



**HAL**  
open science

## A study on the minimum duration of training data to provide a high accuracy forecast for PV generation between two different climatic zones

Ted Soubdhan, Minh-Thang Do, Benoît Robyns

### ► To cite this version:

Ted Soubdhan, Minh-Thang Do, Benoît Robyns. A study on the minimum duration of training data to provide a high accuracy forecast for PV generation between two different climatic zones. *Renewable Energy*, 2016, 85, pp.959-964. 10.1016/j.renene.2015.07.057 . hal-01823260

**HAL Id: hal-01823260**

**<https://hal.science/hal-01823260v1>**

Submitted on 3 Jun 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 A study on the minimum duration of training data  
2 to provide a high accuracy forecast for PV  
3 generation between two different climatic zones

4 Minh Thang DO<sup>1\*</sup>, Ted SOUBDHAN<sup>1</sup>, Benoit ROBYNS<sup>2</sup>

5 <sup>1</sup> Laboratory LARGE, University of Antilles and Guiana, Guadeloupe, France

6 <sup>2</sup> Laboratory L2EP, Ecole des Hautes Etudes d'Ingénieur (HEI), Lille, France

7 \*Corresponding author. Address: University of Antilles and Guiana, 97159 Pointe-à-Pitre cedex, France. E-mail  
8 address: mtdo@univ-ag.fr

9 Abstract:

10 This study focus on the minimum duration of training data required for PV generation forecast.  
11 In order to investigate this issue, the study is implemented on 2 PV installations: the first one  
12 in Guadeloupe represented for tropical climate, the second in Lille represented for temperate  
13 climate; using 3 different forecast models: the Scaled Persistence Model, the Artificial Neural  
14 Network and the Multivariate Polynomial Model. The usual statistical forecasting error  
15 indicators: NMBE, NMAE and NRMSE are computed in order to compare the accuracy of  
16 forecasts.

17 The results show that with the temperate climate such as Lille, a longer training duration is  
18 needed. However, once the model is trained, the performance is better.

19 Keywords: PV forecasting models, neural network, multivariate model, forecasting errors, training  
20 duration.

21 1. Introduction

22 With the support of environment policies and the increase of the fossil fuel price, renewable  
23 energy sources have been growing strongly in the last few years. The electric power produced  
24 by these intermittent sources shows strong variations (sudden and with large amplitude), which  
25 must be always compensated on the grid by others dispatchable sources (Ernst et al., 2009),  
26 (Do et al., 2010), (Do et al., 2011).

27 These large variations can put a pressure on the balance of supply/demand of the power system,  
28 especially in non-interconnected systems, such as island areas. In the case of French islands for  
29 which Island Energy System is the electrical system manager, in order to ensure the stability of  
30 the power system, a ministerial decree of 2008 set the penetration rate at 30%, beyond which  
31 the system operator is permitted to disconnect intermittent energy (CRE, 2009). Note that this  
32 rate of 30% was achieved for a few hours in 2012 in Reunion and in Guadeloupe, resulting in  
33 the disconnection of certain facilities of PV production.

34 The forecast of these fluctuation sources across the concerned islands should allow a better  
35 control of the availability of renewable energy production, and thereby reduce the pressure on  
36 the balance of supply/demand. Moreover associated with power storage unit (batteries or STEP,  
37 for example) will help providing various services to the power system, from ancillary services  
38 (adjusting voltage, frequency) to smoothing peak hours. The forecast of fluctuation sources will  
39 contribute to the optimization of the design and the use of these storage units.

40 Generally, the PV production forecast is classified into 2 categories: the day ahead (DA)  
41 forecast and the hour ahead (HA) forecast, depend on the domain of application (IEA, 2013).

42 - The day ahead (DA) forecast is usually based on the numerical weather prediction models.  
43 They are dynamical equations that predict the evolution of the atmosphere up to several  
44 days ahead from initial conditions. From the forecast of weather condition, the output power

45 of PV can be estimated (Beyer HG et al., 2009), (Lorenz E et al., 2008), (Traunmüller W  
46 & Steinmaurer G., 2010).

47 - The hour ahead (HA) forecast is usually based on stochastic learning techniques. The  
48 underlying assumption of these techniques is that future value of PV production can be  
49 predicted by training the algorithms with historical data (Fernandez-Jimenez et al., 2012),  
50 (Pedro & Coimbra, 2012), (Mandal et al., 2012).

51 To generate a prediction of the PV production using stochastic learning techniques, it is  
52 necessary to have historical data. These data are used for the learning phase of forecast models  
53 of PV production at the studied site (IEA, 2013).

54 In the literature, several studies have been conducted on the methodology for the forecast of PV  
55 generation (Shi et al., 2012), (Krömer et al., 7–11 July 2012), (Lorenz et al., 2012). The authors  
56 generally use two historical years: one year for training the model and another year for the test  
57 phase. In the literature, the analysis of the minimum duration for the phases of test and training  
58 data is little documented.

59 The proposed study will determine the minimum of historical experimental data necessary to  
60 achieve a high accuracy forecast, which can be quantified by a set of several statistical error  
61 indicators. There are some works in other areas that focus on this problem (Fine & Turmon,  
62 1994), (Thirumalainambi, 2003), (Cui et al., 2004) but a study on the prediction of PV  
63 generation, to our knowledge, is not yet performed.

64 There are two major factors that affect the time required for the collection of historical data,  
65 they are the climatic conditions at the PV site and the statistical model used for the forecast.

66 The climatic conditions at the PV site affects the level and the type of data fluctuation. The  
67 more fluctuate the data is, the more they are difficult to predict, i.e.: the error on the forecast is  
68 higher. Therefore, the duration of the data collection is supposed to be longer. In this study, we

69 investigate the influence of climatic conditions on the historical data necessary for a forecast of  
70 PV production by comparing two sites of PV generation with very different climatic conditions:  
71 one in Lille, in northern France and the other one in Guadeloupe, in the Caribbean.

72 The forecasting model also has a large influence on the duration required for the collection of  
73 historical data. There are simple forecasting models that requires a few data for the learning  
74 phase. The more complex the forecasting models is, the more historical data is needed but the  
75 result is often more accurate.

76 In this study, we will compare some of the most popular statistical models in forecasting PV  
77 production: Artificial Neural Network (ANN) (Yona et al., 2007), (Fernandez-Jimenez et al.,  
78 2012), (Mandal et al., 2012), Multivariate Polynomial Model (MPM) (Dazhi, 2012), (IEA,  
79 2008) and the Scaled Persistence Model (IEA, 2013).

## 80 2. Input data and site description

81 In this paper, we will use PV production at time ( $h$ ) and two exogenous inputs: the cloud cover  
82 and the air temperature, also at time ( $h$ ), to forecast the PV production at time ( $h+1$ ).

83 For each hour ( $h$ ) considered as the beginning time of the forecasting, the input vectors are  
84 given by:

$$x(h) = [N(h); T_a(h); P_m(h)] \quad (1)$$

85 Where: -  $N(h)$  is the cloud cover measured at time ( $h$ )  
86 -  $T_a(h)$  is the ambiance temperature measured at time ( $h$ )  
87 -  $P_m(h)$  is the average output power produced by the PV system in the previous  
88 60 minutes respective to the  $h$ -hour

### 89 a. Data measured in Guadeloupe, tropical climate

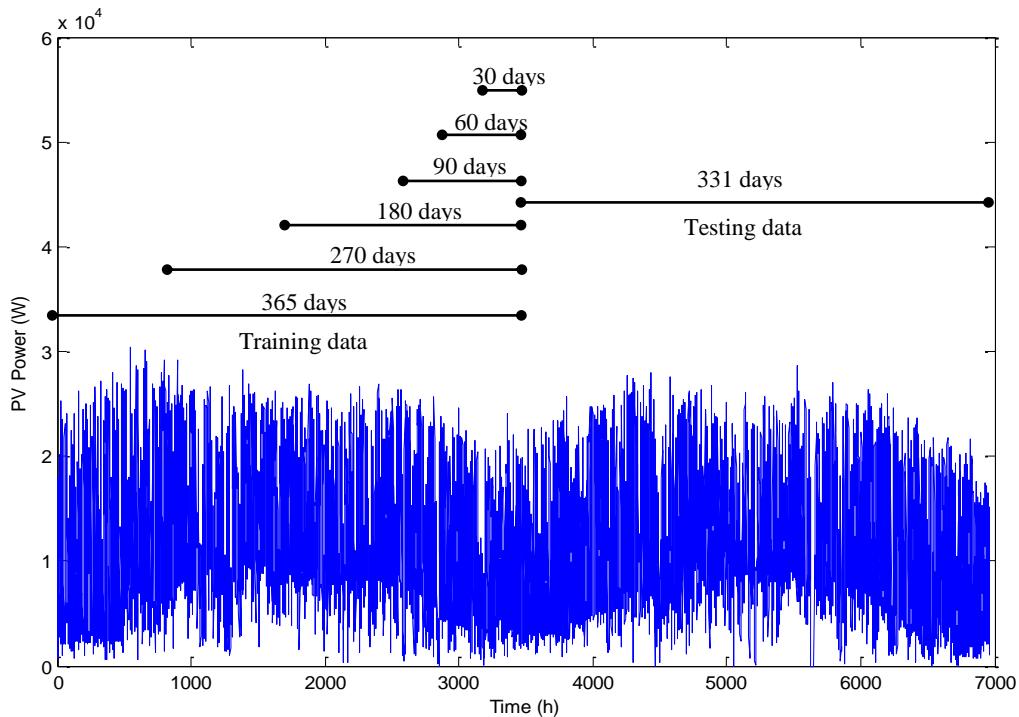
90 The PV system installed in Guadeloupe (16°14'36.0"N 61°33'25.9"W) has 832 membrane PV  
91 modules of 136Wp each, which makes a total of 113kWc.

92 The system consists of two inverters. The data logging of PV output power (in W) is integrated  
 93 in these inverters (collected every 5 min). The two exogenous inputs: cloud cover (in octa) and  
 94 air temperature (in °C) are measured every hour by the weather station at the airport nearby  
 95 (Raizet Airport) by Meteo France. As during the night, the forecast is not necessary because the  
 96 output power of PV is zeros, the study takes into account only the data between 7a.m and 17p.m  
 97 every day.

98 The average output power of PV in every hour is calculated from the measurement every 5  
 99 minutes of the inverters:

$$P_m(h) = \frac{1}{12} \sum_{t=h-55\text{min}}^h P(t) \quad (2)$$

100 The forecast models will be studied with a varying training period from one month to one year  
 101 (365 days) and a testing period of 331 days. The output power of the PV system at time  $(h+1)$ ,  
 102  $P_m(h+1)$  will be predicted from three inputs:  $N(h)$ ,  $T_a(h)$  and  $P_m(h)$ .



103

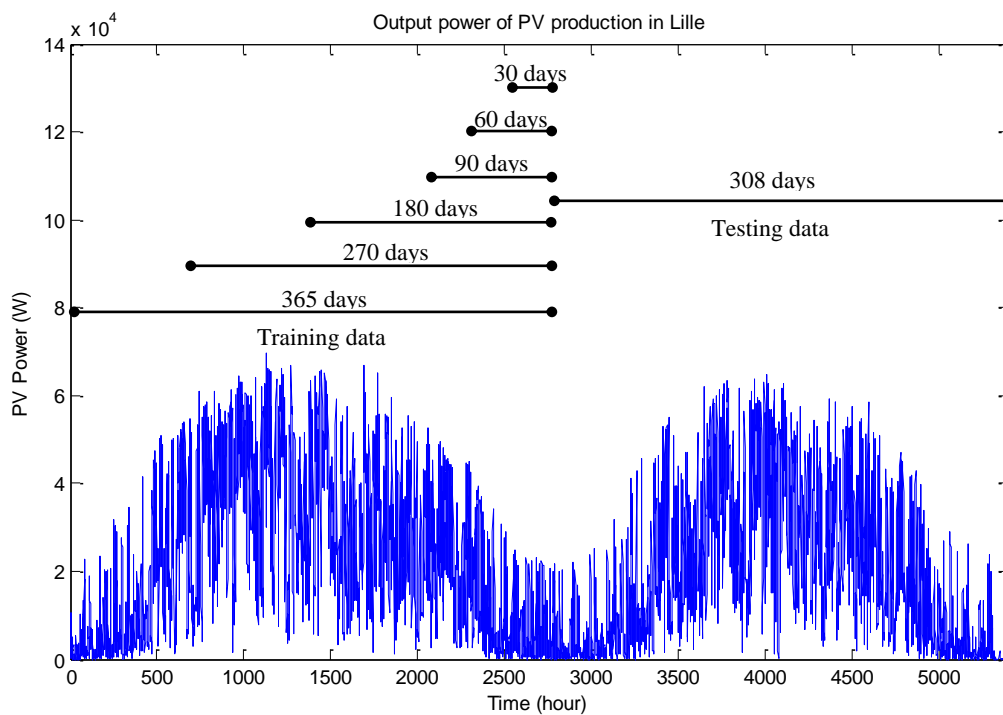
104

Figure 1. Output power for training and testing in Guadeloupe

105 b. Data measured in Lille, temperate climate

106 The installation of a 93kWc PV system in Lille (50°38'58.7"N 3°09'04.3"E) consists of 452  
107 panels. Among these panels, 376 modules have an inclination of 3° and the inclination of the  
108 rest is 60°.

109 The system consists of 31 inverters. The data logging of PV output power (in W) is integrated  
110 in these inverters (collected every 10 min). The two exogenous inputs: cloud cover (in octa)  
111 and ambiance temperature (in °C) are measured every hour by the weather station at the airport  
112 nearby (Lesquin Airport) by Meteo France. As the sun rise is later in Lille, the study takes into  
113 account only the data between 10a.m and 17p.m every day to avoid the zero output power of  
114 PV during the night.



115

116 Figure 2. Output power for training and testing in Lille

117 The average output power of PV in every hour is calculated from the measurement every 10  
118 minutes of the inverters:

$$P_m(h) = \frac{1}{6} \sum_{t=h-50\text{min}}^h P(t) \quad (3)$$

119 The forecast models will be studied with a varying training period from one month to one year  
 120 (365 days) and a testing period of 308 days. The output power of the PV system at time  $(h+1)$ ,  
 121  $P_m(h+1)$  will be predicted from three inputs:  $N(h)$ ,  $T_a(h)$  and  $P_m(h)$ .

122 c. Evaluating the quality of forecast

123 The precision of the forecast will be evaluated using the following statistical error indicators:

124 - Normalized mean bias error (%)

$$NMBE = \frac{1}{M} \frac{\sum_{h=1}^M P_m(h) - \tilde{P}_m(h)}{\max(P_m) - \min(P_m)} \times 100 \quad (4)$$

125

126 - Normalized mean absolute error (%)

$$NMAE = \frac{1}{M} \frac{\sum_{h=1}^M |P_m(h) - \tilde{P}_m(h)|}{\max(P_m) - \min(P_m)} \times 100 \quad (5)$$

127

128 - Normalized root -mean-square error (%)

$$NRMSE = \frac{\sqrt{\frac{1}{M} \sum_{h=1}^M (P_m(h) - \tilde{P}_m(h))^2}}{\max(P_m) - \min(P_m)} \times 100 \quad (6)$$

129 Where: -  $P_m(h)$  is the average output power produced by the PV system at hour  $(h)$

130 -  $\tilde{P}_m(h)$  is the predicted output power of the PV system at hour  $(h)$

131 -  $M$  is the number of hours considered

132 3. The forecasting models

133 In this research, the forecasting models will take into account the data of PV production at time

134  $(h)$ ,  $P_m(h)$  and two exogenous inputs: the cloud cover  $N(h)$  and the air temperature  $T_a(h)$ , also

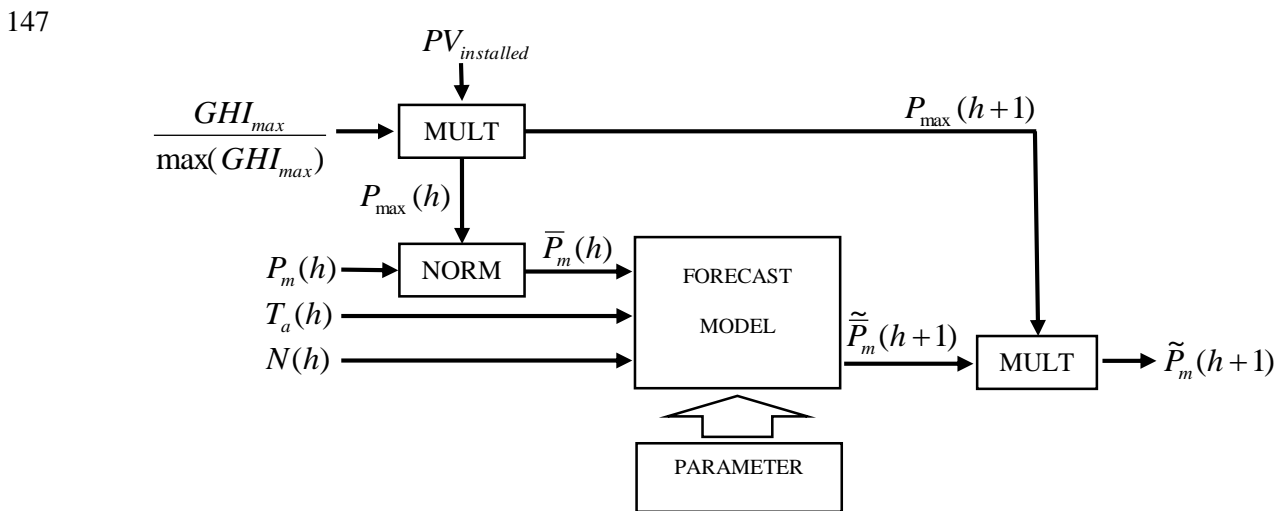


135 at time ( $h$ ), to predict the PV production at time ( $h+1$ ). As the PV production depends on the  
 136 position of the sun in the day while others input variables does not, a normalization is needed  
 137 to eliminate this subordination. The input of the PV production at time ( $h$ ) becomes now  $\bar{P}_m(h)$ ,  
 138 the normalized value of the PV production with respect to the maximum value at time ( $h$ ),  
 139  $P_{max}(h)$ .

140 The absolute value of the PV output power prediction is then obtained by multiplying this  
 141 normalized value with the maximum value of PV production at time ( $h+1$ ),  $P_{max}(h+1)$ . This  
 142 maximum value can be evaluated from the  $GHI_{max}$  curve with the following equation:

$$P_{max} = \frac{GHI_{max}}{\max(GHI_{max})} \cdot PV_{installed} \quad (7)$$

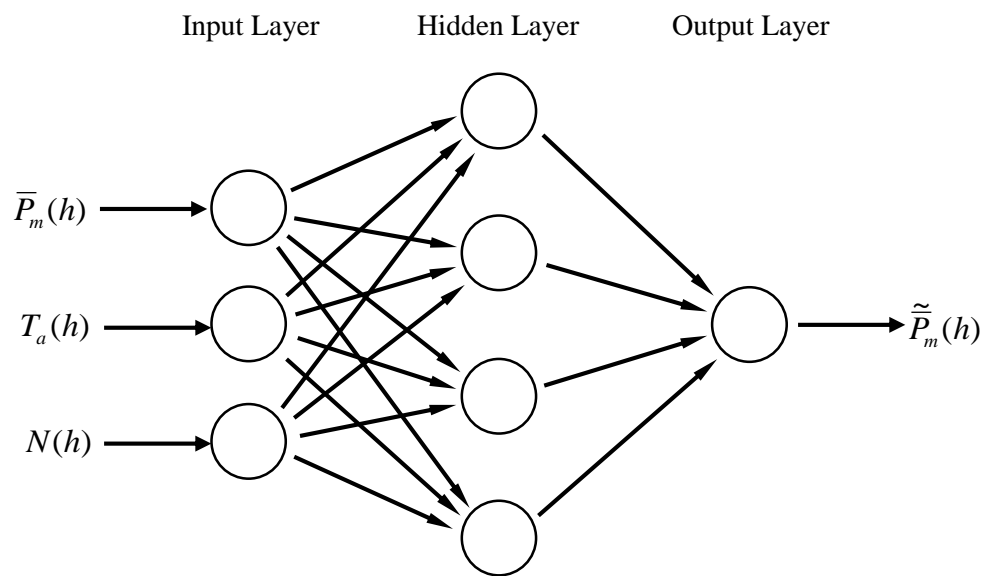
143 The Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received  
 144 from above by a surface horizontal to the ground, which consists of the direct irradiance and  
 145 the diffuse irradiance. The  $GHI_{max}$  is the GHI calculated in the condition of clear sky, using the  
 146 Kasten clear sky models. (Kasten, 1980)



148 Figure 3. General process of the PV output power forecast

149 a. Artificial Neural Network

150 The Artificial Neural Network used in this research is a feedforward neural network for non-  
 151 linear regression with 1 hidden layer. The choice of the number of neurons in the hidden layer  
 152 is a complicated issue. In this paper, a value of 4 is chosen based on several researches in  
 153 literature (Blum, 1992), (Swingler, 1996), (Berry & Linoff, 1997), (Boger, 1997). The Artificial  
 154 Neural Network will be trained by the historical data with the supervised learning technique,  
 155 using the Levenberg-Marquardt algorithm (Seber & Wild., 2003).



156

157 Figure 4. Diagram of the Artificial Neural Network configuration

158 b. Multivariate Polynomial Model

159 The polynomial models are very popular due to their simplicity in form, their well-known and  
 160 understood properties and their flexibility of shapes. Moreover, they are computationally easy  
 161 to use. For these reasons, there are many applications of these models in the forecasting in  
 162 general, and in PV production forecasting in particular (Dazhi, 2012), (IEA, 2008).

163 However, as the model takes into account only one input variable, the precision of forecast is  
 164 limited and therefore the Artificial Neural Network model becomes the most applied model in

165 this field (Pedro & Coimbra, 2012). In this paper, a Multivariate Polynomial Model, which  
 166 takes into account several input variables is proposed to improve the accuracy of the prediction  
 167 of PV output power.

168 The output prediction of the normalized value at time  $(h+1)$  is a multivariate polynomial  
 169 function of the input variables:

$$\begin{aligned} \tilde{P}_m(h+1) = & b_1 \cdot \bar{P}_m(h) \cdot N(h) \cdot T_a(h) + b_2 \cdot \bar{P}_m(h) \cdot N(h) + b_3 \cdot \bar{P}_m(h) \cdot T_a(h) + b_4 \cdot N(h) \cdot T_a(h) \\ & + b_5 \cdot \bar{P}_m(h) + b_6 N(h) + b_7 T_a(h) + b_8 \end{aligned} \quad (8)$$

170 The coefficients of the multivariate polynomial function are estimated from the historical data  
 171 using the Levenberg-Marquardt nonlinear least squares algorithm (Seber & Wild., 2003). This  
 172 algorithm allows to determine the set of parameters  $b$  giving the minimal squares of the  
 173 deviations:

$$S(b) = \sum_{h=1}^M \left[ \bar{P}_m(h) - \tilde{P}_m(h) \right]^2 \quad (9)$$

174 c. Persistence Model

175 The persistence model is based on a simple rule: the output value of the predicted variable at  
 176  $(h+1)$  is equal to its value at  $(h)$ . The advantage of this technique is that it does not need to be  
 177 trained by a series of historical data, however the accuracy of the forecast is not high.

178 In this paper, the Scaled Persistence Model is applied in order to reduce the forecasting error.

179 This model is applied on the normalized value:

$$\tilde{P}_m(h+1) = \bar{P}_m(h) \quad (10)$$

181 4. Results

182 In this section, we will analyze the influence of the duration of training data on the forecasting  
 183 error. The parameters of the forecast model (Artificial Neural Network and Multivariate  
 184 Polynomial Model) will be estimated using the historical data in 30 days, 60 days, 90 days, 180  
 185 days, 270 days and 365 days. Then, the model will be tested on a duration of 331 days  
 186 (Guadeloupe) and 308 days (Lille). Due to its characteristic, the Scaled Persistence Model does  
 187 not require any historical data, therefore, it will be tested directly with the testing data.

188 4.1 Case of Guadeloupe

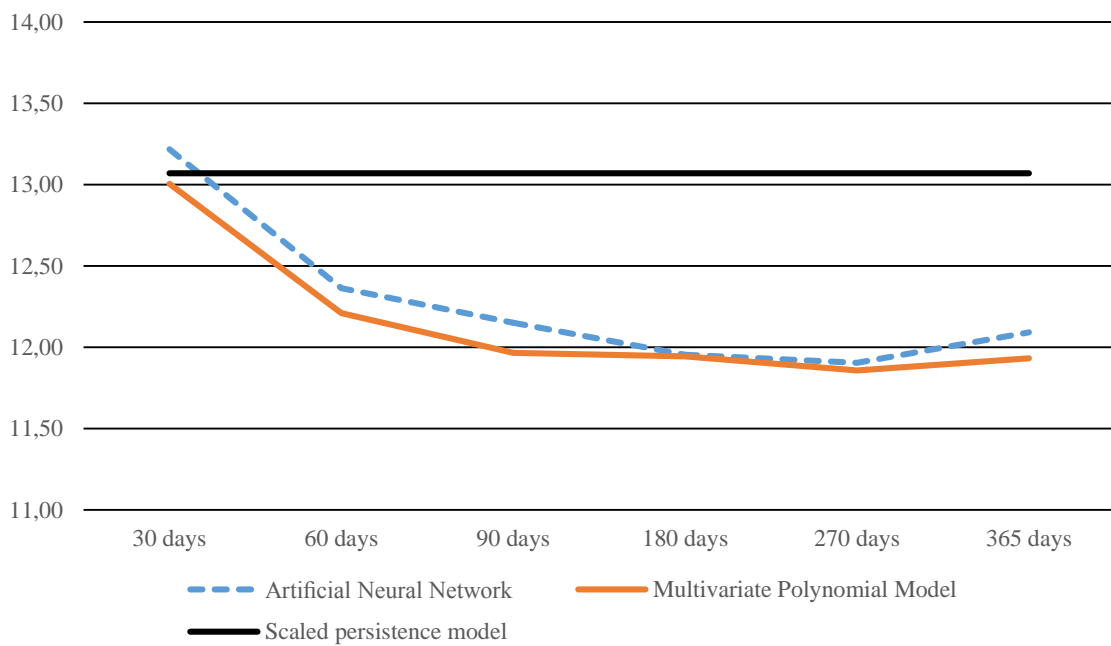
189 The Table 1 and the Figure 5 show the evolution of the forecast error with different training  
 190 duration in Guadeloupe.

191 Table 1. The evolution of the forecast error with different training duration (Guadeloupe)

NMBE			
Duration of training data	Artificial Neural Network	Multivariate Polynomial Model	Scaled Persistence Model
30days x 10 samples	1.74%	-0.87%	0.79%
60days x 10 samples	0.22%	-1.41%	
90days x 10 samples	-0.62%	-0.61%	
180days x 10 samples	-0.30%	-0.34%	
270days x 10 samples	0.27%	0.27%	
365days x 10 samples	0.67%	0.60%	
NMAE			
30days x 10 samples	9.90%	10.06%	9.60%
60days x 10 samples	9.32%	9.51%	
90days x 10 samples	9.23%	9.26%	
180days x 10 samples	9.13%	9.21%	
270days x 10 samples	9.07%	9.10%	
365days x 10 samples	9.22%	9.19%	
NRMSE			

30days x 10 samples	13.22%	13.00%	13.07%
60days x 10 samples	12.36%	12.21%	
90days x 10 samples	12.15%	11.97%	
180days x 10 samples	11.95%	11.94%	
270days x 10 samples	11.90%	11.86%	
365days x 10 samples	12.09%	11.93%	

192



193

194 Figure 5. The evolution of the NRMSE (%) with different training duration (Guadeloupe)

195 The value of the bias error NMBE allows to evaluate the tendency of the prediction model to  
 196 underestimate or overestimate the PV output power. In this study, the NMBE is relatively small  
 197 with all 3 methods and oscillating around the zero. However, the variation of the NMBE does  
 198 not follow any certain rule (Table 1).

199 The NMAE is used to measure how close the forecasts or predictions are to the eventual  
 200 outcomes. As presented in the Table 1, from the training data of more than 1 month, the NMAE  
 201 of the Artificial Neural Network and the Multivariate Polynomial Model is lower than the

202 Scaled Persistence Model. The NMAE of the Multivariate Polynomial Model is slightly higher  
203 than that of the Artificial Neural Network.

204 The Root-Mean-Square Error (RMSE) is a frequently used measure of the differences between  
205 value (Sample and population values) predicted by a model or an estimator and the values  
206 actually observed. Basically, the RMSE represents the sample standard deviation of the  
207 differences between predicted values and observed values. The NRMSE is the RMSE divided  
208 by the range of observed values of the variable being predicted. In literature, the NRMSE is  
209 mostly preferred than the NMAE because this indicator contain both information of the bias  
210 and the variance of the prediction.

211 As the Scaled Persistence Model does not require the historical data, the forecasting error of  
212 this model is usually the highest among the three. However, it is still lower than the forecasting  
213 error of Artificial Neural Network with only 1 month of training data. Generally, if the duration  
214 of the collected data is shorter than 1 month, it is advisable to use the Scaled Persistence Model.  
215 From more than one month of training data, the Artificial Neural Network and the Multivariate  
216 Polynomial Model are more accurate. The forecasting error reduce rapidly from more than 13%  
217 to less than 12% with the training data of 3 months (Multivariate Polynomial Model) or of 6  
218 months (Artificial Neural Network). With the Multivariate Polynomial Model, the difference of  
219 forecast error between the training duration of 3 months, 6 months, 9 months and 1 year in  
220 Guadeloupe is negligible. The training period can be considered as sufficient after 3 months of  
221 data collection.

222 With the training data of 365 days, the forecast error is slightly higher than that given by the  
223 training data of 270 days, which is not coincident with the tendency of reduction of the forecast  
224 error with longer training data. This contradiction could be explained by a hypothesis that the  
225 training data of the period of 3 months which has just been added to the training data may  
226 disturb the forecast model. To verify this hypothesis, another forecast is implemented using the

227 training data from the first 9 months of the training year (see Figure 1). The result shows that  
 228 the forecast error in both models is higher than that using the training data from the last 9 months  
 229 (12.21% with Artificial Neural Network and 12.01% with Multivariate Polynomial Model).  
 230 Therefore, the hypothesis can be confirmed.

#### 231 4.2 Case of Lille

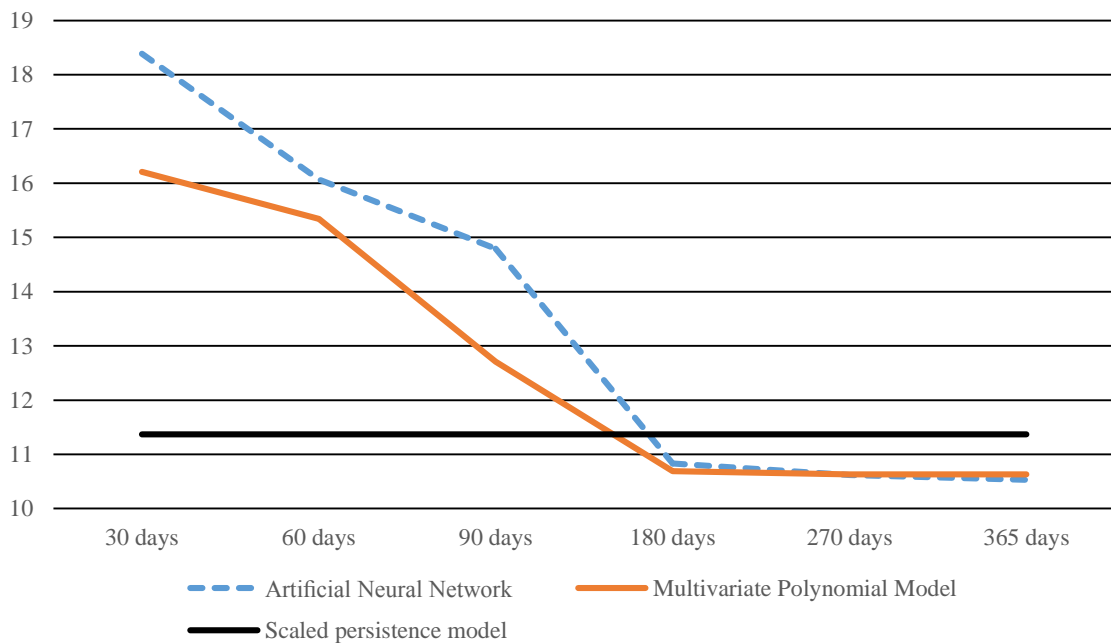
232 The Table 2 and the Figure 6 show the evolution of the forecast error with different training  
 233 duration in Lille.

234 Apart from the Scaled Persistence Model, which does not require the historical data, the forecast  
 235 error of the Artificial Neural Network and the Multivariate Polynomial Model is high with the  
 236 training data from 1 to 3 months due to the seasonal effect of Lille. With this effect, the right  
 237 parameters of the forecast model cannot be evaluate from the training data of 1 to 3 months.  
 238 However, from 6 months and above, the forecast error is reduced significantly (from 12.71%  
 239 to 10.69% with Multivariate Polynomial Model). The reduction starts from 6 months of training  
 240 data, not from 1 year, due to the symmetry of the PV production of Lille in a year (see Figure  
 241 2). The difference of forecast error between the training duration of 6 months, 9 months and 1  
 242 year in Lille is negligible. The training period could be finished after 6 months of data  
 243 collection.

244 Table 2. The evolution of the forecast error with different training duration (Lille)

NMBE			
Duration of training data	Artificial Neural Network	Multivariate Polynomial Model	Scaled Persistence Model
30days x 10 samples	-5.62%	-7.65%	-0.09%
60days x 10 samples	-9.35%	-6.78%	
90days x 10 samples	-4.45%	-4.08%	
180days x 10 samples	-0.24%	-0.23%	

270days x 10 samples	-0.73%	-0.81%	
365days x 10 samples	-0.76%	-0.81%	
NMAE			
30days x 10 samples	13.02%	11.43%	7.44%
60days x 10 samples	12.17%	10.90%	
90days x 10 samples	10.54%	9.36%	
180days x 10 samples	7.63%	7.62%	
270days x 10 samples	7.65%	7.71%	
365days x 10 samples	7.62%	7.75%	
NRMSE			
30days x 10 samples	18.39%	16.21%	11.37%
60days x 10 samples	16.07%	15.34%	
90days x 10 samples	14.79%	12.71%	
180days x 10 samples	10.83%	10.69%	
270days x 10 samples	10.62%	10.63%	
365days x 10 samples	10.53%	10.63%	



245

246

Figure 6. The evolution of the NRMSE (%) with different training duration (Lille)



247 We can observe that the duration of data collection required to have a good forecast in Lille is  
248 higher than the duration of data collection needed in Guadeloupe (6 months compared to 3  
249 months). However, the forecast error in Lille is smaller (10.69% compared to 11.97%).

#### 250 4.3 Analysis and recommendations

251 Among the 3 models, the Scaled Persistence Model does not require the historical data,  
252 therefore the forecast can be obtained directly without training period. However, the quality of  
253 forecast is not high. In the beginning, this model can compete with the others 2 forecast models  
254 but from 2 months of training data (Guadeloupe) and from 6 months of training data (Lille), the  
255 quality of the forecast by these model is much better than that of the Scaled Persistence Model.

256 The Artificial Neural Network and the Multivariate Polynomial Model have the same evolution  
257 of forecast error due to the fact that both of these models use the Levenberg-Marquardt  
258 algorithm as core learning technique. The Multivariate Polynomial Model has a better forecast  
259 quality than the Artificial Neural Network. In Guadeloupe, the Artificial Neural Network needs  
260 6 months of training data to acquire the forecast error lower than 12% while the Multivariate  
261 Polynomial Model needs only 3 months to go under this level. Another advantage of the  
262 Multivariate Polynomial Model is the simplicity and the transparency of the model. This is an  
263 analytical model, it means that the relation between the predicted value and the input variables  
264 can be represented in the form of an equation, different from the Artificial Neural Network,  
265 where this relation can only be represented by a black box.

266 As presented in the Figure 1 and the Figure 2, the PV production of Guadeloupe and Lille has  
267 very different characteristics. In Lille, there is a seasonal effect that there is not in Guadeloupe.  
268 However the PV production of Lille is less fluctuant than that of Guadeloupe. Therefore, to  
269 obtain a good quality forecast, the model of Lille needs a longer duration of training data (6

270 months instead of 3 months). But once the model is trained, it provides a better performance  
271 than that of Guadeloupe.

## 272 5. Conclusions

273 This paper focus on the necessary duration of data collection to have a quality forecast of PV  
274 production for two climatic conditions: a temperate and a tropical weather conditions. In order  
275 to investigate this issue, the study is implemented on 2 PV installations: the first one in  
276 Guadeloupe, the second in Lille; using 3 different forecast models: the Scaled Persistence  
277 Model, the Artificial Neural Network and the Multivariate Polynomial Model.

278 The results show that with the temperate climate such as Lille, where the seasonal effect existed,  
279 a training duration of at least 6 months is needed to acquire a high accuracy forecast for PV  
280 generation instead of 3 months with the tropical climate at Guadeloupe. However, once the  
281 model is trained, the performance is better (error of 10.69% at Lille compared to 11.97% at  
282 Guadeloupe).

283 This research proposes to use the Scaled Persistence Model to forecast the PV output power  
284 during the data collection as this model does not require the historical data. Once the data  
285 collection phase is finished, the Multivariate Polynomial Model could be applied in order to  
286 provide a better accuracy. With the application of PV output forecast, the exploitation of PV  
287 becomes more efficient, reducing the risk of power outages and improving the reliability of  
288 power supply.

## 289 References

290 Berry, M.J.A. & Linoff, G., 1997. *Data Mining Techniques*. NY: John Wiley & Sons.

291 Beyer HG et al., 2009. *Report on Benchmarking of Radiation Products*. Report under contract no.  
292 038665 of MESoR.

293 Blum, A., 1992. *Neural Networks in C++*. NY: Wiley.

294 Boger, Z.a.G.H., 1997. Knowledge extraction from artificial neural network models. In *IEEE Systems,*  
295 *Man, and Cybernetics Conference*. Orlando, FL, 1997.

296 CRE, 2009. *Cahier des charges de l'appel d'offresportant sur des installations au sol de production*  
297 *d'électricitéà partir de l'énergie solaire*. Ministère de l'Ecologie, de l'Energie, du Développement  
298 durable et de l'Aménagement du territoire.

299 Cui, Y.-J., Davis, S., Cheng, C.-K. & Bai, X., 2004. A study of sample size with neural network. In  
300 *Proceedings of 2004 International Conference on Machine Learning and Cybernetics.*, 2004.

301 Dazhi, Y., 2012. *Solar modeling and forecast*. A report submitted to the department of electrical and  
302 computer engineering and the examination committee of national university of Singapore.

303 Do, M.T., Sprooten, J., Clenet, S. & Robyns, B., 2010. Influence of wind turbines on power system  
304 reliability through probabilistic studies. In *Innovative Smart Grid Technologies Conference Europe*  
305 *(ISGT Europe).*, 2010.

306 Do, M.T., Sprooten, J., Clenet, S. & Robyns, B., 2011. Reliability evaluation of power system with  
307 large-scale wind farm integration using first-order reliability method. In *Proceedings of the 2011-14th*  
308 *European Conference on Power Electronics and Applications.*, 2011.

309 Ernst, B., Reyer, F. & Vanzetta, J., 2009. Wind power and photovoltaic prediction tools for balancing  
310 and grid operation. In *CIGRE/IEEE PES Joint Symposium Integration of Wide-Scale Renewable*  
311 *Resources Into the Power Delivery System.*, 2009.

312 Fernandez-Jimenez, L.A. et al., 2012. Short-term power forecasting system for photovoltaic plants.  
313 *Renew. Energy*, pp.311–317.

314 Fine, T.L. & Turmon, M.J., 1994. Sample Size Requirements of Feedforward Neural Network. In  
315 *Advances in Neural Information Processing Systems.*, 1994.

316 IEA, 2008. *Performance prediction of grid connected photovoltaic systems using remote sensing*.

317 IEA, 2013. *Photovoltaic and Solar Forecasting: State of the Art Report*.

318 Kasten, F., 1980. A simple parameterization of two pyrheliometric formulae for determining the Linke  
319 turbidity factor. *Meteorol. Rundsch.*, 33, pp.124–127.

320 Krömer, P. et al., 7–11 July 2012. Evolutionary Prediction of Photovoltaic Power Plant Energy  
321 Production. In *In Proceedings of International Conference on Genetic and Evolutionary Computation*.  
322 Philadelphia, PA, USA, 7–11 July 2012.

323 Lorenz E et al., 2008. Qualified Forecast of Ensemble Power Production by Spatially Dispersed Grid-  
324 Connected PV Systems. In *Proceedings of the 23rd European Photovoltaic Solar Energy Conference  
325 and Exhibition.*, 2008.

326 Lorenz, E., Heinemann, D. & Kurz, C., 2012. Local and regional photovoltaic power prediction for large  
327 scale grid integration: Assessment of a new algorithm for snow detection. *Prog. Photovolt. Res. Appl.*,  
328 pp.760–769.

329 Mandal, P. et al., 2012. Forecasting power output of solar photovoltaic system using wavelet transform  
330 and artificial intelligence techniques. *Procedia Comput. Sci.*, pp.332–337.

331 Pedro, H.T.C. & Coimbra, C.F.M., 2012. Assessment of forecasting techniques for solar power  
332 production with no exogenous inputs. *Solar Energy*, 86, pp.2017–2028.

333 Seber, G.A.F. & Wild., a.C.J., 2003. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience.

334 Shi, J. et al., 2012. Forecasting power output of photovoltaic systems based on weather classification  
335 and support vector machines. *IEEE Trans. Ind. Appl.*, pp.1064–1069.

336 Swingler, K., 1996. *Applying Neural Networks: A Practical Guide*. London: Academic Press.

337 Thirumalainambi, R.a.B.J., 2003. Training data requirement for a neural network to predict aerodynamic  
338 coefficients. In *Proc. SPIE 5102, Independent Component Analyses, Wavelets, and Neural Networks*.  
339 Orlando, FL, 2003.

340 Traunmüller W & Steinmaurer G., 2010. Solar irradiance forecasting, benchmarking of different  
341 techniques and applications of energy meteorology. In *Proceedings of the EuroSun 2010 conference*.  
342 Graz, Austria, 2010.

343 Yona, A. et al., 5–8 November 2007. Application of Neural Network to One-day-ahead 24 hours  
344 Generating Power Forecasting for Photovoltaic System. In *In Proceedings of the International*  
345 *Conference on Intelligent Systems Applications to Power Systems*. Kaohsiung, Taiwan, 5–8 November  
346 2007.