analysis of different global solar radiation forecasting models errors according to the global solar radiation variability.
Different predictions models were performed such as machine learning techniques (Neural Networks, Gaussian processes and support vector machines) in order to forecast the Global Horizontal solar Irradiance (GHI). We also include in this study a simple linear autoregressive (AR) model as well as two naïve models based on persistence of the GHI and persistence of the clear sky index (denoted herein scaled persistence model).
The models are calibrated and tested with data from three French islands: Corsica (42.15°N ; 9.08°E), Guadeloupe (16.25°N ; 61.58°W) and Reunion (21.15°S ; 55.5°E). Guadeloupe and Reunion are located in a subtropical climatic zone whereas Corsica is in a tempered climatic zone hence, the global solar radiation variation differs significantly.
The output error of the different models was quantified by the normalized root mean square error (nRSME).
In order to quantify the influence of the global solar radiation variability on the forecasting models error we performed a classification of typical days. Each class of day is defined by a global solar radiation variability rate. For each class and for each location, forecasting models were performed and the error was quantified.
With this analysis, global solar radiation forecasting models can be selected according to the location, the global solar radiation fluctuations and hence the meteorological conditions.

1.INTRODUCTION

Large and frequent variations of solar radiation can be observed in tropical climates with amplitudes reaching 800 W/m² and occurring within a short time interval, from few seconds to few minutes, according to the geographical location. Such fluctuations can be due for example to the dynamic of clouds which can be very complex and depend on cloud type, size, speed and spatial distribution and, more generally, due to some specific local meteorological conditions.

Thus, solar energy forecasting, a process used to predict the amount of solar energy available in the current and near terms, might be a difficult task. Some of the best predictors found in literature are Autoregressive and moving average (ARMA) [5,7,8], Bayesian inferences [9,10], Markov chains [11], k-Nearest-Neighbors predictors [12] or artificial intelligence techniques as Artificial Neural Network (ANN) [9-11]. Although these methodologies are potentially good in many areas, we observed in our previous studies on global radiation prediction [9,13,14] that the simple model based on the persistence of the clear sky index gives often very good results with acceptable errors [15,16,17] for short term forecasting time horizon (<= 1 hour). The goal of this paper is to determine the influence of solar radiation variability regarding different classes of days on the expected error provided by different forecasting methods that the modeller can possibly implement.

The paper is organized as follows: Section 2 describes the data we have used. Section 3 exposes the classification methodology and the results obtained for the three studied locations. In the two following sections, the forecasting methods are exposed, the forecasting errors for each location and for each class of days are exposed.

2.GLOBAL SOLAR RADIATION DATA

To validate this study, three insular sites where chosen (1 in the northern hemisphere, 1 in the northern tropical zone and 1 in the southern tropical zone). The three Islands are:
-Reunion Island; it exhibits a particular meteorological context dominated by a large diversity of microclimates. Two main regimes of cloudiness are superposed: the clouds driven by the synoptic conditions over the Indian Ocean and the orographic cloud layer generated by the local reliefs. The data used to build

the models are measured at the meteorological station of St Pierre (21°20’S ; 55°29’E, 75m a.s.l) located in the southern part of Reunion Island. Measurements are available on an hourly basis and two years of data (2011 and 2012).

-Guadeloupe Island, we have used a two years database of GHI measured on an hourly basis at the meteorological station of le Raizet (16°26N, 61°24W, 11 m a.s.l.) located in the middle of the island.

-Corsica Island, the data used to build the models, are GHI measured at the meteorological station of Ajaccio (41°55’N, 8°44’E, 4m a.s.l.) and Bastia (42°42’N, 9°27’E, 10m a.s.l.). They are located near the Mediterranean Sea and nearby mountains (1000 m altitude at 40km from the sites). 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 1998 to 2009 (eleven years). As for all experimental acquisitions, missing values are observed, here, this represents less than 2% of the data. A classical cleaning approach is then operated in order to identify and remove this data.

3.CLASSIFICATION OF TYPICAL DAYS

A k-means clustering, or Lloyd's algorithm [2] was applied to the dataset of each location. This method will partition each daily signal of global solar radiation into *k* mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation *n (*in our case a daily global solar radiation signal). Unlike hierarchical clustering, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters. The distinctions mean that k-means clustering is often more suitable than hierarchical clustering for large amounts of data. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized.

k-means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters.

Previous studies [2,6,12,22] for these locations have shown that 4 classes of typical days were generally found:

- Clear sky days
- Mid clear sky days
- corMid cloudy sky days
- Cloudy sky days

We have initialised the k-means algorithm with the assumption of the typical classes mentioned above.

The results for the different locations are shown in the following figures. Table 1 summarize the number of days in each class for the different locations.

Case of Corsica :

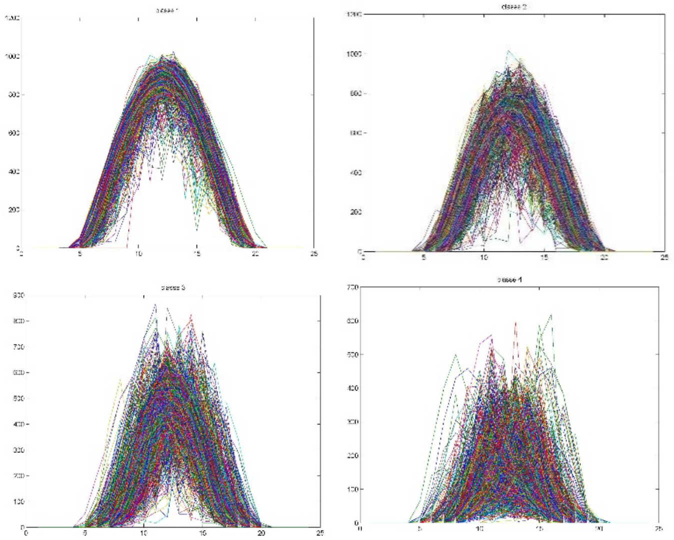

Figure 1 : the four classes of typical days for Corsica Island

Case of Reunion :

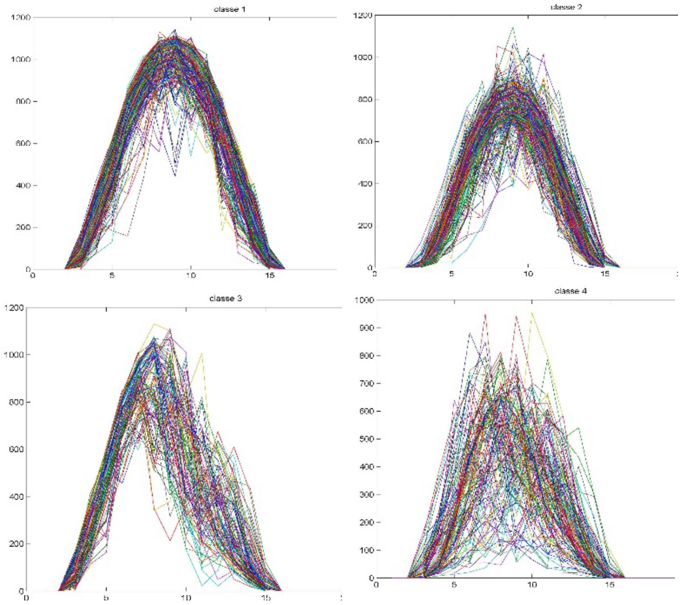

Figure 2 : the four classes of typical days for Réunion Island

Case of Guadeloupe :

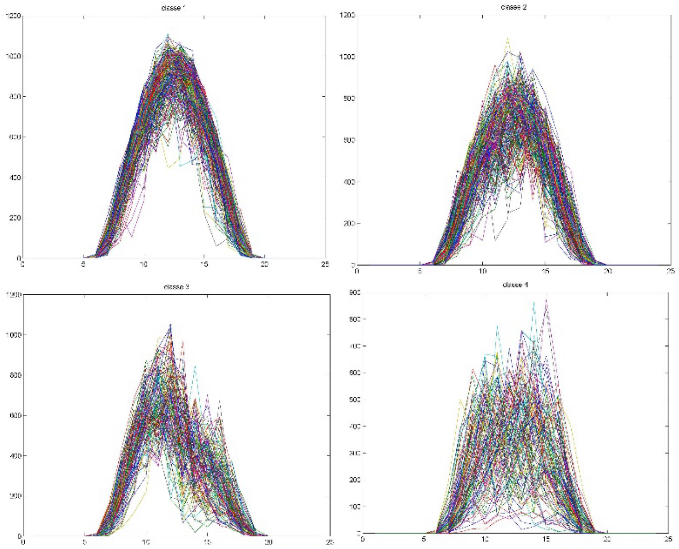

Figure 3 : the four classes of typical days for Guadeloupe Island

		Class 1	Class 2	Class 3	Class 4
Corsica	Number of days	1210	930	1115	760
	percentage	30,14%	23,16%	27,77%	18,93%
Réunion	Number of days	232	262	104	132
	percentage	31,78%	35,89%	14,25%	18,08%
Guadeloupe	Number of days	253	218	138	121
	percentage	34,66%	29,86%	18,90%	16,58%

Table 1: ponderation of each class for the different locations

4.DESCRIPTION OF THE PREDICTION TECHNIQUES

In this section, we present the three different prediction methodologies evaluated in this study: naïve models, linear model and non-linear models.

naïve methods

Two naïve predictors are studied in this work. The first is the simple persistence model defined by the following equation :

$\widehat{I_g}(t+h) = I_g(t)$ Equation 4

It simply states that future values of GHI are equal to GHI observed at time t (i.e. the atmospheric conditions remain unchanged between current time t and future time $t+h$. One way to improve this simple model is to take into account the sun path by using a clear sky model and define persistence on the clear sky index i.e

$\widehat{k_t^*}(t+h) = k_t^*(t)$. The corresponding GHI forecast can be obtained through equation 5.

$\widehat{I_g}(t+h) = I_g(t).\frac{I_{g,clsk(t+h)}}{I_{g,clsk(t)}}$ Equation 5

In the rest of the paper, this persistence on the clear sky index model will be called scaled persistence.

Linear model: autoregressive process (AR)

In an AR model (Chatfield, 2004), the future value of a variable namely $\widehat{k^*}(t+h)$ is assumed to be a linear combination of several past observations as shown in the equation 6.

$\widehat{k^*}(t+h)=\phi_0 + \sum_{i=0}^{p} \phi_{i+1} k^*(t-i) + \epsilon_t$, Equation 6

Where ϵ_t is a white noise with variance σ^2. The model's parameters are the$\{\Phi_i\}_{i=0,1,\cdots p+1}$and p is called order (or autoregressive order) of the model. One key challenge in the building of an AR model is to determine the appropriate model order. Methods based on the autocorrelation coefficients (ACF) and partial autocorrelation coefficients (PACF) analysis are proposed to select the best orders [34]. In this study, the complexity of the model governed by the autoregressive order *p* is determined with the auto mutual information factor.

Neural network models (NN)

A NN with *d* inputs, *m* hidden neurons and a single linear output unit defines a non-linear parameterized mapping from an input vector **x** to an output *y* given by the relationship:

$y = y(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{m} w_j\, f\left(\sum_{i=1}^{d} w_{ji}\, x_i\right).$ Equation 7

Each of the *m* hidden units are related to the tangent hyperbolic function $f(x) = (e^x - e^{-x})/(e^x + e^{-x})$. The parameter vector $\mathbf{w} = (\{w_j\},\{w_{ji}\})$ wich governs the non-linear mapping, is estimated during the training or learning phase. During this phase, the NN is set up using the dataset $\mathcal{D}$. The second phase, called the generalization phase, consists of evaluating on the another dataset $\mathcal{D}_*$, the ability of the NN to give correct outputs when it is confronted with examples that were not seen during the training phase.

For our application, the relationship between the output $\widehat{k^*}(t+h)$and the inputs $\{k^*(t), k^*(t-1), \cdots, k^*(t-p)\}$ has the form given by equation 8.

$\widehat{k^*}(t+h) = \sum_{j=1}^{m} w_j\, f\left(\sum_{i=0}^{p} w_{ji}\, k^*(t-i)\right).$ Equation 8

As shown by equations 7 and 8, the NN model is equivalent to a nonlinear autoregressive (AR) model for time series forecasting problems. As for the AR model, the number of past input values *p* is calculated with the auto mutual information factor. Careful attention must be put on the model structure assumptions. A too complex NN will easily overfit the training data. The NN complexity is in relation with the number of hidden units or conversely the dimension of the vector **w**. Several techniques like pruning [32] or Bayesian regularization [29] can be employed to control the NN complexity. In the present study, the

NN model has been computed with the Matlab© software and its Neural Network toolbox. The Levenberg-Marquardt (approximation to the Newton's method) learning algorithm with a max fail parameter before stopping training equal to 3 was used to estimate the NN model's parameters. The max fail parameter corresponds to a regularization tool limiting the learning steps after a characteristic number of predictions failures and consequently allow to control the model complexity.

Gaussian Process model

Gaussian Processes (GPs) are a relatively recent development in non-linear modelling [25]. GPs are generally stated as kernel-based method. Indeed, it can be shown (Rasmussen, 2006) that, given *n* training samples, the prediction (for an input test vector x_*) can be seen in terms of a linear combination of *n* kernel functions; each one centered on a training point. Therefore, the forecasted clear sky index is given by the equation 9.

$$\widehat{k^*}(\mathrm{t}+\mathrm{h}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\, k_{se}(\mathrm{x_i}, \mathrm{x}_*). \qquad \text{Equation 9}$$

Where k_{se}denotes the squared exponential covariance function $k_{\mathrm{se}}(\mathrm{x_p}, \mathrm{x_q}) = \sigma_{\mathrm{f}}^2 \exp\left[\frac{-(\mathrm{x_p}-\mathrm{x_q})^2}{2\mathrm{l}^2}\right]$ and $\mathbf{x_i}$ is the ith input training vector. σ_f^2 and *l* are called hyperparameters of the covariance function. They control the model complexity and can be learned (or optimized) from the training data at hand [25]. The coefficients α_i are obtained by a matrix multiplication between a covariance matrix (resulting from the application of the covariance function on all the training data points) and the vector of the *n* training output samples **y**.

Support vector machine

Support vector machine is another kernel based machine learning technique used in classification tasks and regression problems [24]. Support vector regression (SVR) is based on the application of support vector machines to regression problem [26]. This method has been successfully applied to time series forecasting tasks. As for the GPs, the prediction calculated by a SVR machine for an input test case x_* is given by equation 10.

$$\hat{y} = \sum_{i=1}^{n} \alpha_i\, k_{rbf}(x_i,\ x_*) + b, \qquad \text{Equation 10}$$

with the commonly used RBF kernel [26] defined by equation 11.

$$k_{rbf}(x_p, x_q) = exp\left[\frac{-(x_p-x_q)^2}{2\sigma^2}\right]. \qquad \text{Equation 11}$$

The parameter b (or bias parameter) is derived from the preceding equation and some specific conditions.

In the case of SVR, the coefficients α_i are related to the difference of two Lagrange multipliers, which are the solutions of a quadratic programming (QP) problem. Unlike NNs, which are confronted with the problem of local minimum, here the problem is strictly convex and the QP problem has a unique solution. In addition, it must be stressed that, not all the training patterns participate to the preceding relationship. Indeed, a convenient choice of a cost function i.e. Vapnik's ε −insentive function [26] in the QP problem allows obtaining a sparse solution. The latter means that only some of the coefficients α_i will be nonzero. The examples that come with non-vanishing coefficients are called Support Vectors. In our work, given the training dataset $\mathcal{D} = \{\mathbf{x}_\mathrm{i}, y_i\}_{\mathrm{i}=1}^{\mathrm{n}}$and a test input vector $\mathbf{x}_*$, we can compute the forecasted clear sky index for a specific horizon *h*:

$$\widehat{k^*}(t+h) = \sum_{i=1}^{n} \alpha_i\, k_{rbf}(\mathbf{x}_i, \mathbf{x}_*) + b. \qquad \text{Equation 12}$$

In the present study, regarding the implementation of the support vector regression, we used the LibSVM library [30]. Like in the NN case, other kinds of support vectors methodologies were tested e.g. the multi-class SVMs [30,33,35]. The corresponding results were systematically worse than those from SVR, thereby, we prefer to not develop it in order to make the paper more readable.

5.RESULTS

We have performed the different forecasting methods and algorithm exposed before to the different data set composed of a given typical class of days for each location.

For each location we have ploted the nRMSE obtained for each typical class from the different forecasting models.

- The Normalized root -mean-square error is computed
-

$$NRMSE = \frac{\sqrt{\frac{1}{M}\sum_{h=1}^{M}\left(\tilde{G}(h)-G(h)\right)^2}}{\max(G)-\min(P)}.$$

Where: - $G(h)$ is global solar radiation measured

- $\tilde{G}(h)$ is the predicted solar radiation predicted

- M is the number of hours considered

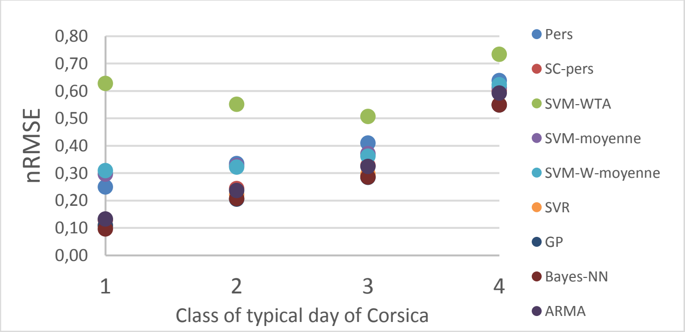

Figure 4: nRMSE obatained for the different class in Corsica

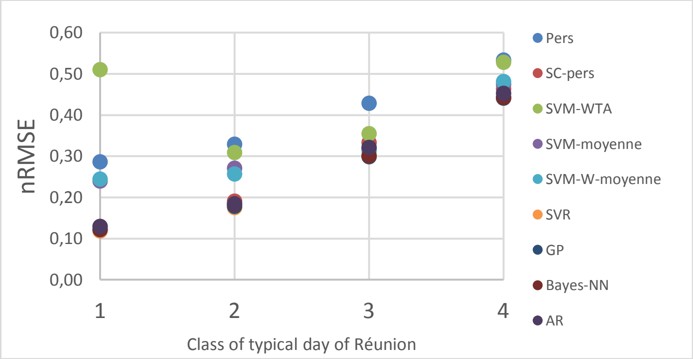

Figure 5: nRMSE obtained for the different class of Reunion

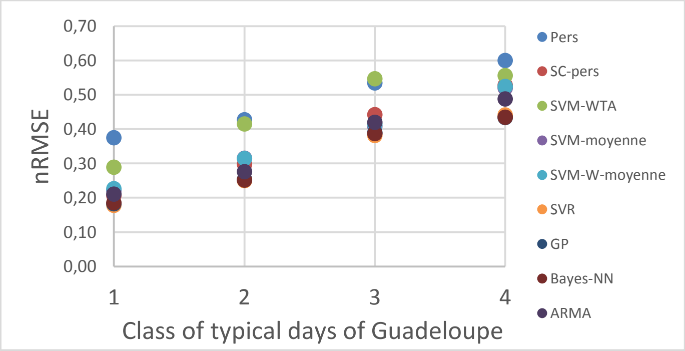

Figure 6 : nRMSE obtained for the typical class of Guadeloupe

The analysis of the figure 4 to 6 shows clearly that the forecasting errors increase with the variability of the considered class of days. This results is verified regardless the forecasting models used and regardless the considered location.

When comparing the results between locations we can observe some differences in the forecasting errors. Indeed in Corsica and Reunion the lowest nRMSE is obtained for the clear sky conditions (class 1) and for the Bayesian NN model. In Guadeloupe we observe the same tendency but the nRMSE is 0,2.

6.CONCLUSION

In this study we have analysed the influence of the global solar variability upon the forecasting error of different models: persistence, scaled persistence, Support vector machine, Gaussian Process, Bayesian neural network and ARMA model.

We have used a global solar radiation data set from three different locations: Reunion Island, Corsica Island and Guadeloupe Island. For each of these location we have performed a classification of typical days using a k-means algorithm. We have then established a data set of 4 classes of typical days for each location.

One of the main result of the study is that the forecasting error, whatever the model used, is much higher (up to 3 to 4 time) when considering cloudy days than considering clear sky days.

.

REFERENCES

[1] Badescu V. Modeling solar radiation at the earth's surface: recent advances. Springer; 2008.
[2] Diagne M, David M, Boland J, Schmutz N, Lauret P. Post-processing of Solar Irradiance Forecasts from WRF Model at Reunion Island. Energy Procedia 2014;57:1364–73. doi:10.1016/j.egypro.2014.10.127.
[3] Bofinger S, Heilscher G. solar electricity forecast : approach and first results, 2006.
[4] Perez R, Hoff T, Dise J, Chalmers D, Kivalov S. The Cost of Mitigating Short-term PV Output Variability. Energy Procedia 2014;57:755–62. doi:10.1016/j.egypro.2014.10.283.
[5] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecast 2006;22:443–73. doi:10.1016/j.ijforecast.2006.01.001.
[6] Soubdhan T, Abadi M, Emilion R. Time Dependent Classification of Solar Radiation Sequences Using Best Information Criterion. Energy Procedia 2014;57:1309–16. doi:10.1016/j.egypro.2014.10.121.
[7] Brockwell PJ, Davis RA. Time series: theory and methods. 2nd ed. New York: Springer-Verlag; 1991.
[8] Bourbonnais R, Terraza M. Analyse des séries temporelles : application à l'économie et à la gestion. 2e éd. Paris: Dunod; 2008.
[9] Lauret P, Fock E, Randrianarivony RN, Manicom-Ramsamy JF. Bayesian neural network approach to short time load forecasting. Energy Convers Manag 2008;49:1156–66.
[10] Lynch SM. Bayesian Statistics. Encycl. Soc. Meas., New York: Elsevier; 2005, p. 135–44.
[11] Diday E, Lemaire J, Pouget J, Testu F. Éléments d'analyse de données. Dunod; 1982.
[12] Voyant C, Paoli C, Muselli M, Nivet M-L. Multi-horizon solar radiation forecasting for Mediterranean locations using time series models. Renew Sustain Energy Rev 2013;28:44–52. doi:10.1016/j.rser.2013.07.058.
[13] Paoli C, Voyant C, Muselli M, Nivet M-L. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol Energy 2010;84:2146–60. doi:10.1016/j.solener.2010.08.011.
[14] Paoli C, Voyant C, Muselli M, Nivet M-L. Solar Radiation Forecasting Using Ad-Hoc Time Series Preprocessing and Neural Networks. Emerg. Intell. Comput. Technol. Appl., vol. 5754, Springer Berlin / Heidelberg; 2009, p. 898–907.
[15] Voyant C, Muselli M, Paoli C, Nivet M-L. Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. Energy 2012;39:341–55. doi:10.1016/j.energy.2012.01.006.
[16] Dambreville R, Blanc P, Chanussot J, Boldo D. Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model. Renew Energy 2014;72:291–300. doi:10.1016/j.renene.2014.07.012.
[17] Kühnert J, Lorenz E, Heinemann D. Chapter 11 - Satellite-Based Irradiance and Power Forecasting for the German Energy Market. In: Kleissl J, editor. Sol. Energy Forecast. Resour. Assess., Boston: Academic Press; 2013, p. 267–97.
[18] Perez R, Kivalov S, Schlemmer J, Hemker Jr. K, Hoff TE. Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance. Sol Energy 2012;86:2170–6. doi:10.1016/j.solener.2012.02.027.
[19] Marquez R, Coimbra CFM. Proposed Metric for Evaluation of Solar Forecasting Models. J Sol Energy Eng 2012;135:011016–011016. doi:10.1115/1.4007496.
[20] Gueymard CA. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renew Sustain Energy Rev 2014;39:1024–34. doi:10.1016/j.rser.2014.07.117.

[21] Diazrobles L, Ortega J, Fu J, Reed G, Chow J, Watson J, et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmos Environ 2008;42:8331–40. doi:10.1016/j.atmosenv.2008.07.020.

[22] T Soubdhan, R Emilion, R Calif Classification of daily solar radiation distributions using a mixture of Dirichlet distributions
Solar energy 83 (7), 1056-1063

[23] Iqdour R, Zeroual A. The MLP Neural Networks for Predicting Wind Speed, Marrakech, Morocco: 2006.

[24] Vapnik V., 1995. The Nature of Statistical Learning Theory. Springer,New York, 1995.

[25] Rasmussen CE, Williams C. Gaussian Processes for Machine Learning. MIT Press; 2006.

[26] Smola A. and Scholkopf B., 2004. A tutorial on support vector regression, Statistics and Computing 14: 199–222

[27] Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M, Cros S, Dumortier D, et al. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens Environ 2004;91:160–74. doi:10.1016/j.rse.2004.02.009.

[28] Kumar U, Jain VK. Time series models (Grey-Markov, Grey Model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India. Energy 2010;35:1709–16. doi:doi: DOI: 10.1016/j.energy.2009.12.021.

[29] Hamilton J. Time series analysis. Princeton N.J.: Princeton University Press; 1994.

[30] Jiang A-H, Huang X-C, Zhang Z-H, Li J, Zhang Z-Y, Hua H-X. Mutual information algorithms. Mech Syst Signal Process 2010;24:2947–60. doi:10.1016/j.ymssp.2010.05.015.

[31] Muzy JF, Bacry E, Baile R, Poggi P. Uncovering latent singularities from multifractal scaling laws in mixed asymptotic regime. Application to turbulence. EPL Europhys Lett 2008;82:60007. doi:10.1209/0295-5075/82/60007.

[32] Lauret P, Diagne M, David M. A Neural Network Post-processing Approach to Improving NWP Solar Radiation Forecasts. Energy Procedia 2014;57:1044–52. doi:10.1016/j.egypro.2014.10.089.

[33] Zhang G. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003;50:159–75. doi:10.1016/S0925-2312(01)00702-0.

[34] Chatfield, C., Time series analysis, an introduction, Chapman & Hall, 2004

[35]Czibula, G., Czibula, I.G., Gaceanu, R.D., 2014. A support vector machine model for intelligent selection of data representations. Appl. Soft Comput. 18, 70–81. doi:10.1016/j.asoc.2014.01.026