Verification of solar irradiance probabilistic forecasts

Philippe Lauret[a,*], Mathieu David[a], Pierre Pinson[b]

[a] *University of La Réunion - PIMENT laboratory, 15, avenue René Cassin, 97715 Saint-Denis*
[b] *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)

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

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

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

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

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

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

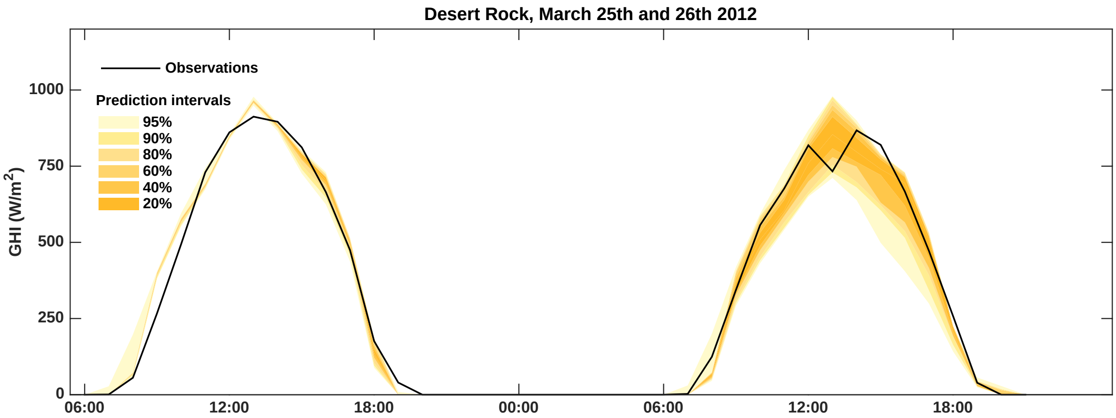


Figure 1: Example of probabilistic solar irradiance forecasts: 2 days of measured GHI at the Desert Rock (NV) and associated day-ahead forecasts with prediction intervals provided by ECMWF-EPS (see section 3).

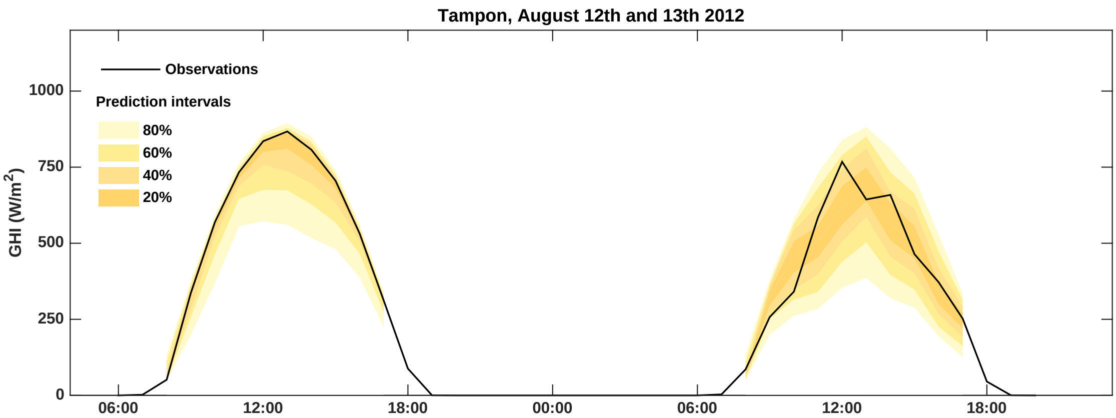


Figure 2: Example of probabilistic solar irradiance forecasts: 2 days of measured GHI at Tampon and associated 1-hour ahead forecasts with prediction intervals generated with the Quantile Random Forest model QRF2 (see section 3).

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

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

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

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

This is why, we think that now is the time to take stock on the evaluation metrics of solar probabilistic forecasts. The objective of this work is therefore to provide the forecasting solar community a comprehensive overview of diagnostic tools and scoring rules that can be used to assess the performance of probabilistic forecasting methods. In particular, we propose an evaluation framework that may help the user to consistently evaluate the quality of the models. In others words, this paper aims at explaining how one should assess the quality of the probabilistic forecasts and how diagnostic tools and scores should be used and interpreted. In addition, we will propose a measure of resolution (through the decomposition of the CRPS) as this attribute is not currently assessed in the literature.

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

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

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

2. Nature of probabilistic forecasts of continuous variables

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

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

A quantile q_τ informs there is a probability τ that the event x materializes below that quantile q_τ. From a set of quantiles, prediction intervals (PIs) can be deduced. PIs define the range of values within which the observation is expected to be with a certain probability i.e. its nominal coverage rate (Pinson et al., 2007). To completely determine a PI, it is necessary to choose the way it should be centered on the probability density function (Pinson et al., 2007). The most common way is to center the PI on the median. Consequently, there is the same probability of risk below and above the median. Therefore, a central PI with a coverage rate of $(1-\alpha)100\%$ is estimated by using the $\alpha/2$ quantile ($\hat{q}_{\tau=\alpha/2}$) as the lower bound and the $(1-\alpha/2)$ quantile ($\hat{q}_{\tau=1-\alpha/2}$) as the upper bound. More precisely, a PI with $(1-\alpha)100\%$ nominal coverage rate is given by

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

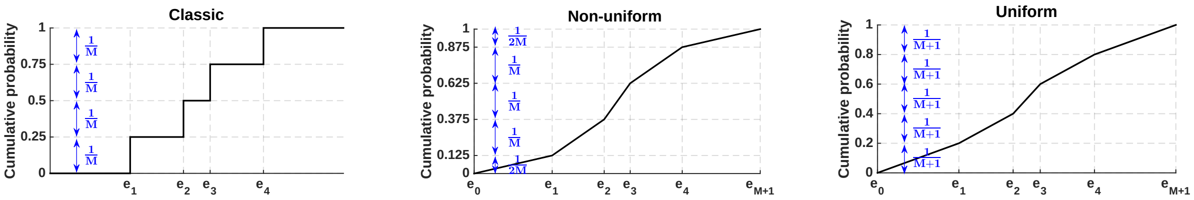
In the realm of weather predictions, three ways to define this cumulative distribution are available: parametric CDFs, discrete estimates of a CDF via a non-parametric method and ensemble forecasts. Parametric CDFs are easy to set up and to assess. Nevertheless, regarding solar forecasts, they are seldom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

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

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

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

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



(a) (b) (c)

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.

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

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

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

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

Thus, when dealing with ensemble forecasts, three ways to build the CDF from the members are available. Unfortunately no definition can be favoured and each CDF construction has its pros and cons. The classic definition is the most used, specifically to compute the Continuous Rank Probability Score (CRPS, see section 5.3.1) with the methodology proposed by (Hersbach, 2000). As this commonly used definition assigns null probabilities to the events that fall outside of the ensemble, it can not be used to derive scores like ignorance (see section 5.3.4). The uniform and the non-uniform definitions requires to arbitrarily fix the boundaries of the CDF. Therefore, they are user dependent. Nevertheless, they allow designing continuous CDF that contains all the possible events. Thus, the procedure used to verify the quality of ensemble forecasts can be exactly the same as for the parametric CDFs or for the predictive distributions summarized by discrete quantiles estimated by some kind of statistical method. Bröcker (2012) showed that the non-uniform definition corresponds to a minimization of the CRPS. But, considering this definition, the optimal shape of the corresponding rank histogram (see section 5.2.2) is not flat. Indeed for this visual verification tool, the height of the first and last ranks should be the half of the other ones. Finally, if the aim is to compare different forecasting models, whatever the chosen definition, the ranking will remain the same. Nevertheless, a unique framework has to be defined to allow the comparison of different works.

3. Illustrative case studies

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

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

3.1. Intraday quantile forecasts

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

Table 1: Main characteristic of the solar measurements

	Desert Rock (USA)	**Tampon** (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^2	1.712 MWh/m^2
Solar variability 1-h ($\sigma\Delta kt^*$)	0.146	0.241
Mean GHI (Testing set)	548 W/m^2	458 W/m^2
Uncertainty component of the CRPS	29.1%	33.1%

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

Table 2: Acronyms related to the four quantile regression models

Quantile regression techniques	**Variant 1**	**Variant 2**
Quantile Regression Forest	QRF1	QRF2
Gradient Boosting	GB1	GB2

3.2. Day-ahead ensemble forecasts

The day-ahead ensemble predictions are provided by the Integrated Forecasting System (IFS) of the European Centre of Medium-Range Weather Forecasts (ECMWF). We will denote these ensemble forecasts as “ECMWF-EPS”. They consist in 50 perturbed members. The temporal resolution is of 3 hours and the spatial resolution is of 0.2° in both longitude and latitude. Consequently, 3h GHI (in Wh/m^2) times series recorded on-site are compared with the nearest ECWMF pixel. In addition, we also propose a post-processed version of the original ECMWF-EPS forecasts. Indeed, the ensemble prediction systems of the NWP models commonly suffer from a lack of spread (Leutbecher and Palmer, 2008). To face this issue, Sperati et al. (2016) proposed a simple approach, named Variance Deficit (VD), to calibrate the ensemble forecasts. Their method spreads the initial ensemble forecasts by correcting their variance. The correction factor is evaluated from a training set. The calibrated ensemble forecasts will be denoted by "ECMWF-EPS + VD".

4. Required properties for a skillful probabilistic system

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

4.1. Reliability

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

4.2. Resolution and sharpness

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

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

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

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

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

5. Presentation and application of the verification tools

5.1. Proposed evaluation framework

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

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

For ensemble forecasts, we propose to use the rank histogram including consistency bars 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 asesss 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
CPRS	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

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

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

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

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

5.2. Diagnostic tools

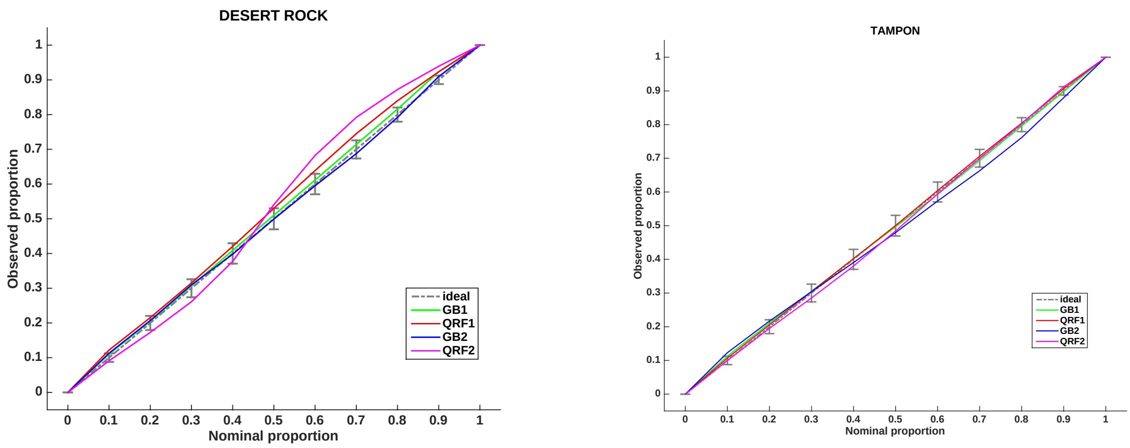
5.2.1. Reliability diagram

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

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

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

Finally, some preceding works (Chu and Coimbra, 2017; Lauret et al., 2017) proposed to evaluate the reliability component of a probabilistic system by calculating the prediction interval coverage probability (PICP) (Khosravi et al., 2013). PICP permits one to assess the empirical coverage probability of the central prediction intervals. However, this metric is not suitable to assess the reliability of probabilistic forecasts because as noted by Pinson et al. (2007), both quantiles that define the prediction interval may be biased. In other 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.



(a) (b)

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.

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

5.2.2. Rank histogram

The rank histogram is a graphical display initially designed for assessing ensemble forecasts (Wilks, 2014). But, it can be extended to quantile forecasts by assuming that all evenly spaced forecasted quantiles form an ensemble. Rank histograms permit to assess the statistical consistency of the ensemble, that is, if the observation can be seen statistically just like another member of the ensemble (Wilks, 2014). A flat rank histogram is a necessary condition for ensemble consistency and shows an appropriate degree of dispersion of the ensemble. Put differently, the flatness of the rank histogram indicates that the ensemble members are statistically indistinguishable from the observations (Wilks, 2014). An under-dispersed ensemble (i.e. ensemble dispersion consistently too small) leads to a U-shape rank histogram and shows that the observation will often be an outlier in the distribution of ensemble members. EPS, such as ECMWF-EPS, are known to suffer from a lack of spread. As a consequence the resulting rank histograms (Figures 5(a) and 6(a)) exhibit a U-shape.

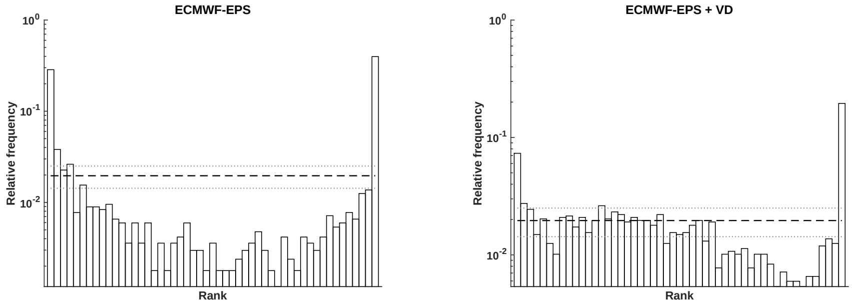
Conversely, an over-dispersed ensemble (i.e. ensemble dispersion consistently too large) gives a hump shape rank histogram and indicates that the observation may too often be in the middle of the ensemble distribution.

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

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

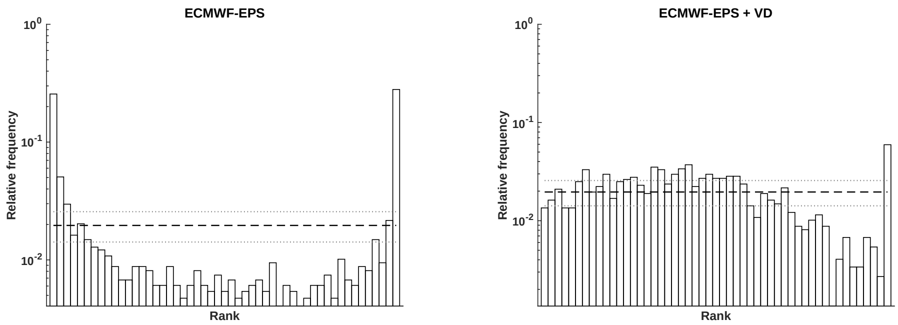
5.2.3. PIT histogram

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



(a) (b)

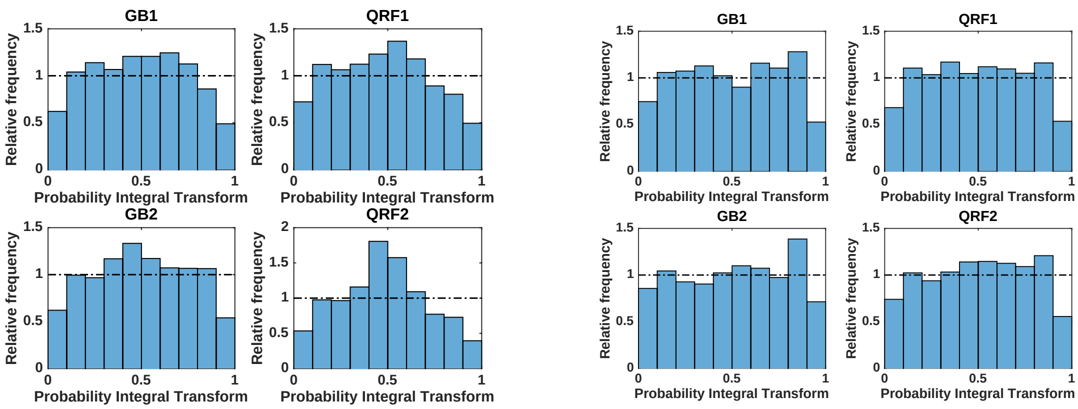
Figure 5: Rank histograms for Desert Rock with consistency band for a 90% confidence level of raw ECMWF-EPS (a) and ECMWF-EPS calibrated with Variance Deficit (VD) method (b).



(a) (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) .

not sufficient condition to state that a forecast is reliable. As for rank histograms, departures from flatness is a sign of conditional biases in the forecasts or over/under-dispersion. 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.



(a) (b)

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.

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

5.2.4. Sharpness diagram

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

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

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

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

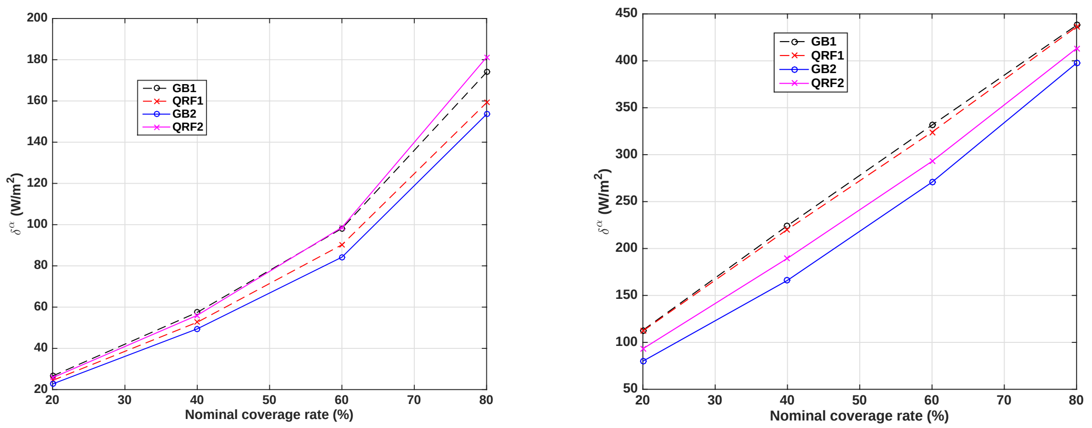
Notice that Gneiting et al. (2007) proposed a diagnostic approach to evaluating probabilistic forecasts that is based on the paradigm of maximizing the sharpness of the predictive distributions subject to calibration. In the proposed evaluation framework,sharpness diagrams take the form of box-plots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

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

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

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

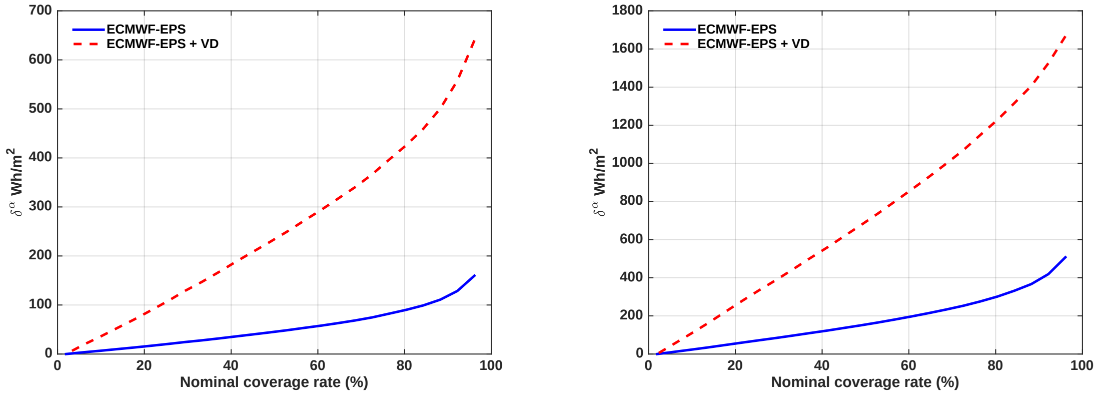
For quantile forecasts, Figures 8(a) and 8(b) plot the $\bar{\delta^{\alpha}}$ diagrams of the four models for different coverage rates. It must be noted that the $\bar{\delta^{\alpha}}$ values have been averaged over all the lead times. One may first notice that prediction intervals are wider for the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



(a) (b)

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.

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



(a) (b)

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).

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

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

5.3. Scores

Numerical scores provide summary measures for the evaluation of the quality of probabilistic forecasts (Gneiting and Raftery, 2007). Scoring rules are based on the predictive distribution of the forecast and on the observed value of the variable of interest. Scores may help to rank competing probabilistic models. Scores are required to be proper (Bröcker and Smith, 2007b; Gneiting and Raftery, 2007). A score is said to be proper if it insures that the perfect forecasts should be given the best score value. If it is not the case, one could then hedge the score, by finding tricks that permit to get better score values without attempting to issue better forecasts. More generally, employing a score that is not proper makes that one can never be sure of the validity of the results from an empirical comparison or benchmarking of rival approaches (Pinson and Tastu, 2014). The scoring rules proposed in this work (CRPS, Ignorance score, Interval score, quantile score) are proper. However, this is not the case of the CWC score discussed in section 1 as demonstrated by (Pinson and Tastu, 2014).

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

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

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

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

The CRPS measures the difference between the predicted and observed cumulative distributions functions (CDF) (Hersbach, 2000). The formulation of the CRPS is

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

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

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

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

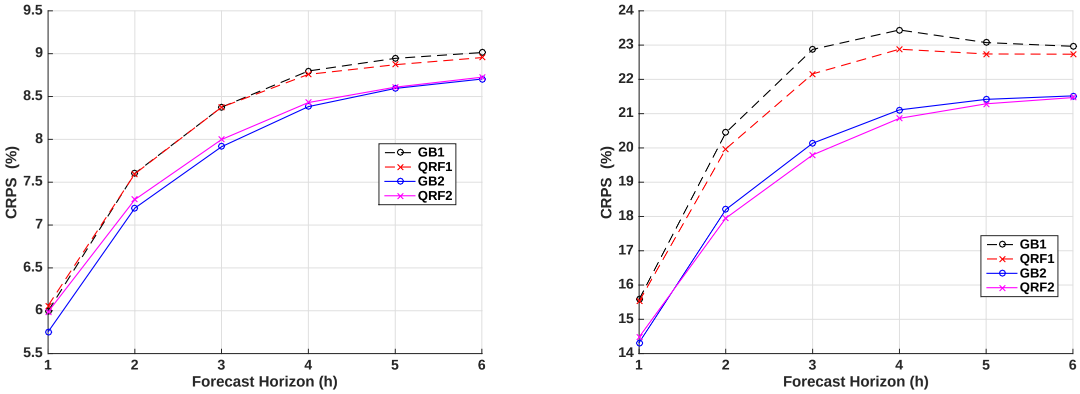
$$\text{CRPS} = RELIABILITY + UNCERTAINTY - RESOLUTION. \tag{7}$$

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

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

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

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



(a) (b)

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.

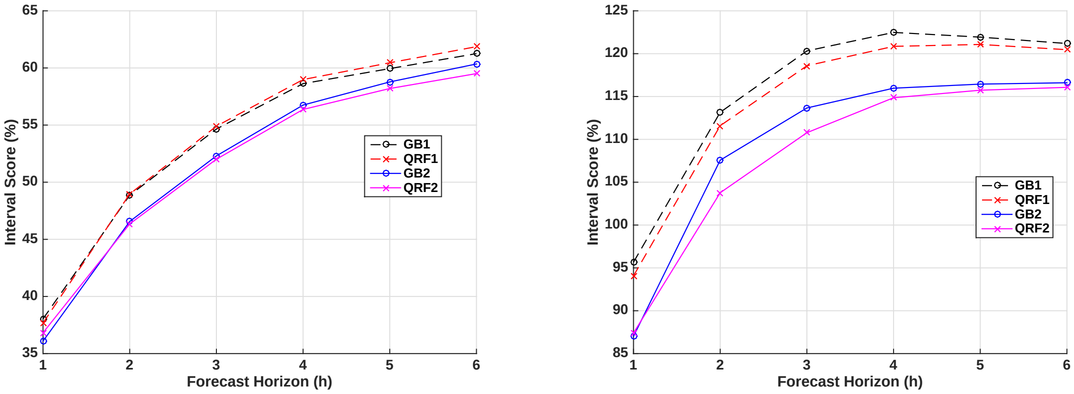
(Coimbra et al., 2013), (Pedro et al., 2018) used the CRPSS to gauge the performance of their probabilistic forecasting models against a reference easy-to-implement method i.e. the persistence ensemble (PeEn). In that case, the CRPSS reads as $\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\text{new method}}}{\mathrm{CRPS}_{\text{PeEn}}}$.

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

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

5.3.2. Interval Score (IS)

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



(a) (b)

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.

defined by

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

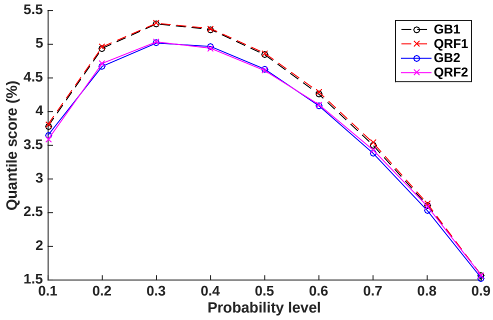
where L^i and U^i represent respectively the $\alpha/2$ lower quantile $\hat{q}_{\tau=\alpha/2}$ and the $1-\alpha/2$ upper quantile $\hat{q}_{\tau=1-\alpha/2}$. As shown by Equation 8, the IS rewards narrow prediction intervals but penalizes (with the penalty term that depends on α) the forecasts for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

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

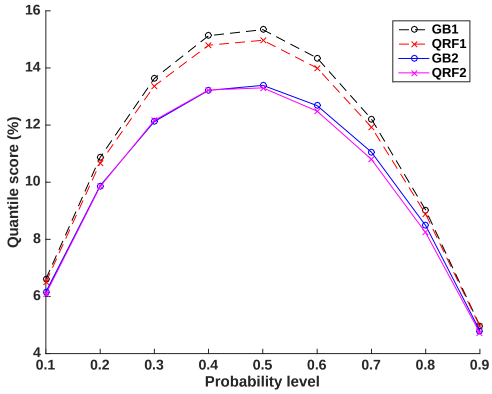
5.3.3. Quantile Score (QS)

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

QS is based on an asymmetric piecewise linear function ψ_τ called the check or pinball loss function. The check function was first defined in the context of quantile regression (Koenker and Bassett, 1978) and is given by



(a)



(b)

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.

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

with τ representing the quantile probabilty level.

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

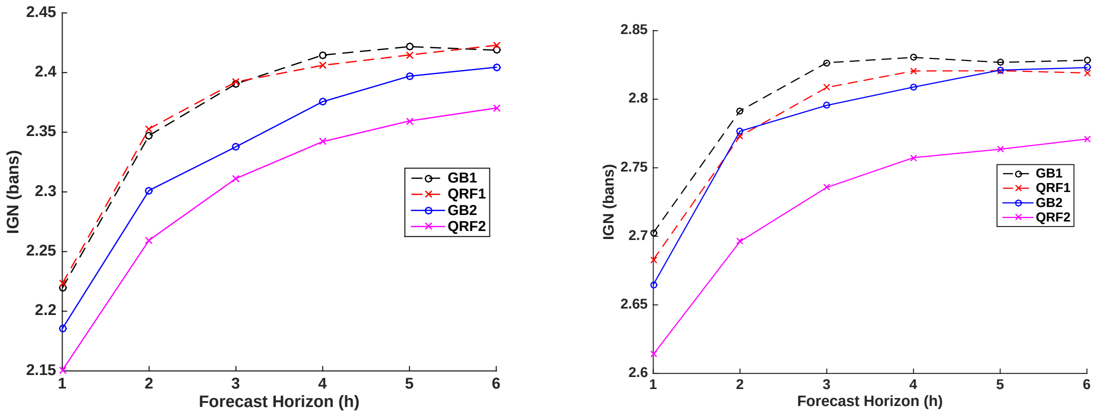
$$QS = \frac{1}{N} \sum_{i=1}^{N} \psi_\tau\big(x^i_{obs} - \hat{q}^i_\tau\big). \tag{10}$$

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

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

5.3.4. Ignorance Score (IGN)

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



(a) (b)

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.

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

$$IGN = -\frac{1}{N}\sum_{i=1}^{N} log\Big(\hat{f}^i(x^i_{obs})\Big). \tag{11}$$

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

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

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

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

$$IGN = log(M + 1) + log\Delta X_k, \tag{12}$$

where

$$\begin{gathered} \Delta X_k = e_{k+1} - e_k \text{ if } 1 < k < M \\ \Delta X_0 = e_1 - e_0 \\ \Delta X_M = e_{M+1} - e_M. \end{gathered} \tag{13}$$

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

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

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

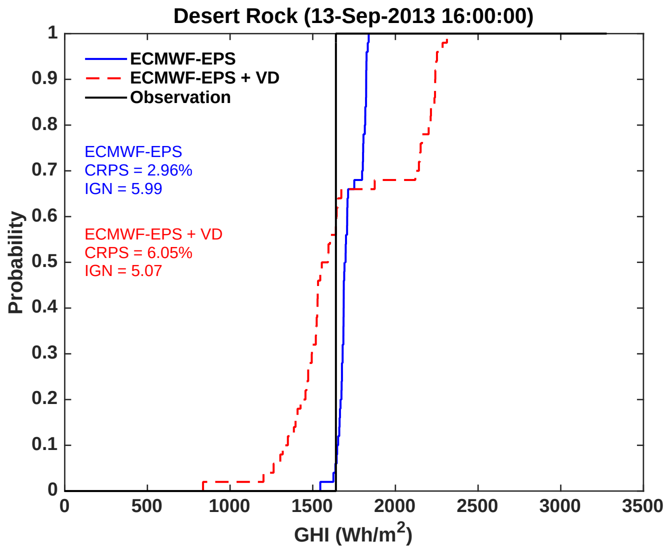


Figure 14: Illustration of the evolution of the CRPS and of the IGN between original and calibrated forecasts: case where these two scores give contradictory information. The CDFs are plotted using the classical definition for ensemble forecasts (see section 2).

6. Conclusions

In this work, we proposed a framework for evaluating solar probabilistic forecasts. Two types of solar probabilistic forecasts namely ensemble forecasts and quantile forecasts were used to illustrate the evaluation framework. This latter is based on visual diagnostic tools and scoring rules originally designed by the weather forecast verification community. For both types of probabilistic forecasts (quantile and ensemble forecasts), we proposed to follow the same approach to assess the quality of the models albeit some diagnostic tools are more appropriate depending on the type of forecast.

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

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

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

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

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

7. Appendices

Appendix A "Uniform" definition of the CDF and PDF derived from an ensemble forecast

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

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

where

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

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

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

Appendix B Hersbach's method to compute the CRPS from ensemble forecasts

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

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

with

$$p_k = \frac{k}{M}. \tag{18}$$

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

Table 7: Determination of α and β

$0 < k < M$	α_k	β_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)	α_k	β_k
$x_{obs} < e_1$	0	$e_1 - x_{obs}$
$x_{obs} > e_M$	$x_{obs} - e_M$	0

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

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

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

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

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

$$\overline{RES} = \overline{UNC} - \overline{CRPS}_{pot}, \tag{22}$$

with

$$\bar{g}_k = \bar{\alpha}_k + \bar{\beta}_k, \tag{23}$$

$$\bar{o}_k = \frac{\bar{\beta}_k}{\bar{\alpha}_k + \bar{\beta}_k}. \tag{24}$$

Appendix C Decomposition of the CRPS through decomposition of the Brier score

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

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

Table 8: Contingency Table for threshold x

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

with

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

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

$$UNC(x) = o(x)[1 - o(x)]. \tag{28}$$

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

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

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

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

with $l_k = n_i + \hat{n}_k$

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

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

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

Appendix D Results of the CRPS decomposition for the intraday models

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

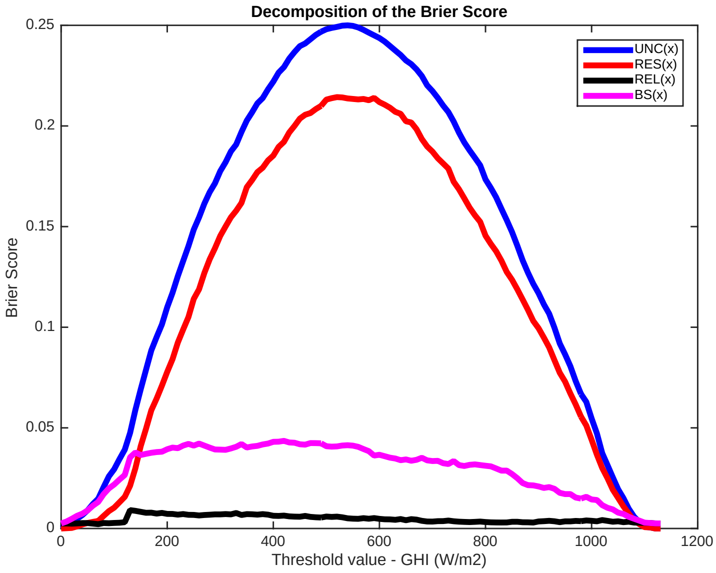


Figure 15: CRPS components through decomposition of the Brier Score (BS) - The area under each curve corresponds to the related CRPS component. Integration of $BS(x)$ for all threshold values x gives the CRPS

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

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

References

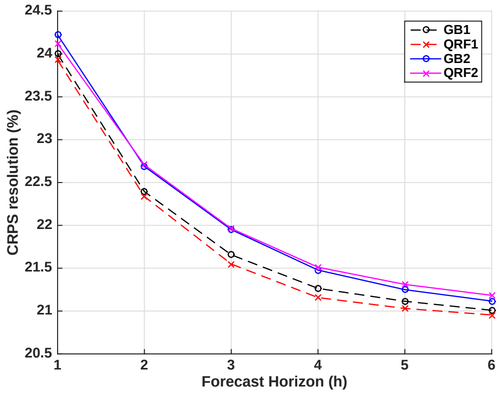
Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term probabilistic solar power forecast. Applied Energy 157, 95–110.

Anderson, J.L., 1996. A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. Journal of Climate 9, 1518–1530.

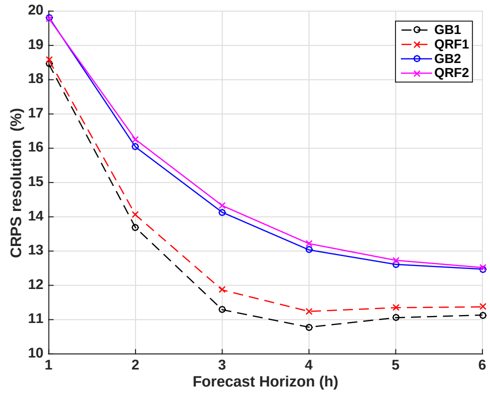
Ben Bouallègue, Z., 2015. Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts. MAUSAN , 541–550.

Bentzien, S., Friederichs, P., 2014. Decomposition and graphical portrayal of the quantile score. Quart. J. Roy. Meteor. Soc. 140, 1924–1934.

Bröcker, J., 2012. Evaluating raw ensembles with the continuous ranked probability score. Quarterly Journal of the Royal Meteorological Society 138, 1611–1617.

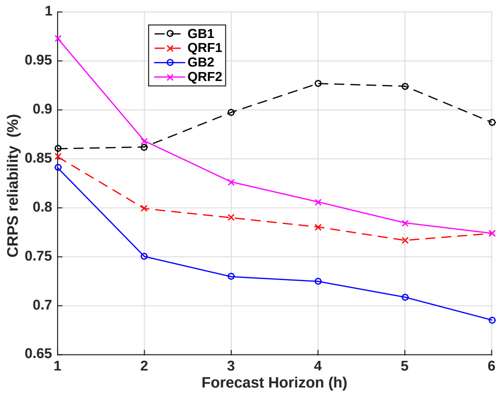


(a)

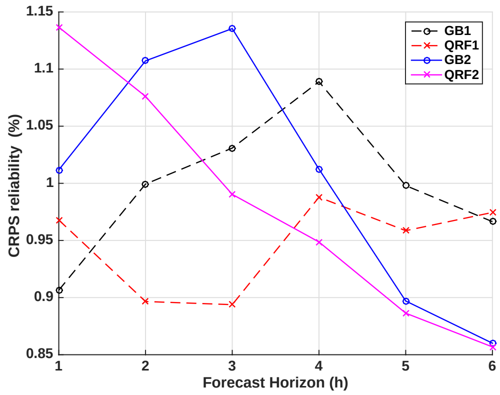


(b)

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.

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

Khosravi, A., Nahavandi, S., Creighton, D., 2013. Prediction Intervals for Short-Term Wind Farm Power Generation Forecasts. IEEE Transactions on Sustainable Energy 4, 602–610.

Koenker, R., Bassett, G., 1978. Regression Quantiles. Econometrica 46, 33–50.

Lauret, P., David, M., Pedro, H., 2017. Probabilistic Solar Forecasting Using Quantile Regression Models. Energies 10, 1591.

Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P., 2015. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Solar Energy 112, 446–457.

Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. Journal of Computational Physics 227, 3515–3539.

Li, K., Wang, R., Lei, H., Zhang, T., Liu, Y., Zheng, X., 2018. Interval prediction of solar power using an Improved Bootstrap method. Solar Energy 159, 97–112.

Lorenz, E., Heinemann, D., 2012. Prediction of solar irradiance and photovoltaic power., in: Comprehensive Renewable Energy. Elsevier, Oxford, UK, pp. 239–292.

Marquez, R., Coimbra, C.F., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Solar Energy 85, 746–756.

van der Meer, D., Widn, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renewable and Sustainable Energy Reviews 81, 1484 – 1512.

Morales, J.M., Conejo, A.J., Madsen, H., Pinson, P., Zugno, M., 2014. Integrating Renewables in Electricity Markets. volume 205 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA.

Murphy, A.H., 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting 8, 281–293.

NCAR-Research applications laboratory, 2015. verification: Weather Forecast Verification Utilities. R package version 1.42.

Pedro, H.T., Coimbra, C.F., 2015. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. Renewable Energy 80, 770–782.

Pedro, H.T., Coimbra, C.F., David, M., Lauret, P., 2018. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renewable Energy 123, 191–203.

Pinson, P., McSharry, P., Madsen, H., 2010. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. Quarterly Journal of the Royal Meteorological Society 136, 77–90.

Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G.N., 2007. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 10, 497–516.

Pinson, P., Reikard, G., Bidlot, J.R., 2012. Probabilistic forecasting of the wave energy flux. Applied Energy 93, 364–370.

Pinson, P., Tastu, J., 2014. Discussion of “Prediction Intervals for Short-Term Wind Farm Generation Forecasts” and “Combined Nonparametric Prediction Intervals for Wind Power Generation”. IEEE Transactions on Sustainable Energy 5, 1019–1020.

Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy 83, 342–349.

Roulston, M., Smith, L., 2002. Evaluating Probabilistic Forecasts Using Information Theory. Monthly Weather Review 130, 1653–1660.

Scolari, E., Sossan, F., Paolone, M., 2016. Irradiance prediction intervals for PV stochastic generation in microgrid applications. Solar Energy 139, 116–129.

Sperati, S., Alessandrini, S., Delle Monache, L., 2016. An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting. Solar Energy 133, 437–450.

Tödter, J., Ahrens, B., 2012. Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. Monthly Weather Review 140, 2005–2017.

Verbois, H., Rusydi, A., Thiery, A., 2018. Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting. Solar Energy 173, 313–327.

Voyant, C., Motte, F., Fouilloy, A., Notton, G., Paoli, C., Nivet, M.L., 2017. Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies. Energy 120, 199–208.

Weijs, S.V., van Nooijen, R., van de Giesen, N., 2010. Kullback–leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. Monthly Weather Review 138, 3387–3399.

Wilks, D.S., 2014. Statistical Methods in the Atmospheric Sciences. An Introduction. Elsevier Science, Burlington.

Winkler, R.L., 1972. A Decision-Theoretic Approach to Interval Estimation. Journal of the American Statistical Association 67, 187–191.

Yang, D., 2019. A universal benchmarking method for probabilistic solar irradiance forecasting. Solar Energy 184, 410–416.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T., Coimbra, C.F., 2018. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar Energy 168, 60–101.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Solar Energy 105, 804–816.