



**HAL**  
open science

## Verification of solar irradiance probabilistic forecasts

Philippe Lauret, Mathieu David, Pierre Pinson

► **To cite this version:**

Philippe Lauret, Mathieu David, Pierre Pinson. Verification of solar irradiance probabilistic forecasts. Solar Energy, 2019, 194, pp.254-271. 10.1016/j.solener.2019.10.041 . hal-02351342

**HAL Id: hal-02351342**

**<https://hal.science/hal-02351342>**

Submitted on 6 Nov 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.

# Verification of solar irradiance probabilistic forecasts

Philippe Lauret<sup>a,\*</sup>, Mathieu David<sup>a</sup>, Pierre Pinson<sup>b</sup>

<sup>a</sup>*University of La Réunion - PIMENT laboratory, 15, avenue René Cassin, 97715 Saint-Denis*

<sup>b</sup>*Technical University of Denmark, Centre for Electric Power and Energy, 2800 Kgs. Lyngby, Denmark*

---

## Abstract

We propose a framework for evaluating the quality of solar irradiance probabilistic forecasts. The verification framework is based on visual diagnostic tools and a set of scoring rules mostly originating from the weather forecast verification community. Two types of probabilistic forecasts are used as a basis to illustrate the application of these verification approaches. The first one consists in ensemble forecasts commonly provided by national or international meteorological centres. The second one originates from statistical methods and produces a set of discrete quantile forecasts, the nominal proportions of which span the unit interval. These probabilistic forecasts are evaluated for two selected sites that experience very different climatic conditions. The first site is located in the continental US while the second one is situated on La Réunion Island. Although visual diagnostic tools can help identify deficiencies in generated forecasts, it is recommended that a set of numerical scores be used to assess the quality of probabilistic forecasts. In particular, the Continuous Ranked Probability Score (CRPS) seems to have all the features needed to evaluate a probabilistic forecasting system and, as such, may become a standard for verifying solar irradiance probabilistic forecasts and by extension probabilistic forecasts of solar power generation.

*Keywords:* probabilistic solar forecasting, evaluation framework, diagnostic tools, scoring rules, CRPS, Ignorance Score

---

## 1. Introduction

Forecasts of solar energy generation are of utmost importance for efficiently integrating solar power generation into existing power grids and to decrease associated costs. Indeed, power production from photovoltaic (PV) or solar thermal plants is highly variable since weather dependent. Therefore, accurate knowledge of the future production from solar power generation capacities is necessary to limit the needs for additional balancing services and potentially storage. Therefore, increasing the value of solar power generation through the improvement of solar irradiance or PV power forecasting models (both usually referred to

---

\*corresponding author

*Email addresses:* philippe.lauret@univ-reunion.fr (Philippe Lauret),  
mathieu.david@univ-reunion.fr (Mathieu David), ppin@elektro.dtu.dk (Pierre Pinson)

9 as “solar forecasting models”) is of paramount importance. In the realm of solar irradiance  
10 forecasting, Global Horizontal Irradiance (GHI) is a prominent key variable. Therefore, this  
11 work will use this variable to illustrate the application of the proposed evaluation framework.

12 Numerous works have been devoted to the development of models that generate point  
13 forecasts of solar power generation, commonly referred to as deterministic forecasts. Some  
14 of these models can be found in (Reikard, 2009; Dambreville et al., 2014; Marquez and  
15 Coimbra, 2011; Coimbra et al., 2013; Huang et al., 2013; Lauret et al., 2015; Voyant et al.,  
16 2017; Pedro and Coimbra, 2015; Lorenz and Heinemann, 2012). Furthermore, error metrics  
17 dedicated to evaluating the accuracy of these deterministic forecasts, like Mean Bias Error  
18 (MBE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) together with  
19 skill-score measures (Hoff et al., 2013; Coimbra et al., 2013), are now quite standard and  
20 well accepted by the solar forecasting community.

21 However, a forecast is inherently uncertain and in a context of decision-making faced by  
22 the grid operator, a point forecast plus an uncertainty (or, better say, prediction) interval is  
23 of genuine added value. Put differently, reliable probabilistic predictions may contribute to a  
24 more efficient integration of intermittent sources in the energy network (Morales et al., 2014).  
25 Contrary to the wind power forecasting community where probabilistic forecasting appears  
26 to be a mature subject (Morales et al., 2014; Iversen et al., 2016; Jung and Broadwater,  
27 2014; Pinson et al., 2007), probabilistic solar forecasting is still in its infancy (Hong et al.,  
28 2016) albeit some recent works (Zamo et al., 2014; Sperati et al., 2016; Alessandrini et al.,  
29 2015; Grantham et al., 2016; Ben Bouallègue, 2015; David et al., 2016; Golestaneh et al.,  
30 2016b) tend to moderate this statement.

31 As mentioned by Pinson et al. (2007), the assessment of probabilistic forecasts is more  
32 complicated than for deterministic ones. Figures 1 and 2 show examples of GHI probabilistic  
33 forecasts. From the visual inspection of Figures 1 and 2, it is quite difficult to state  
34 whether the prediction intervals are good or not. To objectively assess the performance of  
35 probabilistic forecasts and the methods used to generate those, it is necessary to employ  
36 appropriate diagnostic tools and quantitative scores.

37 According to Murphy (1993), goodness of weather forecasts can be characterized by three  
38 types namely consistency, quality and value. Consistency is related to the correspondence  
39 between forecasters’ judgment and their forecasts. Quality refers to the correspondence  
40 between forecasts and the observations and value is linked to the benefit (economical or  
41 others) gained from the use of these probabilistic forecasts in an operational context. In this  
42 work, we concentrate on the assessment of the quality of the models.

43 Several attributes characterize the quality of probabilistic forecasts (Wilks, 2014; Jolliffe  
44 and Stephenson, 2003) but two main properties, i.e. reliability and resolution are used  
45 to measure the quality of the forecasts (Jolliffe and Stephenson, 2003). A third attribute  
46 namely sharpness can be used to evaluate how informative the forecasts are. In the weather  
47 forecasting verification community, several diagnostic tools are used to characterize these  
48 required properties of reliability, resolution and sharpness. One can cite among others the  
49 reliability diagram (Pinson et al., 2010; Wilks, 2014) and rank histogram (Hamill, 2001;  
50 Wilks, 2014) for assessing reliability. Regarding forecasts of continuous variable, there is  
51 currently no visual tool to assess resolution. The sharpness property can be evaluated

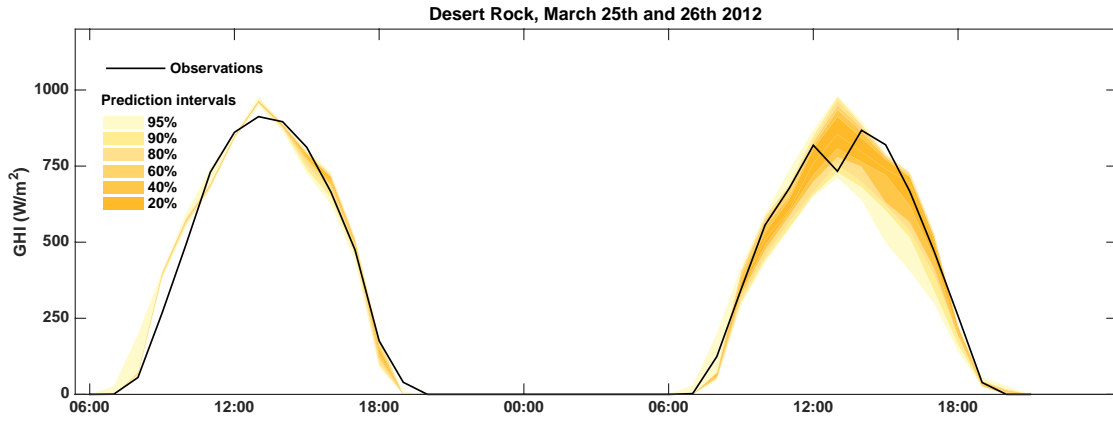


Figure 1: Example of probabilistic solar irradiance forecasts: 2 days of measured GHI at the Desert Rock (NV) and associated day-ahead forecasts with prediction intervals provided by ECMWF-EPS (see section 3).

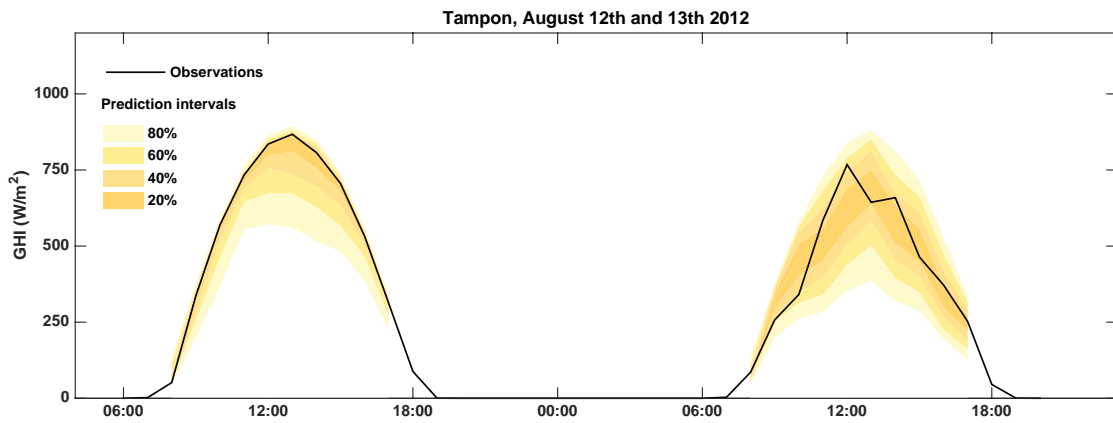


Figure 2: Example of probabilistic solar irradiance forecasts: 2 days of measured GHI at Tampon and associated 1-hour ahead forecasts with prediction intervals generated with the Quantile Random Forest model QRF2 (see section 3).

52 through the use of sharpness diagrams (Pinson et al., 2007; Gneiting et al., 2007).

53 In addition to these tools that permit to visually assess the attributes of a forecasting  
54 system, a metric called continuous ranked probability score (CRPS) (Hersbach, 2000) is  
55 commonly used by the weather forecasting community to objectively quantify the overall skill  
56 of the probabilistic forecasts. The CRPS is a metric capable of addressing both reliability  
57 and resolution simultaneously. Indeed, the CRPS can be decomposed into three components  
58 namely reliability, resolution and uncertainty. This decomposition provides a detailed picture  
59 of the performance of the forecasting methods (Hersbach, 2000) and consequently may help  
60 in the ranking of the probabilistic forecasts. A scoring rule originated from the information  
61 theory called the logarithm or ignorance score metric has also been proposed for assessing  
62 the quality of weather probabilistic forecasts (Roulston and Smith, 2002; Pinson et al., 2012).

63 Although solar probabilistic forecasting is not as mature as wind probabilistic forecasting  
64 (Hong et al., 2016), some recent works (Alessandrini et al., 2015; Sperati et al., 2016; Zamo  
65 et al., 2014; Grantham et al., 2016; David et al., 2016, 2018; Chu and Coimbra, 2017;  
66 Golestaneh et al., 2016b; Verbois et al., 2018) proposed to assess the quality of the models  
67 with some classical diagnostic tools originated from the weather verification community  
68 like rank histogram and reliability diagram. This literature review also revealed that the  
69 CRPS is a commonly used scoring rule. However, in our opinion, most of these works  
70 did not conduct a detailed analysis of how to use and interpret the verification tools. For  
71 instance, the CRPS formula proposed by (Hersbach, 2000) is restricted to ensemble forecasts  
72 but David et al. (2018) and Lauret et al. (2017) used it to compute the CRPS of discrete  
73 quantile forecasts. Moreover, most of the previous works that evaluated the overall skill  
74 of competing methods through the use of the CRPS did not attempt to have a detailed  
75 performance of the methods which is possible from the decomposition of the CRPS into  
76 reliability, resolution and uncertainty. Besides, to our best knowledge, the ignorance score  
77 is not currently used by the solar forecasting community.

78 In addition, other metrics are proposed to assess the properties of prediction intervals  
79 such as Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized  
80 Averaged Width (PINAW) (Khosravi et al., 2013; Chu and Coimbra, 2017; Lauret et al.,  
81 2017). PICP is related to the reliability of the probabilistic forecasts while PINAW gives  
82 a measure of the sharpness of the predictive distributions. However, as discussed below,  
83 these two metrics (PICP and PINAW) are not the most appropriate for measuring the  
84 quality of interval forecasts. It is also worth noting that a metric called coverage width-  
85 based criterion (CWC), which assesses the quality of the prediction intervals by combining  
86 PICP and PINAW has been proposed by (Khosravi et al., 2013). But as demonstrated by  
87 (Pinson and Tastu, 2014), this score can lead to possible misinterpretations of the results.  
88 Unfortunately, some researchers in the solar community (Scolari et al., 2016; Chu et al.,  
89 2015; Li et al., 2018) recently used this metric to assess the quality of their forecasting  
90 models. Furthermore, the CWC score has been recently cited in a reference paper (Yang  
91 et al., 2018) and a review paper (van der Meer et al., 2018).

92 This is why, we think that now is the time to take stock on the evaluation metrics of  
93 solar probabilistic forecasts. The objective of this work is therefore to provide the forecasting  
94 solar community a comprehensive overview of diagnostic tools and scoring rules that can

95 be used to assess the performance of probabilistic forecasting methods. In particular, we  
96 propose an evaluation framework that may help the user to consistently evaluate the quality  
97 of the models. In others words, this paper aims at explaining how one should assess the  
98 quality of the probabilistic forecasts and how diagnostic tools and scores should be used and  
99 interpreted. In addition, we will propose a measure of resolution (through the decomposition  
100 of the CRPS) as this attribute is not currently assessed in the literature.

101 In this paper, two types of GHI probabilistic forecasts are used to illustrate the pro-  
102 posed verification framework. The first one is the ensemble forecast commonly provided  
103 by Ensemble Prediction Systems (EPS) of the Numerical Weather Predictions (NWP) of  
104 meteorological utilities such as ECMWF. The second one, denoted by quantile forecasts,  
105 is based on statistical methods and produces a set of quantiles spanning the unit interval.  
106 Both types generate forecasts represented by predictive distributions that can be modelled  
107 either by a Cumulative distribution function (CDF) or a Probability distribution Function  
108 (PDF).

109 Finally, note that in this paper, we restrict ourselves to the univariate context that corre-  
110 sponds to probabilistic forecasts that do not take into account spatio-temporal dependencies  
111 that are generated by stochastic processes like for instance cloud passing. The interested  
112 reader is referred to (Golestaneh et al., 2016a) who proposed a method to capture the  
113 spatio-temporal correlations in PV forecasts.

114 The remainder of this paper is organized as follows. Section 2 defines the probabilistic  
115 forecast as the estimation of a predictive distribution of the variable of interest (GHI in  
116 our case). Section 3 presents the two sites that will serve as support for the application of  
117 the verification tools on quantile and ensemble forecasts while Section 4 lists the properties  
118 required for skillful probabilistic forecasts. Section 5 presents in details the verification tools  
119 and illustrates their application on quantile and ensemble forecasts. Finally, section 6 gives  
120 some concluding remarks.

## 121 2. Nature of probabilistic forecasts of continuous variables

122 Probabilistic forecasts correspond to the estimation of the statistical distribution of a  
123 future event. Thus, a probabilistic forecast may be defined as a cumulative distribution  
124 function (CDF)  $F$  of a random variable  $X$ , such that  $F(x) = Pr(X \leq x)$ . This CDF can be  
125 summarized by a set of quantiles. The quantile  $q_\tau$ , at probability level  $\tau \in [0, 1]$ , is defined  
126 as follow

$$q_\tau = F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}. \quad (1)$$

127 A quantile  $q_\tau$  informs there is a probability  $\tau$  that the event  $x$  materializes below that  
128 quantile  $q_\tau$ . From a set of quantiles, prediction intervals (PIs) can be deduced. PIs define the  
129 range of values within which the observation is expected to be with a certain probability i.e.  
130 its nominal coverage rate (Pinson et al., 2007). To completely determine a PI, it is necessary  
131 to choose the way it should be centered on the probability density function (Pinson et al.,  
132 2007). The most common way is to center the PI on the median. Consequently, there is  
133 the same probability of risk below and above the median. Therefore, a central PI with a

134 coverage rate of  $(1 - \alpha)100\%$  is estimated by using the  $\alpha/2$  quantile ( $\hat{q}_{\tau=\alpha/2}$ ) as the lower  
 135 bound and the  $(1 - \alpha/2)$  quantile ( $\hat{q}_{\tau=1-\alpha/2}$ ) as the upper bound. More precisely, a PI with  
 136  $(1 - \alpha)100\%$  nominal coverage rate is given by

$$\widehat{PI}_{(1-\alpha)100\%} = [\hat{q}_{\tau=\alpha/2}, \hat{q}_{\tau=1-\alpha/2}]. \quad (2)$$

137 In the realm of weather predictions, three ways to define this cumulative distribution  
 138 are available: parametric CDFs, discrete estimates of a CDF via a non-parametric method  
 139 and ensemble forecasts. Parametric CDFs are easy to set up and to assess. Nevertheless,  
 140 regarding solar forecasts, they are seldom proposed in the literature because they suffer  
 141 from a lack of calibration. Indeed, the distribution of future observations of the solar power  
 142 can not be accurately reproduced by a single probabilistic law. David et al. (2016) gave an  
 143 example with the GARCH model that assumes a Gaussian distribution.

144 An alternative to the parametric approach is the generation of discrete estimates of a  
 145 CDF. This non-parametric method allows defining a predictive CDF without any assumption  
 146 on the distribution of the future event. The forecast is provided as a set of quantiles spanning  
 147 the unit interval. This kind of probabilistic forecast is also called quantile forecasts (Pinson  
 148 et al., 2007). The Global Energy Forecasting Competition 2014 (GEFCOM 2014) (Hong  
 149 et al., 2016) is a good example of this approach. Indeed, the solar forecasts were to be  
 150 expressed in the form of 99 quantiles with various nominal proportions between zero and  
 151 one. Widely used statistical models, like Quantile Regressions (QR) or Gradient Boosting  
 152 Decision Trees (GBDT) can estimate these predictive distributions.

153 The last type corresponds to ensemble forecasts classically generated by Numerical  
 154 Weather Predictions (NWP) models. The distribution of the future event is given by an  
 155 ensemble of members that are not directly linked to the notion of quantiles. For example, in  
 156 the case of a NWP model, an ensemble forecast corresponds to a perturbed set of forecasts  
 157 computed by slightly changing the initial conditions of the control run and of the modeling  
 158 of unresolved phenomena (Leutbecher and Palmer, 2008). This ensemble prediction system  
 159 (EPS) allows representing the uncertainties of the prediction scheme. Nevertheless, ensemble  
 160 forecasts can be seen as discrete estimates of a CDF when they are sorted in ascending  
 161 order. In the literature, different ways to associate these sorted members to cumulative  
 162 probabilities are proposed. Considering  $M$  sorted members of an ensemble  $E = (e_1, \dots, e_M)$ ,  
 163 the most common definition in the domain of weather forecast assessment states that there  
 164 is a probability of  $1/M$  that the observation falls between two consecutive members  $e_j$  and  
 165  $e_{j+1}$  (Anderson, 1996; Hersbach, 2000). If we assign a null probability for future events that  
 166 fall outside the ensemble (i.e.  $x_{obs} < e_1$  or  $x_{obs} > e_M$ ), the predictive distribution can be  
 167 seen as a piecewise constant function

$$\widehat{F}(x) = \sum_{k=1}^M \alpha_k H(x - e_k). \quad (3)$$

168  $H$  is the Heaviside function which is 1 if the argument is positive and zero otherwise.  
 169 The weight  $\alpha_k = 1/M$  corresponds to the jump of probability that happens when  $x = e_k$ .

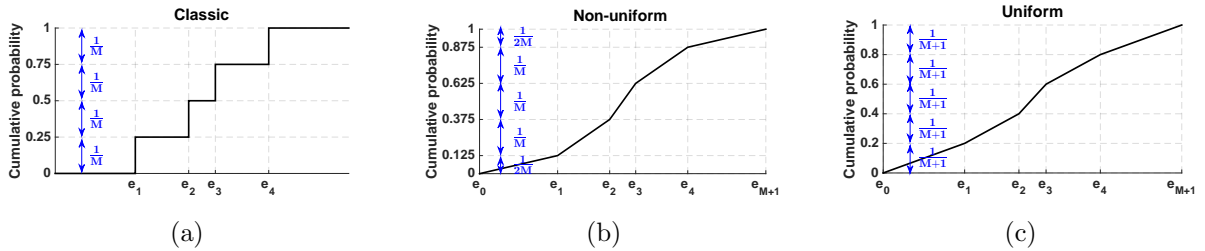


Figure 3: Different definitions of the CDF derived from an ensemble forecast ( $M = 4$ ): (a) classical; (b) non-uniform spacing of the cumulative probabilities and a linear interpolation between the members; (c) uniform spacing and a linear interpolation between the members.

170 Figure 3(a) gives a visual representation of this classical definition of a CDF derived from  
 171 an ensemble with 4 members ( $M = 4$ ).

172 In the case of continuous variable, as the solar irradiance (GHI), the shape of the CDF  
 173 resulting from the preceding definition is obviously not realistic. Several works (Bröcker,  
 174 2012; Roulston and Smith, 2002; Pinson et al., 2010) proposed alternative approaches to  
 175 face this issue. Among others, these alternatives allow defining a continuous predictive  
 176 distribution and non-null probabilities outside the ensemble. We briefly present two other  
 177 ways to build a CDF from an ensemble forecast.

178 First, Bröcker (2012) proposes to preserve a jump of  $1/M$  between two members but to  
 179 assign a probability mass of  $1/2M$  for the events that fall outside of the ensemble. It results  
 180 in a non-uniform partition of the probability space  $[0; 1]$ . Figure 3(b) gives an example of  
 181 this definition for an ensemble with 4 members ( $M = 4$ ) and a linear interpolation between  
 182 the members. The tails of the distributions are bounded by  $e_0$  and  $e_{M+1}$ . The choice of  
 183 these limits are arbitrary. For continuous variables, Roulston and Smith (2002) proposed  
 184 to use the minimum and the maximum of the climatology. Notice that this non-uniform  
 185 definition amounts to consider each ensemble member  $i$  as a quantile with probability level  
 186  $\tau(i) = \frac{i-0.5}{M}$ .

178 The second approach, described by (Pinson et al., 2010; Bröcker, 2012), assigns a prob-  
 179 ability mass of  $1/(M + 1)$  between two members and for the events that fall outside of the  
 180 ensemble. Note that using this definition that an ensemble member can be interpreted as  
 181 a quantile forecast by considering its rank within the ensemble. The probability level  $\tau(i)$   
 182 associated with the member of rank  $i$  is defined as:  $\tau(i) = \frac{i}{M+1}$ . This approach leads to  
 183 an uniform spacing of the cumulative probabilities. Figure 3(c) presents graphically the  
 184 shape of the CDF when considering this last definition and a linear interpolation between  
 185 the members. As for the non-uniform definition, the boundaries of the CDF,  $e_0$  and  $e_{M+1}$ ,  
 186 are arbitrarily chosen (see appendix A for more details).

178 Thus, when dealing with ensemble forecasts, three ways to build the CDF from the mem-  
 179 bers are available. Unfortunately no definition can be favoured and each CDF construction  
 180 has its pros and cons. The classic definition is the most used, specifically to compute the  
 181 Continuous Rank Probability Score (CRPS, see section 5.3.1) with the methodology pro-  
 182 posed by (Hersbach, 2000). As this commonly used definition assigns null probabilities to  
 183



201 the events that fall outside of the ensemble, it can not be used to derive scores like ignorance  
202 (see section 5.3.4). The uniform and the non-uniform definitions requires to arbitrarily fix  
203 the boundaries of the CDF. Therefore, they are user dependent. Nevertheless, they allow  
204 designing continuous CDF that contains all the possible events. Thus, the procedure used to  
205 verify the quality of ensemble forecasts can be exactly the same as for the parametric CDFs  
206 or for the predictive distributions summarized by discrete quantiles estimated by some kind  
207 of statistical method. Bröcker (2012) showed that the non-uniform definition corresponds  
208 to a minimization of the CRPS. But, considering this definition, the optimal shape of the  
209 corresponding rank histogram (see section 5.2.2) is not flat. Indeed for this visual verifica-  
210 tion tool, the height of the first and last ranks should be the half of the other ones. Finally,  
211 if the aim is to compare different forecasting models, whatever the chosen definition, the  
212 ranking will remain the same. Nevertheless, a unique framework has to be defined to allow  
213 the comparison of different works.

### 214 3. Illustrative case studies

215 Two sites will serve as benchmarks for the application of the different tools and scores  
216 described below. The first site, Desert Rock (USA), has an arid climate with a very sunny  
217 and stable sky. The second site, Tampon (Réunion island), is located in a tropical island  
218 and experiences a very variable sky. The experimental dataset corresponds to two consec-  
219 utive years of recorded data of global horizontal irradiance (GHI). Table 1 gives detailed  
220 information about the data. The solar variability, quantified by the standard deviation of  
221 the changes in the clear sky index  $\sigma\Delta kt^*$  (Hoff and Perez, 2012), is the main difference  
222 between the two considered locations. We intentionally chose these two sites. Indeed, the  
223 solar variability is a key factor in the accuracy of deterministic forecasts. The higher the  
224 variability, the less accurate the forecasts are (Lauret et al., 2015). Finally, to build some  
225 of the models used in this work, we used the first year of data (2012) as training set and  
226 the second year of data (2013) as testing set. Therefore, all the metrics and visual tools  
227 presented hereafter are derived from the testing set.

228 Two forecasting time horizons will be addressed in this work. First, intra-day forecasts  
229 with lead times ranging from 1 to 6 hours will be appraised. These forecast are provided  
230 by state of the art forecasting models that generate predictive distributions from a set  
231 of quantiles spanning the unit interval. Second, day-ahead probabilistic forecasts will be  
232 studied. Generated by Numerical Weather Predictions (NWP) models, they are provided  
233 as ensemble forecasts.

#### 234 3.1. Intraday quantile forecasts

235 Regarding intraday quantile forecasts, the quality of four state-of-the-art probabilistic  
236 models will be appraised. In this paper, we will not give the details of the implementation  
237 of these models as they have already been described in previous works (David et al., 2018;  
238 Pedro et al., 2018). In addition, we recall that the goal here is to illustrate the application  
239 of the proposed evaluation framework and not to have a detailed evaluation of these models.

Table 1: Main characteristic of the solar measurements

	<b>Desert Rock</b> (USA)	<b>Tampon</b> (Réunion)
Provider	SURFRAD	PIMENT
Position	36.6N, 116.0W	21.3S, 55.5E
Elevation	1007m	550m
Cimate type	Arid	Insular tropic
Period of record	2012-2013	2012-2013
Annual solar irradiation	2.105 MWh/m <sup>2</sup>	1.712 MWh/m <sup>2</sup>
Solar variability 1-h ( $\sigma\Delta kt^*$ )	0.146	0.241
Mean GHI (Testing set)	548 W/m <sup>2</sup>	458 W/m <sup>2</sup>
Uncertainty component of the CRPS	29.1%	33.1%

240 The selected models are based on two quantile regression techniques namely the quantile  
241 regression forest (QRF) and the Gradient Boosting (GB) techniques. Briefly, the proposed  
242 techniques estimate directly the set of quantiles from a regression model  $Y = f(X)$  that  
243 relates the response variable  $Y$  (here GHI for lead time  $h = 1, 2, \dots, 6$  hours) to a set of  
244 predictor variables ( $X$ ). Two variants of regression models with different sets of predictor  
245 variables are built. For the first variant described in (Lauret et al., 2017), the vector of  
246 explanatory variables  $X$  consists of the actual measurement plus five past ground measure-  
247 ments while the second one takes as additional inputs two geometrical solar features related  
248 to the course of the sun in the sky namely the cosine of the zenith angle ( $\cos(SZA)$ ) and  
249 the cosine of the hour angle ( $\cos(HA)$ ). The adding of the two variables originates from  
250 the following reasons. First, some authors (Grantham et al., 2016; Lorenz and Heinemann,  
251 2012) showed a clear dependency of the forecasting error in relation to  $SZA$ . Second, we  
252 expect that the hour angle will bring some information regarding the asymmetry of the sky  
253 conditions between mornings and afternoons. This may be hold particularly for site like Le  
254 Tampon that experiences such a dichotomy between mornings and afternoons. Table 2 lists  
255 the acronyms of the resulting four quantile regression models.

Table 2: Acronyms related to the four quantile regression models

<b>Quantile regression techniques</b>	<b>Variant 1</b>	<b>Variant 2</b>
Quantile Regression Forest	QRF1	QRF2
Gradient Boosting	GB1	GB2

### 256 3.2. Day-ahead ensemble forecasts

257 The day-ahead ensemble predictions are provided by the Integrated Forecasting System  
258 (IFS) of the European Centre of Medium-Range Weather Forecasts (ECMWF). We will  
259 denote these ensemble forecasts as “ECMWF-EPS”. They consist in 50 perturbed members.  
260 The temporal resolution is of 3 hours and the spatial resolution is of  $0.2^\circ$  in both longitude

261 and latitude. Consequently, 3h GHI (in  $Wh/m^2$ ) times series recorded on-site are compared  
262 with the nearest ECWMF pixel. In addition, we also propose a post-processed version of  
263 the original ECMWF-EPS forecasts. Indeed, the ensemble prediction systems of the NWP  
264 models commonly suffer from a lack of spread (Leutbecher and Palmer, 2008). To face  
265 this issue, Sperati et al. (2016) proposed a simple approach, named Variance Deficit (VD),  
266 to calibrate the ensemble forecasts. Their method spreads the initial ensemble forecasts  
267 by correcting their variance. The correction factor is evaluated from a training set. The  
268 calibrated ensemble forecasts will be denoted by “ECMWF-EPS + VD”.

## 269 4. Required properties for a skillful probabilistic system

270 As mentioned in the introduction, two main attributes (reliability and resolution) char-  
271 acterize the quality of probabilistic forecasts (Pinson et al., 2007). The evaluation of these  
272 two attributes can be complemented by a sharpness assessment.

### 273 4.1. Reliability

274 Reliability or calibration refers to the statistical consistency between the forecasts and  
275 the observations. In other terms, the nominal coverage rate of the prediction intervals should  
276 be equal to the empirical one (e.g. a 90% PI should cover 90% of the observations). The  
277 reliability property is an important prerequisite as non reliable forecasts would lead to a  
278 systematic bias in subsequent decision-making processes (Pinson et al., 2007).

### 279 4.2. Resolution and sharpness

280 Resolution measures the capacity of a forecasting model to issue forecasts that are case-  
281 dependent. This important property, which is not easy to catch, is commonly not considered  
282 by the solar forecasting community. To understand concretely what resolution is, we will  
283 first define the climatological forecast (i.e. climatology). Imagine a distribution built from  
284 all the available past data of the parameter to forecast. The climatological forecast uses  
285 this unique distribution to forecast any future events. A high resolution forecasting system  
286 generates forecasts that differ from the climatology and, as a consequence, forecasts that are  
287 significantly different from each other. Climatological forecasts are perfectly reliable though  
288 having no resolution. Consequently, a skillful probabilistic forecasting system should issue  
289 reliable forecasts and with high resolution.

290 Sharpness evaluates how informative the forecasts are. Practically, sharpness refers to  
291 the concentration of the predictive distributions (Pinson et al., 2007; Gneiting et al., 2007)  
292 and can be measured by the average width of the prediction intervals. Unlike the two  
293 previous attributes, sharpness is a function of the forecasts only and does not depend on  
294 the observations. Consequently, a forecasting system can produce sharp forecasts yet being  
295 useless if those probabilistic forecasts are not reliable.

296 Unlike resolution and reliability, the sharpness property can be intuitively assessed. As  
297 an example, the first day of Figure 1 well illustrates an extremely sharp forecasts with  
298 narrow prediction intervals. Conversely, the second day of Figure 2 shows a example of a  
299 low sharpness forecast with large predictions intervals.

300 It must be emphasized here that these two components (sharpness and resolution) have  
301 different interpretations according a meteorologist’s point of view or a statistician’s point  
302 of view. In the meteorological literature (Wilks, 2014; Jolliffe and Stephenson, 2003), the  
303 sharpness property refers to the ability of a forecasting system to generate forecasts that are  
304 able to deviate from the climatological value of the variable to predict (also called predictand)  
305 whereas from a statistical point of view the sharpness property relates to the concentration  
306 of the predictive distributions (Pinson et al., 2007; Gneiting et al., 2007).

307 Similarly, from a meteorological point of view, resolution measures the ability of a fore-  
308 casting system to produce predictive distributions conditioned by the value of the predictand  
309 (i.e. forecasts that are case-dependent) (Pinson et al., 2007). From a statistical point of  
310 view, resolution amounts to evaluate the capacity of the forecast system to produce different  
311 density forecasts depending on the forecast conditions (i.e. the predictive distributions are  
312 not only conditioned by the value of the predictand) (Pinson et al., 2007). For instance, the  
313 prediction intervals may exhibit increasing widths with increasing forecast horizon. Also,  
314 regarding the solar irradiance (GHI), the width of the PIs may vary according the sun’s  
315 position in the sky - see for the instance the work of (Grantham et al., 2016). In this work,  
316 we will not provide such a conditional assessment. Instead, we will propose a measure of  
317 resolution through the decomposition of the CRPS. From a meteorological perspective, it is  
318 also worth noting that, for perfectly reliable forecasts, sharpness is identical to resolution.  
319 In this work, we will clearly distinguish the definition of sharpness and resolution. That is to  
320 say, sharpness will refer to the concentration of the prediction intervals while resolution will  
321 quantify the ability of the forecasting system to generate conditional predictive distributions.  
322 Finally, it must be noted that reliability can be improved by means of statistical techniques  
323 also called calibration techniques (Gneiting et al., 2005), whereas this is not possible for  
324 resolution.

## 325 **5. Presentation and application of the verification tools**

### 326 *5.1. Proposed evaluation framework*

327 Diagnostic tools are used to visually assess the quality of probabilistic forecasts, while  
328 numerical scores are used to quantify the skills of a forecasting system and to rank competing  
329 prediction methods. Tables 3 and 4 summarize the diagnostic tools and scoring rules used to  
330 evaluate probabilistic forecasts generated either by ensemble methods or quantile techniques.  
331 Regarding pros and cons, and also the most common approaches already used in other fields  
332 (i.e. weather forecast verification and wind power forecasting), we propose to differentiate  
333 the methodologies and the tools to assess the quality of quantile forecasts and ensemble  
334 prediction systems (EPS).

335 Considering quantile forecasts, we advise to visually assess the quality of the forecasts  
336 using reliability diagrams with consistency bars. Then, to use the CRPS and its related  
337 decomposition as described in appendix C to quantify the overall performance of the methods  
338 and to measure the reliability and the resolution components.

339 For ensemble forecasts, we propose to use the rank histogram including consistency bars  
340 and the CRPS as defined by (Hersbach, 2000) (see appendix B) to respectively qualify and

Table 3: Visual diagnostic tools.

Diagnostic tool	Initially designed for	Pros	Cons	Remarks
Reliability Diagram (RD)	Reliability assessment of quantile forecasts	-Departure from perfect reliability (ideal diagonal line) easily visualized - Easy to build	Finiteness of the data and possible presence of serial correlation in sequence of observations/forecasts can cause deviations from the ideal line even for reliable forecasts. This issue can be mitigated by plotting RD with consistency bars	Can be used for Ensemble if members are assigned specific probability levels (uniform/non uniform CDF - see section 2)
Rank Histogram (RH)	Reliability assessment of Ensemble forecasts	- Easy to build - Statistical consistency of the ensemble quickly checked (flat RH) - Easy detection of deficiencies in ensemble calibration such as bias, under or over-dispersion	- As for RD, sensitivity to the finiteness of the data (Need to draw RH with consistency bars) - Caution: a flat RH does not imply a reliable forecast	Can be extended to quantile forecasts if quantiles are evenly spaced
PIT histogram (PIT)	Reliability assessment of quantile forecasts	- Departure from perfect reliability easily assessed - Calibration of predictive CDF easily checked (flat PIT histogram) - Like RH, easy detection of calibration deficiencies	- Subject to the finiteness of the data (plot with consistency bars advised) - Need to specify the number of histograms bins. - Require the computation of the predictive CDF - Interpolation needed between the discrete quantiles to estimate the value the CDF attains at the observation. - As for RH, a flat PIT is not a sufficient condition to state that a forecast is reliable	Can be used for Ensemble (uniform CDF)
Sharpness diagram	Ensemble and quantile forecasts	- Easy to build - Sharpness is an intuitive property that permits to assess the concentration of the predictive distributions.	- Sharpness diagrams must be interpreted with care because they are only relevant if the associated forecasts are reliable. - Sharpness can only contribute to a qualitative evaluation of the probabilistic forecasts.	- Even if narrow PIs are preferred, sharpness cannot be seen as a property to verify the quality of probabilistic forecasts but more like the consequence of a high resolution. - Can be used for Ensemble (uniform/non uniform CDF)

Table 4: Scoring Rules

Scores	Pros	Cons	Remarks
CRPS	CRPS has the same dimension as the variable to predict and can be normalized. Therefore, it permits comparisons between different datasets. For deterministic forecasts, CRPS turns to be the MAE (Mean Absolute Error). Thus, the performance of a probabilistic method can be compared against a deterministic one. Decomposition of the CRPS into reliability and resolution provides additional insight into the performance of a probabilistic model. As a non-local score, CRPS is a robust score.	No analytic formulae except for specific distributions (Gaussian, Student's t, ...) - See R package scoringRules for details. CRPS averages over the complete range of forecast thresholds. Consequently, deficiencies in different parts of the distributions (e.g. the tails of the distribution) can be hidden.	Specific formulae for Ensemble forecasts proposed by Herbasch (see Appendix A). Can be calculated through numerical integration (see Equation 6) but requires interpolation of uniform/non uniform CDF. Can be also computed through integration of the Brier Score (see Appendix C)
Ignorance Score	Easy to compute especially for Ensemble forecasts.	No detailed information regarding the performance of a forecasting system as IGN cannot be decomposed into reliability and resolution. No sites' comparisons can be carried out as IGN cannot be normalized. As a local score, and as such, sensitive to the form of the predictive PDF, IGN is less robust than the CRPS. IGN cannot be applied to predictive PDF with null probabilities.	Specific formula for Ensemble forecasts proposed by Roulston assuming a linear interpolation of the CDF between the members (see Equation 12). Otherwise, requires computation of the predictive PDF to estimate the value the PDF attains at the observation (requires interpolation of uniform/non uniform CDF).
Quantile Score	QS permits to obtain detailed information about the forecast quality of specific quantiles that are of great interest for the user. QS can be decomposed into reliability and resolution.	Score restricted to a specific quantile. Cannot be used to rank different forecast methods considering their overall performance.	QS can reveal deficiencies in different parts of the predictive distribution (e.g. tails of the distribution)
Interval Score	Very easy to compute. IS has the same dimension as the variable to predict and can be normalized.	Cannot be decomposed into reliability and resolution.	Designed specifically for interval forecasts

341 quantify the performances of the EPS. Indeed, these two tools does not require additional  
342 assumptions (i.e. to define the nature of the distribution and its boundaries) and they are  
343 already widely used.

344 For both type of forecasts, ignorance score (IGN), interval score (IS), quantile score (QS)  
345 and sharpness diagrams can complement the characterization of the forecasting methods.  
346 However, sharpness diagrams must be interpreted with care because they are only relevant  
347 if the associated forecasts are reliable.

348 Finally, if interval score, quantile score and sharpness diagrams are computed for ensem-  
349 ble forecasts, it is important to clearly indicate the assumption done to obtain the quantiles  
350 (e.g. uniform or non-uniform spacing).

351 In the following sections, we will present in detail the verification tools. Throughout the  
352 description, we will provide illustrations of the application of these tools to quantile and  
353 ensemble forecasts.

## 354 *5.2. Diagnostic tools*

### 355 *5.2.1. Reliability diagram*

356 The reliability diagram is a graphical verification display used to evaluate the reliability  
357 of the probabilistic forecasts. In this paper, we follow the methodology defined by (Pinson  
358 et al., 2010) that is especially designed for predictive distributions summarized by quantile  
359 forecasts. More precisely, quantile forecasts are reliable if their nominal proportions are equal  
360 to the proportions of the observed value. It means that, over an evaluation set of significant  
361 size, (statistically) the difference between observed and nominal probabilities should be as  
362 small as possible (Pinson et al., 2010). Notice that for ensemble forecasts, the uniform CDF  
363 or non uniform CDF (see section 2) must be chosen before applying this methodology.

364 This representation is attractive since the deviations from perfect reliability (i.e. the  
365 diagonal line) can be easily visualized (Pinson et al., 2010). Nonetheless, due to the finite  
366 sample of pairs of observation/forecast and also due to possibly serial correlation in the  
367 sequence of forecasts and observations, it is possible that observed proportions are not  
368 exactly along the diagonal, even if the forecasts are perfectly reliable. (Pinson et al., 2010).  
369 In other words, reliability diagrams can be misinterpreted since even for perfectly reliable  
370 forecasts, deviations from the ideal diagonal case can be observed.

371 To deal with the issue of limited number of pairs of observation/forecast, Bröcker and  
372 Smith (2007a) built reliability diagrams with consistency bars. In addition, Pinson et al.  
373 (2010) have proposed consistency bars taking into account the combined effect of serial cor-  
374 relation and limited data. Interpretation of reliability diagrams with consistency bars is  
375 that one cannot reject the hypothesis of the quantile forecasts being reliable if the observed  
376 proportions lie within the consistency bars. In practice, adding consistency bars to the rela-  
377 bility diagrams may reinforce the user's (possibly subjective) judgment about the reliability  
378 of the different models.

379 Finally, some preceding works (Chu and Coimbra, 2017; Lauret et al., 2017) proposed  
380 to evaluate the reliability component of a probabilistic system by calculating the prediction  
381 interval coverage probability (PICP) (Khosravi et al., 2013). PICP permits one to assess  
382 the empirical coverage probability of the central prediction intervals. However, this metric

383 is not suitable to assess the reliability of probabilistic forecasts because as noted by Pinson  
 384 et al. (2007), both quantiles that define the prediction interval may be biased. In other  
 385 words, PICP it is not sufficient to check if the nominal coverage of the intervals is respected.  
 It is also necessary to verify that both quantiles defining the PI are unbiased.

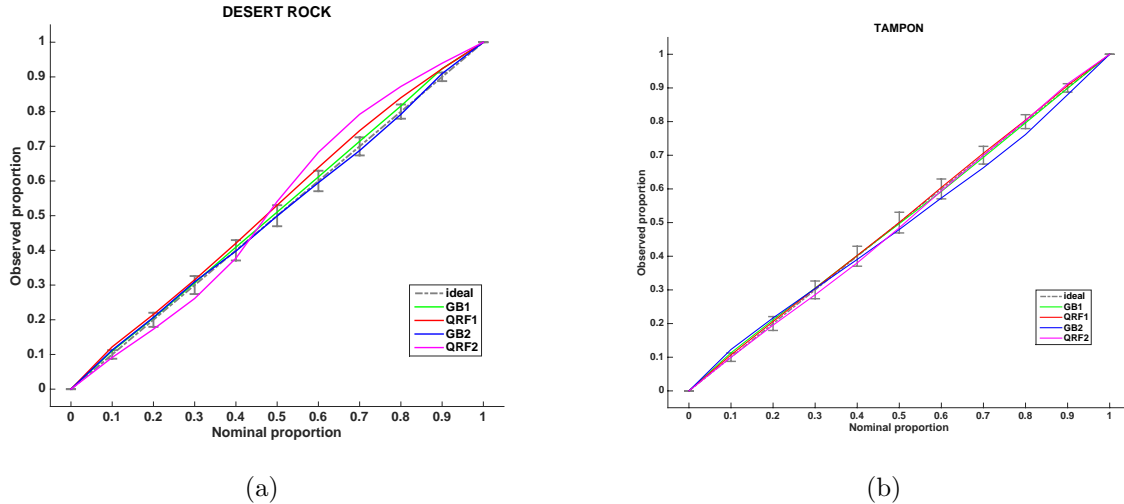


Figure 4: Reliability diagrams related to the intra-day quantile forecasts. (a) Site of Desert Rock (b) Site of Le Tampon. Consistency bars for a 90% confidence level around the ideal line are individually computed for each nominal proportion.

386 In order to visually assess the reliability of quantile forecasts, Figures 4(a) and 4(b) plot  
 387 the reliability diagrams (averaged over all the forecasting horizons) for the two selected sites.  
 388 Consistency bars for a 90% confidence level are individually computed for each nominal pro-  
 389 portion. From the visual inspection of the reliability diagrams of Desert Rock, one can  
 390 possibly state that the GB1 and GB2 models are reliable as the observed proportions of all  
 391 quantiles lie within the consistency bars. Conversely, for QRF1 and QRF2 models, observed  
 392 proportions of some quantiles lie outside the consistency bars. In particular, quantile fore-  
 393 casts generated by the QRF2 model should not be considered reliable. In addition, notice  
 394 the particular signature of the QRF2 model that corresponds to an over dispersed predic-  
 395 tive distribution (i.e. an underconfident model). For the site of Le Tampon, it seems that,  
 396 except the GB2 model, all the other models lead to possible reliable quantile forecasts since  
 397 all of their observed proportions lie within the consistency bars. At this stage, the visual  
 398 reliability assessment related to Le Tampon is not conclusive. This is why we recommend  
 399 in a second step the use of proper score like the CRPS (and its related decomposition) to  
 400 quantify objectively the performance of the methods. This will permit a clear cut ranking  
 401 of the different models.  
 402

### 403 5.2.2. Rank histogram

404 The rank histogram is a graphical display initially designed for assessing ensemble fore-  
 405 casts (Wilks, 2014). But, it can be extended to quantile forecasts by assuming that all

406 evenly spaced forecasted quantiles form an ensemble. Rank histograms permit to assess the  
407 statistical consistency of the ensemble, that is, if the observation can be seen statistically  
408 just like another member of the ensemble (Wilks, 2014). A flat rank histogram is a neces-  
409 sary condition for ensemble consistency and shows an appropriate degree of dispersion of  
410 the ensemble. Put differently, the flatness of the rank histogram indicates that the ensemble  
411 members are statistically indistinguishable from the observations (Wilks, 2014). An under-  
412 dispersed ensemble (i.e. ensemble dispersion consistently too small) leads to a U-shape rank  
413 histogram and shows that the observation will often be an outlier in the distribution of  
414 ensemble members. EPS, such as ECMWF-EPS, are known to suffer from a lack of spread.  
415 As a consequence the resulting rank histograms (Figures 5(a) and 6(a)) exhibit a U-shape.

416 Conversely, an over-dispersed ensemble (i.e. ensemble dispersion consistently too large)  
417 gives a hump shape rank histogram and indicates that the observation may too often be in  
418 the middle of the ensemble distribution.

419 In addition, rank histograms can also detect deficiencies in ensemble calibration or relia-  
420 bility (Wilks, 2014). For instance, some unconditional biases can be revealed by asymmetric  
421 (triangle shape) rank histograms. Furthermore, overpopulation of the smallest (resp. high-  
422 est) ranks will correspond to an overforecasting (resp. underforecasting) bias. Such a bias  
423 can be observed in figures 5(b) and 6(b). Indeed the calibration with the VD method reduces  
424 the under-dispersion but an overforecasting bias appears for both sites as a large number of  
425 the smallest ranks remain above the consistency bars. It must be stressed that one should  
426 be cautious when analyzing rank histograms. Indeed, as shown by (Hamill, 2001), a perfect  
427 flat rank histogram does not state that the corresponding forecast is reliable. Further, when  
428 the number of observations is limited, consistency bars can also be calculated with the pro-  
429 cedure proposed by (Bröcker and Smith, 2007a). To build a rank histogram, it is necessary  
430 to find the rank of the observations when pooled within the ordered ensemble and then plot  
431 the histogram of the ranks. For an ensemble of  $M$  members, the number of ranks of the  
432 histogram is  $M + 1$ . The histogram of verification ranks will be uniform with theoretical  
433 relative frequency of  $\frac{1}{M+1}$  if the consistency condition is met.

434 Finally, the two case studies (Figures 5 and 6) show that forecasts calibrated with the  
435 VD method are more reliable than the original ones. But as a large part of the ranks falls  
436 outside of the consistency bars the resulting forecasts can not be considered reliable.

### 437 5.2.3. PIT histogram

438 Although being at this stage redundant with the reliability diagram, we also present here  
439 the PIT histograms in order to discuss possible issues related with the use of this graphical  
440 tool. PIT histograms may help to assess the calibration property by verifying whether the  
441 observations can be seen as random samples of the predictive distributions (Gneiting et al.,  
442 2007). PIT histograms assess calibration of cumulative predictive distributions checking  
443 whether the observations can be considered as random samples of these distributions. Con-  
444 trary to rank histograms, PIT histograms require the computation of the predictive CDF.  
445 The PIT is the value that the predictive CDF has for a particular observation. PIT values  
446 can be calculated over a testing set of observations and one can then plot the histogram  
447 of the PIT values. Similarly to rank histograms, a flat PIT histogram is a necessary but



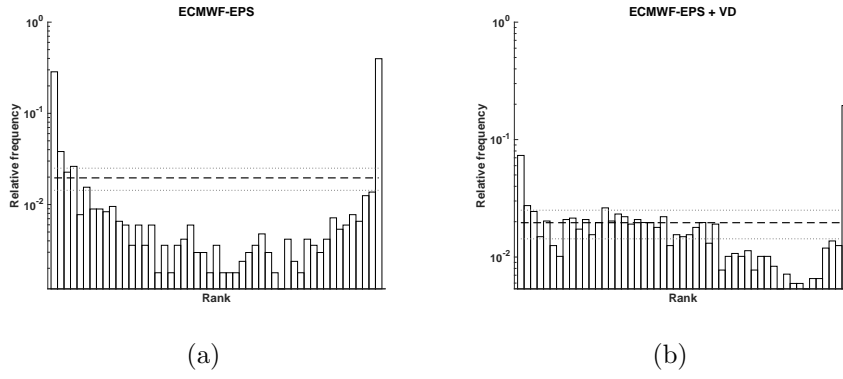


Figure 5: Rank histograms for Desert Rock with consistency band for a 90% confidence level of raw ECMWF-EPS (a) and ECMWF-EPS calibrated with Variance Deficit (VD) method (b).

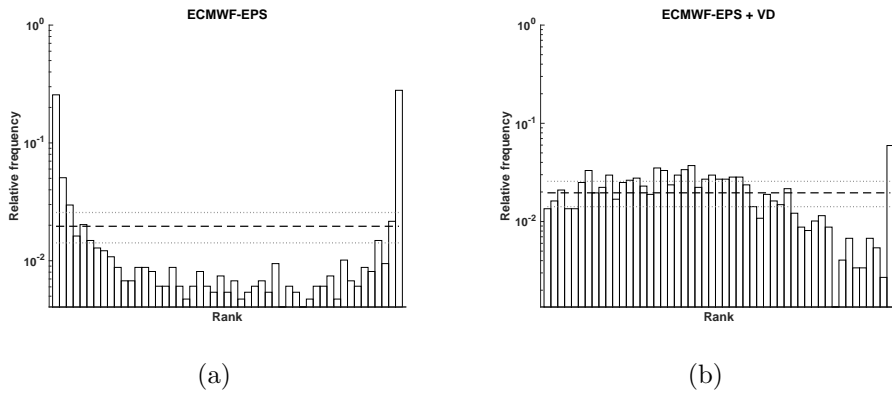


Figure 6: Rank histograms for Le Tampon with consistency band for a 90% confidence level of raw ECMWF-EPS (a) and ECMWF-EPS calibrated with Variance Deficit (VD) method (b) .

448 not sufficient condition to state that a forecast is reliable. As for rank histograms, departures  
 449 from flatness is a sign of conditional biases in the forecasts or over/under-dispersion.  
 450 Like rank histograms, consistency bars can be added to PIT histograms to see how much deviation from the ideal uniform line can be seen as acceptable, in view of sample size.

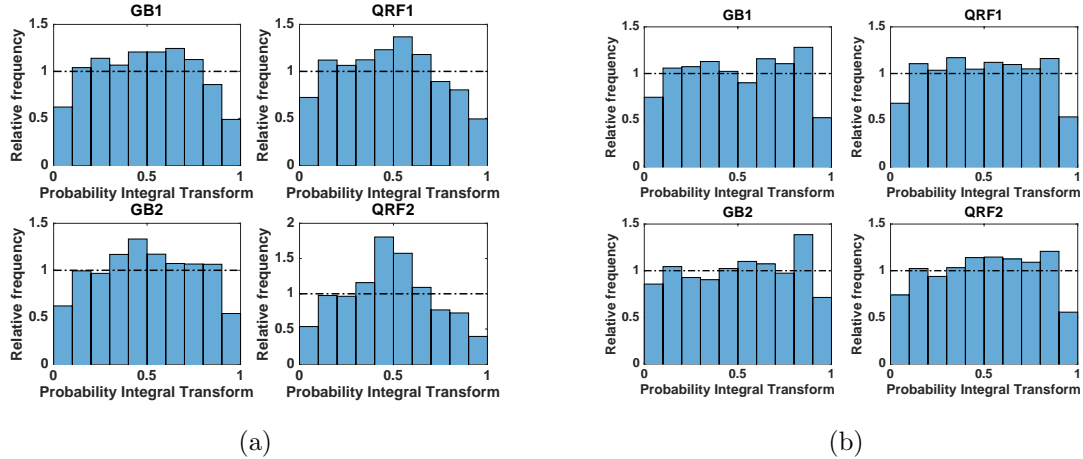


Figure 7: Assessment of the reliability of the intra-day quantile forecasts with PIT diagrams, (a) Site of Desert Rock (b) Site of Le Tampon.

451 Figure 7 shows the PIT histograms (averaged over all the lead times) related to the two  
 452 sites. Following the preceding reliability analysis which possibly stated that, except the GB2  
 453 model, all models were reliable for the site of Le Tampon (see Figure 4(b)), one may expect  
 454 corresponding flat PIT histograms for the GB1, QRF1 and QRF2 models (Figure 7(b)).  
 455 However, this is not the case. We suspect that this may come from the fact that one needs to  
 456 specify the number of histograms bins to plot the PIT histogram. In addition, interpolation  
 457 is needed between the discrete quantiles to estimate the value the CDF attains at the  
 458 observation. This may motivate the choice of reliability diagrams against PIT histograms  
 459 for assessing calibration. However, it is worth noting that, in accordance with the reliability  
 460 diagram, the PIT histogram of the QRF2 method for Desert Rock confirms that this model  
 461 corresponds to an over-dispersed forecasting system (i.e. too wide predictive distributions).  
 462

#### 463 5.2.4. Sharpness diagram

464 A probabilistic forecast is sharp if prediction intervals are shorter on average than pre-  
 465 diction intervals derived from naïve methods, such as climatology or persistence.

466 Similarly to Pinson et al. (2007), we propose to assess the sharpness of the predictive  
 467 distributions by calculating the mean size of the central prediction intervals denoted by  $\bar{\delta}^\alpha$   
 468 for different nominal coverage rates  $(1 - \alpha)\%$ .

469 This leads to a graphical verification display called  $\delta$ -diagrams. For an evaluation set of  
 470  $N$  forecasts,  $\bar{\delta}^\alpha$  is given by

$$\bar{\delta}^\alpha = \frac{1}{N} \sum_{i=1}^N (\hat{q}_{\tau=1-\alpha/2} - \hat{q}_{\tau=\alpha/2}). \quad (4)$$

471 Notice that Gneiting et al. (2007) proposed a diagnostic approach to evaluating proba-  
 472 bilistic forecasts that is based on the paradigm of maximizing the sharpness of the predictive  
 473 distributions subject to calibration. In the proposed evaluation framework, sharpness dia-  
 474 grams take the form of box-plots of the width of the prediction intervals.

475 As mentioned above, some researchers in the solar forecasting community used the  
 476 PINAW metric to measure sharpness. This metric is the average width of the  $(1 - \alpha)100\%$   
 477 prediction interval normalized by the mean of variable  $x$  to predict (e.g. here GHI) for  
 478 a testing set of  $N$  pairs of forecasts/observations. For a specific nominal coverage rate  
 479  $(1 - \alpha)100\%$ , PINAW reads as

$$\text{PINAW}(\alpha) = \frac{\sum_{i=1}^N (\hat{q}_{\tau=1-\alpha/2} - \hat{q}_{\tau=\alpha/2})}{\sum_{i=1}^N x}. \quad (5)$$

480 However, even if it can be interesting to compare the performance of forecasting methods  
 481 at different locations, it must be stressed that the sharpness is a property of the forecasts only  
 482 and as such can not depend on the mean of the observations.

483 For quantile forecasts, Figures 8(a) and 8(b) plot the  $\bar{\delta}^\alpha$  diagrams of the four models  
 484 for different coverage rates. It must be noted that the  $\bar{\delta}^\alpha$  values have been averaged over  
 485 all the lead times. One may first notice that prediction intervals are wider for the site of  
 486 Le Tampon than for Desert Rock. As discussed in (Lauret et al., 2017), the variable sky  
 487 conditions experienced by the site of Le Tampon have an impact on the shape of the predic-  
 488 tive distributions. Conversely, the site of Desert Rock that experiences higher occurrences  
 of clear and stable skies exhibits narrower prediction intervals.

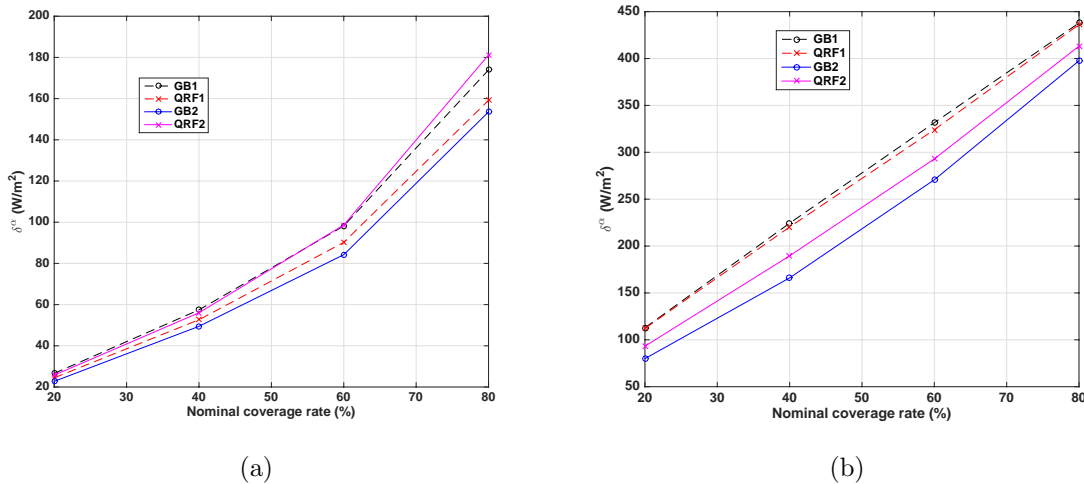


Figure 8: Sharpness diagrams of intra-day quantile forecasts for coverage rates ranging from 20% to 80%  
 (a) Site of Desert Rock (b) Site of Le Tampon.

489 For both sites, it appears that the GB2 model leads to the lowest  $\bar{\delta}^\alpha$  values for all the  
 490 forecasting horizons albeit the difference with the other models is less pronounced for the site  
 491 of Desert Rock. At this point, the sharpness evaluation may favor the GB2 model for both  
 492

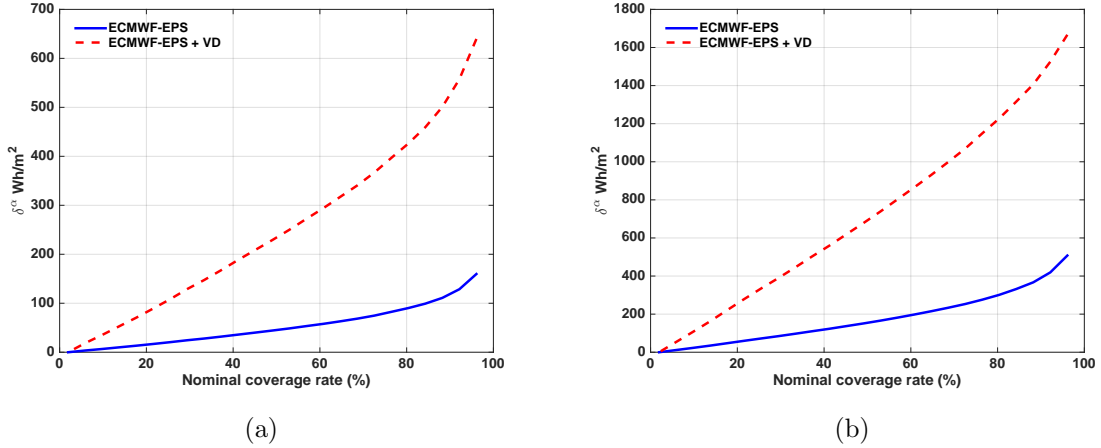


Figure 9: Sharpness diagrams for coverage rates ranging from 0% to 100% of ECMWF-EPS and ECMWF-EPS + VD for Desert Rock (a) and for Le Tampon (b).

593 sites. However, while the GB2 model may possibly generate reliable forecasts for the Desert  
 594 Rock site, this may not be the case for Le Tampon site. If one attempts to select the best  
 595 approach for both sites by combining the two previous separate reliability and sharpness  
 596 assessments, the picture is less clear. Hence evaluating separately reliability and sharpness  
 597 and drawing conclusions on the sole examination of either one of these diagnostic tools may  
 598 be misleading.

599 Regarding ensemble forecasts, as none of the ensemble forecasts are reliable (see 5.2.2,  
 600 there is normally no need to lead further investigations about the sharpness of the prediction  
 601 intervals. Indeed, a comparison of the sharpness of the forecasts could lead to a misunder-  
 602 standing. Nevertheless, we do it for this study case to illustrate this issue. Figure 9 shows  
 603 sharpness diagrams for coverage rates ranging from 0% to 100%, for the two sites and for  
 604 the two considered ensemble forecasts. To compute the mean size of the central prediction  
 605 interval  $\delta^\alpha$ , we assume an uniform spacing of the quantiles derived from the ensemble (see  
 606 section 2). As shown by Figure 9, predictions intervals (PIs) of original ECMWF-EPS fore-  
 607 casts are narrower than the calibrated ones. This is the consequence of the under-dispersion  
 608 and therefore of the low reliability of the ECMWF-EPS forecasts. So, in this case, even  
 609 if narrow PIs are preferred, sharpness diagrams should not be used as criteria to assess the  
 610 quality of the forecasts. In the next section, we will use the CRPS and its related decompo-  
 611 sition into reliability and resolution in an attempt to assess objectively and quantitatively  
 612 the properties required for a skillfull probabilistic system.

### 513 5.3. Scores

514 Numerical scores provide summary measures for the evaluation of the quality of prob-  
 515 abilistic forecasts (Gneiting and Raftery, 2007). Scoring rules are based on the predictive  
 516 distribution of the forecast and on the observed value of the variable of interest. Scores  
 517 may help to rank competing probabilistic models. Scores are required to be proper (Bröcker  
 518 and Smith, 2007b; Gneiting and Raftery, 2007). A score is said to be proper if it insures

519 that the perfect forecasts should be given the best score value. If it is not the case, one  
520 could then hedge the score, by finding tricks that permit to get better score values without  
521 attempting to issue better forecasts. More generally, employing a score that is not proper  
522 makes that one can never be sure of the validity of the results from an empirical comparison  
523 or benchmarking of rival approaches (Pinson and Tastu, 2014). The scoring rules proposed  
524 in this work (CRPS, Ignorance score, Interval score, quantile score) are proper. However,  
525 this is not the case of the CWC score discussed in section 1 as demonstrated by (Pinson and  
526 Tastu, 2014).

527 In addition to the property of propriety, a score can be local or non-local. A score is said  
528 to be local if it depends only on the value of the predictive distribution at the observation,  
529 not on other features of the functional form of the predictive PDF.

530 While different proper scores have been proposed in the literature (Bröcker and Smith,  
531 2007b; Gneiting and Raftery, 2007), we focus here on proper scoring rules for probabilistic  
532 forecasts of continuous variables and particularly on the following scores: CRPS, Interval  
533 score, quantile score and Ignorance Score.

534 Finally, it must be noted that, in the following, the different figures plot the relative counter-  
535 parts of the CRPS, Interval Score and Quantile Score. These relative metrics are normalized  
536 by dividing the absolute values by the mean of the GHI for the considered testing period  
537 (see Table 1).

### 538 5.3.1. Continuous Rank Probability Score (CRPS) and its decomposition

539 The CRPS measures the difference between the predicted and observed cumulative dis-  
540 tributions functions (CDF) (Hersbach, 2000). The formulation of the CRPS is

$$CRPS = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{+\infty} \left[ \hat{F}_{fcst}^i(x) - F_{x_{obs}}^i(x) \right]^2 dx, \quad (6)$$

541 where  $\hat{F}_{fcst}(x)$  is the predictive CDF of the variable of interest  $x$  (e.g. GHI) and  $F_{x_{obs}}(x)$  is a  
542 cumulative-probability step function that jumps from 0 to 1 at the point where the forecast  
543 variable  $x$  equals the observation  $x_{obs}$  (i.e.  $F_{x_{obs}}(x) = 1_{\{x \geq x_{obs}\}}$ ). The squared difference  
544 between the two CDFs is averaged over the  $N$  forecast/observation pairs. The CRPS score  
545 rewards concentration of probability around the step function located at the observed value  
546 (Wilks, 2014). In other words, the CRPS penalizes lack of resolution of the predictive  
547 distributions as well as biased forecasts. In addition, for deterministic forecasts, the CRPS  
548 turns to be the MAE (Mean Absolute Error). This fact permits to compare directly the  
549 performance of a probabilistic model against a deterministic one or equivalently evaluate  
550 the added value brought by a probabilistic approach (Ben Bouallègue, 2015). Notice that  
551 the CRPS is negatively oriented (smaller values are better) and the same dimension as the  
552 forecasted variable.

553 For ensemble forecasts, Hersbach (2000) proposed a method to compute the CRPS using  
554 the classical definition of the CDF (see section 2 and figure 3(a)). In the realm of weather  
555 predictions, his method is widely used and at least embedded in one R-package (NCAR-  
556 Research applications laboratory, 2015). Appendix B summarizes the Hersbach's method  
557 to compute the CRPS for ensemble forecasts.

558 As mentioned above and as a proper score (Gneiting and Raftery, 2007), CRPS can be  
 559 further partitioned into the two main attributes of probabilistic forecasts namely reliability  
 560 and resolution. The decomposition of the CRPS leads to

$$\text{CRPS} = \text{RELIABILITY} + \text{UNCERTAINTY} - \text{RESOLUTION}. \quad (7)$$

561 The reliability term provides an estimation of the forecast biases while the resolution  
 562 term quantifies the improvement that results from issuing probability forecasts that are case  
 563 dependent. The uncertainty term cannot be modified by the forecast system and depends  
 564 only on the observations variability (Wilks, 2014). As the CRPS is negatively oriented, the  
 565 goal of a forecast system is to minimize (resp. maximize) as much as possible the reliability  
 566 term (resp. the resolution term). This decomposition of the CRPS may lead to a detailed  
 567 picture of the performance of the forecasting methods.

568 Regarding the calculation of these different terms, two possibilities exist. The first one  
 569 is based on the work of (Hersbach, 2000) and as such best suited for ensemble forecasts rep-  
 570 resented by the classical definition of the CDF. Appendix B gives the formulae to calculate  
 571 the three terms. The second possibility makes use of the fact that CRPS is the integral of  
 572 the Brier Score over all the predictand thresholds. The Brier score is a proper score used  
 573 to evaluate probabilistic forecasts of binary predictands (Wilks, 2014). Appendix C gives  
 574 all the details regarding this second method. As the CRPS has the same unit as the vari-  
 575 able to predict, it can be normalized by the mean (e.g. mean GHI) or the maximum (e.g.  
 576 installed capacity) of the variable to forecast. The normalized CRPS permits to carry out  
 577 comparisons between different datasets (e.g. different locations).

578 Figures 10(a) and 10(b) plot the relative CRPS of the quantile forecasts in relation  
 579 with the forecast horizon for the two considered sites. As expected, the performance of  
 580 the models decreases as the lead-time increases (i.e. the lower the CRPS, the better the  
 581 model). One also may note that the site of Le Tampon, which experiences variable sky  
 582 conditions compared to Desert Rock, yields higher CRPS values. The interested reader is  
 583 referred to (Lauret et al., 2017) where more details are given regarding the impact of the  
 584 sky conditions on the quality of the probabilistic forecasts. As shown by Figures 10(a) and  
 585 10(b), the two non linear models that include the two solar geometric predictors namely  
 586 zenith angle and hour angle (i.e. GB2 and QRF2 models) perform clearly better than the  
 587 variant 1 models regardless the site. Thus, it appears that adding the two solar geometric  
 588 variables brings a clear improvement and especially for a site like Le Tampon which is known  
 589 to experience a morning/afternoon sky asymmetry. Unlike the previous separate analysis of  
 590 reliability and sharpness, CRPS establishes a clear-cut ranking of the models. However, some  
 591 inconsistencies appear with the reliability analysis which showed that the the QRF2 model  
 592 (resp. the GB2 model) was non reliable for Desert Rock (resp. for Le Tampon). Therefore,  
 593 in order to gain a better understanding of the CRPS results, we use the decomposition of  
 594 the CRPS depicted in Appendix C. This decomposition, detailed in Appendix D, shows  
 595 that the reliability component makes a small contribution to the CRPS and that the higher  
 596 quality of the variant 2 models comes from the resolution attribute.

597 We close this subsection related to the CRPS with the CRPS skill score (CRPSS). In a  
 598 similar manner that scores have been proposed to evaluate the skill of deterministic forecasts

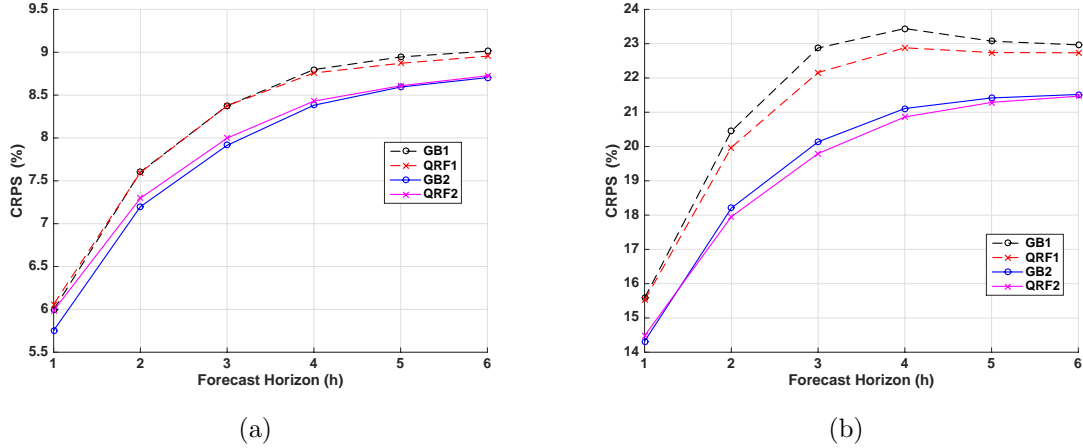


Figure 10: Relative (in % of mean GHI) CRPS of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. The CRPS metric clearly shows the superiority of the variant 2 (GB2 and QRF2) models and particularly for Le Tampon.

599 (Coimbra et al., 2013), (Pedro et al., 2018) used the CRPSS to gauge the performance of  
 600 their probabilistic forecasting models against a reference easy-to-implement method i.e. the  
 601 persistence ensemble (PeEn). In that case, the CRPSS reads as  $CRPSS = 1 - \frac{CRPS_{new\_method}}{CRPS_{PeEn}}$ .

602 In this study, as our primary goal is to verify solar irradiance probabilistic forecasts and  
 603 not to compare and rank forecasting models, we do not detail the implementation of the  
 604 PeEn model. The interested reader should refer to (Pedro et al., 2018). However, as noted  
 605 by (Yang, 2019), the previous definition of the CRPSS may lead to some misinterpretations  
 606 of the skill score as the CRPS of the PeEn model varies according to certain parameters  
 607 (e.g. number of members of the ensemble, forecast lead time, etc.). To address this issue,  
 608 Yang (2019) proposed, instead of PeEn, a new baseline model called the complete-history  
 609 PeEn (CHPeEn) model that gives a nearly constant CRPS.

610 Another way to avoid a CRPSS that depends on the implementation of the reference  
 611 model, and to benefit from the decomposition of the CRPS mentioned above, is to use the  
 612 uncertainty part of the CRPS as the baseline value. The uncertainty component corresponds  
 613 to the CRPS of the climatology and is only sensitive to the observations variability and  
 614 therefore, for a given location and temporal resolution of the data, does not depend on any  
 615 other kind of parameters. Notice that, for meteorologists, when computing skill scores, the  
 616 baseline model is commonly climatology.

### 617 5.3.2. Interval Score (IS)

618 Following Winkler (1972), Gneiting and Raftery (2007) proposed a proper score to specif-  
 619 ically assess the quality of central  $(1 - \alpha)100\%$  prediction interval forecasts. This scoring  
 620 rule called Interval Score (IS), averaged over the  $N$  pairs of forecasts and observations, is

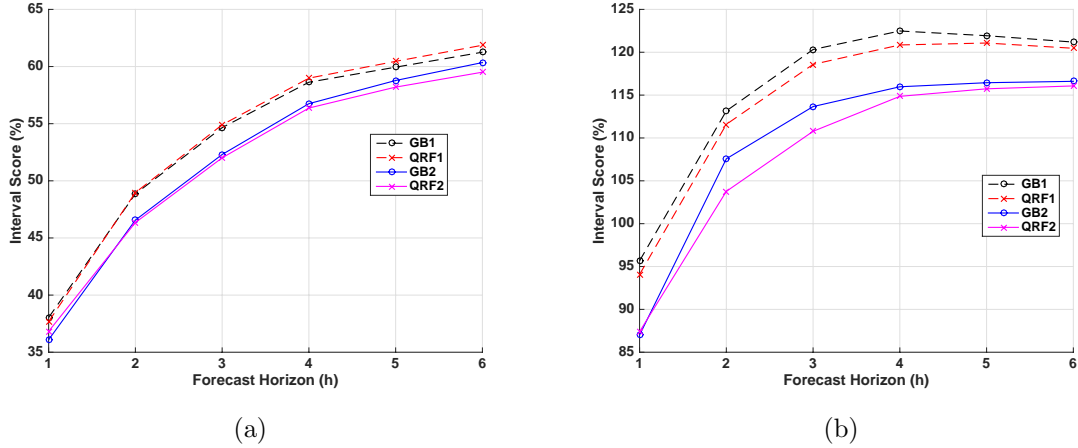


Figure 11: Relative (in % of mean GHI) Interval Score ( $IS_{0.2}$ ) (for 80% central prediction interval) of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. This simple and very easy-to-compute scoring rule shows also that the variant 2 models outperform the variant 1 models.

621 defined by

$$IS_{\alpha} = \frac{1}{N} \sum_{i=1}^N \left( U^i - L^i \right) + \frac{2}{\alpha} \left( L^i - x_{obs}^i \right) 1_{x_{obs}^i < L^i} + \frac{2}{\alpha} \left( x_{obs}^i - U^i \right) 1_{x_{obs}^i > U^i}, \quad (8)$$

622 where  $L^i$  and  $U^i$  represent respectively the  $\alpha/2$  lower quantile  $\hat{q}_{\tau=\alpha/2}$  and the  $1 - \alpha/2$  upper  
 623 quantile  $\hat{q}_{\tau=1-\alpha/2}$ . As shown by Equation 8, the IS rewards narrow prediction intervals but  
 624 penalizes (with the penalty term that depends on  $\alpha$ ) the forecasts for which the observation  
 625  $x_{obs}$  is outside the interval.

626 Figure 11 shows the IS score for the 80% central prediction interval. Again, variant 2  
 627 models perform better than the other models. In our opinion, this easy-to-calculate score  
 628 can advantageously complete the set of proper scores available to the user.

### 629 5.3.3. Quantile Score (QS)

630 Some users may be interested by the performance of some specific quantiles ( e.g. over-  
 631 forecasting or underforecasting) and particularly those related to the tails of the predictive  
 632 distribution. Quantile Score (QS) permits to obtain detailed information about the fore-  
 633 cast quality at specific probability levels. As noted by (Bentzien and Friederichs, 2014), the  
 634 CRPS averages over the complete range of forecast thresholds through integration of the  
 635 Brier Score (see Appendix C). As a consequence, deficiencies in different parts of the distri-  
 636 bution, e.g. the tails of the distribution, might be hidden. Bentzien and Friederichs (2014)  
 637 recommend to extend the verification framework by calculating QS for different probability  
 638 levels. Notice also that, Bentzien and Friederichs (2014) proposed a decomposition of the  
 639 QS into its reliability and resolution components.

640 QS is based on an asymmetric piecewise linear function  $\psi_{\tau}$  called the check or pinball loss  
 641 function. The check function was first defined in the context of quantile regression (Koenker  
 642 and Bassett, 1978) and is given by



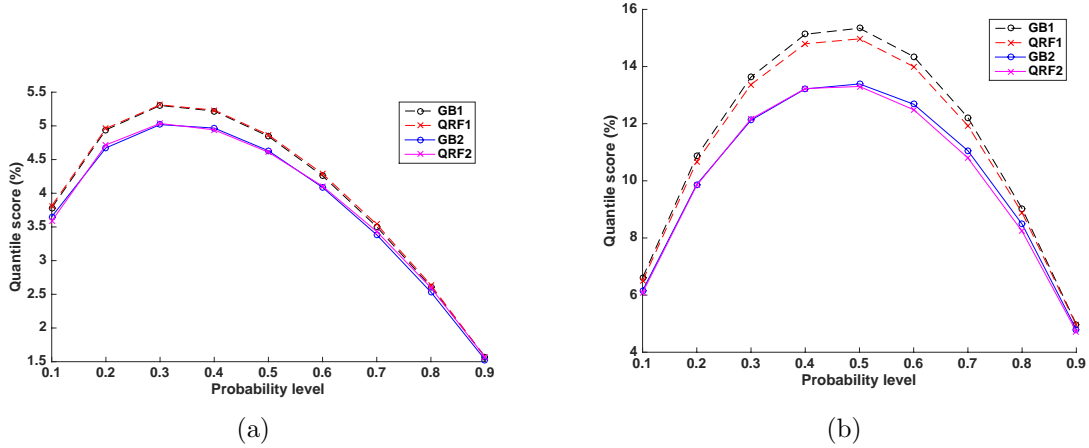


Figure 12: Relative (in % of mean GHI) Quantile Score of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. QS permits to assess the performance of specific quantiles. For Desert Rock, the lowest quantiles are more penalized than the highest ones while for Le Tampon the intermediate quantiles exhibit higher scores.

$$\psi_{\tau}(u) = \begin{cases} \tau u & \text{if } u \geq 0 \\ (\tau - 1)u & \text{if } u < 0, \end{cases} \quad (9)$$

643 with  $\tau$  representing the quantile probability level.

644 QS is given by the mean of the check function applied to the  $N$  pairs of observations  $x_{obs}^i$   
645 and quantile forecasts for a specific probability level  $\tau$ ,  $\hat{q}_{\tau}^i$ . QS reads as

$$QS = \frac{1}{N} \sum_{i=1}^N \psi_{\tau}(x_{obs}^i - \hat{q}_{\tau}^i). \quad (10)$$

646 QS is negatively oriented (i.e. the lower, the better). Finally, notice that Bröcker (2012)  
647 showed that the CRPS can be seen as a weighted sum of quantiles scores applied to the  
648 quantiles derived from the non-uniform CDF.

649 Figure 12 plots the quantile score in relation with the probability levels ranging from  
650 0.1 to 0.9. Again, this detailed analysis of the performance of the models favors the variant  
651 2 models (and particularly for Le Tampon site). Figure 12(b) reveals a symmetric pattern  
652 and shows that the highest quantiles and lowest quantiles are rather well estimated for Le  
653 Tampon. Conversely, regarding the site of Desert Rock, an asymmetric pattern is observed  
654 as the lowest quantiles are more penalized. This is possibly due to the high occurrences of  
655 clear skies experienced by Desert Rock.

#### 656 5.3.4. Ignorance Score (IGN)

657 Initially proposed by (Good, 1952), this score is cited under various names: log score  
658 (Gneiting and Raftery, 2007), divergence (Weijs et al., 2010) or ignorance score (Roulston

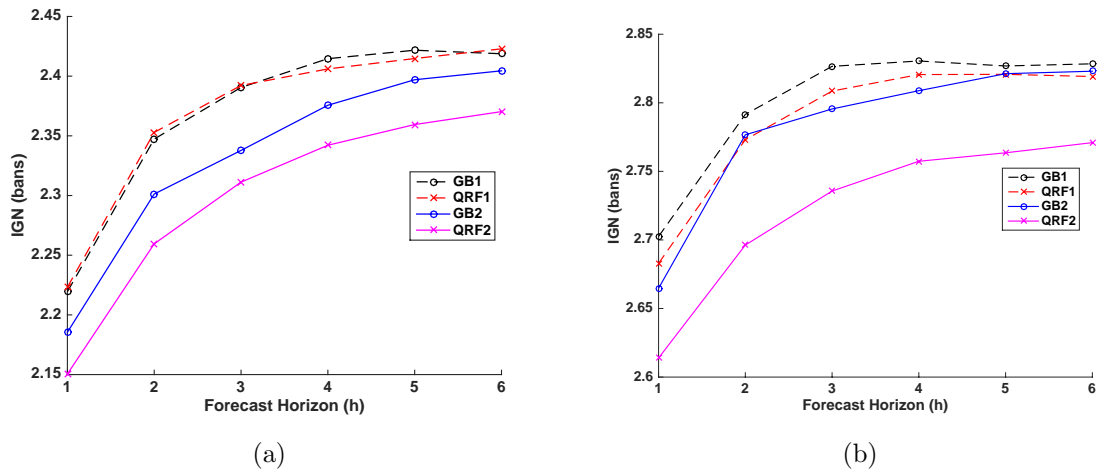


Figure 13: Ignorance Score of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. The IGN score favors clearly the QRF2 model. Notice that the unit of this score is the bans and therefore cannot be normalized by the mean of the irradiance of the testing period.

659 and Smith, 2002). Considering  $N$  verification pairs of probabilistic forecasts given by their  
 660 PDF  $\hat{f}^i(x)$  and outcomes  $x_{obs}^i$ , the ignorance (IGN) is defined as follow

$$IGN = -\frac{1}{N} \sum_{i=1}^N \log(\hat{f}^i(x_{obs}^i)). \quad (11)$$

661 This strictly proper score is appealing because it gathers interesting properties like ad-  
 662 ditivity and locality (i.e. the score depends “only on the value of the probabilistic forecast  
 663 at the verification” (Bröcker and Smith, 2007b)). Like the CRPS, the IGN is a negatively  
 664 oriented score (smaller values are better). Based on the log function, this score is strongly  
 665 affected by the large errors, when the observations fall far away from the highest forecasted  
 666 probabilities. Equation 11 provides a simple way to compute the ignorance score from con-  
 667 tinuous PDFs of parametric distributions or from predictive distributions (i.e. derived from  
 668 discrete estimates, see section 2).

669 Notice that (Tödter and Ahrens, 2012) proposed a generalization of the IGN with an  
 670 approach similar to Hersbach’s work (Hersbach, 2000) about the CRPS. They introduced  
 671 a non-local version of the IGN for binary events and a new score called the Continuous  
 672 Ranked Ignorance score (CRIGN) by analogy to the CRPS. For ensemble forecasts, no clear  
 673 definition of the CDF to use to compute these non-local scores is provided. Thus, the CRIGN  
 674 will not be addressed in this work.

675 Regarding quantile forecasts, Figure 13 plots the ignorance score of the four models.  
 676 This scoring rule confirms the superiority of the variant 2 models although the QRF2 model  
 677 appears to be the best performer. For this particular application, the ignorance score can  
 678 complement the CRPS analysis and may increase the user’s confidence to select the QRF2  
 679 method.

680 Considering ensemble forecasts, Roulston and Smith (2002) proposed a simple approach

681 to compute the IGN. They used the “uniform” definition of the CDF derived from an  
682 ensemble forecast (see section 2 and figure 3(c)) combined with a linear interpolation of  
683 the probabilities between two consecutive members. Then, they applied Equation 11 to the  
684 corresponding PDF that is the first derivative of the CDF (see appendix A for more details).  
685 Thus, the ignorance score of an outcome  $x_{obs}$  that lies between two consecutive members  
686  $[e_k; e_{k+1}]$  of an ensemble forecast with  $M$  members is given by Equation 12. We propose here  
687 a slightly different formulation of the IGN defined in the article of (Roulston and Smith,  
688 2002). They defined the IGN using the binary logarithm (or log base 2) classically proposed  
689 by the field of information theory. We prefer here to use the common logarithm function (or  
690 log base 10) to coincide with the general framework of the IGN (see Equation 11) mainly  
691 used in the literature. For ensemble forecasts, IGN is given by

$$IGN = \log(M + 1) + \log \Delta X_k, \quad (12)$$

692 where

$$\begin{aligned} \Delta X_k &= e_{k+1} - e_k \text{ if } 1 < k < M \\ \Delta X_0 &= e_1 - e_0 \\ \Delta X_M &= e_{M+1} - e_M. \end{aligned} \quad (13)$$

693  $[e_0; e_{M+1}]$  is the a priori interval on which the outcome  $x_{obs}$  is expected to be. Roulston  
694 and Smith (2002) proposed to use the minimum and the maximum of the climatology as  
695 boundaries of this interval. One can notice that this formulation of the IGN assigns the  
696 highest probabilities to the smallest differences between consecutive members. For a verifi-  
697 cation dataset of  $N$  forecast-realization pairs, the ignorance score corresponds obviously to  
698 the arithmetical mean as in Equation 11. Notice that, unlike the CRPS, the ignorance score  
699 cannot be decomposed into reliability, resolution and uncertainty.

700 In what follows, we show that the IGN score, as a local score, can be a less robust score  
701 than the CRPS. Tables 5 and 6 give the IGN, the CRPS and its decomposition for the tested  
702 ensemble forecasts. For Le Tampon and regarding both scores, the calibration brings an  
703 improvement. The decomposition of the CRPS highlights that the calibration increases the  
704 reliability but reduces the resolution. Regarding the site of Desert Rock, the two scores give  
705 an opposite ranking. The IGN assigns a better score to the calibrated ensemble. Conversely,  
706 the CRPS better rates the initial ECMWF forecasts. The decomposition of the CRPS shows  
707 that the increase in reliability, resulting from the calibration, does not counter-balance the  
708 reduction in resolution. Figure 14 illustrates this difference of scoring for a clear sky that has  
709 been forecasted and occurred. The original ECMWF forecast (blue line) already contains  
710 the observation (black line) and the associated CDF is very sharp. So, the IGN and the  
711 CRPS are already relatively low. The VD method (red dashed line) spreads the CDF and  
712 the observation falls close to the median of the calibrated CDF where the probability mass  
713 is the highest. As it is a local score that depends only on the probability at the observation,  
714 the IGN is slightly improved. Conversely, the CRPS, which takes into account the spread  
715 of the CDF, increases significantly. Considering the large number of clear sky conditions  
716 that are forecasted and observed at Desert Rock, the results obtained for this specific case

Table 5: Scores for Desert Rock

	CRPS (%)	CRPS decomposition (%)			IGN
		Reliability	Resolution	Uncertainty	
ECMWF-EPS	6.97	1.77	37.9	43.1	9.67
ECMWF-EPS + VD	7.37	0.97	36.7	43.1	7.84

Table 6: Scores for Le Tampon

	CRPS (%)	CRPS decomposition (%)			IGN
		Reliability	Resolution	Uncertainty	
ECMWF-EPS	25.1	6.03	23.5	42.6	9.13
ECMWF-EPS + VD	23.1	2.41	21.9	42.6	7.89

717 can be extended to a whole year. We can conclude that the VD calibration method spreads  
718 blindly the ECMWF forecasts, even when it is not necessary. As it is a local score, the IGN  
719 is not able to catch and to quantify such a behavior of forecasting models. Consequently, it  
720 seems less robust than the CRPS.

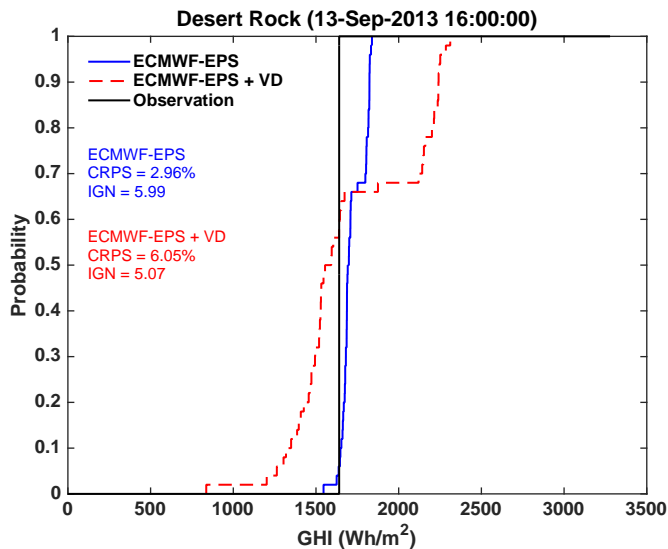


Figure 14: Illustration of the evolution of the CRPS and of the IGN between original and calibrated forecasts: case where these two scores give contradictory information. The CDFs are plotted using the classical definition for ensemble forecasts (see section 2).

## 721 6. Conclusions

722 In this work, we proposed a framework for evaluating solar probabilistic forecasts. Two  
723 types of solar probabilistic forecasts namely ensemble forecasts and quantile forecasts were

724 used to illustrate the evaluation framework. This latter is based on visual diagnostic tools  
725 and scoring rules originally designed by the weather forecast verification community. For  
726 both types of probabilistic forecasts (quantile and ensemble forecasts), we proposed to follow  
727 the same approach to assess the quality of the models albeit some diagnostic tools are more  
728 appropriate depending on the type of forecast.

729 The proposed approach consists in first evaluating the reliability attribute. Graphical  
730 displays such as reliability diagrams and rank histograms with consistency bars, respectively  
731 for quantile forecasts and ensemble forecasts, are efficient, easy-to-build graphical tools ded-  
732 icated to this purpose. Once the reliability attribute checked, a sharpness analysis can be  
733 conducted. However, in our opinion, even if sharpness is an intuitive property that can be  
734 visually assessed with diagrams, it can only contribute to a qualitative evaluation of the  
735 forecasting methods. More generally, visual diagnostic tools cannot allow one to objectively  
736 conclude on a higher quality of a given model. Therefore, we recommend to systematically  
737 compute an overall score i.e. the CRPS which, in our opinion, might be a standard in assess-  
738 ing probabilistic forecasts of continuous variable. This proper score allows ranking  
739 models and its relative counterpart (i.e. CRPS normalized by the mean irradiance) permit  
740 to carry out sites' comparisons. Furthermore, the decomposition of the CRPS into reliability  
741 and resolution may provide additional insight into the performance of a forecasting system.

742 Also, we recommend to complement the CRPS scoring rule with a set of proper scores like  
743 interval score, ignorance score and quantile score. For instance, quantile score may provide  
744 detailed performance of the models at specific parts of the predictive distributions. Re-  
745 garding the ignorance score, although it can advantageously complement the CRPS results,  
746 attention should be paid to its use, as its locality makes it less robust than the CRPS.

747 Finally, when dealing with ensemble forecasts, dedicated verification tools, such as rank  
748 histograms and the CRPS proposed by (Hersbach, 2000), can be used without any additional  
749 assumptions. Indeed, they assume a classical definition of the underlying CDF and it is  
750 not necessary to define the CDF boundaries. However, care must be taken while deriving  
751 quantiles, prediction intervals and associated metrics from ensembles. As several possibilities  
752 are available, it is important to clearly state which one is used (e.g. uniform or non-uniform  
753 spacing). The authors of this paper have a preference for the uniform spacing because it  
754 defines the quantiles such that the members of the ensemble can be seen as a predictive  
755 distribution.

756 In terms of perspectives, applications related for example to energy management system  
757 or simply micro-grids should greatly benefit from the evaluation framework proposed in  
758 this work. More precisely, the verification tools (and particularly scoring rules like CRPS)  
759 should help selecting the best probabilistic forecasts in order to optimize the operation of  
760 the energy management system and consequently increase the economical benefit of the  
761 associated energy systems.

762 This work focused on the forecasting of the solar irradiance. However, the proposed  
763 methodology and associated tools can be extended to the evaluation of probabilistic forecasts  
764 of solar power generation.

765 **7. Appendices**

766 **Appendix A “Uniform” definition of the CDF and PDF derived from an en-**  
 767 **semble forecast**

768 Let  $E = (e_1, \dots, e_M)$  be an ensemble forecast with  $M$  members  $e_k$ ,  $k = 1, \dots, M$ . The  
 769 uniform definition of the resulting Cumulative Distribution Function (CDF) assigns a prob-  
 770 ability mass of  $1/(M + 1)$  between two consecutive members and for the events that fall  
 771 outside of the ensemble range. The tails of the CDF are bounded by  $e_0$  and  $e_{M+1}$  (see  
 772 figure 3(c)). Considering a linear interpolation between the consecutive members and the  
 773 two limits defined above, the analytic formulation of the CDF  $\hat{F}_k(x)$  corresponding to the  
 774 “uniform” definition is

$$\hat{F}_k(x) = \frac{x + (k\Delta X_k - e_k)}{(M + 1)\Delta X_k}, \quad (14)$$

775 where

$$\Delta X_k = e_{k+1} - e_k \text{ with } k = 0, \dots, M. \quad (15)$$

776 The corresponding Probability Density Function (PDF)  $\hat{f}_k(x)$  is the first derivative of  
 777 the CDF defined above i.e.

$$\hat{f}_k(x) = \frac{d\hat{F}_k(x)}{dx} = \frac{1}{(M + 1)\Delta X_k}. \quad (16)$$

778 **Appendix B Hersbach’s method to compute the CRPS from ensemble fore-**  
 779 **casts**

780 Here, we reproduce the methodology proposed by (Hersbach, 2000) to compute the CRPS  
 781 and its decomposition. Let  $E = (e_1, \dots, e_M)$  be an ensemble forecast with  $M$  members  $e_k$ ,  
 782  $k = 1, \dots, M$  and  $x_{obs}$  the observation. It is important to notice that Hersbach assumes a  
 783 classical definition of the CDF obtained from the ensemble (see figure 3(a)). Thus, the CRPS  
 784 could be seen as the sum of areas defined by the members  $E$ , the square of their associated  
 785 cumulative probability  $p_k$  and the position of the observation  $x_{obs}$ . One then have

$$CRPS = \sum_{k=0}^M \alpha_k p_k^2 + \beta_k (1 - p_k)^2, \quad (17)$$

786 with

$$p_k = \frac{k}{M}. \quad (18)$$

787 The values of  $\alpha$  and  $\beta$  are determined with the position of the observation  $x_{obs}$  when  
 788 pooled within the sorted members. Table 7 gives the values of  $\alpha$  and  $\beta$  for all the possible  
 789 cases. Some care must be taken for  $k = 0$  and  $k = M$ . Indeed, the corresponding intervals  
 790 (i.e.  $(-\infty, e_1]$  and  $[e_M, +\infty)$ ) contribute to the CRPS only if the observation falls outside

Table 7: Determination of  $\alpha$  and  $\beta$ 

$0 < k < M$	$\alpha_k$	$\beta_k$
$x_{obs} > e_{k+1}$	$e_{k+1} - e_k$	0
$e_{k+1} > x_{obs} > e_k$	$x_{obs} - e_k$	$e_{k+1} - x_{obs}$
$x_{obs} < e_k$	0	$e_{k+1} - e_k$
$k = 1, M$ (Outliers)	$\alpha_k$	$\beta_k$
$x_{obs} < e_1$	0	$e_1 - x_{obs}$
$x_{obs} > e_M$	$x_{obs} - e_M$	0

791 the range of the ensemble (see second part of table 7 about the outliers). Finally, considering  
792 a verification dataset of  $N$  forecast-realization pairs, the overall  $\overline{CRPS}$  corresponds to the  
793 mean of the CRPS obtained for each individual forecast i.e.  $\overline{CRPS} = \frac{1}{N} \sum_{i=1}^N CRPS_i$ .

794 Considering ensemble forecasts, the decomposition of the CRPS has no sense for a single  
795 forecast-realization pair. Indeed, such case has null uncertainty and resolution. Therefore,  
796 the decomposition of the  $\overline{CRPS}$  proposed by Hersbach is based on the mean values  $\bar{\alpha}_k =$   
797  $\frac{1}{N} \sum_{i=1}^N \alpha_k^i$  and  $\bar{\beta}_k = \frac{1}{N} \sum_{i=1}^N \beta_k^i$ . The components of the CRPS are

$$\overline{REL} = \sum_{k=0}^M \bar{g}_k [\bar{o}_k - p_k]^2, \quad (19)$$

$$\overline{UNC} = \frac{\sum_{i=1}^N \sum_{j=1}^i |x_{obs}^i - x_{obs}^j|}{N^2}, \quad (20)$$

$$\overline{CRPS}_{pot} = \sum_{k=0}^M \bar{g}_k \bar{o}_k (1 - \bar{o}_k), \quad (21)$$

$$\overline{RES} = \overline{UNC} - \overline{CRPS}_{pot}, \quad (22)$$

801 with

$$\bar{g}_k = \bar{\alpha}_k + \bar{\beta}_k, \quad (23)$$

$$\bar{o}_k = \frac{\bar{\beta}_k}{\bar{\alpha}_k + \bar{\beta}_k}. \quad (24)$$

## 803 Appendix C Decomposition of the CRPS through decomposition of the Brier 804 score

805 Hersbach (2000) showed that the CRPS can be calculated through the integration of the  
806 Brier Score over all possible values of the predictand. The Brier Score (BS) is a scoring  
807 rule used for the prediction of the occurrence of a specific event. Usually, such an event is  
808 characterized by a threshold value  $x$ . The event happened if  $x_{obs} \leq x$  and not happened if  
809  $x_{obs} > x$ . One can then have

$$CRPS = \int BS(x)dx = \int REL(x)dx - \int RES(x)dx + \int UNC(x)dx, \quad (25)$$

Table 8: Contingency Table for threshold  $x$ 

Probability $p_k$	Event occurred		Event not occurred	
	$x_{obs} \leq x$		$x_{obs} > x$	
0	$n_0$		$\hat{n}_0$	
...	...		...	
$i$	$n_k$		$\hat{n}_i$	
...	...		...	
1	$n_M$		$\hat{n}_M$	

810 with

$$REL(x) = \sum_{k=0}^M g_k(x) [o_k(x) - p_k]^2, \quad (26)$$

$$RES(x) = \sum_{k=0}^M g_k(x) [o_k(x) - o(x)]^2, \quad (27)$$

$$UNC(x) = o(x) [1 - o(x)]. \quad (28)$$

813 In our case, the integration over  $x$  of the different components ranges for values of GHI from  
814 0 to the maximum of the climatology.

815 For each value of the predictand  $x$ , terms necessary to compute the Brier Score compo-  
816 nents can be calculated from a 2x2 contingency table (see Table 8). In other words, the joint  
817 distribution of forecasts and observations for  $M + 1$  forecast probabilities can be summarized  
818 in a  $(M + 1) \times 2$  contingency table.

819 The total number of pairs of forecasts/observations  $N$  (i.e. the sample size) is given by  
820  $N = \sum_{k=0}^M n_k + \sum_{k=0}^M \hat{n}_k$ .

$$g_k(x) = \frac{l_k}{N}, \quad (29)$$

821 with  $l_k = n_k + \hat{n}_k$

$$o_k(x) = \frac{n_k}{l_k} \quad (30)$$

$$o(x) = \sum_{k=0}^M g_k(x) o_k(x). \quad (31)$$

823 Figure 15 shows the components of the CPRS through the decomposition of the Brier  
824 Score.

## 825 Appendix D Results of the CRPS decomposition for the intraday models

826 First, it should be noted that the uncertainty part is given in Table 1. Figure 16 shows  
827 the resolution part of the CRPS which confirms the lack of resolution of the different models  
828 as the forecast horizon increases. Regarding resolution, the statements made regarding the



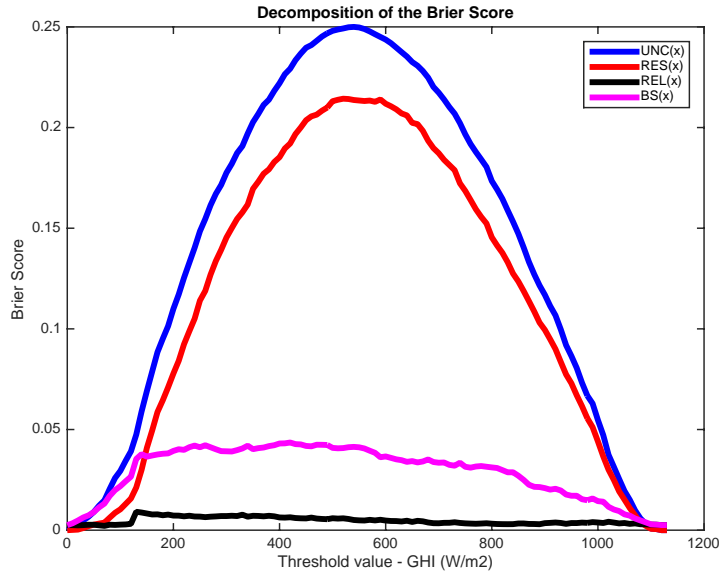


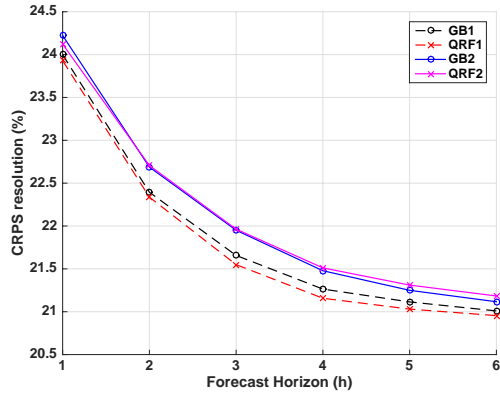
Figure 15: CRPS components through decomposition of the Brier Score (BS) - The area under each curve corresponds to the related CRPS component. Integration of  $BS(x)$  for all threshold values  $x$  gives the CRPS

829 CRPS still hold i.e. the two non-linear models (GB2 and QRF2) that include the solar  
 830 geometric predictors lead to better resolution.

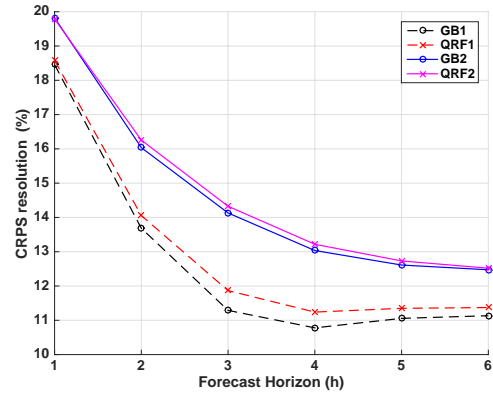
831 Figure 17 plots the reliability component of the CRPS. Surprisingly, the reliability do  
 832 not show a tendency to increase with the lead time. Indeed, we expect the reliability term  
 833 to increase with increasing forecast horizon (we recall that the reliability term is negatively  
 834 oriented i.e. a lower reliability value corresponds to a more reliable forecasts). However,  
 835 in agreement with the reliability assessment, the GB2 model exhibits the lowest reliability  
 836 for the site of Desert Rock while for Le Tampon, low reliability values are obtained with  
 837 the QRF1 model. Nonetheless, it must be noted that the reliability component weakly  
 838 contributes to the CRPS and that the higher quality of the probabilistic forecasts generated  
 839 by the variant 2 models originates from the resolution attribute.

## 840 References

- 841 Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term  
 842 probabilistic solar power forecast. *Applied Energy* 157, 95–110.
- 843 Anderson, J.L., 1996. A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model  
 844 Integrations. *Journal of Climate* 9, 1518–1530.
- 845 Ben Bouallègue, Z., 2015. Assessment and added value estimation of an ensemble approach with a focus on  
 846 global radiation forecasts. *MAUSAN* , 541–550.
- 847 Bentzien, S., Friederichs, P., 2014. Decomposition and graphical portrayal of the quantile score. *Quart. J.*  
 848 *Roy. Meteor. Soc.* 140, 1924–1934.
- 849 Bröcker, J., 2012. Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal*  
 850 *of the Royal Meteorological Society* 138, 1611–1617.

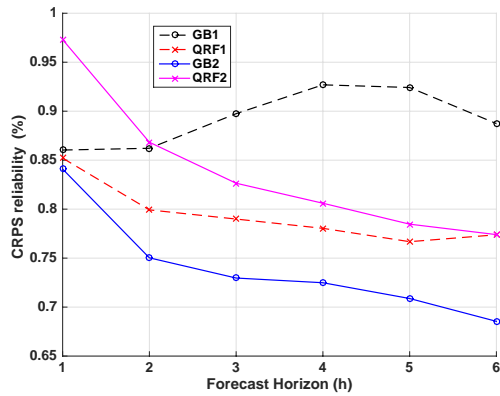


(a)

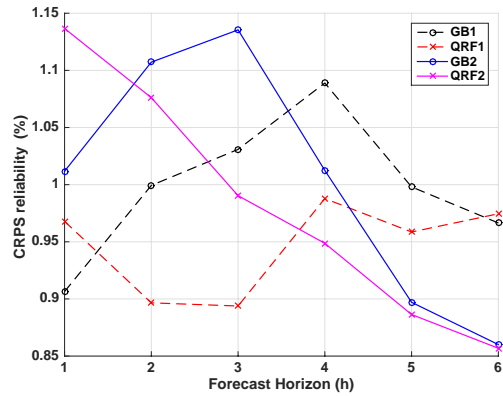


(b)

Figure 16: Relative (in % of mean GHI) resolution component of the CRPS of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. As expected, resolution decreases with increasing lead time. The variant 2 models lead to better resolution.



(a)



(b)

Figure 17: Relative (in % of mean GHI) reliability Component of the CRPS of the different intraday methods (a) Site of Desert Rock (b) Site of Le Tampon. The reliability component weakly contributes to the CRPS.

- 851 Bröcker, J., Smith, L.A., 2007a. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*  
852 22, 651–661.
- 853 Bröcker, J., Smith, L.A., 2007b. Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather*  
854 *and Forecasting* 22, 382–388.
- 855 Chu, Y., Coimbra, C.F., 2017. Short-term probabilistic forecasts for Direct Normal Irradiance. *Renewable*  
856 *Energy* 101, 526–536.
- 857 Chu, Y., Li, M., Pedro, H.T., Coimbra, C.F., 2015. Real-time prediction intervals for intra-hour DNI  
858 forecasts. *Renewable Energy* 83, 234–244.
- 859 Coimbra, C.F., Kleissl, J., Marquez, R., 2013. Overview of Solar-Forecasting Methods and a Metric for  
860 Accuracy Evaluation, in: *Solar Energy Forecasting and Resource Assessment*. Elsevier, pp. 171–194.
- 861 Dambreville, R., Blanc, P., Chanussot, J., Boldo, D., 2014. Very short term forecasting of the Global  
862 Horizontal Irradiance using a spatio-temporal autoregressive model. *Renewable Energy* 72, 291–300.
- 863 David, M., Mazorra Aguiar, L., Lauret, P., 2018. Comparison of intraday probabilistic forecasting of solar  
864 irradiance using only endogenous data. *International Journal of Forecasting* 34, 529–547.
- 865 David, M., Ramahatana, F., Trombe, P., Lauret, P., 2016. Probabilistic forecasting of the solar irradiance  
866 with recursive ARMA and GARCH models. *Solar Energy* 133, 55–72.
- 867 Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. *Journal*  
868 *of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 243–268.
- 869 Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the*  
870 *American Statistical Association* 102, 359–378.
- 871 Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated Probabilistic Forecasting Using  
872 Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review* 133,  
873 1098–1118.
- 874 Golestaneh, F., Gooi, H.B., Pinson, P., 2016a. Generation and evaluation of spacetime trajectories of  
875 photovoltaic power. *Applied Energy* 176, 80 – 91.
- 876 Golestaneh, F., Pinson, P., Gooi, H.B., 2016b. Very Short-Term Nonparametric Probabilistic Forecasting of  
877 Renewable Energy Generation With Application to Solar Energy. *IEEE Transactions on Power Systems*  
878 31, 3850–3863.
- 879 Good, I.J., 1952. Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* 14,  
880 107–114.
- 881 Grantham, A., Gel, Y.R., Boland, J., 2016. Nonparametric short-term probabilistic forecasting for solar  
882 radiation. *Solar Energy* 133, 465–475.
- 883 Hamill, T.M., 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather*  
884 *Review* 129, 550–560.
- 885 Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction  
886 Systems. *Weather and Forecasting* 15, 559–570.
- 887 Hoff, T.E., Perez, R., 2012. Modeling PV fleet output variability. *Solar Energy* 86, 2177–2189.
- 888 Hoff, T.E., Perez, R., Kleissl, J., Renne, D., Stein, J., 2013. Reporting of irradiance modeling relative  
889 prediction errors: Reporting of irradiance modeling relative prediction errors. *Progress in Photovoltaics:*  
890 *Research and Applications* 21, 1514–1519.
- 891 Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy fore-  
892 casting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*  
893 32, 896–913.
- 894 Huang, J., Korolkiewicz, M., Agrawal, M., Boland, J., 2013. Forecasting solar radiation on an hourly time  
895 scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. *Solar Energy* 87, 136–149.
- 896 Iversen, E.B., Morales, J.M., Møller, J.K., Madsen, H., 2016. Short-term probabilistic forecasting of wind  
897 speed using stochastic differential equations. *International Journal of Forecasting* 32, 981–990.
- 898 Jolliffe, I., Stephenson, D., 2003. *Forecast Verification. A practitioner’s guide in atmospheric science*. Wiley,  
899 Chichester, England.
- 900 Jung, J., Broadwater, R.P., 2014. Current status and future advances for wind speed and power forecasting.  
901 *Renewable and Sustainable Energy Reviews* 31, 762–777.

- 902 Khosravi, A., Nahavandi, S., Creighton, D., 2013. Prediction Intervals for Short-Term Wind Farm Power  
903 Generation Forecasts. *IEEE Transactions on Sustainable Energy* 4, 602–610.
- 904 Koenker, R., Bassett, G., 1978. Regression Quantiles. *Econometrica* 46, 33–50.
- 905 Lauret, P., David, M., Pedro, H., 2017. Probabilistic Solar Forecasting Using Quantile Regression Models.  
906 *Energies* 10, 1591.
- 907 Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P., 2015. A benchmarking of machine learning  
908 techniques for solar radiation forecasting in an insular context. *Solar Energy* 112, 446–457.
- 909 Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. *Journal of Computational Physics* 227, 3515–  
910 3539.
- 911 Li, K., Wang, R., Lei, H., Zhang, T., Liu, Y., Zheng, X., 2018. Interval prediction of solar power using an  
912 Improved Bootstrap method. *Solar Energy* 159, 97–112.
- 913 Lorenz, E., Heinemann, D., 2012. Prediction of solar irradiance and photovoltaic power., in: *Comprehensive  
914 Renewable Energy*. Elsevier, Oxford, UK, pp. 239–292.
- 915 Marquez, R., Coimbra, C.F., 2011. Forecasting of global and direct solar irradiance using stochastic learning  
916 methods, ground experiments and the NWS database. *Solar Energy* 85, 746–756.
- 917 van der Meer, D., Widn, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic  
918 power production and electricity consumption. *Renewable and Sustainable Energy Reviews* 81, 1484 –  
919 1512.
- 920 Morales, J.M., Conejo, A.J., Madsen, H., Pinson, P., Zugno, M., 2014. Integrating Renewables in Electricity  
921 Markets. volume 205 of *International Series in Operations Research & Management Science*. Springer  
922 US, Boston, MA.
- 923 Murphy, A.H., 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting.  
924 *Weather and Forecasting* 8, 281–293.
- 925 NCAR-Research applications laboratory, 2015. verification: Weather Forecast Verification Utilities. R  
926 package version 1.42.
- 927 Pedro, H.T., Coimbra, C.F., 2015. Nearest-neighbor methodology for prediction of intra-hour global hori-  
928 zontal and direct normal irradiances. *Renewable Energy* 80, 770–782.
- 929 Pedro, H.T., Coimbra, C.F., David, M., Lauret, P., 2018. Assessment of machine learning techniques for  
930 deterministic and probabilistic intra-hour solar forecasts. *Renewable Energy* 123, 191–203.
- 931 Pinson, P., McSharry, P., Madsen, H., 2010. Reliability diagrams for non-parametric density forecasts of  
932 continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological  
933 Society* 136, 77–90.
- 934 Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G.N., 2007. Non-parametric probabilistic  
935 forecasts of wind power: required properties and evaluation. *Wind Energy* 10, 497–516.
- 936 Pinson, P., Reikard, G., Bidlot, J.R., 2012. Probabilistic forecasting of the wave energy flux. *Applied Energy*  
937 93, 364–370.
- 938 Pinson, P., Tastu, J., 2014. Discussion of “Prediction Intervals for Short-Term Wind Farm Generation Fore-  
939 casts” and “Combined Nonparametric Prediction Intervals for Wind Power Generation”. *IEEE Transac-  
940 tions on Sustainable Energy* 5, 1019–1020.
- 941 Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts.  
942 *Solar Energy* 83, 342–349.
- 943 Roulston, M., Smith, L., 2002. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly  
944 Weather Review* 130, 1653–1660.
- 945 Scolari, E., Sossan, F., Paolone, M., 2016. Irradiance prediction intervals for PV stochastic generation in  
946 microgrid applications. *Solar Energy* 139, 116–129.
- 947 Sperati, S., Alessandrini, S., Delle Monache, L., 2016. An application of the ECMWF Ensemble Prediction  
948 System for short-term solar power forecasting. *Solar Energy* 133, 437–450.
- 949 Tödter, J., Ahrens, B., 2012. Generalization of the Ignorance Score: Continuous Ranked Version and Its  
950 Decomposition. *Monthly Weather Review* 140, 2005–2017.
- 951 Verbois, H., Rusydi, A., Thiery, A., 2018. Probabilistic forecasting of day-ahead solar irradiance using  
952 quantile gradient boosting. *Solar Energy* 173, 313–327.

- 953 Voyant, C., Motte, F., Fouilloy, A., Notton, G., Paoli, C., Nivet, M.L., 2017. Forecasting method for global  
954 radiation time series without training phase: Comparison with other well-known prediction methodologies.  
955 Energy 120, 199–208.
- 956 Weijs, S.V., van Nooijen, R., van de Giesen, N., 2010. Kullback–leibler divergence as a forecast skill score  
957 with classic reliability–resolution–uncertainty decomposition. Monthly Weather Review 138, 3387–3399.
- 958 Wilks, D.S., 2014. Statistical Methods in the Atmospheric Sciences. An Introduction. Elsevier Science,  
959 Burlington.
- 960 Winkler, R.L., 1972. A Decision-Theoretic Approach to Interval Estimation. Journal of the American  
961 Statistical Association 67, 187–191.
- 962 Yang, D., 2019. A universal benchmarking method for probabilistic solar irradiance forecasting. Solar  
963 Energy 184, 410–416.
- 964 Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T., Coimbra, C.F., 2018. History and trends in solar  
965 irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar  
966 Energy 168, 60–101.
- 967 Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods  
968 for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily  
969 production. Solar Energy 105, 804–816.