Article

Forecast horizon and Solar variability influences on the performances of Multiscale Hybrid Forecast Model (MHFM)

Stéphanie Monjoly †, Rudy Calif †, 0000-0001-5514-5790, Maina André †, Ted Soubdhan †

[1] EA 4935 LaRGE, Laboratoire de Recherche en Géosciences et Énergies, Université des Antilles, 97170 P-á-P, France; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: Rudy Calif, rudy.calif@univ-ag.fr; Tel.: +59-059-048-3112

Abstract: In this paper, the forecast horizon and solar variability influences on MHFM model based on multiscale decomposition, AR and NN models, are studied. This article follows the works published in [1] showing the performance of the MHFM using 3 multiscale decomposition methods and a forecast horizon equal to 1 hour. Several forecast horizon strategies and his influence on the MHFM performances are investigated. We show that the best strategy for a rRMSE variying from 4.43% to 10.24% is obtained for forecast horizons from 5 minutes to 6 hours. In a second part, the solar variability influence on the MHFM is studied. A classification based on a shows that the best performance of MHFM is obtained for clear sky days with a rRMSE of 2.91% and worst for cloudy sky days with a rRMSE of 6.73%.

Keywords: Hybrid Forecast Model; Forecast Horizon; Daily Global Solar Radiation Clustering; Fuzzy c-means; Variability Characterization

1. Introduction

The electrical output from solar resources is a major issue, particulary for island such as Guadeloupe Archipelago due to its non-interconnected electrical network. To improve the integration of this kind of energy in the electric network, in the preceeding work [1], a hybrid forecast model based on multiscale decomopsition methods, AR and NN models is proposed. Three multiscale decomposition methods have been tested (Empirical Mode Decomposition EMD, Ensemble Empirical Mode Decomposition EEMD and Wavelet Decomposition WD). This previous study aimed to demonstrate the robustness and the efficient of the model and had been performed only for a 1h forecast horizon. Our objective here is to pursue this work evaluating the MHFM performances versus the forecast horizon strategy and the solar variability. Classically the forecast horizons can be divided in 3 categories [2] from 5 minutes to 2h intra-hour horizon, from 1h to 6h intra-day horizon and from 1 day to 3 days day-ahead horizon. Whatever the forecast models or the methods presented in the literature, the results are often established for several horizons and generally, the RMSE error increases with the horizon. This article makes an additional study by lingering over the way we determine the forecast horizon. Indeed, two strategies for the forecast horizon determination, are proposed. We show that following to the strategie adopted the results change. In addition, the solar variability influence on the model performances, is studied. For that, a classification of GHI measurements is proposed. Several classification techniques are devoted to global solar radiation data : Soubdhan et al. [3] used the k-means method on the clearness index PDF and found 4 classes, Badosa et al. [4] used the k-means algorithm to solar radiation classification based on 3 parameters (clearness index, Total Accumulated Relative Change and Total Relative Change) they found 5 classes. Soubdhan et al.[5] used the Dirichlet Distribution on the clearness index computed with Global solar radiation measured in Guadeloupe. They found 4 classes. Muselli et al.[6] applied Ward aggregation classification method

 doi:10.20944/preprints201809.0089.v1

on daily global irradiation on horizontal and tilted plane in Corsica and found 3 classes. Maafi and Harrouni[7] used fractal dimension and clearness index on 2 sites in Algeria and find 3 classes. Fuzzy c-means algorithm was recently used by Benmouiza et al [8] to hourly solar radiation classification. In this study, this Fuzzy C-Means (FCM) classification technique is used. FCM algorithm is a powerful tool which presents more precise results comparing with k-means algorithm [8,9]. It has been also chosen because it introduces the fuzziness for the belongingness of each object and can retain more information of the data set than the hard k-means clustering algorithm [10]. Fuzzy c-means algorithm was recently used by Benmouiza et al [8] to hourly solar radiation classification. Furthermore, to confirm the obtained classification a variability study based on metric of variability is given. Several studies and methods of variability characterization were carried out [11–16]. The paper is built as follows: section 2 describes the dataset considered and the data pre-processing, section 3 presents briefly the MHFM methodology (see [1] for a full description) and section 4 gives the metrics used to evaluate the MHFM performances. Section 5 describes the two strategies applied to determine the forecast horizon influence on the MHFM, the results obtained are presented in section 6. Then, the section 7 presents the FCM algorithm used to establish the influence of variability on the MHFM. Section 8 relates results of the daily classification, the study of the GHI variability according to the classes obtained and the GHI variability influence on the MHFM.

2. Data Set description

The global solar radiation GHI_{mes} data were measured at Petit-Canal (16°23′ N latitude and 61°24′ W longitude) located at Guadeloupe Island. The data are sampled at 1 second continuously during the year 2012 from January to December. Guadeloupe is subject to many passing clouds resulting in high variability of global solar radiation measured.

The original solar radiation series is not stationary, to perform the hybrid model a detrended time series must be used. Traditionally, the temporal trend is removed in GHI_{mes} considering clear sky index [17,18]. The detrended time series are obtained by clear sky index described by the following equation:

$$K_c = \frac{GHI_m}{GHI_{clear}} \tag{1}$$

where GHI is the Global Horizontal Irradiance, index m refers to the measured GHI index and $clear$ refers to theoretical clear sky irradiance computed by the Kasten clear sky model[19].

3. Hybrid Forecast Model methodology

The hybrid model presented in [1] is based on a multiscale decomposition method. The hybrid model is performed to the Global solar radiation forecasting and its process is divised in several steps. We recall brief these steps :

- Step 1 : detrend the data estimating the clear sky index.(section 2)
- Step 2 : decompose the K_c signal using a multiscale decomposition method (Empirical Mode Decomposition, Ensemble Empirical Mode Decomposition or Wavelet Decomposition). The decomposition components obtained will be forecast separately using AR or NN model according to the time scale of component considered. Components with slow fluctuations (representing long time) scale are predict using the AR model and the components with fast fluctuations (representing short time scale) with the NN model [1].
- Step 3 : Forecast the short time scales components with a NN model and the long time scales components with a AR model. Each multiscale decomposition component was predicted.
- Step 4 : Summing all the component forecasting to obtain the predicted K_c time series.
- Step 5 : rebuild the Global solar radiation signal using the Kasten Clear sky model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

4. Evaluation of Forecast Performances

The Hybrid forecast model performance can be evaluated using the following classical statistical performance indicators

- Relative Mean Biais Error $(rMBE)$

$$rMBE = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{p}_i - o_i}{\frac{1}{N}\sum_{i=1}^{N} o_i} \tag{2}$$

- Relative Mean Absolute Error $(rMAE)$

$$rMAE = \frac{\frac{1}{N}\sum_{i=1}^{N} |\hat{p}_i - o_i|}{\frac{1}{N}\sum_{i=1}^{N} o_i} \tag{3}$$

- Relative Root Mean Square Error $(rRMSE)$

$$rRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{p}_i - o_i)^2}}{\frac{1}{N}\sum_{i=1}^{N} o_i} \tag{4}$$

where o_i is the observed value of GHI and $\hat{p}_i$ is the forecast value of GHI and N the number of point in the dataset for the considered period.

- Skill s : compare the model performance with a reference model [20]. In this study we compare the proposed model with the persistence model applying the skill parameter proposed by Coimbra et al.[21]:

$$s = (1 - \frac{RMSE_{model}}{RMSE_{SC_{pers}}}).100 \tag{5}$$

where index SC_{pers} refers to the scaled persistence reference model define by Eq. 6.

$$\widehat{K_c}(t+h) = K_c(t) \tag{6}$$

The corresponding GHI forecast is obtained using Eq. 7:

$$\widehat{GHI}(t+h) = GHI(t) \times GHI_{clear}(t+h)/GHI_{clear}(t) \tag{7}$$

5. Forecast horizon Influence on Hybrid Forecast Model Performances

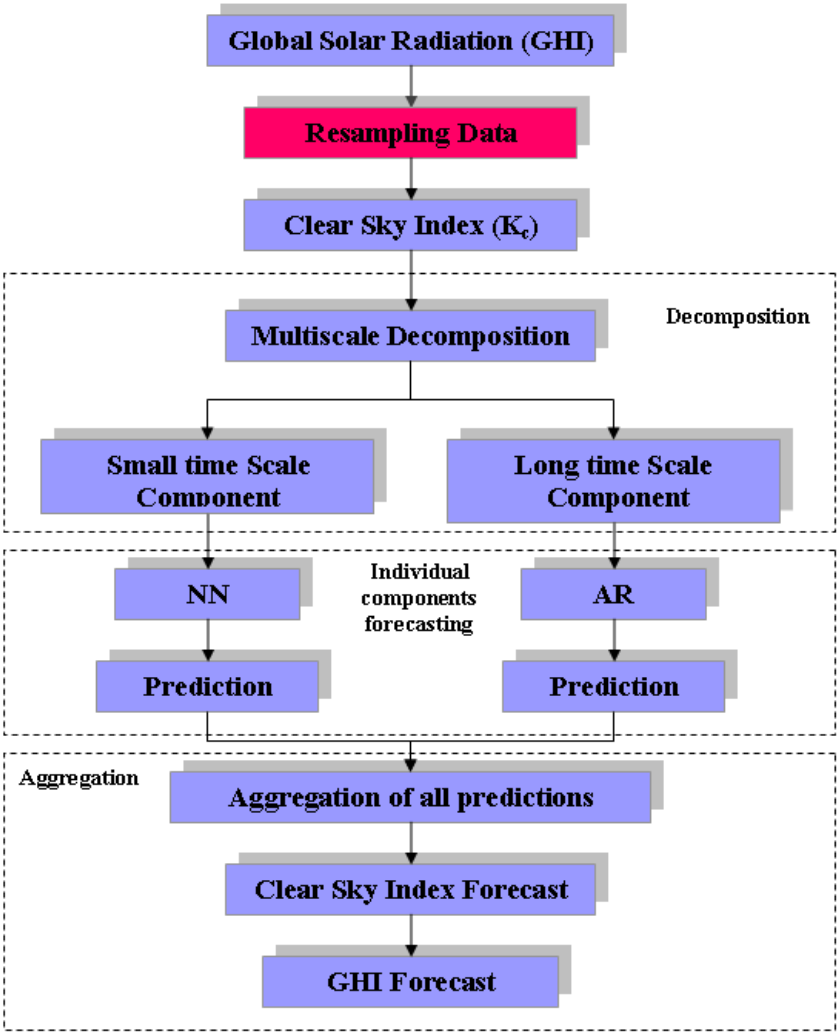
In [1] the MHFM performances are studied only for a forecast horizon corresponding to $\tau = 1$ hour. In this work, the forecast horizon influence on the propose hybrid model, is studied. According to categorization given in [2] we focus on intra-hour and intra-day forecast horizons i.e.from 5 minutes to 6 hours. Two strategies based on the data rich in order to evaluated the MHFM performances, are implemented.

5.1. Strategy 1 :sampling data T_r= forecast horizon τ

The first strategy consists to resampling the data in order to time sampling T_r equal to the time horizon τ. In this case the model predicts directly the next point. Classically, this strategy is used in the solar forecasting field. Figure.1 illustrates the MHFM flowchart including the resampling step before the multiscale decomposition step. Figure 2 briefly illustrates the individual components forecasting process for example, $T_r = \tau = 1$ hour. This strategy is operated as follows for each decomposition components after resampling the data set with time sampling $T_r = \tau$:

- To determine during the AR and NN learning phase, the number of input data selected with the AIC and BIC criterion for the AR model and the mutual information for the NN model [1,18]. The learning phase concerns the half of dataset, i.e 6 months,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1



MHFM flowchart adapted to the strategy 1, the fuchsia case indicates the step added.

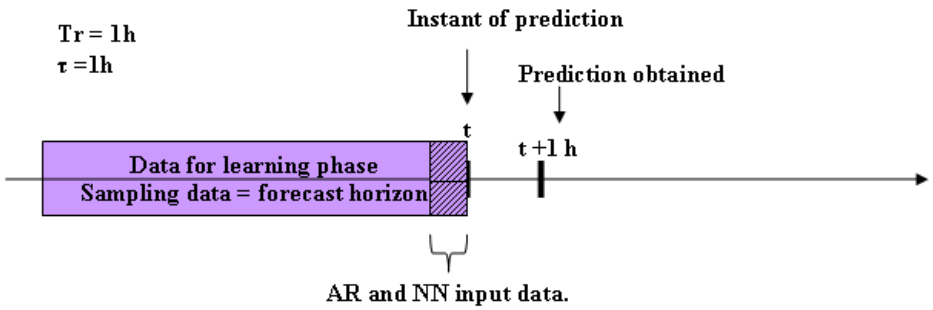
Figure 1

- To provide a forecasting at $t + \tau$ of 6 months data test (different from data of learning phase) according to this strategy it consists in predicting the next point.
- To repeat the process for each τ value.

5.2. Strategy 2 :sampling data $T_r \neq$ forecast horizon τ

In the second strategy, the time sampling is different to the time horizon ($T_r \neq \tau$). Generally, in statistics, the quantity of data can be of great importance, for representative results. This can be considered, in part , having a long period of measurements or in other part, having data with high frequency sampling. The goal, here, is to verify data rich on the MHFM performances; intuitively, we could think that the kind of data (with high frequency sampling) may cause an impact to the model performance. The second strategy consists to verify this assumption. Unlike the first strategy there is no additional step, the data sampling $T_r = 5$ minutes is the same for all the considered forecast horizon τ. Figure 3 illustrates the MHFM flowchart with the second strategy and Figure 4 briefly represent an example of the individual components forecasting phase for a τ =1 hour. This strategy is operated as follows for each decomposition components :

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1



AR and NN diagram used in the strategy 1 : example for $T_r = \tau = 1h$

Figure 2

- To select the number of input data determined with the AIC and BIC criterion for the AR model and the mutual information for the NN model [1,18] using 6 months data set with sampling time $T_r = 5$ minutes. The number of input found will be valid for all the the considered forecast horizon τ.
- To provide a forecasting at $t + \tau$ of 6 months data test, the data test sampling time is also 5 minutes and the model allows us to obtain every 5 minutes the forecast at $t + \tau$.

In summary, this approach performs the AR and NN learning phase with 5 minutes data sampling and provides a forecasting at $t + \tau$, with the time horizon $5 \leqslant \tau \leqslant 360$ *min* (6*h*).

6. Results-Discussion

In this section, the results obtained for each strategy proposed in the preceeding section, are presented. Considering the two strategies set out in order to obtain forecasts for several horizons, we apply decomposition methods on data with different sampling. In the strategy 1 the sampling time is equal to the forecast horizon, for example, for a horizon of 5 minutes, data sampled at 5 minutes are used and for a horizon of 2 hours, data are sampled at 2 hours. The initial dataset is sampled at 1 seconde. Moreover the time horizon is longer the smaller the dataset. This will reduce the number of decomposition components. Indeed the longer the dataset is, the greater the number of components of the decomposition. In the EMD case the number N of intrinsic mode functions (IMF) is proportional to the length of the data set [22] :

$$N = \frac{T}{n\Delta t} \tag{8}$$

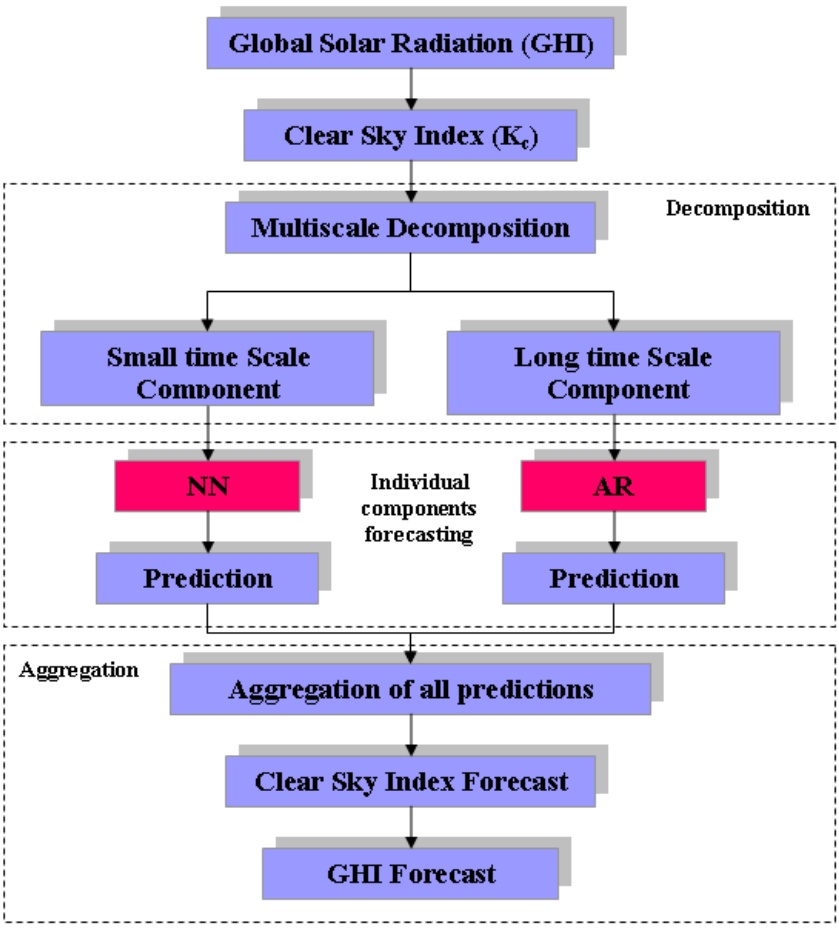
where T represents the total data length, Δt represents the digitizing rate and n represents the minimum number of Δt needed to define the frequency accurately.

In this work, time horizons of 5, 10, 15, 30 minutes, 1h, 2h, 4h and 6h, are tested.

6.1. Results : Strategy 1: Sampling Data = Forecast Horizon

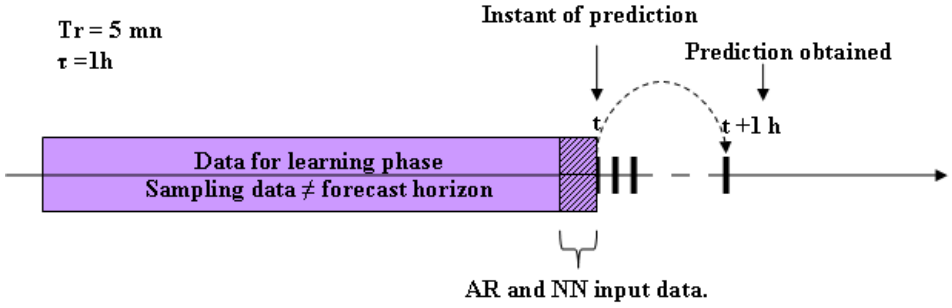
Table 1 shows the Hybrid model forecasting performances using the first strategy. We can note that whatever the Multiscale decomposition-hybrid model chosen the rRMSE error increases with the forecast horizon as illustrated in Figure 5. This increase is nonlinear and seems follow a logarithmic tendency. As shown in [1], the best results are obtained with the WD-Hybrid model (rRMSE varying between 4.41% and 11.42 %). The skill parameters which allow to compare the hybrid model performances to the persistence model varying between 78.58% and 85.54 %, highlighting the best performances of proposed model (comparatively to the persistence model), for all the forecast horizons.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1



Hybrid model flowchart adapted to the strategy 2, the fuchsia case indicate the Step undergoing a modification

Figure 3



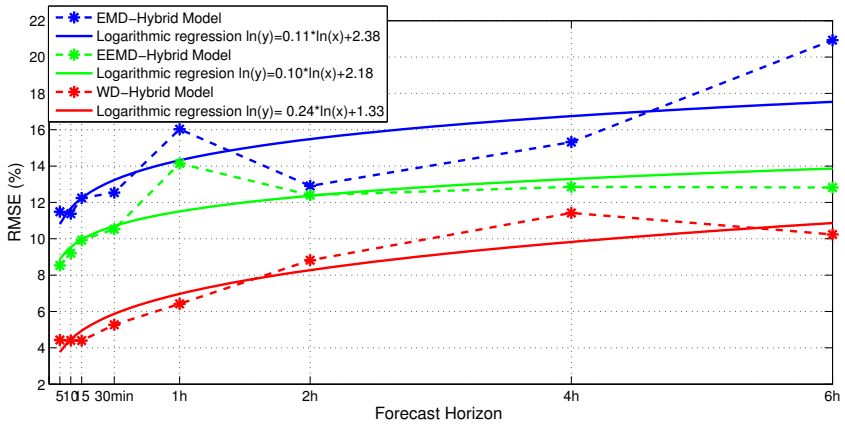
AR and NN flowchart used in the strategy 2 predicting directly the considered forecast horizon

Figure 4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

Table 1. The MHFM performances according to the Forecasting Horizon (Strategy 1).

Models		Forecast Horizon							
		5 min	10 min	15 min	30 min	1h	2h	4h	6h
EMD-Hybrid Model	rMBE (%)	0.17	-0.25	0.23	0.12	-0.32	1.36	1.03	1.99
	rMAE (%)	7.42	7.73	8.30	8.54	11.43	9.00	9.87	13.16
	rRMSE (%)	11.49	11.38	12.24	12.54	16.03	12.91	15.32	20.93
	Skill (%)	44.54	50.22	46.82	51.43	56.30	77.24	79.43	70.45
EEMD-Hybrid Model	rMBE (%)	-0.02	-0.12	-0.27	0.25	-0.60	0.23	0.63	0.12
	rMAE (%)	5.14	5.80	6.42	7.21	8.90	8.20	8.56	8.45
	rRMSE (%)	8.53	9.20	9.92	10.53	14.14	12.40	12.86	12.82
	Skill (%)	58.83	59.72	57.10	59.21	61.45	78.15	83.31	81.91
WD-Hybrid Model	rMBE (%)	0.12	0.01	-0.04	0.11	0.074	0.38	0.29	0.42
	rMAE (%)	2.84	3.02	3.04	3.62	4.17	6.11	8.14	7.37
	rRMSE (%)	**4.43**	**4.41**	**3.04**	**5.27**	**6.42**	**8.82**	**11.42**	**10.24**
	Skill (%)	**78.58**	**80.69**	**80.73**	**79.57**	**82.50**	**84.45**	**85.18**	**85.54**



.

Figure 5. Strategy 1 RMSE error (dotted line) and an attempt of modeling by a logarithmic regression (solid line); in the logarithmic regression equation y refers to the RMSE and x refers to the forecast horizon τ

6.2. Results: Strategy 2: Sampling Data $T_r \neq$ Forecast Horizon τ

We have seen that the previous strategy added an additional stage in the hybrid model process for each horizon considered. The objective of the second strategy is to verify if the hybrid model forecast performances are also good by using data high frequency sampling (T_r = 5 minutes for all the horizons). This approach uses 5 minutes data sampling to forecast the GHI at several horizon τ. The AR and NN model learning phase perform with 5 minutes data sampling and the forecast is directly obtained every 5 minutes at $t + \tau$. With this approach we remark that the WD-Hybrid Model obtain the best results(rRMSE varying from 4.43% to 22.33% and skill parameter varying from 78.58% to -7.81%). Figure.6 represents the rRMSE versus the forecast horizon compared to the associated logarithmic regression. We had make the assumption that enrich the learning data set would improve the model performance but the WD-hybrid model errors obtained using the second strategy are higher than those obtained by the first strategy. Finally, To verify if the hybrid model become more complexe with the strategy 2 we use it with a sampling data varying from 5 to 15 minutes. This complexity would arise from the fact that it is supplied with data high frequency sampling data whatever the chosen horizon. Figure7 represents the rRMSE versus the forecast horizon for different data sampling and shows that the learning data set sampling has an influence on the hybrid model performances and suggests that the first strategy is the more efficient. Indeed, High frequency data doesn't seem

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

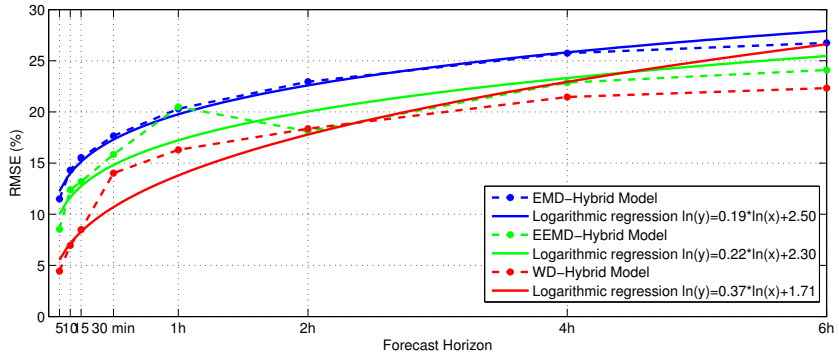


Figure 6. Strategy 2 rRMSE error (dotted line) and their logaritmic regression (solid line), in the logarithmic regression equation y refers to the RMSE and x refers to the forecast horizon τ

improve the MHFM performances. These lower performances are certainly due to the fact that we impose the model to predict phenomena observed at small scale (via decomposition components) while these ones are not observable on the chosen time horizon. Use data sampling close or equal to the time horizon (strategy 1) allows to consider only the observable phenomenon for this horizon and thus to obtain better results.

Table 2. Forecasting performance of the Hybrid model according to the Forecasting Horizon with a data sampling = 5min (Strategy 2)

Models		Forecast Horizon							
		5 min	10 min	15 min	30 min	1h	2h	4h	6h
EMD-Hybrid Model	rMBE (%)	0.17	0.013	0.07	-0.33	0.48	-0.14	1.36	-0.14
	rMAE (%)	7.12	9.39	10.12	11.52	13.51	15.66	17.61	18.88
	rRMSE (%)	11.49	14.30	15.53	17.65	20.30	22.95	25.74	26.75
	Skill (%)	44.54	30.85	24.91	14.68	1.86	-10.95	-24.29	-29.13
EEMD-Hybrid Model	rMBE (%)	-0.02	-0.38	-0.53	0.31	-1.15	-0.49	0.27	0.33
	rMAE (%)	5.174	7.59	8.20	9.92	11.78	13.61	15.61	16.71
	rRMSE (%)	8.53	12.38	13.19	15.84	18.10	20.49	22.87	24.10
	Skill (%)	58.83	40.16	36.23	23.40	12.49	0.96	-10.64	-16.54
WD-Hybrid Model	rMBE (%)	0.12	0.18	0.12	0.24	-0.22	0.17	0.26	0.15
	rMAE (%)	2.84	4.63	5.79	8.82	10.64	12.06	14.48	15.14
	rRMSE (%)	**4.43**	**6.95**	**8.49**	**14.02**	**16.30**	**18.36**	**21.46**	**22.33**
	Skill (%)	**78.58**	**66.38**	**58.98**	**32.20**	**21.20**	**11.24**	**-3.67**	**-7.81**

Afterward, we decide to study the influence of another parameter on the MHFM performances. Guadeloupe island is located in the intertropical zone and is subject to many passing clouds, what entails a GHI high variability. In the following section, we study the influence of the GHI variability on the MHFM performances through a classification of the type of days.

7. Solar variability Influence on MHFM performances

7.1. Clearness index

To evaluate the solar variability influence on the performances of the MHFM, a classification of daily Solar is proposed. Each defined day class allows to highlight a type of variability. To perform the classification, the clearness index noted K_t is used. This parameter removes the effect of daily solar trend and normalizes variability to unity. The clearness index K_t usually employed for solar radiation clustering [3–5,23].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

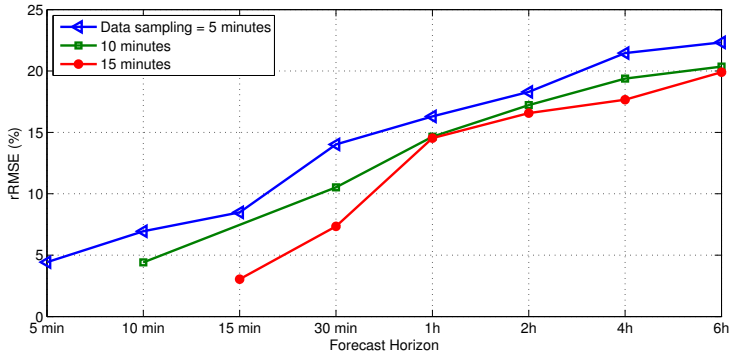


Figure 7. WD-Hybrid model rRMSE over the forecast horizon for different data sampling

$$K_t = \frac{GHI_{mes}}{GHI_{extra}} \tag{9}$$

Where GHI_{mes} refers to measured global solar radiation and GHI_{extra} refers to extraterrestrial radiation estimated according to the Kasten model [19,24]. We notice that the values of K_t are bounded between

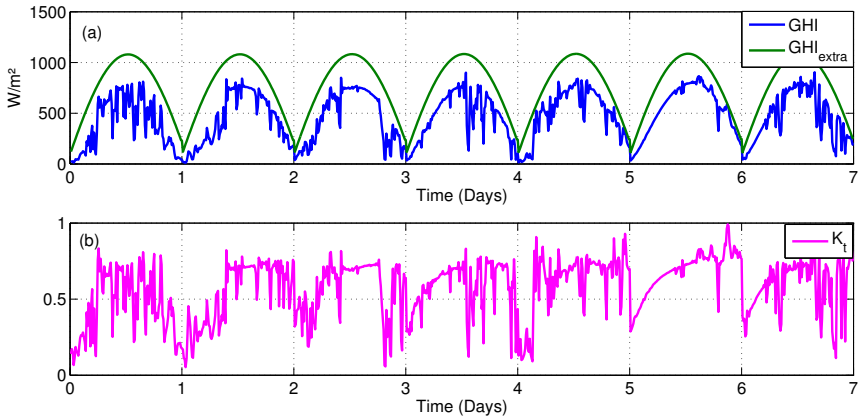


Figure 8. (a) 7 days GHI data and the corresponding theoritical extraterestrial model (GHI_{extra}). (b) K_t signal obtained

0 and 1. This characteristic allows a uniform classification of all daily sequences of data.

7.2. Fuzzy-C means clustering algorithm

The Fuzzy C-means clustering is an iterative method to classify individuals (or samples) in C classes. It was introduced by Ruspini [25], and later extended by Dunn and Bezdek [26,27]. It determines the centers of the classes and generates the matrix to estimate the membership of individuals, to one of the predefined classes. The main purpose of this method is to minimize a cost function, which is usually chosen to be the total distance between each sample to the center of each class [27,28].

$$J(U,V) = \sum_{k=0}^{n} \sum_{i=1}^{C} \mu_{ik}^{m} \|x_{ik} - v_i\| \tag{10}$$

With n the total number of samples, C is the predefined number of classes,x_k is the vector representing the k^{th} individual, v_i is the vector representing the center of the i^{th} class and μ_{ik} is the degree of membership of the k^{th} individual in the i^{th} class. The matrix U contains the coefficients μ_{ik}.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

V is the matrix containing the center of the C classes v_i and m is a constant greater than 1 (generally $m = 2$). By differentiating the function $J(U, V)$ according to v_i, keeping U constant and according to μ_{ik}, keeping V constant, the following equations are obtained [27]:

$$\mu_{ik} = \frac{1}{\text{ffl} \sum_{j=1}^{C} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{(m-1)}}} \tag{11}$$

$$V_i = \frac{\sum_{j=1}^{n} (\mu_{ik})^m x_k}{\sum_{j=1}^{n} (\mu_{ik})^m} \tag{12}$$

In Eq.10 and Eq.11 the symbol $\|.\|$ represent the Euclidian distance.

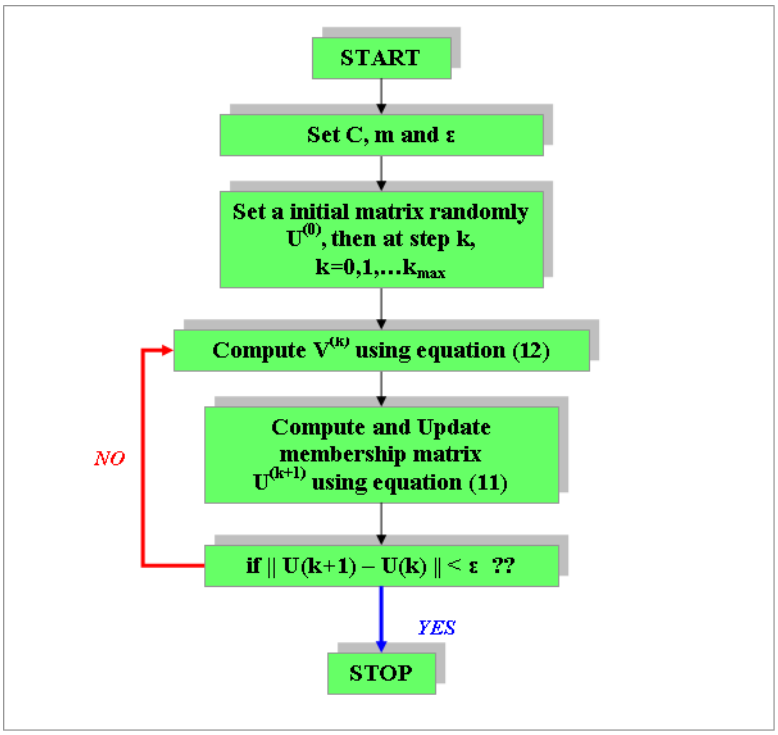


Figure 9. Fuzzy-C-means algorithm flowchart

Validity Criterion

We use a validity criterion to determine the optimal numbers of classes and "fuzzy factor value". It was defined as a fuzzy clustering validity function noted S, which measures the overall average and the separation of a fuzzy-C partition [29]. S can be explicitly written as :

$$S = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n} \mu_{ij}^m \|V_i - X_j\|^2}{n * (d_{min})^2} \tag{13}$$

where d_{min} represents the minimum euclidian distance between cluster centroids i.e:

$$d_{min} = \min_{i,j} \|V_i - Vj\| \tag{14}$$

The class number (or "fuzzy factor value") is optimal for the smallest value of S.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

8. Results-Discussion

In this section, we present the results of the fuzzy-c means clustering, then we perform a variability analysis for each identified class of days and finally we determine the MHFM performances for each class of days.

8.1. Daily solar radiation classification results

In this study, the classification method is applied to 366 days of Global solar radiation sampled at 5 minutes. The clearness index K_t defines in Eq.9 is computed for each day in orde to estimate the K_t histograms. These histograms are considered as the fuzzy c-means algorithm inputs. The fuzzy C means clustering algorithm was tested for several numbers of classes $C = 2, 3, 4, 5, 6$. The validation criterion, given in Eq.13, allows to define that the optimum number of classes is 4 : the S value corresponding to each number of cluster, is drawn up in Table 3. Soubdhan et al.[5] also made a classification of the GHI measured in Guadeloupe. This classification method permitted to classify the K_t histograms using a Dirichlet Mixture PDF. Our results have highlighted 4 classes, in agreement with those obtained by Soubdhan et al. in [5].

Table 3. validity criterion

Number of Cluster C	2	3	4	5	6
Validity Criteria S	0.43	0.37	**0.33**	0.51	0.57

The 4 classes obtained are clear sky day, intermittent clear sky day, cloudy sky day and intermittent cloudy sky day.

Table 4. Allocation of days according to their classes in 2012

	Clear Sky	Intermittent Clear Sky	Cloudy Sky	Intermittent Cloudy Sky
Number of day	109	84	62	111
In percentage	29	24	17	30

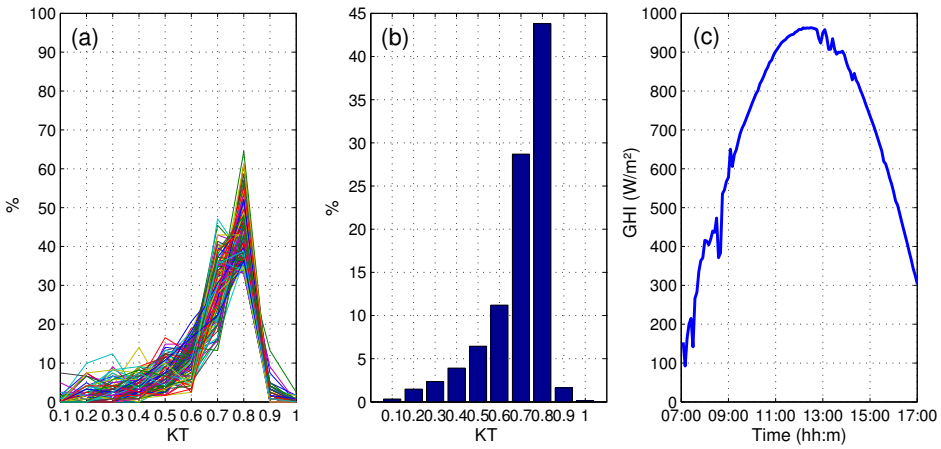


Figure 10. Class representing clear sky day event : a) Clearness index of 109 clear sky days b) Average K_t histogram of the 109 clear sky days c)a clear sky day GHI example

- Clear sky day (CS)

  In 2012 clear sky day events represent 29% GHI days of the year (Table 4). This type of day have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

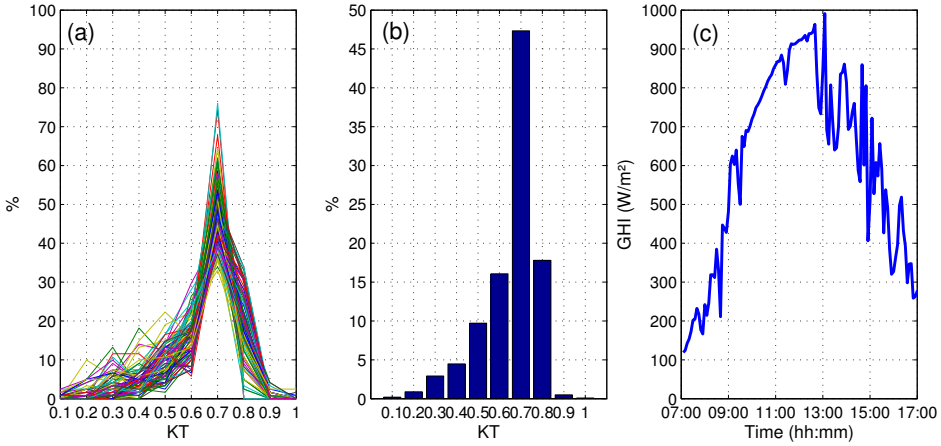


Figure 11. Class representing Intermittent clear sky day event : a) Clearness index of 84 intermittent clear sky days b) Average K_t histogram of the 84 intermittent clear sky days c)a intermittent clear sky day GHI example

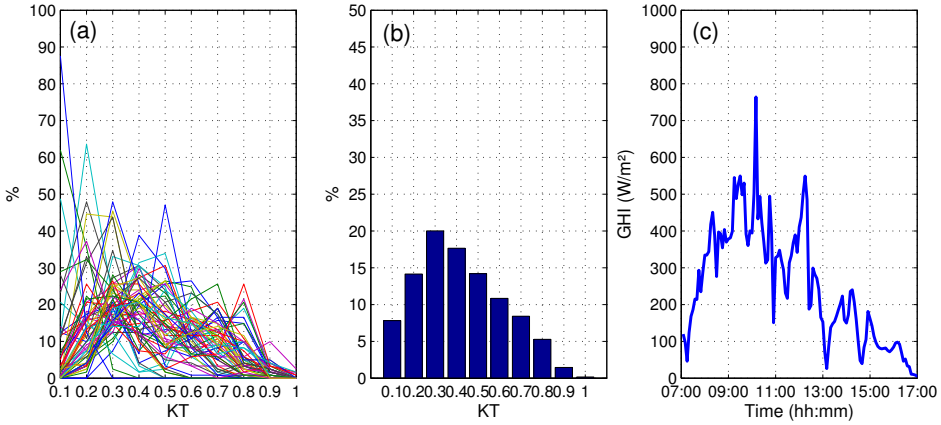


Figure 12. Class representing cloudy sky day event : a) Clearness index of 62 cloudy sky days b) Average K_t histogram of the 62 cloudy sky days c)a cloudy sky day GHI example

very few cloudy passages (Figure. 10.c). The clear sky day K_t distribution having a maximum of occurrence value(44%) around $K_t = 0.8$ (Figure. 10.a). Figure. 10.b show that around 86% of K_t values $\in [0.5; 1]$.

- Intermittent clear sky day (ICS)
  This second class represent 24% of events (Table 4). This type of day have an important solar radiation but the cloudy passages are frequent (Figure. 11.c). The K_t distribution having a maximum occurrence value (47%) around $K_t = 0.7$ and around 80% of K_t values $\in [0.5; 1]$ (Figure. 11.b).
- Cloudy Sky day(ClS)
  The cloudy sky day have important slow cloudy passages (Figure 12.c). This is the class is the least represented with 17% of the 2012 year days (Table 4). The K_t distribution having a maximum occurrence (20%)around $K_t = 0.3$ and around 73% of K_t values $\in [0; 0.5]$ (Figure 12.b).
- Intermittent cloudy sky day(IClS)
  For an intermittent cloudy sky day, important cloudy passages are observed (Figure. 13.c). In this class we have 111 days or 30% of the 2012 year (Table 4).The K_t distribution having a maximum occurrence (25%) around $K_t = 0.7$ and around 63% of K_t values $\in [0.5; 1]$ (Figure. 13.b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

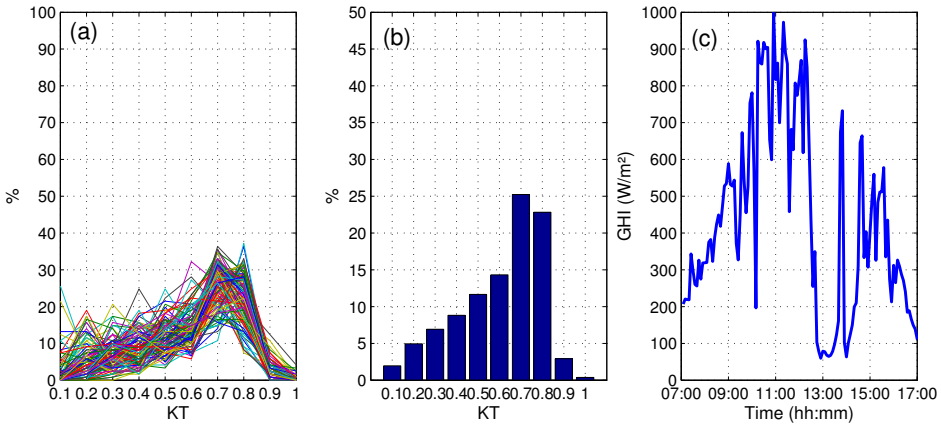


Figure 13. Class representing intermittent cloudy sky day event : a) Clearness index of 111 intermittent cloudy sky days b) Average K_t histogram of the 111 intermittent cloudy sky days c)a intermittent cloudy sky day GHI example

In the next section we study the global solar radiation variability. We analyse the fluctuation behavior at different time scale using the Validity Score (VS) parameter [15]. This parameter is based on the amplitude of fluctuations.

8.2. Variability characterization

The objective is to study the amplitude of fluctuations for different timescales to characterize the dynamic of each class of days. To achieve this we used the variability score. Lave et al.[15] define the variability score as the maximum value of ramp rate magnitude (RR_0) times ramp rate probability (Eq.15). The variability score is determined using the cumulative distribution function of ramp rates using a given timescale [16].

$$VS(\Delta t) = 100 * max(RR_0.P(|RR_{\Delta t}| > RR_0)) \tag{15}$$

where Δt represents the time scale, RR_0 and $RR_{\Delta t}$ are a percent of Standard Test Conditions (STC) irradiance $= 1000W.m^{-2}$. The probability $P(|RR_{\Delta t}| > RR_0)$ represents the fraction of time where the absolute value of $RR_{\Delta t}$ is higher than RR_0. Like in [15] we choose to use the moving averages definition of ramp rates $RR_{\Delta t}$ given by Kleissl [30]:

$$RR_{\Delta t}(t) = \frac{1}{\Delta t}(\sum_{t}^{t+\Delta t} GHI - \sum_{t-\Delta t}^{t} GHI) \tag{16}$$

Firstly we present in Figure. 14 the Δt-cumulative distribution of GHI ramp rate for each class. These figures allow to have an information on the amplitude of the fluctuations and their probability for each type of days. In Figure. 14.a) we can see that 35% of ramp rates for the clear sky days were larger than $50W.m^{-2}$. This probability achieve 40% for the intermittent clear sky days, 50% for the cloudy sky days and 55% for the intermittent cloudy sky days. When we increases Δt the probability to be larger than $50W.m^{-2}$ increase too (Figure. 14.b, Figure. 14.c and Figure. 14.d). We also noticed that when Δt increase the behavior of the Cloudy Sky cumulative distribution of GHI ramp rate change, its achieve the lowest probability to have an amplitude of fluctuations higher than $50W.m^{-2}$ when Δt= 20 min. That means the cloudy sky days have a low dynamic. Whatever the observed amplitude of fluctuations, the intermittent cloudy sky days always have the highest probability to be higer than RR_0. For Δt= 15 and 20 min, when RR_0 reach 25% of STC ($250W.m^{-2}$), the cloudy sky day and intermittent cloudy sky days cumulative distribition (respectively clear sky days and intermittent clear sky days cumulative distribution) eventually have the same behavior.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

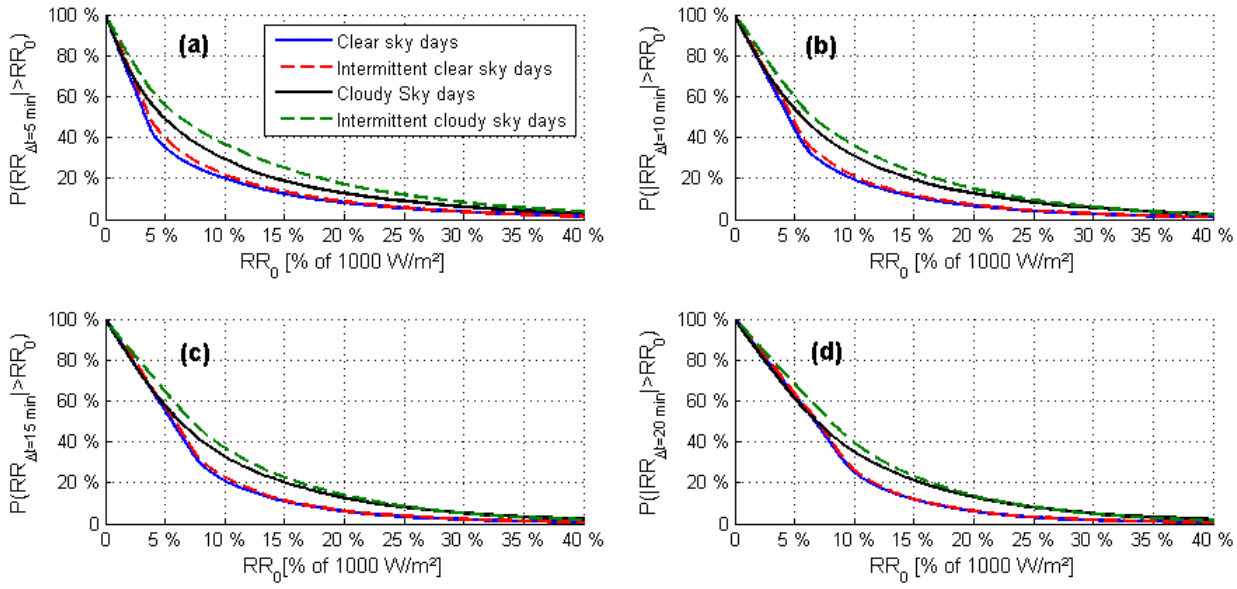


Figure 14. Cumulative distribution of GHI ramp rate for $\Delta t = 5min$ (a),$\Delta t = 10min$ (b), $\Delta t = 15min$ (c) and $\Delta t = 20min$ (d)

Then we present the Varaibility Score over timescale Δt (Figure. 15). We show that for each class VS increase over Δt. The intermittent cloudy sky days is the most variable and the clear sky day the less variable. In [15] Lave et al. made the same observation. They determine the VS over Δt for 10 sites in USA, chose to considerate Δt= 1s, 10s, 30s, 60s and 3600s and demonstrate that VS increase over Δt.

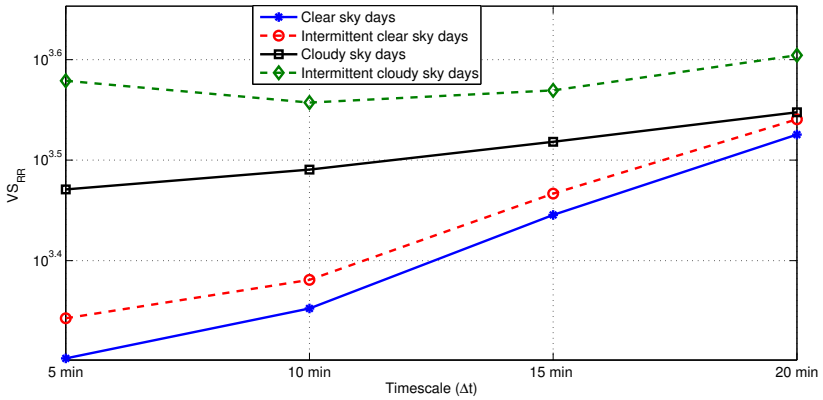


Figure 15. Variability score VS versus timescale Δt for each class in 2012 in semi-log representation

This variability characterization will have allowed to approve the robustness of the clustering method. Indeed, we found 4 type of days with 4 different variability whether in term of amplitude of the fluctuations that of dynamics of day. In the next section we presents the MHFM performances according to the type of days.

8.3. Variability influence on Hybrid forecast model performances

We applied the first strategy presented in section 5-1. We used 6 months the data available for the learning phase of the hybrid model then predicted 6 other months. The errors were calculated for every day predicted independently. Then, through the classification made beforehand we grouped the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2018 doi:10.20944/preprints201809.0089.v1

days of the same class and established an average of the error obtained. These errors are summarized in Table 5 for a forecast horizon of 5 minutes. The best performances are obtained with the WD-Hybrid model. The clear sky days are the best predicted ($rRMSE = 2.91\%$), their low variability facilitate their forecast by the model. We note in the previous section that the most variable days is the intermittent cloudy sky days but the MHFM performances for this type of days ($rRMSE = 5.48\%$) is better than this obtained for cloudy sky days ($rRMSE = 6.73\%$). The variability of GHI signal is not the only factor being able to influence the forecast error. The cloudy sky days is the second type of days the most variable and the only type of days which have a very low GHI (generally $< 600W.m^{-2}$). Moreover, the cloudy sky day $rMBE$ obtained with the WD-Hybrid Model is the highest ($rMBE = 0.32\%$). The 3 others type of day being the most represented (Table 4), during NN and AR the learning phase the hybrid model rarely meet the cloudy sky days GHI signal and finally overestimate their forecasts. All these observations show that it is the combination of high variability and low GHI value which could justify the least good performances of the model on the cloudy sky days class. Nevertheless, the robustness of MHFM allows to limit their effects on the quality of the forecast to obtain a rRMSE lower than 7%.

Table 5. Forecasting performances of MHFM for the 4 specific class of days

		EMD-Hybrid Model	EEMD-Hybrid Model	WD-Hybrid Model
Clear Sky	rMBE (%)	0.21	-0.20	**0.02**
	rMAE (%)	5.14	3.70	**2.00**
	rRMSE (%)	7.47	5.84	**2.91**
Intermittent Clear Sky	rMBE (%)	**0.07**	-0.51	0.08
	rMAE (%)	6.00	4.39	**2.35**
	rRMSE (%)	8.76	6.95	**3.35**
Cloudy Sky	rMBE (%)	**0.11**	-0.21	0.32
	rMAE (%)	10.14	7.71	**4.72**
	rRMSE (%)	14.81	11.76	**6.73**
Intermittent Cloudy Sky	rMBE (%)	**0.15**	-0.19	0.17
	rMAE (%)	9.32	7.42	**3.81**
	rRMSE (%)	13.36	11.16	**5.48**

9. Conclusion

The goal of this study concerns the influence of time sampling combined to the forecast horizon and the solar variability on the hybrid forecast model performances. Two strategies for the forecast at several horizon are proposed, the first based on the resampling data in order that the data sampling equal to the forecast horizon and the second on order that all the horizon are obtained with the same data sampling (5 minutes). Like in [1] we tested the hybrid model for 3 multiscale decomposition methods (EMD, EEMD and WD). With the first strategy, whatever the Multiscale decomposition-hybrid model chosen the rRMSE error increases with the forecast horizon and the best results are obtained with the WD-Hybrid model (rRMSE varying between 4.41% and 11.42 %). With the second strategy the AR and NN learning phase performs with 5 minutes data sampling and provides a forecasting at $t + \tau$. The rRMSE error increases with the forecast horizon and the WD-Hybrid Model obtain a rRMSE varying from 4.43% to 22.33%. Like demonstrated by Monjoly et al. in [1] the best results are obtained with the WD-Hybrid model. The forecast horizon strategy based on the resampling data (first strategy) is the most efficient. Solar variability influence are determined using a daily classication of the GHI based on clearness index and fuzzy c-means algorithms. We obtain 4 classes according to previuos studies : clear sky day, intermittent clear sky day, cloudy sky and intermittent cloudy sky. Then we describe the variability characterization of each class based on the variability score : as expected, the intermittent cloudy sky days is the most variable and the clear sky day the less variable for exemple 35% of ramp rates for the clear sky days were larger than $50W.m^{-2}$. This

probability achieve 40% for the intermittent clear sky days, 50% for the cloudy sky days and 55% for the intermittent cloudy sky days. Nevertheless the hybrid forecast performances obtained with the most variable class (Intermittent cloudy sky) aren't the worst this reveals the weakness of the hybrid model for these cases qualified to extreme events. Consequently, in future works, the hybrid model should be improve in this way.. We note that the variability of GHI signal is not the only parameter being able to influence the forecast, the daily GHI profile is an other one. Indeed, the cloudy sky day is the only class with a low GHI profile, its the second class the most variable and it's the type of day having the highest rRMSE (6.73%).

Author Contributions: Stéphanie Monjoly contributed to perform the analyzes and the writing of the manuscript. Rudy Calif contributed to the supervision and the general design of the manuscript. Maïna André et Ted Soubdhan contributed to the review and editing of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

1. Monjoly, S.; André, M.; Calif, R.; Soubdhan, T. Hourly forecasting Global Solar radiation based on Multiscale decomposition methods - A hybrid Approach. *Energy.* **2017**, *119*, 288–298.
2. Diagne, M.; David, M.; Lauret, P.; Boland, J.; Schmutz, N. Review of Solar irradiance forecasting methods and a proposition for small scale insular grids. *Renewable and Sustainable energy reviews* **2013**, *27*, 65–76.
3. Soubdhan, T.; Abadi, M.; Emilion, R. Time Dependent Classification of Solar Radiation Sequences Using Best Information Criterion. *Energy Procedia.* **2014**, *57*, 1309–1316.
4. Badosa, J.; Haeffelin, M.; Chepfer, H. Scales of spatial and temporal variation of solar irradiance on Reunion tropical island. *Solar Energy.* **2013**, *88*, 42–56.
5. Soubdhan, T.; Emilion, R.; Calif, R. Classification of daily solar radiation distributions using a mixture of Dirichlet distributions. *Solar Energy.* **2009**, *83*, 1056–1063.
6. Muselli, M.; Poggi, P.; Notton, G.; Louche, A. Classification of typical Meteorological Days from Global Irradiation Records ad Comparison Between Two Mediterranean Coastal Sites in Corsica Island. *Energy Conversion and Management* **2000**, *41*, 1043–1063.
7. Maafi, A.; Harrouni, S. Preliminary results of the fractal classification of daily solar irradiances. *Solar Energy* **2003**, *75*, 53–61.
8. Benmouiza, K.; Tadj, M.; Cheknane, A. Classification of Hourly Solar Radiation using Fuzzy-C means algorithm for optimal stand-alone PV system sizing. *Electrical power and Energy Systems* **2016**, *82*, 233–241.
9. Mingoti, S.A.; Lima, J.O. Comparing SOM neural network and with fuzzy c-means , k-means and traditional hieralchical clustering algorithm. *Eur J Oper Res* **2006**, *174*, 1742–1759.
10. Lu, Y.; Ma, T.; Yin, C.; Xie, X.; Tian, W.; Zhong, W. Implementation of the fuzzy C-means clustering Algorithm in Meteorological Data. *International Journal of Database Theory and Application* **2013**, *6*, 1–18.
11. Perez, R.; Kivalov, S.; Schlemmer, J.; Hemker.Jr, K.; Hoff, T. Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance. *Solar Energy* **2012**, *86*, 2170–2176.
12. Perez, R.; Kivalov, S.; Schlemmer, J.; Hemker.Jr, K.; Hoff, T. Parametrization of site-specific short-term irradiance variability. *Solar Energy* **2011**, *85*, 1343–1353.
13. Hoff, T.; Perez, R. Quantifying PV Power Output Variability. *Solar Energy* **2010**, *84*, 1782–1793.
14. Lave, M.; Kleissl, J.; Stein, J. A wavelet-based variability model (WVM) for solar PV power plants. *IEEE Trans Sustain Energy* **2012**, *99*, 1–9.
15. Lave, M.; Reno, M.J.; Broderick, R.J. Characterizing local High-frequency solar variability and its impact to distribution studies. *Solar Energy* **2015**, *118*, 327–337.
16. Gagné, A.; Turcotte, D.; Goswamy, N.; Poissant, Y. High resolution characterisation of solar variability for two sites in Eastern Canada. *Solar Energy* **2016**, *137*, 46–54.
17. Dambreville, R.; Blanc, P.; Chanussot, J.; Boldo, D. Very short term forecasting of the global horizontal irradiance using a spatio- temporal autoregressive model. *Renewable Energy.* **2014**, *72*, 291–300.
18. Lauret, P.; Voyant, C.; Soubdhan, T.; Mathieu, D.; P.Poggi. A benchmarking of machine learning technique for solar radiation forecasting in an insular context. *Solar Energy.* **2015**, *112*, 446–457.

19. Kasten, F. A simple parametrization of two pyrheliometric formulae for detremining the linke turbidity factor. *Meteorol. Rundsch.* **1980**, pp. 124–127.
20. Bacher, P.; Madsen, H.; Nielsen, H.A. Online short-term solar power forecasting. *Solar Energy* **2009**, *83*, 1772–1783.
21. Coimbra, C.F.M.; Kleissl, J.; Marquez, R. Overview of Solar Forecasting Method and a metric for accuracy evaluation. In *Solar Energy Forecasting and Resource Assessment, Elsevier*; J.Kleissl., Ed.; 2013; pp. 171–194.
22. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.; Shih, H.H.; Q. Zheng, N.C.Y.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert sprectrum for nonlinear and non-stationnary time series analysis. *Proc. R. Soc. London.* **1998**, *454*, 903–995.
23. Jiménez-Pérez, P.; Mora-López, L. Modelling and Forecasting hourly Global Solar Radiation using Clustering and Classification techniques. *Solar Energy* **2016**, *135*, 682–691.
24. Kasten, F. The link Turbidity Factor Based on Improved Values of integral Rayleigh Optical thickness. *Solar Energy* **1996**, *56*, 239–244.
25. Ruspini, E.H. Numerical Methods for fuzzy clustering. *Information Sciences.* **1970**, *2*.
26. Dunn, J.C. A fuzzy relative of ISODATA process and its use in clustering compact and well separated cluster. *Journal Cybernetics.* **1974**, *3*, 32–57.
27. JC Bezdek. *Pattern recognition with fuzzy objective function algorithms Kluwer Academic Publishers, Norwell*; 1981.
28. Bezdek, J.; Ehrlich, R. Numerical Methods for fuzzy clustering. *Computer and Geosciences.* **1984**, *10*, 191–203.
29. Xie, X.; Beni, G. A validity measure for fuzzy clustering. *EEEI Trans. Pattern Anal Machine Intell* **1991**, *8*, 841–847.
30. J Kleissl. *Solar Energy Forecasting and Resource Assessment. Academic Press*; 2013.

Sample Availability: Samples of the compounds ..... are available from the authors.