Added-value of Ensemble Prediction System on the quality of solar irradiance probabilistic forecasts

Josselin Le Gal La Salle[a], Jordi Badosa[b], Mathieu David[a], Pierre Pinson[c], Philippe Lauret[a]

[a] *Université de la Réunion - Laboratoire de Physique et ingénierie mathématique pour l'énergie, l'environnement et le bâtiment (PIMENT), 15 avenue René Cassin, 97715, Saint-Denis Cedex 9, La Réunion, France*

[b] *LMD-Laboratoire de météorologie dynamique, Palaiseau, France*

[c] *DTU - Technical University of Denmark*

Abstract

Accurate solar forecasts is one of the most effective solution to enhance grid operations. As the solar resource is intrinsically uncertain, a growing interest for solar probabilistic forecasts is observed in the solar research community. In this work, we compare two approaches for the generation of day-ahead solar irradiance probabilistic forecasts. The first class of models termed as deterministic-based models generates probabilistic forecasts from a deterministic value of the irradiance predicted by a Numerical Weather Prediction (NWP) model. The second type of models denoted by ensemble-based models issues probabilistic forecasts through the calibration of an Ensemble Prediction System (EPS) or from information (such as mean and variance) derived from the ensemble. The verification of the probabilistic forecasts is made using a sound framework. The Continuous Ranked Probability Score is a numerical score used to assess the overall performance of the different models. The decomposition of the CRPS into reliability and resolution provides a further detailed insight into the quality of the probabilistic forecasts. In addition, a new diagnostic tool which evaluates the contribution of the statistical moments of the forecast distributions to the CRPS is proposed. This tool denoted by MC-CRPS allows identifying the characteristics of an ensemble that have an impact on the quality of the probabilistic forecasts. The assessment of the different models is done on several sites experiencing very different climatic conditions. The gain in forecast quality measured by the CRPS ranges from 4% to 16% depending on the site. Results show a general superior performance of ensemble-based models but this statement needs to be tempered for sites that experience highly variable sky conditions.

Keywords: `Day-ahead solar irradiance probabilistic forecast`, `Ensemble prediction system`, `Non parametric methods`, `Ensemble calibration`, `CRPS`

Contents

[⋆] Fully documented templates are available in the elsarticle package on CTAN.

1. Introduction

Operations of electrical power systems are becoming more challenging as the share of solar energy increases. In particular, due to the intrinsic variability of the solar resource, high penetration of solar power generation into the electrical grid may put in danger the grid supply-demand balance. Energy storage systems (EES) are one of the means used to ensure the grid stability. ~~However, unlike EES,~~ Accurate PV power forecasting is a cost-effective way to dimension and operate ESS optimally. Consequently, PV power forecasts facilitate the large-scale integration of solar energy into the grid. In addition, for energy trading, accurate PV power forecasts are also required because penalties in proportion with the forecast errors are applied.

In this study, however, we focus on the global horizontal solar irradiance (GHI) forecasts instead of PV power forecasts. The present work constitutes thus a first step in assessing the contribution of the proposed methodologies for improving the quality of the PV power forecasts and of their potential gain for improved grid operations. ~~since solar power generation is highly correlated to the GHI.~~ Day-ahead GHI forecasts are treated here as they have been considered essential to secure the power grid [1]. Moreover, we propose to work on probabilistic forecasting in order to estimate the uncertainty associated to day-ahead GHI forecasts. This additional knowledge permits for instance grid operators to improve their decisions regarding the grid operations. The interested reader can refer to [2] or [3] to understand the benefits of a probabilistic forecast against a deterministic one.

Day-ahead GHI forecasts are classically generated by Numerical Weather Predictions models (NWPs). For instance, The Integrated Forecasting System (IFS) model of the European Centre of Medium-Range Weather Forecasts (ECMWF) provides day-ahead GHI forecasts [4]. The forecasts can take either the form of a deterministic forecast or an ensemble forecast denoted by the term Ensemble Prediction System (EPS). EPS consists in a set of several perturbed forecasts of irradiance, each representing a possible future state of the atmosphere. If an EPS gives an important information about the uncertainty associated to a forecast, it requires a high computational cost. Thus, the added value of EPS for probabilistic forecasting needs to be determined to justify their computation.

We propose below to conduct a bibliographic survey related to day-ahead solar forecasts with a special emphasis on the use of NWP outputs to generate probabilistic forecasts. One of the first approach used to generate day-ahead probabilistic irradiance forecasts was proposed by Lorenz et al. [5]. In this work, a Gaussian distribution of the error of the ECMWF-IFS deterministic irradiance forecast was used to generate prediction intervals. Alessandrini et al. [6] developed an analog statistical method approach applied to a set of explanatory weather variables (GHI, cloud cover, air temperature, etc.) provided by the NWP Regional Atmospheric Modeling System (RAMS) to generate probabilistic PV power forecasts for three solar farms located in Italy. Zamo et al. [7] proposed two statistical approaches to generating probabilistic forecasts of daily PV production from information provided by Météo France's EPS, PEARP. The first approach makes use of the PEARP control member as unique input to quantile regression methods while the second one averages the set of quantiles calculated from each of the 35 members of the PEARP ensemble. Bacher et al. [8] used a weighted quantile regression (WQR) technique to compute up to 24h ahead probabilistic PV forecasts. In addition to lagged PV measurements, the WQR model used also a NWP-based GHI deterministic forecast. Lauret et al. [9] used the IFS model to produce quantile forecasts of solar irradiance and Iversen et al. [10] introduces the idea of modeling uncertainty by stochastic differential equations from a NWP-based deterministic forecast provided by the Danish Meteorological Institute. Bakker et al. [11] proposed a comparison of seven statistical regression models to issue GHI probabilistic forecasts from the deterministic numerical weather prediction (NWP) model HARMONIE-AROME (HA) and the atmospheric composition model CAMS.

It must be noted that the above cited works make use of deterministic information extracted from NWP models to generate probabilistic forecasts with the help of statistical techniques like quantile regression or analog ensemble. Others authors like Sperati et al. [12] proceeded differently. In their work, Sperati et al. [12] generated up to 72h probabilistic forecasts from the raw EPS provided by the ECMWF. In this study, two post-processing methods (also called calibration techniques) applied to the initial raw ensemble were used to further improve the quality of the probabilistic forecasts. Massidda and Marrocu [13] went a little bit further and proposed a methodology to combine ECMWF ensemble and the high-resolution IFS deterministic forecast.

If we extend our bibliographic survey to the probabilistic predictions of other weather variables such as wind, temperature or precipitation, more publications can be found on how to use information from NWP models to generate probabilistic forecasts. For example, Pinson [14] and Pinson and Madsen [15] suggested a framework for the calibration of wind ensemble forecasts. Junk et al. [16] proposed an original calibration model for wind-speed forecasting applied to ECMWF-EPS based on the combination between Nonhomogeneous Gaussian Regression and Analog Ensemble Models. Likewise, Hamill and Whitaker [17] suggested an adaptation of the analog ensemble technique for the calibration of ensemble precipitation forecast, using the statistical moments of the distribution such as mean and spread of the members as predictors.

Wilks [18], followed in his methodology by Williams et al. [19], compared several post-processing techniques of weather EPS forecasts, such as ensemble dressing, Logistic Regression, Nonhomogeneous Gaussian Regression (NGR) and Rank-Histogram recalibration. The reader can refer to [20] and [21] ~~[22] and [23]~~ for more details regarding the parametric calibration of ensemble forecasts with techniques like NGR with a special emphasis on the choice of the type of the parametric distribution used by the regression technique. Finally, the interested reader should consult the following reference book : [24], who proposed a summary of the common probabilistic forecasting ensemble-based models with their respective pros and cons.

Based on this bibliographic survey, two different approaches for day-ahead GHI probabilistic forecasting with the help of NWP models can be identified, which we denoted here by approaches 1 and 2 :

1. Approach 1 referred herein as *deterministic-based models* : the probabilistic forecast is computed from deterministic NWP predictors with the help of statistical methods. Linear Quantile Regression and Analog Ensemble techniques are particularly attractive to implement this methodology.
2. Approach 2 referred herein as *ensemble-based models* : the estimation of the forecast is made through the calibration of an EPS or from information (for example mean or spread) inferred from the ensemble. For instance, calibration techniques like Nonhomogeneous Regression can be used to improve the raw ensemble EPS. Also, methods based on Linear Quantile Regression and Analog methods can be used to produce probabilistic forecasts from the mean and spread of the ensemble.

It must be stressed however that, to the best of our knowledge, no previous works have been dedicated to the comparison of the two approaches and particularly in the realm of solar probabilistic forecasts. In this work, our main goal is therefore to assess the relative merits of each approach for day-ahead GHI probabilistic forecasts. Besides, we would like to highlight the possible added-value brought by EPS for probabilistic forecasting. Indeed, it is well known that the generation of such ensemble necessitates high computing capacities compared to a single deterministic forecast that is fed into a statistical method to produce the probabilistic forecasts. More precisely, it should be noted that the calculation cost is not the same to produce only the control member of EPS or the whole set of members.

To understand the benefits associated with the usage of EPS, we propose in this paper a a sound and consistent methodology to evaluate the respective contribution of each approach. First, the quality appraisal of the different models will be made according the verification framework proposed by Lauret et al. [25]. This framework (which is not consistently proposed in the literature) is based on visual diagnostic tools and numerical scores like the Continuous ranked Probability Score (CRPS) which permits to objectively rank the competing forecasting methods. However, this classical verification framework is not sufficient to completely explain the contribution of the statistical moments of the forecast distributions to the forecast quality. That is why we propose in a second step a new tool that evaluates the accuracy of all moments of the forecast distribution and its contribution to the CRPS score. We hope that this new diagnostic tool will provide a more in-depth understanding of the performance of each approach. To this end, we evaluate models that generate day-ahead GHI probabilistic forecasts on 3 sites that experience different sky conditions. The probabilistic models are built :

1. With only the control member of the EPS as a deterministic predictor (deterministic-based approach),
2. With a deterministic predictor inferred from the whole set of EPS's members The first statistical moment (mean of the members) can be such a deterministic predictor (ensemble-based approach),

3. With several predictors inferred from the ensemble like the mean and the variance of the ensemble (ensemble-based approach).

We propose the following structure for the paper. Section 2 introduces the different forecasting models while section 3 briefly presents the diagnostic tools used for the verification of probabilistic forecasts. Section 4 presents the case studies and details the data used to evaluate the different probabilistic models. Section 5 provides a detailed assessment of the performance of the different methods. Finally, a discussion will be conducted in section 6, trying to understand the pros and cons of each forecasting methods and the factors impacting the forecast quality.

2. Building probabilistic forecasts

Regarding probabilistic forecasts of continuous predictand like GHI, a probability statement i.e. either a Probability Distribution Function (PDF) f or a Cumulative Distribution Function (CDF) F encodes the uncertainty of the forecast. In this work, three ways to estimate this CDF or PDF are considered: parametric PDFs, discrete quantile estimates of a CDF via a non-parametric method and CDF derived from EPS.

Regarding this last case, EPS can be seen as discrete estimates of a CDF when they are sorted in ascending order. Lauret et al. [25] discussed three ways to associate these sorted members to cumulative probabilities. In this work, we chose the uniform distribution which consists in a~~n~~ uniform spacing of the members and a linear interpolation between the members. More precisely, this choice assigns a probability mass of $1/(M+1)$ between two members and for events that fall outside of the ensemble. Using this definition, the i^{th} ensemble member can be interpreted as a quantile forecast with a probability level equal to $\tau = \frac{i}{M+1}$. Put differently, the ensemble forecasts are in the form of 51 equally spaced quantiles with probability levels $\tau = \frac{1}{52}, \frac{2}{52}, \cdots, \frac{51}{52}$. This construction is illustrated in Figure 1, for an EPS with 4 members. In the following, we present first the different statistical techniques used to estimate the uncertainty of the forecasts. Secondly, we detail the two approaches introduced in section 1.

2.1. Statistical techniques used to generate probabilistic forecasts

2.1.1. The linear quantile regession (LQR) technique

This method estimates the quantiles of the cumulative distribution function F of some response variable Y (also called predictand) by assuming a linear relationship between the quantiles of Y, namely (q_τ(and a set of explanatory variables X (called predictors):

$$q_\tau = \beta_\tau \; X + \epsilon, \tag{1}$$

where β_τ is a vector of parameters to optimize for each probability level τ and ϵ represents a random error term.

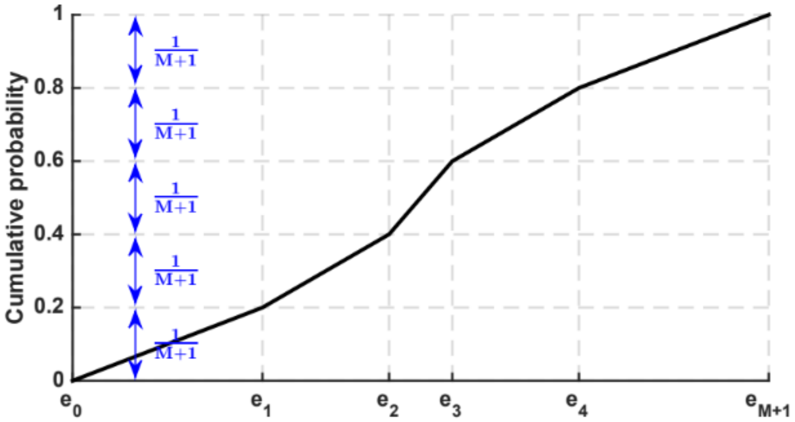


Figure 1: Illustration of an uniform construction of a CDF from an ensemble of $M = 4$ members. The tails of the CDF are bounded by e_0 and e_{M+1} which correspond to the minimum and the maximum of the climatology.

Following Koenker [26], the ~~quantiles $q_\tau = F^{-1}(\tau)$ can be estimated as the solution of the optimization problem :~~

$$\hat{q}_\tau = \arg\min_{\beta} \sum_{i=1}^{N} \Psi_\tau \left(Y_i - q_\tau\right), \tag{2}$$

vector $\hat{\boldsymbol{\beta}}_\tau$ that defines each quantile is obtained as the solution of the following minimization problem:

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{N} \Psi_\tau \left(Y_i - \beta X_i\right). \tag{3}$$

where N is the number of pairs of observed predictand Y_i, set of predictors X_i taken from the training set. $\Psi_\tau(u)$ is the quantile loss function defined as :

$$\Psi_\tau(u) = \begin{cases} u\tau & \text{if } u \geq 0, \\ u(\tau - 1) & \text{if } u < 0, \end{cases} \tag{4}$$

with τ representing the quantile probability level. Hence, in quantile regression, the quantiles are estimated by applying asymmetric weights to the mean absolute error.
Thus, the quantity $\hat{q}_\tau = \hat{\beta}_\tau X$ is the estimation of the τ^{th} quantile obtained by the LQR method. ~~The pairs of observed predictand and the set of predictors, Y_i and X_i, for the estimation of the $\hat{\beta}_\tau$ parameters are taken from the training set.~~

It must be noted that the quantile regression method estimates each quantile separately (i.e. the minimization of the quantile loss function is made for each τ separately). As a consequence, one can obtain quantile regression curves that may intersect, i.e $\hat{q}_{\tau 1} > \hat{q}_{\tau 2}$ when $\tau_1 < \tau_2$. To avoid this issue during the model fitting, we used the rearrangement method described by Chernozhukov et al. [27].

Figure 2 ~~plots an example of the observed irradiance versus day-ahead predicted irradiance and shows that uncertainty depends on the level of the forecasted irradiance. The lines correspond to~~ shows some quantiles estimates of the CDF of ~~day-ahead~~ the predictand Y (here GHI) as a function of the day-ahead forecasted GHI. Hence, in this case, the preditor X is the predicted irradiance ~~and the response variable Y whose quantiles have to be estimated is the day-ahead GHI.~~ which will be represented in this work either by the ECMWF control member or the mean of the ensemble (see Table 2 below). This example shows that the forecast uncertainty depends on the level of the predicted irradiance. More precisely, and as shown by Figure 2, the dispersion of points is lower for values of predicted irradiance close to 0 W/m^2 and greater for values between 40 and 100 W/m^2.

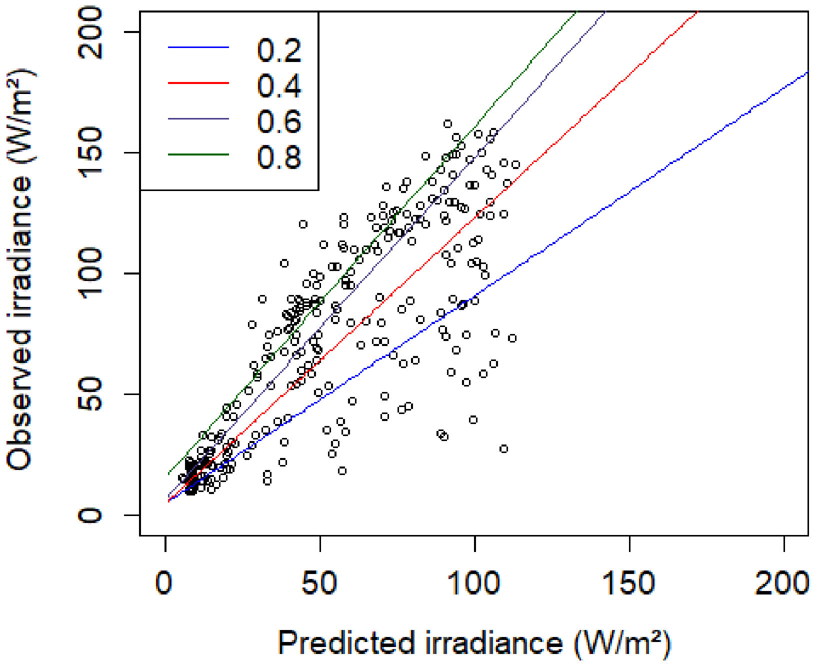


Figure 2: Observed GHI vs. the predicted day-ahead GHI. The lines are the estimates of the quantiles with probability levels of 0.2, 0.4, 0.6 and 0.8. Data are from the training period of Hawaii. Observed and predicted GHI are averaged on the 3-hour window ~~[5h-8h]~~ [17h-20h] local time.

2.1.2. The Analog Ensemble (AnEn) technique

The analog ensemble technique is now quite a standard in the energy meteorology forecasting community ~~[28, 29, 17, 30]~~ [29, 17]. Similarly to the LQR method, the analog technique is a non-parametric method that can be used to estimate the predictive CDF of the predictand.

Considering a training set of N ordered pairs of GHI observations/GHI forecasts $(Y_i, \hat{Y}_i)_{i=1,\cdots,N}$

~~calculated over a training period and the set of the corresponding observations $\{Y_i\}_{i=1,\cdots,N}$ sorted in a similar manner as for the forecasts~~, the procedure for determining the forecast CDF is as follows:

1. For a new forecast taken from a testing set, calculate its distance from every past forecast and find the rank R of the past forecast that is closest to the new forecast.
2. Form an ensemble by selecting the $2\alpha + 1$ past training observations Y_k having their ranks k inside the interval $[R - \alpha, R + \alpha]$.
3. Compute the predictive CDF at a specific value y of the predictand using the following equation:

$$\hat{F}(xy) = P(XY \leq xy) = \frac{1}{2\alpha + 1} \sum_{k=1}^{2\alpha+1} H(xy - Y_k), \tag{5}$$

where Y is the random value related to the predictand (here GHI) and H is the Heaviside or step function. The effectiveness of the method is strongly dependent on the value of α. It is proposed here to take $\alpha = 0.02N$. This choice has been motivated by a preliminary study made on the training period. Appendix D details the selection of the optimal value of α. Finally, as for the linear quantile regresssion, notice that the GHI forecasts used in the *AnEn* technique will be given either by the ECMWF control member or the mean of the ensemble (see Table 2 below).

2.1.3. The Nonhomogeneous truncated Gaussian Regression technique (t_NGR)

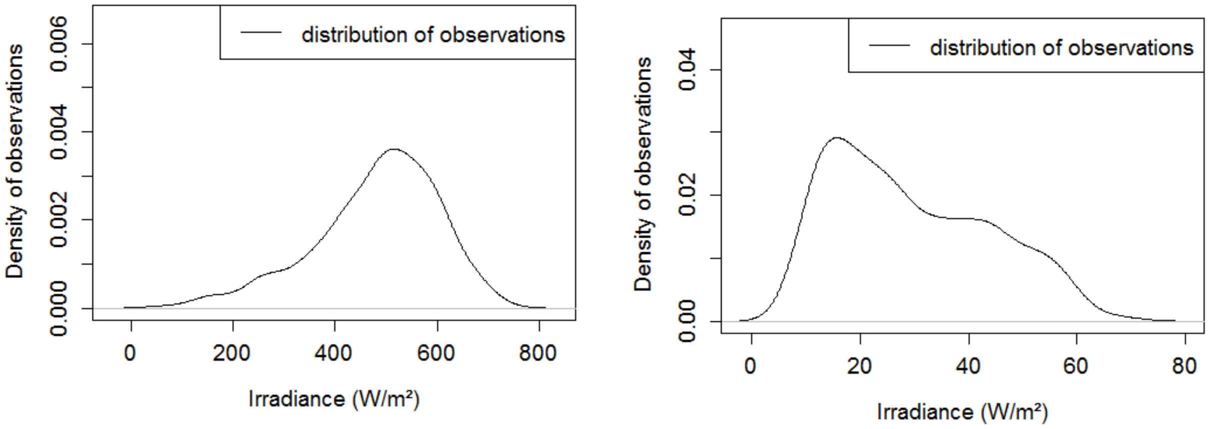
The NGR technique also called in some studies "Ensemble Model Output Statistics" (EMOS) has been introduced by Gneiting et al. [20] for probabilistic forecasting of weather variables. This technique is dedicated to the post-processing of ensemble forecasts produced by an EPS. The NGR technique builds the predictive PDF of the predictand Y from a normal PDF. As such, this kind of model can be termed as a parametric model. The predictive pdf $\hat{f}$ estimated by the NGR method is given by:

$$\hat{f} \sim \mathcal{N}(a + \sum_{k=1}^{M} (b_k m X_k), c + dS^2), \tag{6}$$

where M is the number of members, ~~X_k~~ m_k is the k^{th} member and S^2 is the variance of the ensemble members distribution. The free parameters a, $b_1, \cdots, b_M$, c and d are determined with the help of an optimization procedure. In this work, and following Gneiting et al. [20], these parameters are calculated by minimizing (over a training period) an evaluation metric for probabilistic forecasts called CRPS (see section 3.3 for details regarding CRPS). Furthermore, as GHI is a necessarily positive quantity, we propose, in this work, a variant of the NGR technique namely a truncated version (at 0) of the nonhomogeneous gaussian regression. In the following, the corresponding model is denoted as *t_NGR*.

2.1.4. The Nonhomogeneous Regression of Generalized Extreme Value technique (NR_GEV)

One can question the choice of a Gaussian distribution in the t_NGR technique. Indeed, the distributions of observations for a fixed forecasting level are actually non-Gaussian. Two examples for the studied sites are presented in Figure 3.



(a) 6th decile for Hawaii
levels of forecasting from 484 to 584 W/m^2

(b) 1st decile for Desert Rock
levels of forecasting from 0 to 38 W/m^2

Figure 3: Example of distributions of observations for a fixed forecasting level

On these specific examples, the distributions of observations are clearly non-Gaussian and the consideration of other types of distributions may improve the skills of the forecast.

As pointed out in [21] and [22], other types of parametric distributions can be used to deal with this issue. Here, a Non homogeneous Regression approach with Generalized Extreme Value distributions is proposed to estimate the PDF of the predictand Y. The PDF of a generalized Extreme value distribution for a specific value y of the predictand GHI is defined as :

$$\hat{f}(y) = \begin{cases} \frac{1}{\sigma}\Big[1+\xi(\frac{y-\mu}{\sigma})\Big]^{(-\frac{1}{\xi})-1} exp\Big(-\big[1+\xi(\frac{y-\mu}{\sigma})\big]^{-\frac{1}{\xi}}\Big) & \xi \neq 0, \\ \frac{1}{\sigma}exp(-\frac{y-\mu}{\sigma})exp\Big[-exp(-\frac{y-\mu}{\sigma})\Big] & \xi = 0. \end{cases} \tag{7}$$

The parameters μ, σ and ξ are to be determined by optimizing the CRPS over the training period. We followed the framework of [31] and [32] to set these coefficients. Following this procedure, the mean μ and the scale parameter σ of the final distributions are determined by linear regression, and depends only on variables inferred from the EPS. The mean is a linear combination of the mean of the members and the fraction of members which predict exactly zero. The scale parameter σ depends on the "Gini's mean difference" (a measure of the variability closely related to the spread of the members, see [33] for details). Notice that the shape parameter is taken as a constant. Thus, the optimization of CRPS determine the linear regression coefficients of μ, the linear regression coefficients of σ and the shape parameter ξ. Notice that the two techniques namely t_NGR discussed above and NR_GEV discussed here are part of a family of parametric methods named Nonhomogeneous Regression (NR).

2.2. *Obtaining probabilistic forecasts from deterministic forecasts (Deterministic-based approach)*

~~The~~ Some of the techniques presented in section 2.1, namely the Linear Quantile Regression (LQR) and the Analog Ensemble(AnEn) techniques, are capable of generating a probabilistic forecast from a deterministic predictor.

In our study, and regarding the deterministic-based approach, the control member of ECMWF-EPS is the predictor variable X of the LQR technique and it will be the forecast used in the AnEn procedure. The corresponding probabilistic models are denoted respectively as LQR_c and $AnEn_c$ in the following.

2.3. *Obtaining probabilistic forecasts from ensemble forecasts (Ensemble-based approach)*

2.3.1. *From raw output of EPS*

Given a raw ensemble forecast of M members $\{\cancel{X}m_i\}_{i=1,\cdots,M}$, it seems natural to define directly a forecast CDF from this EPS as illustrated in Figure 1. Notice that this definition corresponds to the "uniform" definition of a CDF derived from an ensemble" discussed in Lauret et al. [25].

2.3.2. *From information extracted from an EPS*

An EPS differs from a deterministic forecast by the multiplicity of predictors. In this work, we propose to assess the quality of two variants of probabilistic models built with information extracted from an EPS.

The first variant will make use of the mean of the ensemble members of the EPS. The ~~usage~~ use of the mean of members as a deterministic predictor is justified by Table 1. For the all the considered sites depicted in Table 3, Table 1 lists, in addition to the Root Mean Square Error (RMSE) of the control member, the RMSEs of ~~two~~ three deterministic predictors extracted from an EPS. ~~compared to the RMSE of the control member of the EPS.~~ As shown by Table 1, the mean of all the members turns out to be the best predictor for deterministic forecasting. Hence, ~~T~~to quantify the improvement brought by the first moment estimation (i.e. the mean), two models denoted by LQR_m and $AnEn_m$ based respectively on the LQR and AnEn techniques will be evaluated.

The second variant will include, in addition to the mean of the members, the spread (i.e. the variance) of the members of the EPS. The t_NGR and the NR_GEV models described in sections 2.1.3 and 2.1.4 use the first and second moment of the EPS distribution to build the predictive distributions. Furthermore, we also propose to use the LQR technique with a vector X of predictors given by

$$X = [\mu, S^2], \tag{8}$$

Site	HAW	DR	SP	PAL	TIR	LAN
Any perturbed member	138	75.3	126.5	97.7	110.7	98.8
Control member	135	72.8	91.9	102.9	100.8	93.2
Mean of the members	129.7	67.9	113.8	81.8	92.6	84.3
Median of the members	133.9	69.4	115.5	84.1	94.7	85.6

Table 1: RMSE of ~~3~~ 4 deterministic forecasts that can be inferred from an EPS: any of the 50 perturbed ~~exchangeable~~ members of ECMWF ensemble forecast, the control member (unperturbed) ~~and~~ , the mean of the members and the median of the members. See Table 3 for the signification of the acronyms of the different sites.

Approach	Deterministic-based		Ensemble-based				
Predictors	Control member		Mean of members		Mean and spread of members		
~~Model~~ Technique	AnEn	LQR	AnEn	LQR	LQR	NR	
Model Abbreviation	$AnEn_c$	LQR_c	$AnEn_m$	LQR_m	LQR_s	t_NGR	NR_GEV

Table 2: Summary of all considered forecasting models with AnEn: Analog Ensemble, LQR: Linear Quantile Regression, NR: Nonhomogeneous Regression

where μ represents the mean of members and S^2 the variance of the ensemble. This method will be referred in this study as LQR_s.

Finally, Table 2 summarizes the different probabilistic models that will be evaluated in this study.

3. Verification of the probabilistic forecasts

In this section, we detail some of the verification tools proposed by Lauret et al. [25] that will be applied to assess the quality of GHI probabilistic forecasts. Following this work, ~~we will use first rank histograms to evaluate visually the reliability of the proposed forecasting techniques. Second,~~ we will rely on a quantitative score namely the continuous ranked probability score (CRPS) and its related skill score (CRPSS) to rank objectively the different methods. Moreover, and based on the recommendations of [25], we will provide the decomposition of the CRPS into the main attributes that affect the quality of the forecasts. In addition to this decomposition, it is worth noting that we will propose in this work a new way to have detailed insight into the performance of the methods. This new methodology is based on the contribution of the moments (mean, variance, etc.) of the forecast distribution to the CRPS (see section 3.4 below).

3.1. Attributes for a skillful probabilistic model

We recall here briefly the two main attributes that characterize the quality of the probabilistic models namely reliability and resolution [34, 35]. Reliability or calibration evaluates the statistical consistency between the forecasts and the observations. In the case of a continuous variable like GHI, a high reliability is obtained if predictive distributions and distributions of observations agree. Resolution refers to the ability of the probabilistic model to discriminate among different forecast situations. More precisely, the more distinct the observed frequency distributions for various forecast situations are from the full climatological distribution, the more resolution the forecast model has. A high quality probabilistic model should issue reliable forecasts with high resolution. In other words, high reliability is a necessary but not a sufficient condition for a high quality probabilistic forecast. The forecast should also exhibit high resolution. For instance, climatological forecasts are perfectly reliable but exhibit no resolution.

3.2. ~~Rank histogram for reliability assessment~~

~~Following the recommendations of Lauret et al. [25], rank histogram (RH) has been chosen for the visual assessment of reliability of the different probabilistic models. Initially designed for assessing reliability of EPS forecasts [34], rank histograms can be used here since the quantile forecasts produced by the proposed statistical methods are evenly spaced.~~[1] ~~Over an evaluation set, RH plots the histogram of the ranks of the observations when pooled within the ordered set of forecasts. Theoretically, the statistical consistency between forecasts and observations is met if the histogram of the ranks is uniform with relative frequency of $\frac{1}{M+1}$ (in our case, we recall that $M = 51$). Put differently, a flat or uniform RH is an indication for statistical consistency i.e. the forecasts are statistically indistinguishable from the observations. However, as suggested by Lauret et al. [25], we plot the RH with consistency bars in order to deal with the issues of the finiteness of the data and possible presence of serial correlation in the sequence of observation/forecast pairs. Indeed, these issues can cause deviations from the ideal flat line even for reliable forecasts [36]. In other words, a forecast can be stated as reliable if the histogram of the ranks remains inside the consistency bars. In the case where statistical consistency is not verified, the different possible other interpretations of a RH are given below. A U-shape RH corresponds to an over-confident probabilistic model (i.e. under-dispersion of the set of forecasts) meaning that the observation is often an outlier in the distribution of forecasts. Conversely, a RH with hump shape means an under-confident model (i.e. distribution of forecasts consistently too large). It indicates that the observation may too often be in the middle of the set of forecasts. Also, asymmetric (or triangle shape) RHs is an indication of unconditional forecast biases Furthermore, overpopulation of the smallest (resp. highest) ranks will correspond to an overforecasting (resp. underforecasting) bias.~~

3.3. CRPS

In the verification framework proposed by Lauret et al. [25], the authors recommend the computation of a score like the Continuous Ranked Probability Score (CRPS) to evaluate the overall quality of the probabilistic models. We recall here the definition of the CRPS.

[1] For quantile forecasts, reliability diagrams constitute the natural visual diagnostic tool to evaluate the calibration property - See [25] for details.

3.3.1. Definition

The CRPS measures the difference between the predicted and observed cumulative distributions functions (CDF) [37]. The CRPS reads as

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\left[\hat{F}^{i}_{fcst}(xy) - F^{i}_{xy_{obs}}(xy)\right]^{2} dxy, \tag{9}$$

where $\hat{F}_{fcst}(xy)$ is the predictive CDF of the predictand xY (here GHI) and $F_{xy_{obs}}(xy)$ is a cumulative-probability step function that jumps from 0 to 1 at the point where the value of the predictand xy equals the observation xy_{obs} (i.e. $F_{xy_{obs}}(xy) = 1_{\{xy \geq xy_{obs}\}}$). 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 [34]. In other words, the CRPS penalizes lack of resolution of the predictive distributions as well as biased forecasts. Notice that the CRPS is negatively oriented (smaller values are better) and it has the same dimension as the forecasted variable. CRPS is a proper score meaning that it obtains the best expected value when the forecast distribution is equal to the true distribution of probability of the observation. Besides, using proper scoring rules allows the decomposition of the score into the two important attributes of the quality of a forecasting probabilistic model namely resolution and reliability. This permits to understand more precisely the characteristics of the quality of the forecast.

3.3.2. CRPS Skill Score

In a similar manner that scores are used to assess the forecast skill of deterministic forecasts [38], [39] used the CRPS Skill Score (CRPSS) to gauge the quality of their probabilistic forecasting models against a reference method. The CRPSS metric (in %) reads as

$$\mathrm{CRPSS} = 100 \times \left(1 - \frac{\mathrm{CRPS_m}}{\mathrm{CRPS_r}}\right), \tag{10}$$

where $\mathrm{CRPS_r}$ denotes the CRPS of the reference method and $\mathrm{CRPS_m}$ refers to the model under evaluation (see Table 2). A negative value of CRPSS indicates that the probabilistic method fails to outperform the reference model, while a positive value of CRPSS means that the forecasting method improves on the reference model. Further, the higher the CRPSS, the better the improvement. In this work, and following the recommendations of [40], the raw output of the ECMWF-EPS constitutes the reference benchmark model.

3.3.3. Decomposition of the CRPS

The decomposition of the CRPS is given by :

$$SCRPS = REL - RES + UNC, \tag{11}$$

where REL, RES and UNC are respectively the reliability part, the resolution part and the uncertainty part of the CRPS. The interested reader is referred to [25] for details regarding the computation of the different components of the CRPS.

In addition to reliability and resolution, the uncertainty term accounts for the variability of the observations. It is an indication of the difficulty to forecast the variable of interest and cannot be modified by the forecasting model. It is also worth noting that the uncertainty part UNC corresponds to the score of the climatology. For scores like CRPS that are negatively oriented, the goal of a forecasting model is to minimize (resp. maximize) as much as possible the reliability term (resp. the resolution term). In fact, a forecasting model with a high resolution term means that the model has captured the maximum of the variability present in the data (which variability is measured by the uncertainty term).

3.4. Contributions of the statistical moments of the forecast distribution to the CRPS

In this study, a new methodology for a better understanding of the skills of a probabilistic forecast in relation with the CRPS score is developed. The main idea is to assess separately the contribution of the statistical moments (mean, variance, etc.) of the predictive distributions to the CRPS and consequently to the quality of a probabilistic forecasting model. The principle of the method is to create two virtual forecasts which show the contribution of the statistical moments of the actual forecast to the CRPS. Let us illustrate the methodology with 3 forecast PDFs ~~see also(~~depicted in Figure 4 ~~)~~. f represents the actual forecast PDF and f_{m1} and f_{m2} the associated virtual PDF forecasts.

The first virtual forecast f_{m1} is derived from the first moment (mean) of the actual forecast f. Let m_1 be the first moment of f and δ the Dirac distribution (corresponding to the dotted vertical in Figure 4), the PDF of f_{m1} is thereby defined by:

$$f_{m1}(\cancel{x}y) \equiv \delta(\cancel{x}y - m_1). \tag{12}$$

Notice that this definition implies that the second, third and further moments of f_{m1} are equal to 0.

The second virtual forecast f_{m2} is given by a Gaussian distribution with first and second moments equal to those of f. Let m_2 be the second moment of f, f_{m2} is defined as:

$$f_{m2} \sim \mathcal{N}(m1, m2). \tag{13}$$

Being a Gaussian distribution, the third, fourth and further moments of f_{m2} are equal to 0.

The contribution of the statistical moments of the distribution to the CRPS is computed as follows. First, the CRPS of each forecast namely $CRPS_f$, $CRPS_{fm1}$ and $CRPS_{fm2}$ are averaged over the N forecast/observation pairs. This leads to the corresponding values $CRPS$, $CRPS_{m1}$ and $CRPS_{m2}$. Second, the difference $G_2 = CRPS_{m1} - CRPS_{m2}$ and $G_+ = CRPS_{m2} - CRPS$ are calculated. Note that one can therefore rewrite the CRPS as:

$$CRPS = CRPS_{m1} - G_2 - G_+. \tag{14}$$

Notice that the $CRPS_{m1}$ of the determistic foreacst f_{m1} is actually its Mean Absolute Error (MAE) (see [37] for details).

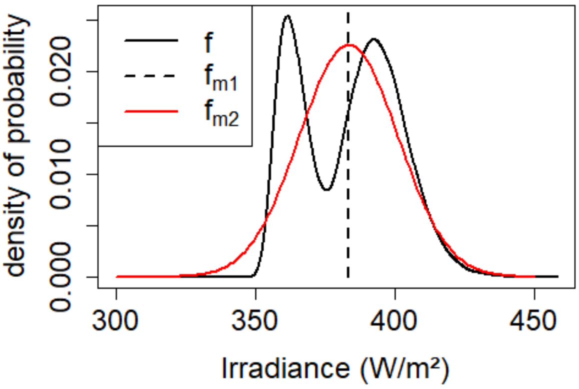


Figure 4: Illustration of the virtual forecasts f_{m1} and f_{m2} related to the forecast PDF f

G_2 is the measure of the gain in CRPS or equivalently in forecast quality that results from the additional information brought by the second moment of the distribution. G_+ represents the gain resulting from the other statistical moments. G_2 is assumed to be positive. If it is found negative, then the probabilistic forecast has no added value compared to a deterministic forecast, as the CRPS of the probabilistic forecast would be higher than the CRPS of the deterministic one ($CRPS_{m1}$), thus denoting a loss of quality of the probabilistic forecast. In the other hand, G_+ is generally positive. It can be null or negative if the forecast distribution ~~is less adapted~~ obtains a higher CRPS score than a Gaussian distribution defined by $\mathcal{N}(m1, m2)$. This would indicate that the forecast distribution is less suitable than a Gaussian distribution .

In section 5.5 below, we propose to present this diagnostic tool under the form of a bar-plot, where $CRPS$, G_+ and G_2 are stacked in this order. G_2 is denoted by the pink part of the bar, G_+ by the green part and $CRPS$ by the blue part. Notice that a black line on the top of the blue part is used to better highlight the value of the CRPS and a dotted black line indicates $CRPS_{m1}$. In the following, we refer to this diagnostic tool based on the contribution of the moments of the forecast distributions to the CRPS as “MC-CRPS”.

4. ~~Data~~ Case studies

~~Three~~ Six sites are chosen to test the selected models. The first one, Desert Rock, which is part of the SURFRAD network, is located in an arid area. It experiences a high occurrence of clear skies and consequently a very low variability. The two other sites, the airport of Hawaii, where the NREL set up a radiometric network, and Saint-Pierre, which is located on the coastal part of the island La Réunion, are insular sites. Both present a high yearly solar irradiation but also an important variability due to frequent partly cloudy skies. These differences between the two types of sites will permit testing the models under different sky conditions. For an extensive study on the multiple factors that impact the climatology and sky conditions in the specific case of Saint-Pierre and La Réunion, see Badosa et al. [41] or Kalecinski [42]. As the aforementioned sites exhibit a similar level of irradiation, three other BSRN sites namely Palaiseau, Tiruvallur and Langley complete this table. As seen, the chosen sites experience different levels of annual solar irradiation. This adding is also an attempt to have a list of sites representative of the various climates around the world. The main characteristics of these ~~three~~ six sites are given in Table 3. The solar variability, presented in the last line of Table 3, is defined as the standard deviation of the changes in the clear sky index [43].

4.1. Measurements

The measured data used in this work are global horizontal irradiance (GHI) time series recorded at the ~~three~~ six considered sites. These datasets have been prepared for previous works related to the development and the benchmarking of probabilistic solar forecasts [44, 45]. They correspond to two years of data divided in a training set (the first year) and test set (the second year). As the ensemble forecasts used here are provided with a 3-hour time step, the recorded time series, initially formatted with a 1-hour granularity, were averaged with a 3-hour time step. A quality check and several test were performed on the recorded GHI time series. The results are given in Appendix A.

	Desert Rock (USA)	**Hawaii (USA)**	**Saint-Pierre (Reunion)**
Acronym	DR	HAW	SP
Provider	SURFRAD	NREL	PIMENT
Position	36.6N, 119.0W	21.3N, 158.1W	21.3S, 55.5E
Elevation (m)	1007	11	75
Climate type	Desert	Insular tropic	Insular tropic
Years of record	2012 - 2013	2010-2011	2012 - 2013
Annual solar irradiation (MWh/m^2)	2.105	1.969	2.053
Solar variability 1-h $(\sigma\Delta kt^*_{1hour})$	0.146	0.209	0.241
	Palaiseau (France)	**Tiruvallur (India)**	**Langley (USA)**
Acronym	PAL	TIR	LAN
Provider	BSRN	BSRN	BSRN
Position	48.7N, 2.2E	13.1N, 80.0E	37.1N, 76.4W.
Elevation (m)	156	36	3
Climate type	Mild oceanic	Monsoon	Humid
Years of record	2016-2017	2018-2019	2015-2016
Annual solar irradiation (MWh/m^2)	1.172	1.835	1.685
Solar variability 1-h $(\sigma\Delta kt^*_{1hour})$	0.281	0.190	0.186

Table 3: Main characteristics of time series of recorded global horizontal irradiance (GHI) used to test the models

4.2. Forecasts

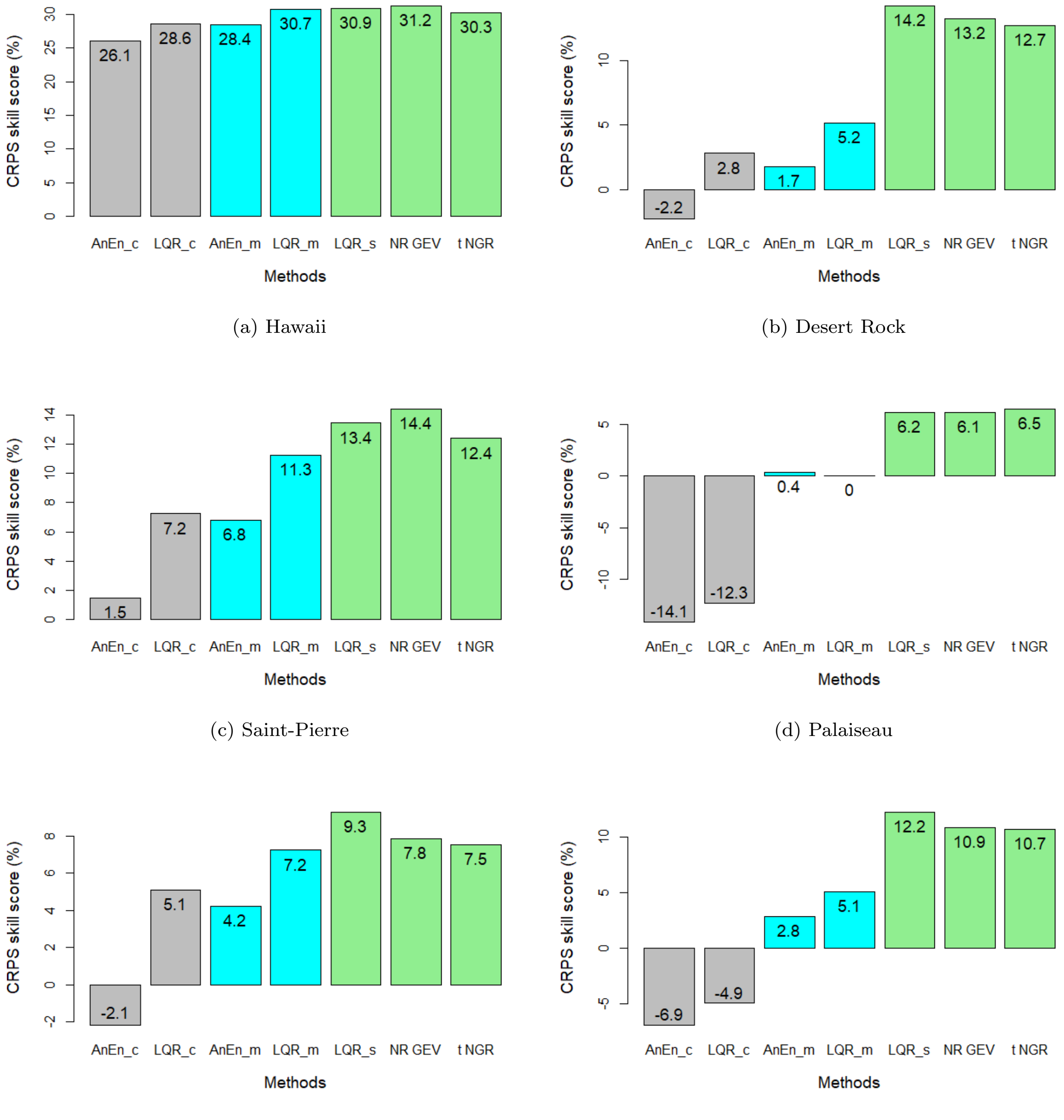
The initial day-ahead ensemble forecasts, covering the same period as the measurements, are provided by the European Centre of Medium-Range Weather Forecasts (ECMWF). They correspond to 50 perturbed members and a control run (unperturbed member) [4]. This leads to a total of $M = 51$ members. The EPS is released by ECMWF at 12:00 for the 72 next hours with a 3-hours timestep which allows it to be used for day-ahead scheduling or trading purposes.

5. Results

Based on the verification framework proposed by Lauret et al. [25], ~~we assess first visually the reliability property with the help of rank histograms. Second,~~ the overall performance of the different probabilistic methods is measured by the CRPS and the CRPSS. Detailed insight in the quality of the models is obtained through the decomposition of the CRPS and the new "MC-CRPS" method. Notice that this section is dedicated to the presentation of the main results of the study. The next section will be devoted to in-depth discussion related to the pros and cons of each approach and the added-value brought by the MC-CRPS methodology.

5.1. ~~Reliability assessment~~

~~Appendix B gives the rank histograms (RHs) of the different probabilistic models. As shown by Figure B.17, raw ensembles (regardless the site under study) exhibit a characteristic U-shape that corresponds to over-confident models. This under-dispersed character confirms the need for calibration procedures that can be implemented through ensemble-based approach. Appendix C gives more details about the necessity of calibration of raw EPS forecasts. Compared to raw EPS forecasts, both approaches improve the reliability of the forecast as population (or equivalently the relative frequency) of the two extreme ranks of the corresponding RHs has clearly diminished. However, the visual analysis of the RHs cannot lead to the conclusion that a particular model is reliable because, whatever the site, the frequencies of some ranks are outside the consistency bars. An in-depth analysis by site shows that for Hawaii the RHs tend to be uniform except for the t_NGR calibration method. For almost all the models, one can notice an asymmetric shape (overpopulation of the right part of the RH) for St Pierre which is a sign of an under-forecasting bias. RHs of Desert Rock exhibit no predominant shape. Some conditional biases can be detected for the $AnEn_c$ and $AnEn_m$.Finally, notice that, contrary to the t_NGR method,the NR_GEV model leads to better calibrated forecasts albeit the extreme ranks for Desert Rock and St Pierre are still overpopulated. Therefore, it appears that for parametric approaches like t_NGR and NR_GEV, the choice of the underlying distribution has an impact on the reliability of the generated forecasts. At this point, the reliability assessment based on RHs may appear not conclusive but let us stress that this kind of tool brings only a qualitative diagnostic. The decomposition of the CRPS depicted in section 5.3 will try to shed more light on the impact of reliability on the quality of each probabilistic models.~~



(a) Hawaii

(b) Desert Rock

(c) Saint-Pierre

(d) Palaiseau

(e) Tiruvallur

(f) Langley

Figure 5: CRPS Skill Score of all models for the six considered sites. Grey : deterministic-based approach, Cyan : ensemble-based approach using the mean of the members, Green : ensemble-based approach using mean and standard deviation of the members.

5.2. Overall performance of the methods

Table 4 lists the CRPS obtained by the different methods. However , in order to better highlight the relative merits of each approach, Figure 5 shows the CRPS skill scores of all the forecasting models. Let us recall that positive values of skill scores mean that the model outperforms the reference model (here the raw ECMWF-EPS) while negative values reveal that the quality of the evaluated model is worse than the reference one.

As shown by Figure 5, regardless the site under study, the highest CRPS skill scores are obtained by the ensemble-based approach (represented by the cyan and green bars). Conversely, except the case of Hawaii, the deterministic-based approach (grey bars) yields lower or even negative skill scores. These negative CRPSS values indicate that the deterministic-based models do not always achieve to increase the quality of the raw ensemble forecasts (see for example Palaiseau and Langley).

A deeper look into the performance of the ensemble-based approach shows that models using the mean and the standard deviation of the ensemble members (green bars) exhibit a better forecast skill than models using only the mean of the members (cyan bars) albeit the improvement is less pronounced for Hawaii. Overall, the model with the highest skill score appears to be either LQR_s or NR_GEV. Regarding the latter, it may suggest that a judicious choice of the underlying PDF (see Equation 7) used by a calibration technique like Nonhomogenous Regression (NR) can further improve the quality of the probabilistic forecasts.

Finally, in order to quantify the relative improvement provided by the ensemble-based approach over the deterministic-based approach, we calculate the gain in CRPS based on the CRPS values of the best performer of each approach. It appears that the level of improvement is very dependent on the studied site. It is slight for Hawaii and Tiruvallur (4%), becomes larger for Saint-Pierre (approximately 8%) and quite significant for Desert Rock (approximately 12%), Langley and Palaiseau (approximately 16%).

~~First, one can state that all probabilistic models improves on the raw EPS and that the gain in quality is more pronounced for the sites of Saint-Pierre and Hawaii which experience variable sky conditions. Second, based on the CRPS results, regardless the considered site, it appears that models issued from the ensemble-based approach outperform those from the deterministic-based approach. More precisely, deterministic-based models that use the control member of the EPS are the worst ones followed by the ensemble-based methods based uniquely on the mean of the members. Further, the t_NGR and LQR_s models are fairly comparable in terms of CRPS, and the NR_GEV model turns out to be the best forecasting method. It must be noted also that the LQR technique shows a general superiority compared to the $AnEn$ technique.~~

5.3. Detailed insight through the decomposition of the CRPS

Table 4 also provides the decomposition of the CRPS into reliability and resolution of the different forecasting methods. As mentioned previously, a forecast should exhibit a small reliability term and a large resolution term. It is worth mentioning first that all models significantly decreases the reliability component of the raw EPS forecasts and that the level of improvement strongly depends on the reliability of the initial raw ensemble. ~~except for~~

	Site	HAW	DR	SP	PAL	TIR	LAN
CRPS (W/m^2)	raw Ensemble	**67.7**	29.4	**59.4**	38.6	46.8	40.0
	$AnEn_c$	50.1	**30.1**	58.5	**44.0**	**47.8**	**42.8**
	LQR_c	48.4	28.6	55.1	43.3	44.4	42.0
	$AnEn_m$	48.5	28.9	55.3	38.4	44.9	38.9
	LQR_m	46.9	27.9	52.7	38.6	43.4	38.0
	LQR_s	46.8	**25.2**	51.4	**36.2**	**42.5**	**35.2**
	t_NGR	47.2	25.7	52.0	**36.2**	43.3	35.8
	NR_GEV	**46.6**	25.5	**50.8**	**36.2**	43.2	35.7
Reliability (W/m^2)	raw Ensemble	23.2	8.4	13.4	7.5	11.5	8.2
	$AnEn_c$	4.2	4.8	6.6	4.9	7.2	4.8
	LQR_c	4.4	5.3	7.1	5.4	6.7	5.3
	$AnEn_m$	4.1	4.7	6.2	4.9	7.9	4.5
	LQR_m	4.4	5.7	7.0	5.7	8.2	5.0
	LQR_s	4.5	5.9	7.6	5.3	8.2	5.4
	t_NGR	4.7	6.5	8.4	5.4	8.0	5.7
	NR_GEV	4.1	6.2	7.2	5.4	7.8	5.8
Resolution (W/m^2)	raw Ensemble	113.3	154.0	126.7	95.4	125.7	122.5
	$AnEn_c$	111.9	149.7	120.8	87.4	120.4	116.4
	LQR_c	113.9	151.7	124.7	88.6	123.3	117.7
	$AnEn_m$	113.4	150.8	123.5	93.0	124.0	120.0
	LQR_m	115.3	152.8	127.0	93.6	125.8	121.4
	LQR_s	115.5	155.6	128.9	95.6	126.8	124.7
	t_NGR	115.3	155.7	129.0	95.8	125.8	124.3
	NR_GEV	115.3	155.6	129.1	95.6	125.7	124.5
Uncertainty (W/m^2)	All Models	157.8	175.0	172.7	126.5	161.0	154.4

Table 4: CRPS and its components reliability, resolution and uncertainty of all considered models for the 6 sites. Cyan : deterministic-based approach, Green~~,~~ : ensemble-based approach. Red values indicate the worst CRPSs while the black bold ones show the best CRPSs.

~~the Desert Rock site where the improvement in reliability is less pronounced.~~ Second, it can be noted that the reliability of all calibrated forecasts is fairly comparable. In addition, regardless the site, it appears that, overall, the ensemble-based approach does not significantly improve reliability compared to the deterministic-based approach. Looking in more details, models based on the $AnEn$ technique often appears to generate the most reliable forecasts while the t_NGR model generally provides the less reliable forecasts. Also, in the case of Non homogeneous calibration technique, GEV distributions seem to be more suitable than Gaussian distributions, since NR_GEV is slightly more reliable than the t_NGR model.

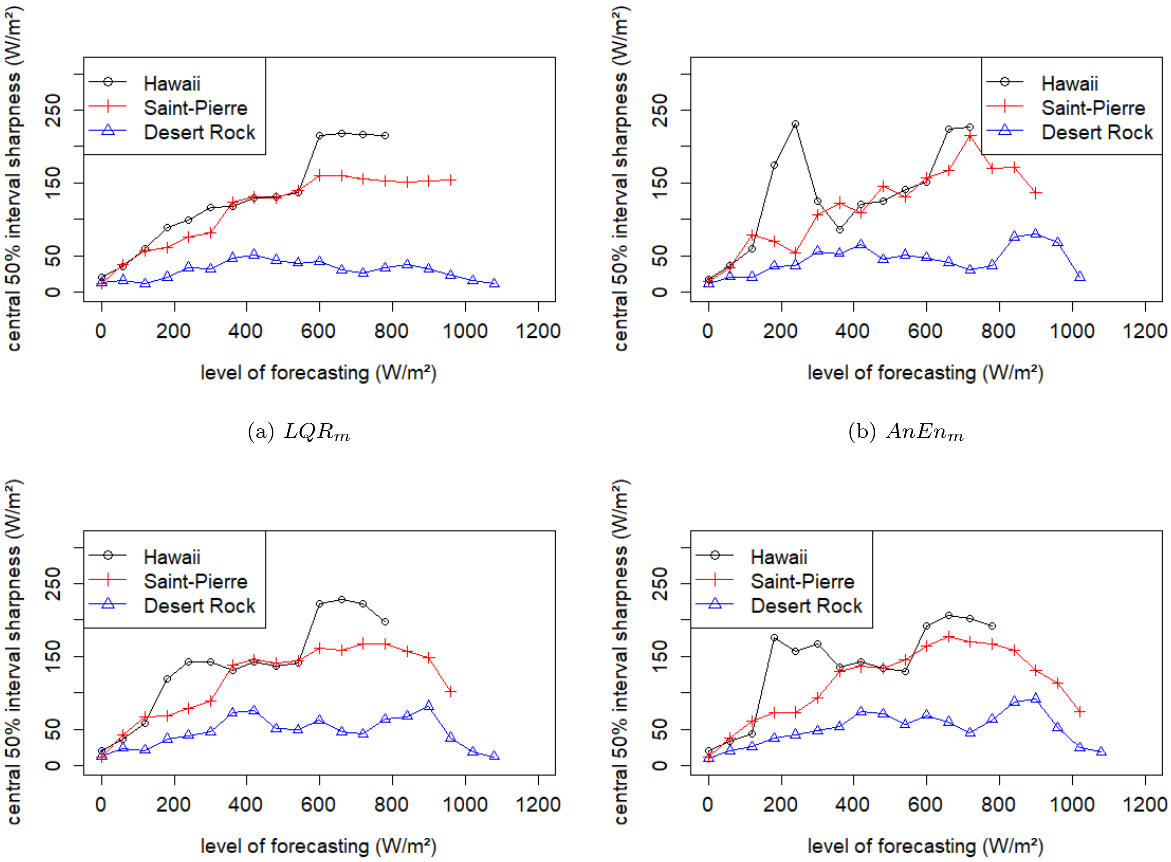
Regarding the resolution component, it must be noted first that the deterministic-based approach fails to improve the resolution of the raw Ensemble. Conversely ~~to reliability~~, resolution increases with the ensemble-based approach, and particularly when the spread of EPS members is taken as as input of the models i.e. case of the LQR_s,t_NGR and NR_GEV models. Put differently, these results suggest that ensemble-based approach uniquely improves the resolution (and not the reliability component) of the forecasting models. ~~Finally, notice that models based on the $AnEn$ technique fail to outperform the resolution of the raw forecasts.~~ Finally, one can state that the decomposition of CRPS given in Table 4 reveals that the difference in quality of the probabilistic forecasts is mainly explained by the resolution component, whereas reliability is fairly comparable.

5.4. ~~*Comparison between sites*~~

~~As shown by Table 4, all forecasts get significantly better scores in Desert Rock. Last line of Table 3 lists a high solar variability in Saint-Pierre and Hawaii and a low variability in Desert Rock. Note that the two sites with high variability have in common to be insular and very mountainous. Lauret et al. [46] found a link between solar variability and the accuracy of deterministic forecasting methods, and our results suggest that a link may also exist between sky conditions experienced by a site and quality of probabilistic forecasts.~~

~~The decomposition of CRPS given in Table 4 reveals that the difference between sites is mainly explained by the resolution, whereas the reliability score is fairly comparable.~~

~~To understand why resolution differs significantly between the 3 sites, we plot the width of the 50% central prediction interval of forecasts generated by 4 models in Figure 6. This is a measure of the sharpness of the forecasts. As the contribution of the reliability part is small whatever the forecasting method, one can hypothesise here that sharpness relates to resolution. Figure 6 suggests that forecasting schemes need to keep large distributions for Saint-Pierre and Hawaii, to not deteriorate the reliability. This explains the difference in resolution. As this observation can be made for all forecasting schemes, we can conclude that sky-conditions are possibly the reasons of these differences.~~



(a) LQR_m

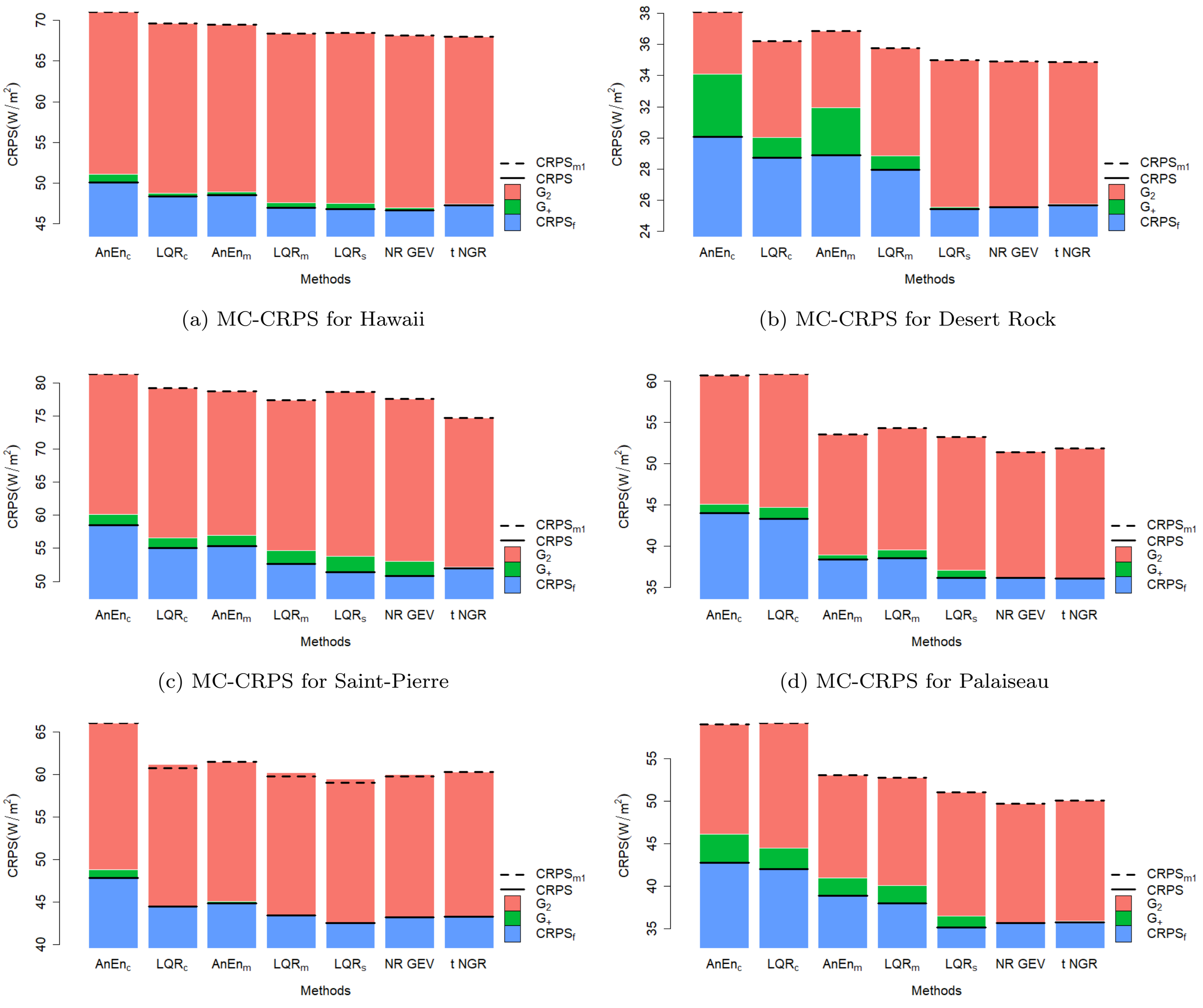
(b) $AnEn_m$

(c) LQR_s

(d) t_NGR

Figure 6: ~~Sharpness of several models for different forecasting levels given by the mean width of central 50% interval for the three different studied sites~~

5.5. Detailed insight through the CRPS Moments-Contributions



(a) MC-CRPS for Hawaii

(b) MC-CRPS for Desert Rock

(c) MC-CRPS for Saint-Pierre

(d) MC-CRPS for Palaiseau

(e) MC-CRPS for Tiruvallur

(f) MC-CRPS for Langley

Figure 7: MC-CRPS of the ~~three~~ six sites and all forecasting models. A black line is used to better highlight the value of the CRPS and a dotted black line indicates the value of $CRPS_{m1}$.

Figure 7 shows the results of the MC-CRPS introduced in section 3.4. ~~In general,~~ As seen, the final CRPS values, ~~the scores~~ of the forecasting models ~~in Hawaii~~ occurs to be ~~strongly~~ dependent on their respective $CRPS_{m1}$ values. ~~, as G_2 is fairly comparable for all models and the gain G_+ low for all models.~~ In particular, models from the ensemble-based approach appear to have best $CRPS_{m1}$ than models from the deterministic-based approach (see for instance the case of Langley). This means that the aggregation of members improves the estimation of the first moment. Among the ensemble-based models, except for the cases of Hawaii and Tiruvallur, the superiority of the LQR_s, NR_GEV and $tNGR$ models using the mean and standard deviation of the ensemble members can be mainly explained by a greater contribution of G_2. Thus the spread of EPS members is effective and improves more importantly G_2 than $CRPS_{m1}$. Further, the best performer among the three aforementioned models is finally determined by G_+. This highlights the importance of the choice of the distribution in the non homogenous regression calibration framework. ~~For Desert Rock, the gain G_2 increases for LQR_s, t_NGR and NR_GEV. This fact may explain their overall superiority. In Saint-Pierre, $CRPS_{m1}$ is worse for deterministic-based models ($AnEn_c$ and LQR_c) and G_2 is the highest for LQR_s and NR_GEV models. For all sites except Desert Rock, the two worst models in terms of $CRPS_{m1}$ are $AnEn_c$ and LQR_c from the deterministic-based approach.~~

~~Except for Hawaii, G_2 increases only for models LQR_s, t_NGR and NR_GEV. Thus the spread of EPS members is effective for Desert Rock and Saint-Pierre, and improves more importantly G_2 than $CRPS_{m1}$. In terms of CRPS, notice finally that~~ For example, overall, NR_GEV performs better than the t_NGR model because of G_+. Let us stress that the choice of strictly truncated Gaussian distributions in the implementation of a NGR technique forces G_+ to be very close to 0 in the MC-CRPS. Hence, the benefits of GEV distributions compared to Gaussian distributions are highlighted by the MC-CRPS method.

6. Discussion

In this section, we try to give more clues regarding the merits of each proposed approach. Also, a discussion related to the advantages brought by the MC-CRPS is proposed.

6.1. Deterministic-based approach versus ensemble-based approach

Let us recall that the deterministic-based approach uses ~~an~~ a unique deterministic predictor while the ensemble-based approach makes use of the information conveyed by the ensemble. Therefore, the main weakness of deterministic-based approach is the lack of information feeding the models. Since the distribution needs to be completely determined from one single deterministic predictor, the spread and the possible skewness and kurtosis of the forecasting distribution need to be only inferred from this single predictor. Conversely, the benefits gained from the multiplicity of predictors provided by the ensemble-based approach need to be significant to justify the computation of the EPS. Two types of benefits can be discussed.

First, the aggregation of predictors leads to a better estimation of the first moment. This is visible in Figure 7 where models issued from the ensemble-based approach gets ~~significantly~~ better $CRPS_{m1}$ than models from the deterministic-based approach. It is clear that a gain in the estimation of the first moment can be obtained by the substitution of the control member by the mean of all members. However, models belonging to the ensemble-based approach are not always better than $AnEn_m$ and LQR_m. It means that the superiority of t_NGR and LQR_s models cannot be explained by a better estimation of the mean value of the probabilistic forecast distribution.

Second, regarding the determination of the second moment, the uncertainty is already carried by the level of forecasting of the mean of EPS members. These variables are dependent, as shown in Appendix C (the standard deviations of the observations clearly depends on the level of forecasting). Hence, using the spread of the members of EPS as input of the forecasting models can only be justified if it brings an extra-information on the uncertainty. It is assumed that the spread of the members is higher if the uncertainty is so. Indeed it indicates if slight errors in the initial conditions could lead to great differences in the final state of the atmosphere.

Thus, it appears necessary to investigate on the quantity of information actually provided by the spread of the members. In order to do this, the correlation between the standard deviation of the observations and the spread of the members has been studied. This has been made for a fixed level of forecasting, in order to remove the dependency between uncertainty and level of forecasting. Then an average over all levels of forecasting has been calculated to produce Figure 8. This kind of plot is of great utility to know the added value of the standard deviation of the EPS forecast members. If the dependence between the spread of the members and the uncertainty of the forecast for a fixed level of forecasting is strong, then a large improvement can be expected for calibration models using the spread of the members as an input, compared to simpler models.

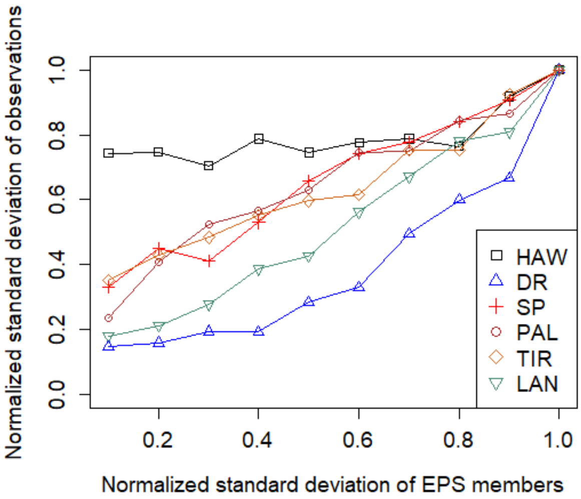


Figure 8: Standard deviation of observations vs. standard deviation of the EPS members (raw ECMWF ensembles). Normalization of the standard deviation has been done by dividing the standard deviations by the maximum of the standard deviation for each site.

As shown by Figure 8, the amount of new information given by the spread of the members is very dependent on the studied site. When for Hawaii, the correlation between the standard deviation of the observations and the spread of the members is almost null, it is quite significant for the ~~two~~ other sites and especially for Langley and Desert Rock. A link can be established between this finding and Table 4 which shows that the success of taking into account the spread of members in the forecasting models depends on the site (it is clearly less valuable in Hawaii than in other sites, and it is particularly successful in Desert Rock and Langley). It is also consistent with Figure 7 where G_2 is significantly higher in Desert

Rock for LQR_s, t_NGR and NR_GEV models.

~~The success of the ensemble-based models LQR_s, t_NGR and NR_GEV compared to deterministic-based forecasting models may justify the usage of EPS for probabilistic forecasting despite the high computation requirements needed to generate the EPS. However the level of improvement is very dependant on the studied site. For each site, the ratio of the CRPS score of the best ensemble-based forecasting model over the CRPS score of the best deterministic-based forecasting model has been calculated to define this improvement. It is slight for Hawaii (4%), becomes larger for Saint-Pierre (approximately 8%) and major for Desert Rock (approximately 12%)~~

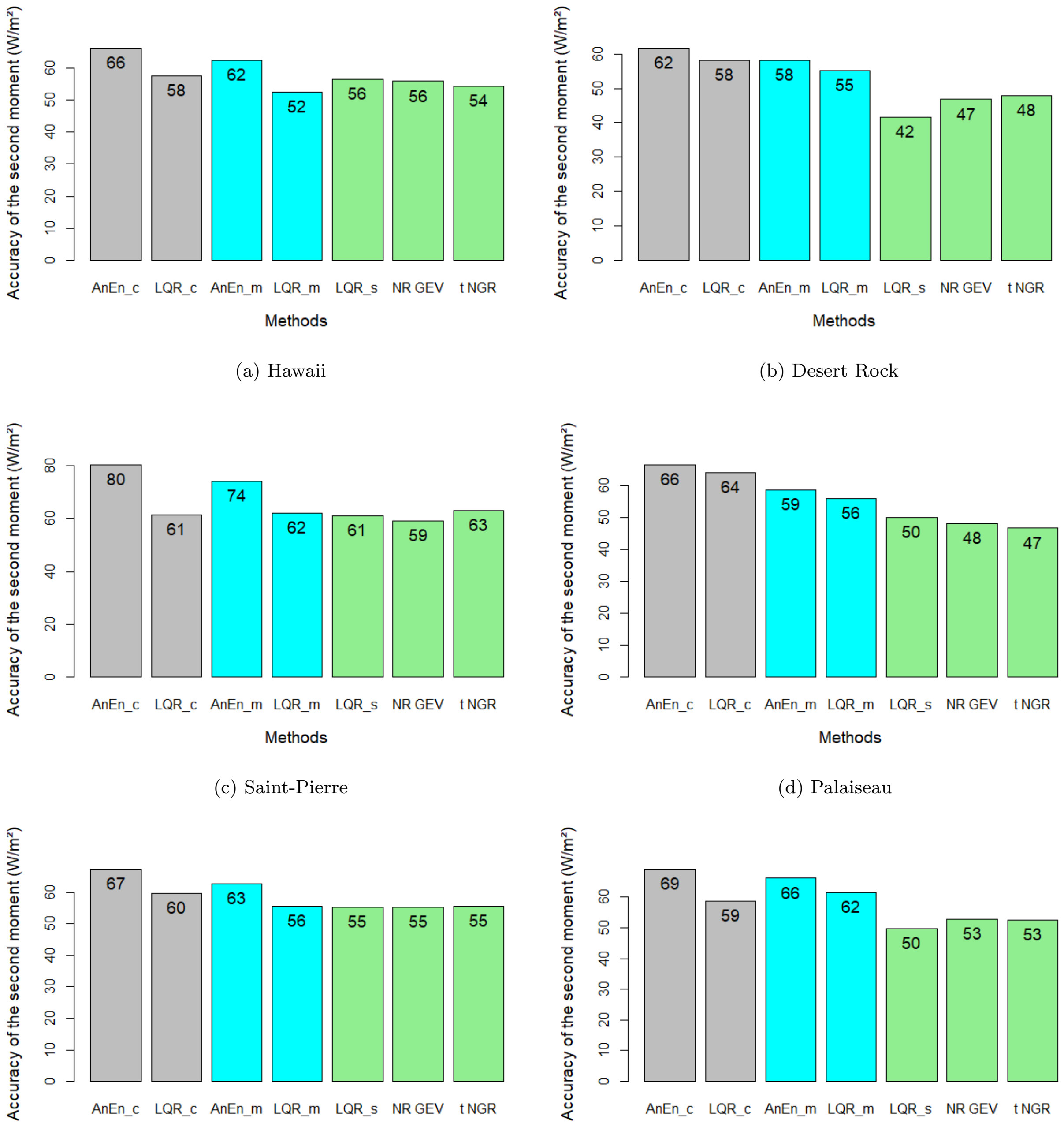
6.2. Discussion related to CRPS Moments-Contributions

In order to consolidate the results obtained in Figure 7, a complete analysis of the statistical moments of the probability distributions produced by the forecasting methods has been conducted. This kind of study is traditionally done to assess the strengths and weaknesses of a forecasting model. Although the deterministic measure of a statistical moment is not a proper scoring rule, it is of great interest to use it to understand the behaviour of the forecasting models.

First, an evaluation of the accuracy of the first moment has been conducted. A good forecasting model should have the ability to give a mean value of the forecasting distributions as close as possible to the mean of the observation values. A measure of this ability can be obtained by calculating the Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) of the mean of the forecasting distributions[2]. In this study, the MAE has been chosen as it is exactly the definition of $CRPS_{m1}$ introduced in section 3.4 (see [37] for details). Figure 7 gives therefore the results related to the accuracy of the first moment of the distributions.

Second, a probabilistic forecast also provides an estimation on the level of uncertainty, which is reflected by the spread of the forecasting distribution (i.e. the second statistical moment). Some works have been specifically dedicated to the assessment of the accuracy of the spread of the predictive distributions. Among others, one can cite the studies related to the spread-skill relationship (see [48] or [49]). These works are guided by the idea that the variance of a probabilistic forecast should be larger if the uncertainty of the forecast is so. Fortin et al. [50] proposed a criterion for the evaluation of the accuracy of the second moment of the distributions. This criterion is based on the fact that statistical consistency requires that the spread of the forecasting distributions should be equal to the RMSE of the mean of the forecast. Following [50], spread is calculated as the square root of the mean of the variances of the forecasting distributions. The accuracy of the second moment is therefore measured by calculating the RMSE of the differences between spread and RMSE of the mean of the distributions (i.e. $RMSE_M$). Figure 9 plots the RMSE of the difference $(spread - RMSE_M)$, computed over the evaluation period.

[2] RMSE and MAE are common metrics used to assess the accuracy of deterministic forecasts [47]



(a) Hawaii

(b) Desert Rock

(c) Saint-Pierre

(d) Palaiseau

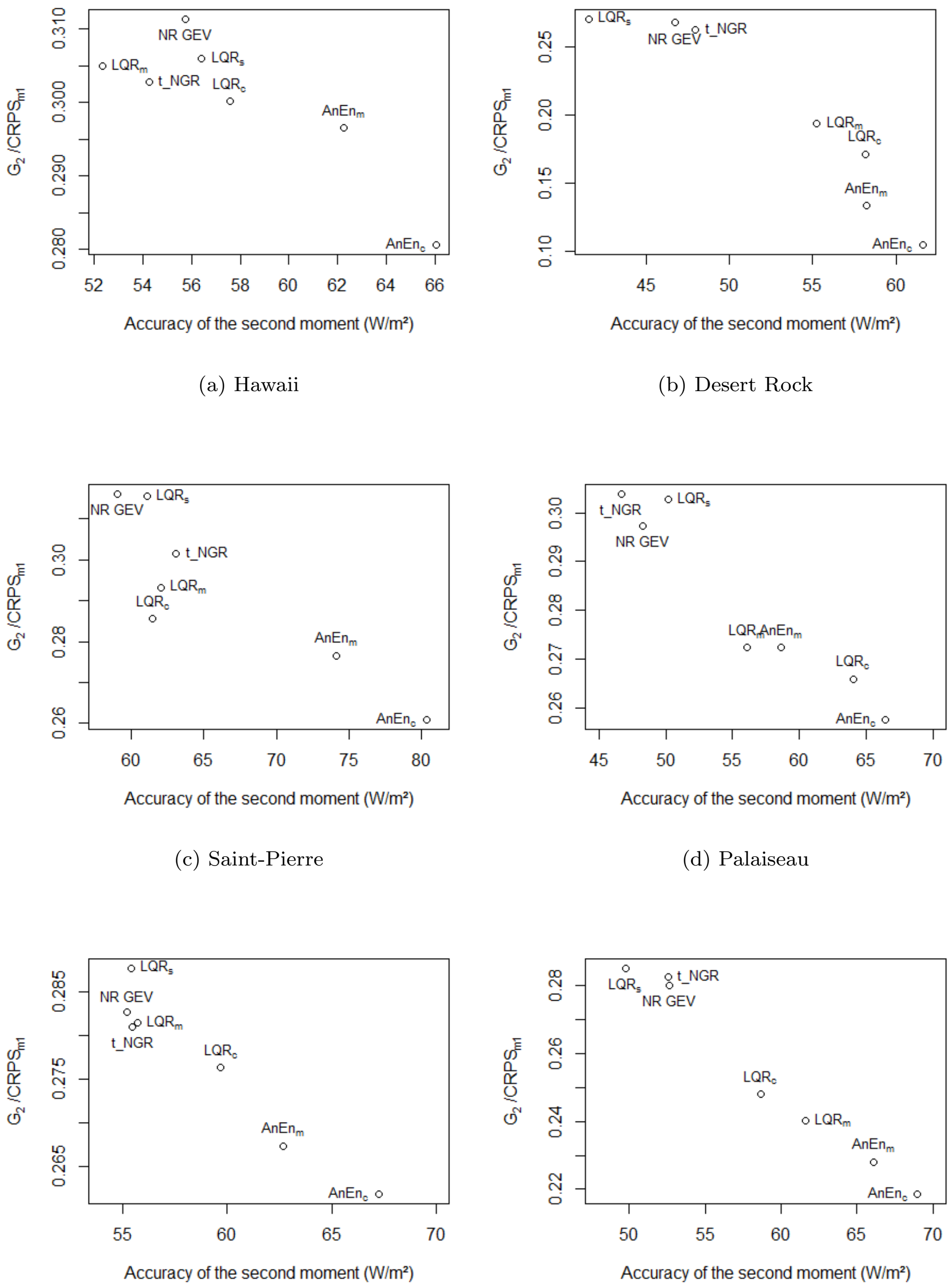
(e) Tiruvallur

(f) Langley

Figure 9: Accuracy of the second moment for the ~~three~~ six studied sites and all forecasting models

Conversely to the first moment, the accuracy of the second moment gradually improves when the information taken by the forecasting model is more complete. Using the mean of members instead of the control member increases the second moment accuracy. Taking into account the spread of the EPS improves further the accuracy by approximately the same extent (except for Hawaii, for the reasons discussed in section 6.1). Nevertheless, this improvement depends on the site and sky conditions. As shown by Figure 9, the accuracy of the second moment for Hawaii is almost equal for each model. It is consistent with the results depicted in Figure 8, showing that the information of the second moment of the EPS distribution in ~~Saint-Pierre~~ Langley and Desert Rock is ~~more~~ the most valuable ~~than~~ , as opposed to the information of Hawaii EPS distribution.

The accuracy of the second moment can be linked to the gain G_2 introduced in the MC-CRPS (see section 3.4). The correlation between these two values is highlighted in Figure 10, which shows the ratio $G_2/CRPS_{m1}$ versus the accuracy of the second moment.



(a) Hawaii

(b) Desert Rock

(c) Saint-Pierre

(d) Palaiseau

(e) Tiruvallur

(f) Langley

Figure 10: Link between G_2 and the accuracy of the second moment.

To sum up, the great advantage of the MC-CRPS is to reconcile the score of a probabilistic forecasting model and the explanation of its performance by examining the accuracy of the moment-based distributions.

Moreover, the link between the calibration of the moments and the score is highlighted, because the contribution of the accuracy of the moments to the score is quantified. Here, in the proposed new diagnostic tool MC-CRPS, the accuracy of the statistical moments of the forecasting distributions is quantified by the proper score itself. This diagnostic tool is complementary of the decomposition discussed in section 3.3.3, i.e. reliability and resolution of f_{m1} and f_{m2} can also be computed and studied. The MC-CRPS diagnostic tool also highlights the benefits of probabilistic forecasting, as the comparison between $CRPS_{m1}$ and $CRPS$ provides a measure of the quality difference between deterministic and probabilistic forecasting.

7. Conclusions

Based on the two types of forecasts i.e deterministic or ensemble forecast (denoted by the term EPS for ensemble prediction system) issued by the meteorological centre ECWMF, two approaches ~~to~~ for generating day-ahead solar irradiance probabilistic forecasts were proposed. The first approach creates probabilistic forecasts from the deterministic day-ahead GHI predictor while the second one generates probabilistic forecasts from the calibration of the EPS or from information inferred from the ensemble.

The goal of this work was to quantify the possible added-value of the EPS on the quality of the forecasts. ~~Three~~ Six sites experiencing different sky conditions were chosen for the appraisal of the different probabilistic models. Quality of the different probabilistic models have been evaluated with common diagnostic tools such as ~~Rank Histograms,~~ the CRPS and its decomposition. A new diagnostic tool called MC-CRPS has also been introduced. It consists in the measure of the contribution of each statistical moment of the forecasting distributions to the CRPS.

Overall, models ~~based on~~ adopting the ensemble-based approach have been found to issue probabilistic forecasts with better quality than the ones based on the deterministic-based approach. The ~~level of improvement~~ gain in quality, based on the CRPS metric, ~~depends on the site and varies~~ ranges from 4 % up to 16 %. ~~It has also been demonstrated that the choice of a good distribution for parametric models is essential. Generalized Extreme Value distribution has been found a better candidate than Gaussian distribution for improving the quality of the probabilistic forecasts. It is shown also that the sky conditions experienced by a site largely impact the skills of the probabilistic forecasting models.~~

One other important contribution of this work is the new diagnostic tool related to the CRPS score based on the moments of the ensemble distribution called MC-CRPS. This MC-CRPS tool allowed to identify two characteristics of EPS that have an impact on the quality of probabilistic forecasts. First, the aggregation of deterministic predictors of the ensemble leads to an improvement of the estimation of the first moment and thus, raises the overall quality of a probabilistic forecast. Second, ~~depending on the sky conditions of the site,~~ the spread of the EPS members turns to be be a good predictor that permits to enhance the estimation of the second moment of the forecasting distributions. Finally, in terms of forecast quality, it can be concluded that using an EPS (which requires high computing capacities) to produce day-ahead GHI probabilistic forecasts ~~is worthwhile~~ should be favored compared to a deterministic (less demanding) approach. This work opens the way to the assessment of the forecast value of each approach i.e. the benefit (economical or others) gained from the use of these probabilistic forecasts in an operational context.

References

[1] M. Pierro, M. De Felice, E. Maggioni, D. Mosere, A. Perottoc, Residual load probabilistic forecast for reserve assessment: a real case study, Renewable Energy 125 (2019) 99–110. doi:10.1016/j.renene.2019.12.056.

[2] Y. Zhu, Z. Toth, R. Wobus, D. Richardson, K. Mylne, The economic value of ensemble-based weather forecasts, Bulletin of the American Meteorological Society 83 (2002) 73–83.

[3] R. Buizza, The value of probabilistic prediction, Atmospheric Science Letters 9 (2008) 36–42. doi:10.1002/asl.170.

[4] M. Leutbecher, T. N. Palmer, Ensemble forecasting, Journal of Computational Physics 227 (2008) 3515–3539. doi:10.1016/j.jcp.2007.02.014.

[5] E. Lorenz, J. Hurka, D. Heinemann, H. Beyer, Irradiance forecasting for the power prediction of grid-connected photovoltaic systems, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2 (2009) 2–10.

[6] S. Alessandrini, L. Delle Monache, S. Sperati, G. Cervone, An analog ensemble for short-term probabilistic solar power forecast, Applied Energy 157 (2015) 95–110. doi:10.1016/j.apenergy.2015.08.011.

[7] M. Zamo, O. Mestre, A. P., O. Pannekoucke, A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part ii: Probabilistic forecast of daily production, Solar Energy 105 (2014) 804–816.

[8] P. Bacher, H. Madsen, H. A. Nielsen, Online short-term solar power forecasting, Solar Energy 83 (2009) 1772–1783. URL: http://linkinghub.elsevier.com/retrieve/pii/S0038092X09001364. doi:10.1016/j.solener.2009.05.016.

[9] P. Lauret, M. David, H. Pedro, Probabilistic solar forecasting using quantile regression models, Energies 10 (2017) 1591. doi:10.3390/en10101591.

[10] E. B. Iversen, J. M. Morales, J. K. Møller, H. Madsen, Probabilistic forecasts of solar irradiance by stochastic differential equations, Environmetrics 25 (2014) 152–164. doi:10.1002/env.2267.

[11] K. Bakker, K. Whan, W. Knap, M. Schmeits, Comparison of statistical post-processing methods for probabilistic nwp forecasts of solar radiation, Solar Energy 191 (2019) 138–150. doi:10.1016/j.solener.2019.08.044.

[12] S. Sperati, S. Alessandrini, L. Delle Monache, An application of the ecmwf ensemble prediction system for short-term solar power forecasting, Solar Energy 133 (2016) 437–450. doi:10.1016/j.solener.2016.04.016.

[13] L. Massidda, M. Marrocu, Quantile regression post-processing of weather forecast for short-term solar power probabilistic forecasting, Energies 11 (2018) 1763. doi:10.3390/en11071763.

[14] P. Pinson, Adaptive calibration of (u,v)-wind ensemble forecasts, Quarterly Journal of the Royal Meteorological Society 138 (2012) 1273–1284. doi:10.1002/qj.1873.

[15] P. Pinson, H. Madsen, Ensemble-based probabilistic forecasting at horns rev, Wind Energy 12 (2009) 137–155. doi:10.1002/we.309.

[16] C. Junk, L. Delle Monache, S. Alessandrini, Analog-based ensemble model output statistics, Monthly Weather Review 143 (2015) 2909–2917. doi:10.1175/MWR-D-15-0095.1.

[17] T. Hamill, J. Whitaker, Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application, Monthly Weather Review 134 (2006) 3209–3229. doi:10.1175/MWR3237.1.

[18] D. S. Wilks, Comparison of ensemble-mos methods in the lorenz '96 setting, Meteorological Applications 13 (2006) 243. doi:10.1017/S1350482706002192.

[19] R. M. Williams, C. A. T. Ferro, F. Kwasniok, A comparison of ensemble post-processing methods for extreme events, Quarterly Journal of the Royal Meteorological Society 140 (2014) 1112–1120. doi:10.1002/qj.2198.

[20] T. Gneiting, A. E. Raftery, A. H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation, Monthly Weather Review 133 (2005) 1098–1118. doi:10.1175/MWR2904.1.

[21] S. Lerch, L. Thorarinsdottir, T., Comparison of non-homogeneous regression models for probabilistic wind speed forecasting, Tellus A: Dynamic Meteorology and Oceanography 65 (2013) 21206. doi:10.3402/tellusa.v65i0.21206.

[22] S. Baran, S. Lerch, Combining predictive distributions for the statistical post-processing of ensemble forecasts, International Journal of Forecasting 34 (2018) 477–496. doi:10.1016/j.ijforecast.2018.01.005.

[23] S. Baran, S. Lerch, Log-normal distribution based emos models for probabilistic wind speed forecasting, Quarterly Journal of the Royal Meteorological Society 141 (2015) 2289–2299. doi:10.1002/qj.2521.

[24] S. Vannitsem, D. Wilks, J. Messner, Statistical Postprocessing of Ensemble Forecasts, Elsevier, 2018.

[25] P. Lauret, M. David, P. Pinson, Verification of solar irradiance probabilistic forecasts, Solar Energy 194 (2019) 254–271. doi:10.1016/j.solener.2019.10.041.

[26] D. V. Koenker, R., Confidence intervals for regression quantiles, Journal of the Royal Statistical Society 36 (1994) 383–393.

[27] V. Chernozhukov, I. Fernández-Val, A. Galicho, Quantile and probability curves without crossing, Econometrica 78 (2010) 1093–1125. doi:10.3982/ECTA7880.

[28] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R. Hyndman, Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, International Journal of Forecasting 32 (2016) 896–913. doi:10.1016/j.ijforecast.2016.02.001.

[29] L. Delle Monache, F. A. Eckel, D. Rife, B. Nagarajan, K. Searight, Probabilistic weather prediction with an analog ensemble, Monthly Weather Review 141 (2013) 3498–3516. doi:10.1175/MWR-D-12-00281.1.

[30] F. Calderon, J. Le Gal La Salle, J. Badosa, P. Lauret, A. Migan, V. Bourdin, Uncertainty estimation for deterministic solar irradiance forecasts based on analogs ensembles, Renewable Energy (2020, to be submitted).

[31] M. Scheuerer, Probabilistic quantitative precipitation forecasting using ensemble model output statistics: Probabilistic precipitation forecasting using emos, Quarterly Journal of the Royal Meteorological Society 140 (2014) 1086–1096. doi:10.1002/qj.2183.

[32] R. Yuen, S. Baran, C. Fraley, T. Gneiting, S. Lerch, M. Scheuerer, T. Thorarinsdottir, ensembleMOS: Ensemble Model Output Statistics, 2018. URL: https://CRAN.R-project.org/package=ensembleMOS, r package version 0.8.2.

[33] S. Yitzhaki, Gini's mean difference: a superior measure of variability for non-normal distributions, Metron - International Journal of Statistics 61 (2003) 285–316.

[34] D. Wilks, Statistical Methods in the Atmospheric Sciences, Academic Press, 2014.

[35] I. T. Jolliffe, D. B. Stephenson, Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, 2003.

[36] P. Pinson, P. McSharry, H. Madsen, Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation, Quarterly Journal of the Royal Meteorological Society 136 (2010) 77–90. URL: http://doi.wiley.com/10.1002/qj.559. doi:10.1002/qj.559.

[37] H. Hersbach, Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting 15 (2000) 559–570. URL: http://journals.ametsoc.org/doi/abs/10.1175/1520-0434%282000%29015%3C0559%3ADOTCRP%3E2.0.CO%3B2. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

[38] C. F. Coimbra, J. Kleissl, R. Marquez, Overview of Solar-Forecasting Methods and a Metric for Accuracy Evaluation, in: Solar Energy Forecasting and Resource Assessment, Elsevier, 2013, pp. 171–194. URL: http://linkinghub.elsevier.com/retrieve/pii/B9780123971777000085.

[39] H. T. Pedro, C. F. Coimbra, M. David, P. Lauret, Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts, Renewable Energy 123 (2018) 191–203. URL: http://linkinghub.elsevier.com/retrieve/pii/S0960148118301423. doi:10.1016/j.renene.2018.02.006.

[40] K. Doubleday, V. V. S. Hernandez], B.-M. Hodge, Benchmark probabilistic solar forecasts: Characteristics and recommendations, Solar Energy 206 (2020) 52 – 67. URL: http://www.sciencedirect.com/science/article/pii/S0038092X20305429. doi:https://doi.org/10.1016/j.solener.2020.05.051.

[41] J. Badosa, M. Haeffelin, H. Chepfer, Scales of spatial and temporal variation of solar irradiance on reunion tropical island, Solar Energy 88 (2013) 42–56. doi:10.1016/j.solener.2012.11.007.

[42] N. Kalecinski, Processus de formation et d'étalement des nuages sur l'Ile de la Réunion: caractérisation à partir de données issues d'observations satellite, sol et du modèle numérique de prévision AROME; application à la prévision des énergies solaires., Ph.D. thesis, Ecole Polytechnique, 2015.

[43] T. E. Hoff, R. Perez, Modeling PV fleet output variability, Solar Energy 86 (2012) 2177–2189. URL: http://linkinghub.elsevier.com/retrieve/pii/S0038092X11004154. doi:10.1016/j.solener.2011.11.005.

[44] M. David, F. Ramahatana, P. Trombe, P. Lauret, Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models, Solar Energy 133 (2016) 55–72. URL: http://linkinghub.elsevier.com/retrieve/pii/S0038092X16300172. doi:10.1016/j.solener.2016.03.064.

[45] M. David, L. Mazorra Aguiar, P. Lauret, Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data, International Journal of Forecasting 34 (2018) 529–547. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169207018300384. doi:10.1016/j.ijforecast.2018.02.003.

[46] P. Lauret, R. Perez, L. Mazorra Aguiar, E. Tapachès, H. Diagne, M. David, Characterization of the intraday variability regime of solar irradiation of climatically distinct locations, Solar Energy 125 (2016) 99–110. URL: https://linkinghub.elsevier.com/retrieve/pii/S0038092X15006490. doi:10.1016/j.solener.2015.11.032.

[47] T. E. Hoff, R. Perez, J. Kleissl, D. Renne, J. Stein, Reporting of irradiance modeling relative prediction errors, Progress in Photovoltaics: Research and Applications 21 (2013) 1514–1519. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2225. doi:10.1002/pip.2225. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.2225.

[48] J. S. Whitaker, A. F. Loughe, The relationship between ensemble spread and ensemble mean skill, Monthly Weather Review 126 (1998) 3292–3302. doi:10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

[49] T. M. Hopson, Assessing the ensemble spread–error relationship, Monthly Weather Review 142 (2014) 1125–1142. doi:10.1175/MWR-D-12-00111.1.

[50] V. Fortin, M. Abaza, F. Anctil, R. Turcotte, Why should ensemble spread match the rmse of the ensemble mean?, Journal of Hydrometeorology 15 (2014) 1708–1713. doi:10.1175/JHM-D-14-0008.1.

[51] C. N. Long, E. G. Dutton, BSRN Global Network recommended QC tests, V 2.0, Technical Report, PANGAEA, 2010. URL: https://epic.awi.de/id/eprint/30083/1/BSRN_recommended_QC_tests_V2.pdf.

[52] R. E. Bird, R. L. Hulstrom, Simplified Clear Sky Model for Direct and Diffuse Insolation on Horizontal Surfaces, Technical Report, Solar Energy Research Institute, Golden, CO, 1981.

Appendix A. Data quality check

A quality check has been conducted for the observation data of each of the ~~three~~ six studied sites. As the decomposition of irradiance into diffuse and direct has not been measured, the exhaustive set of BSRN recommended quality checks could not been conducted (see [51]), but only the first plot. It consists in the plot of measured irradiance versus solar zenith angle. The rarely reached limit is plotted in dashed line and the physical possible limit is plotted in solid line. The second check is a frequency histogram of the clear-sky index (k*) for each site. k* is defined as:

$$k* = \frac{Irradiance}{ClearSky\ Irradiance} \tag{A.1}$$

where the clear-sky irradiance is calculated with the Bird clear-sky model [52]. The maximum of the observed frequency is supposed to be at $k* = 1$. The third check is a plot of the k*, only for clear-sky days. The morning data is reported by black dots and afternoon data by red dots. From this plot, it is possible to see if clear-sky irradiances are well-reported by the measurement data. If not, the line drawn by the dots is not straight. To extract clear-sky days from the data, the process proposed in Badosa et al. [41] has been followed. The last figure is a plot of the k* for each hour and day of the year. It allows to detect if systematical biases exist at some days/hours of the year. It also allows to easily detect missing data.

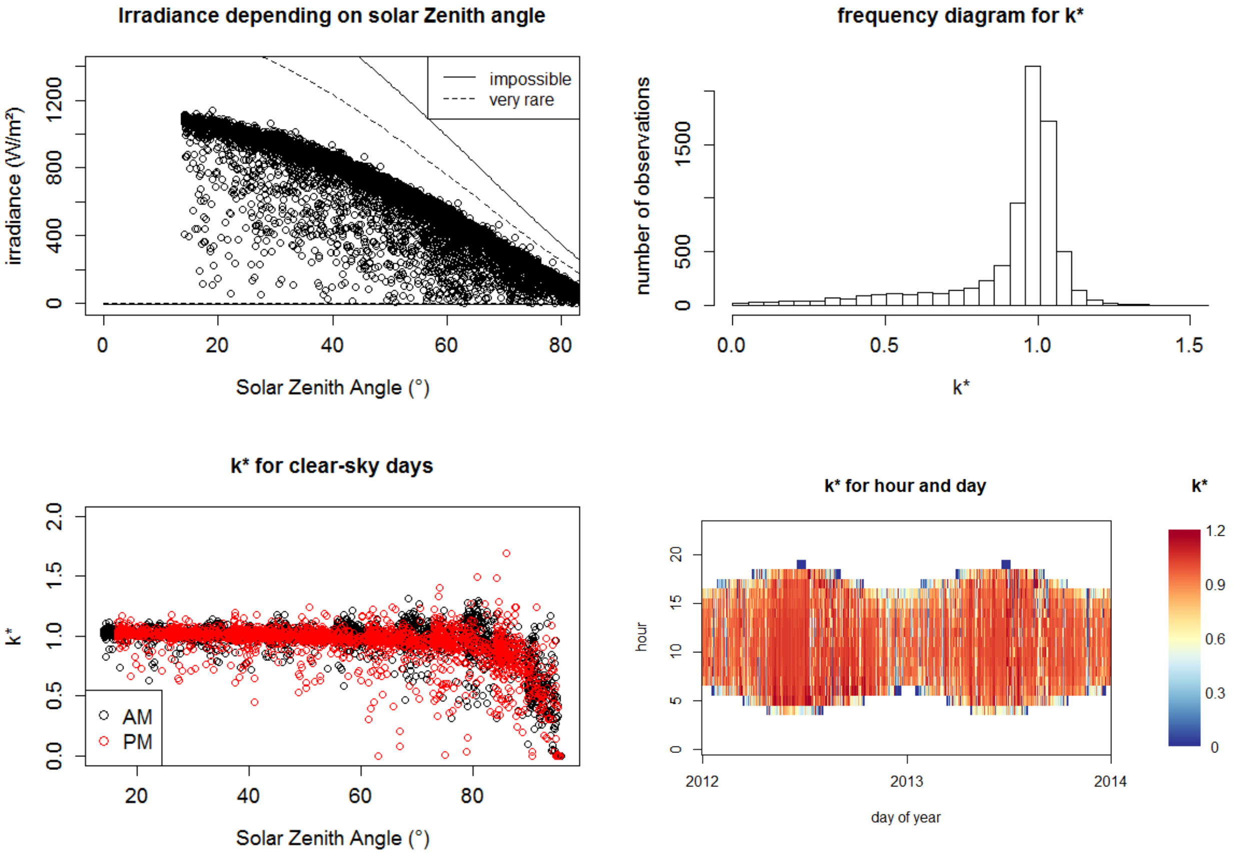


Figure A.11: Desert Rock

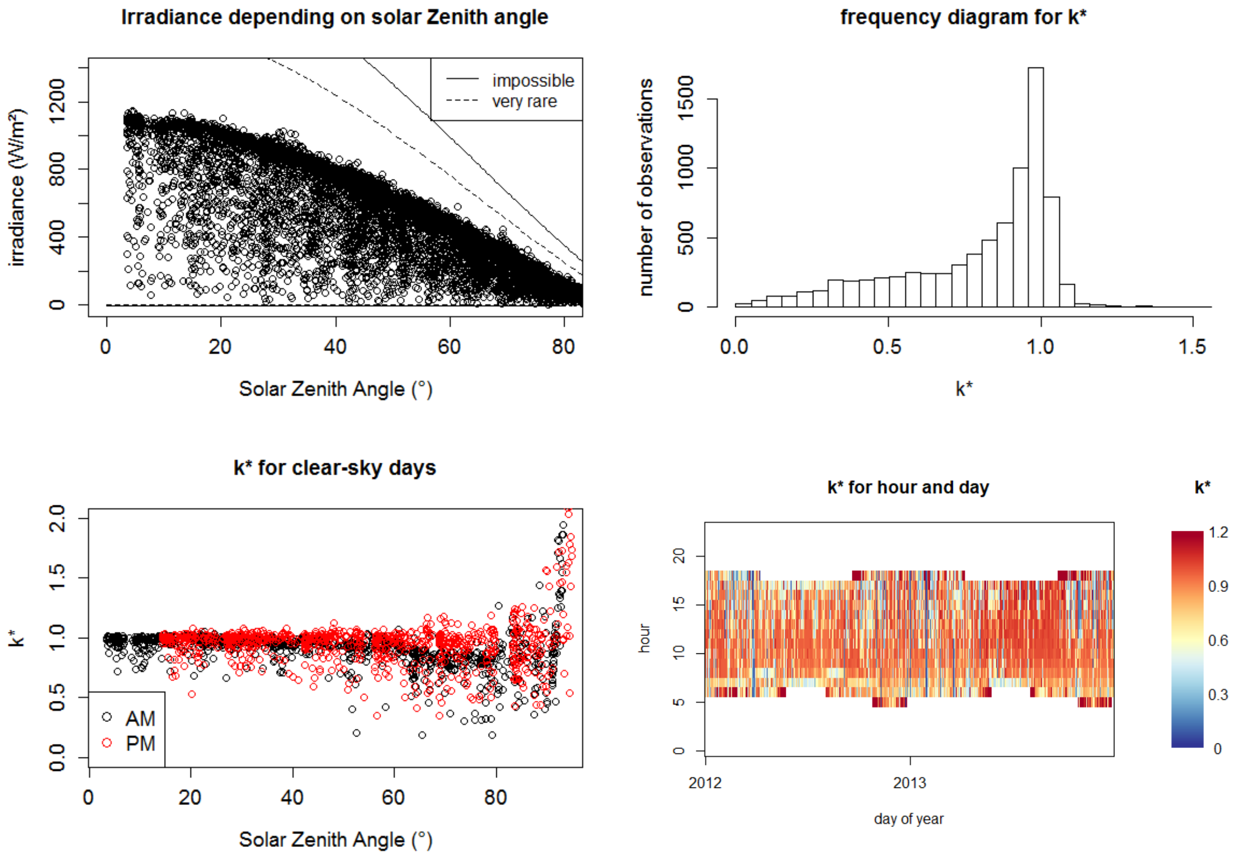


Figure A.12: Saint-Pierre

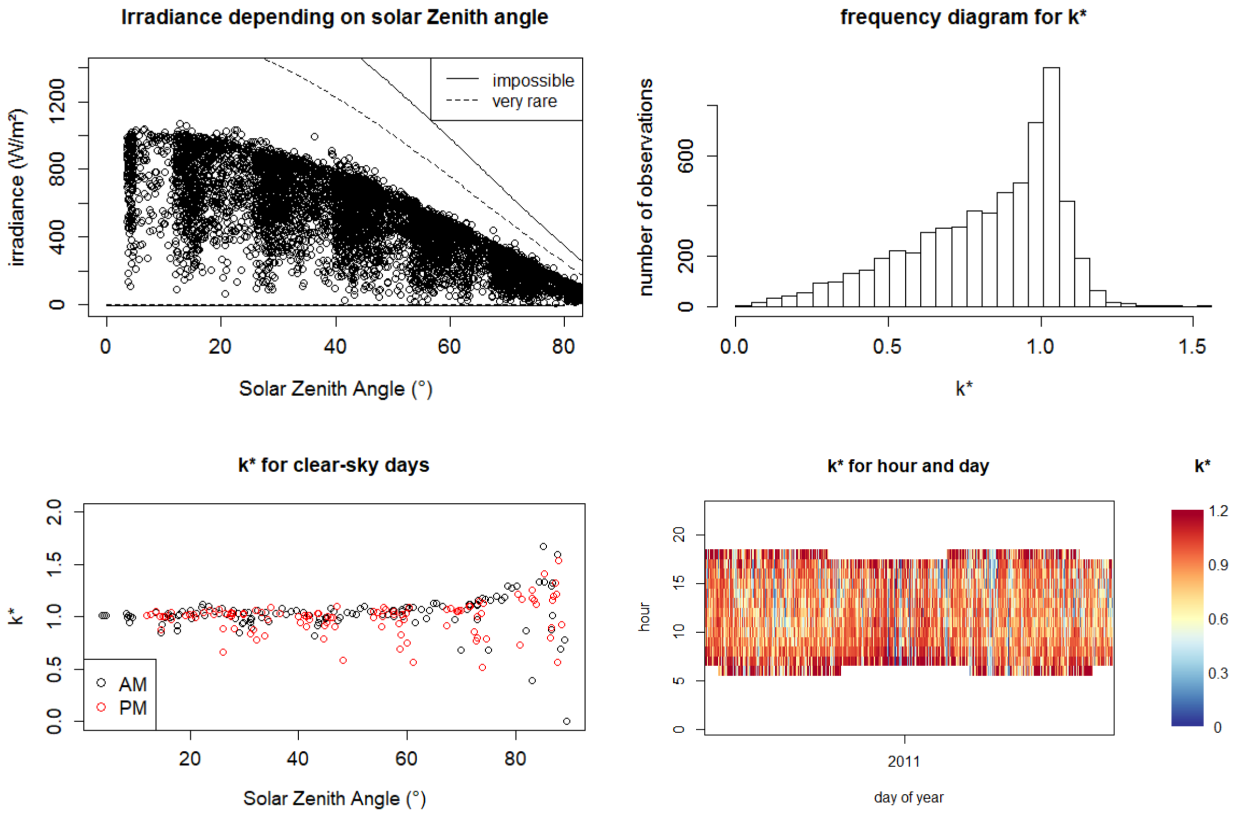


Figure A.13: Hawaii

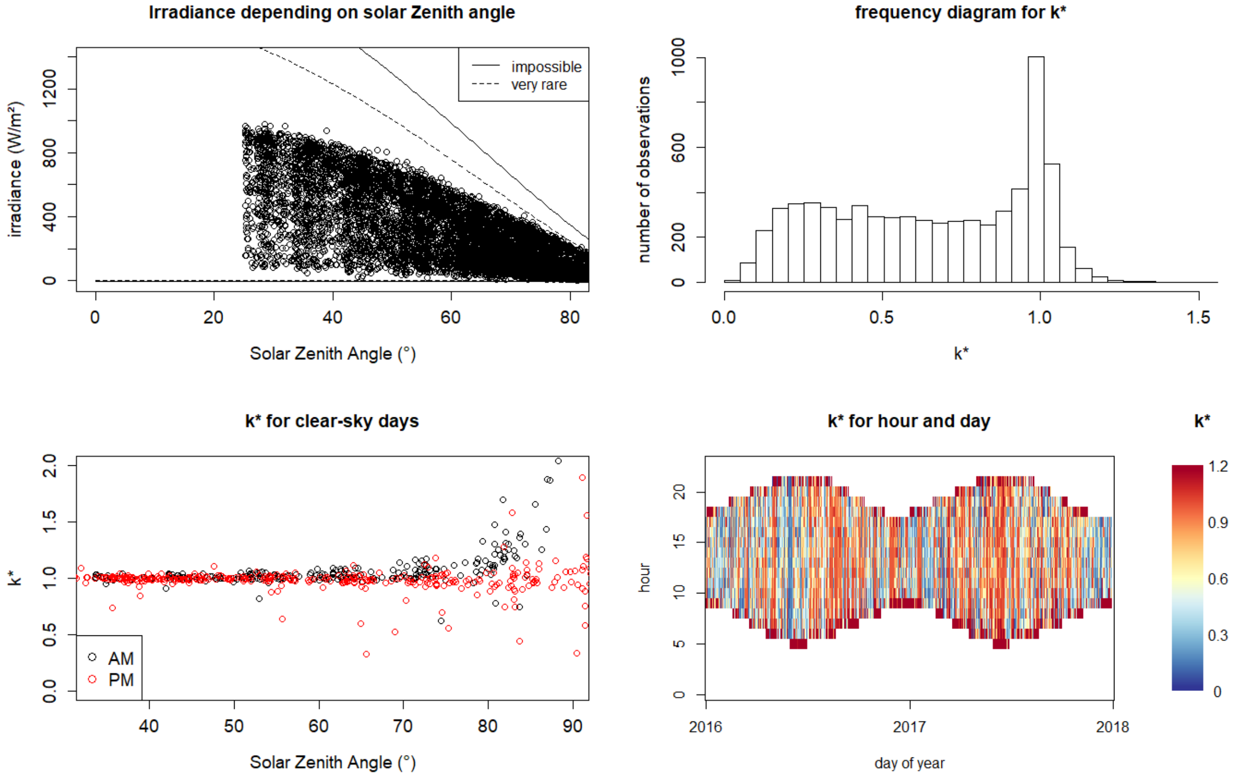


Figure A.14: Palaiseau

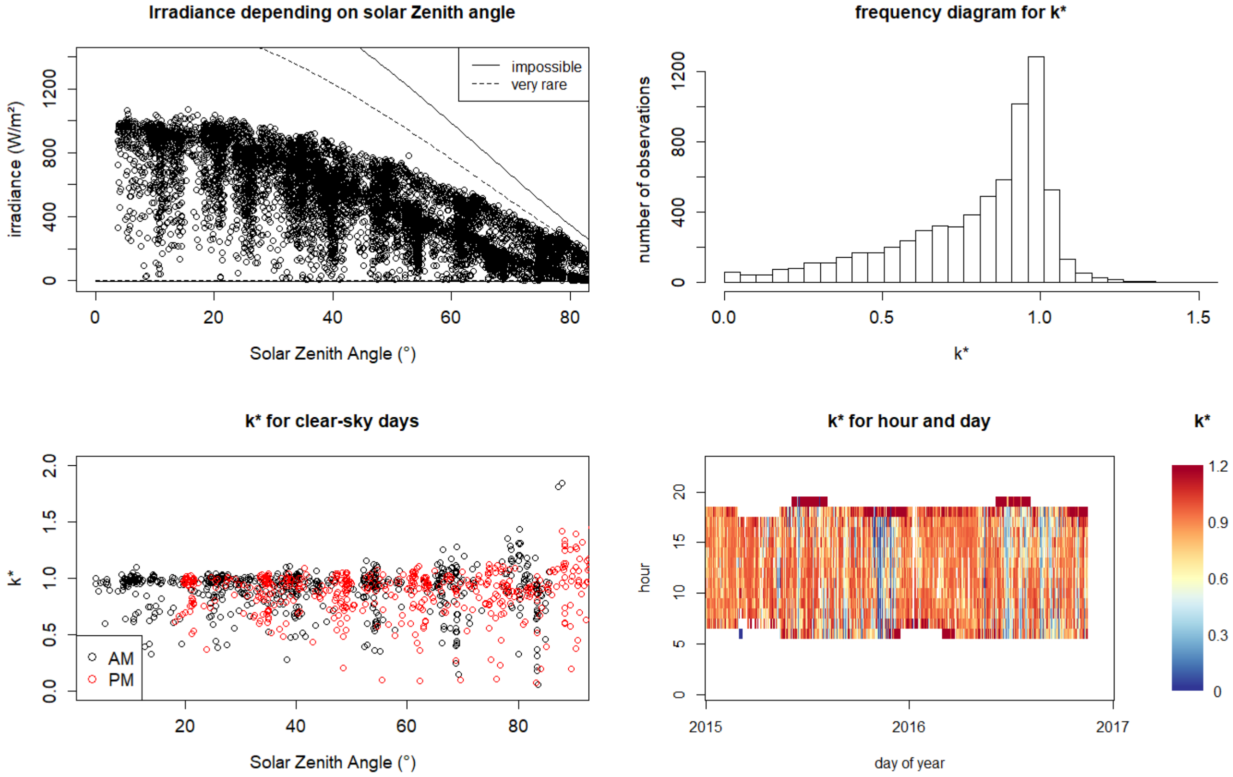


Figure A.15: Tiruvallur

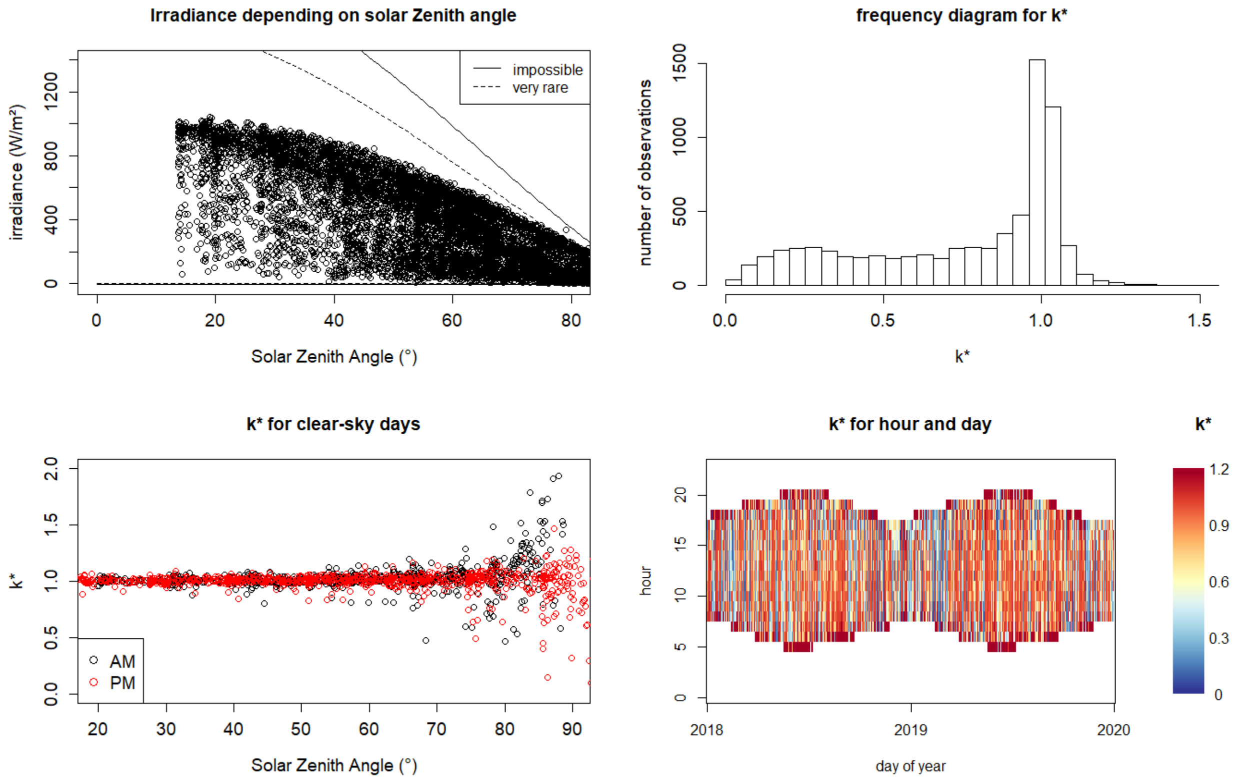
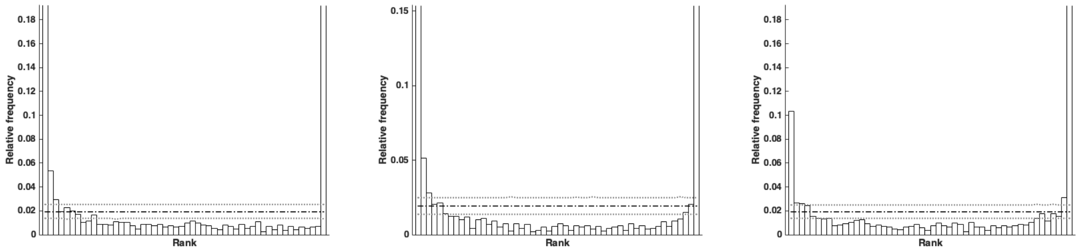


Figure A.16: Langley

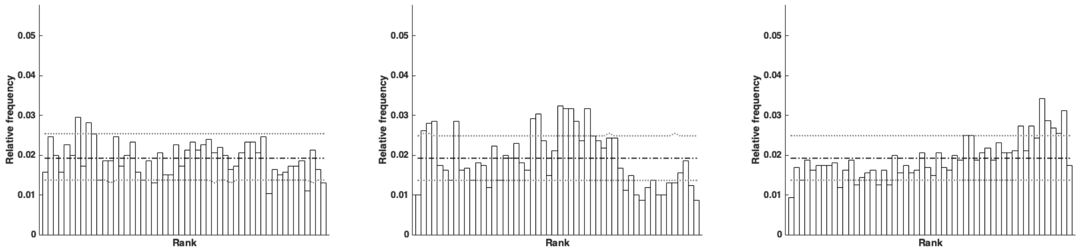
No major issues have been detected concerning the ~~three~~ six studied sites. ~~The only site with possible issues is the site of Saint-Pierre. It is noted that the 31/12/2012 for Saint-Pierre is the only day without data. Moreover, it is possible to see some disturbances at the very beginning and ending of the day (i.e. for low solar zenital angle).~~ For some sites (Tiruvallur, Langley, Saint-Pierre), it is possible to guess that some reflexions occur for extreme hours and some seasons. This leads to the phenomenon of overirradiance where ~~(~~k^* can easily reach a value of 4~~)~~

Appendix B. ~~Rank histograms~~



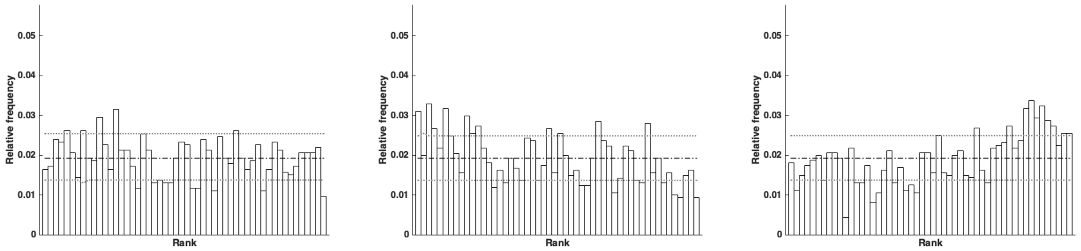
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.17: Rank histograms for raw ensemble forecasts



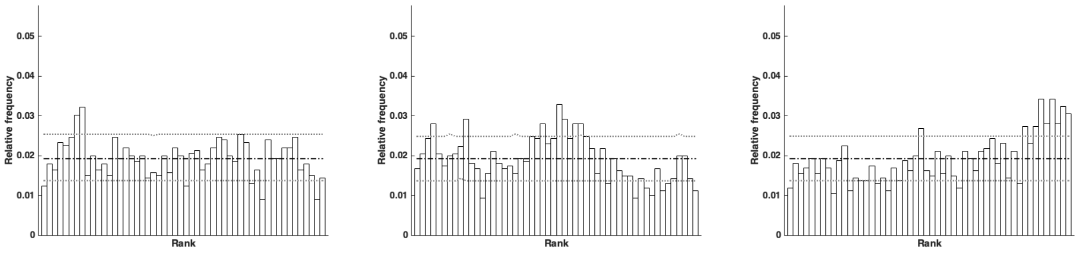
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.18: Rank histograms for $AnEn_c$



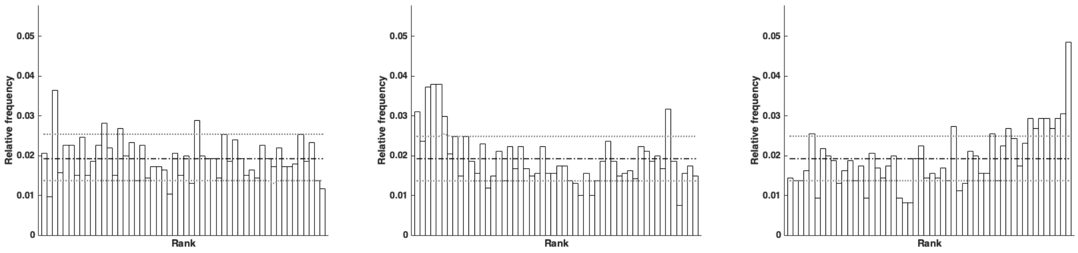
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.19: Rank histograms for LQR_c



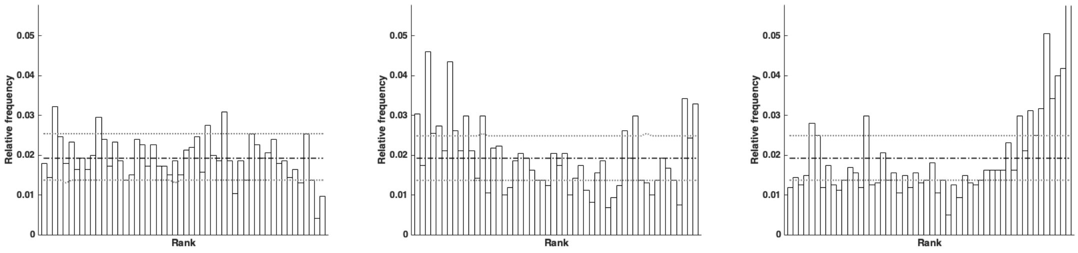
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.20: Rank histograms for $AnEn_m$



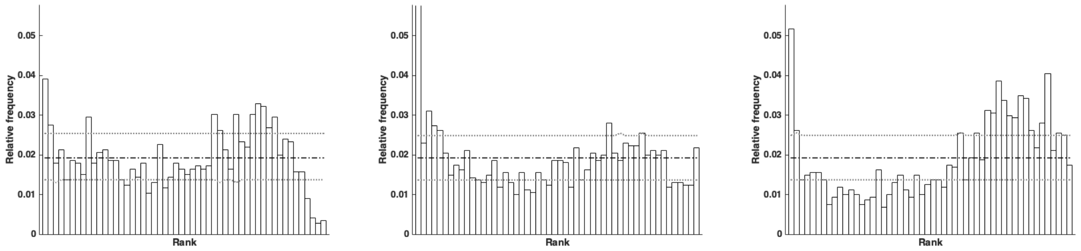
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.21: Rank histograms for LQR_m



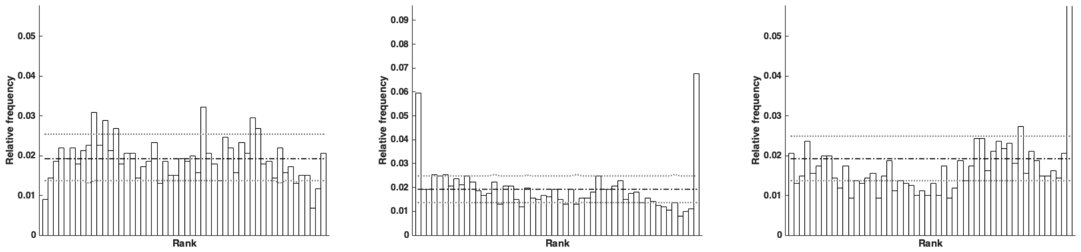
(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.22: Rank histograms for LQR_s



(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.23: Rank histograms for *NGR*



(a) Hawaii (b) Desert Rock (c) Saint-Pierre

Figure B.24: Rank histograms for *NR_GEV*

Appendix C. Bias and standard deviation of EPS members distribution and observations for the ~~three~~ six sites

The definition of the probabilistic forecast presented in section 2.3.1 is often under-dispersive, and consequently obtains poor scores. The associated rank histograms usually get characteristic U-shapes, with overpopulated extreme ranks. In this section, we attempt to demonstrate why a calibration procedure is needed for raw forecasts. To this end, a comparison between members distributions and observation distributions depending on the level of forecasting has been conducted for the 2 first statistical moments. These plots show clearly under-dispersive raw ensembles. The standard deviations need to be corrected. The discrepancy between distributions of members and observations indicates a statistical inconsistency between observations and forecasts, and therefore a bad reliability, and justifies the use of calibration models.

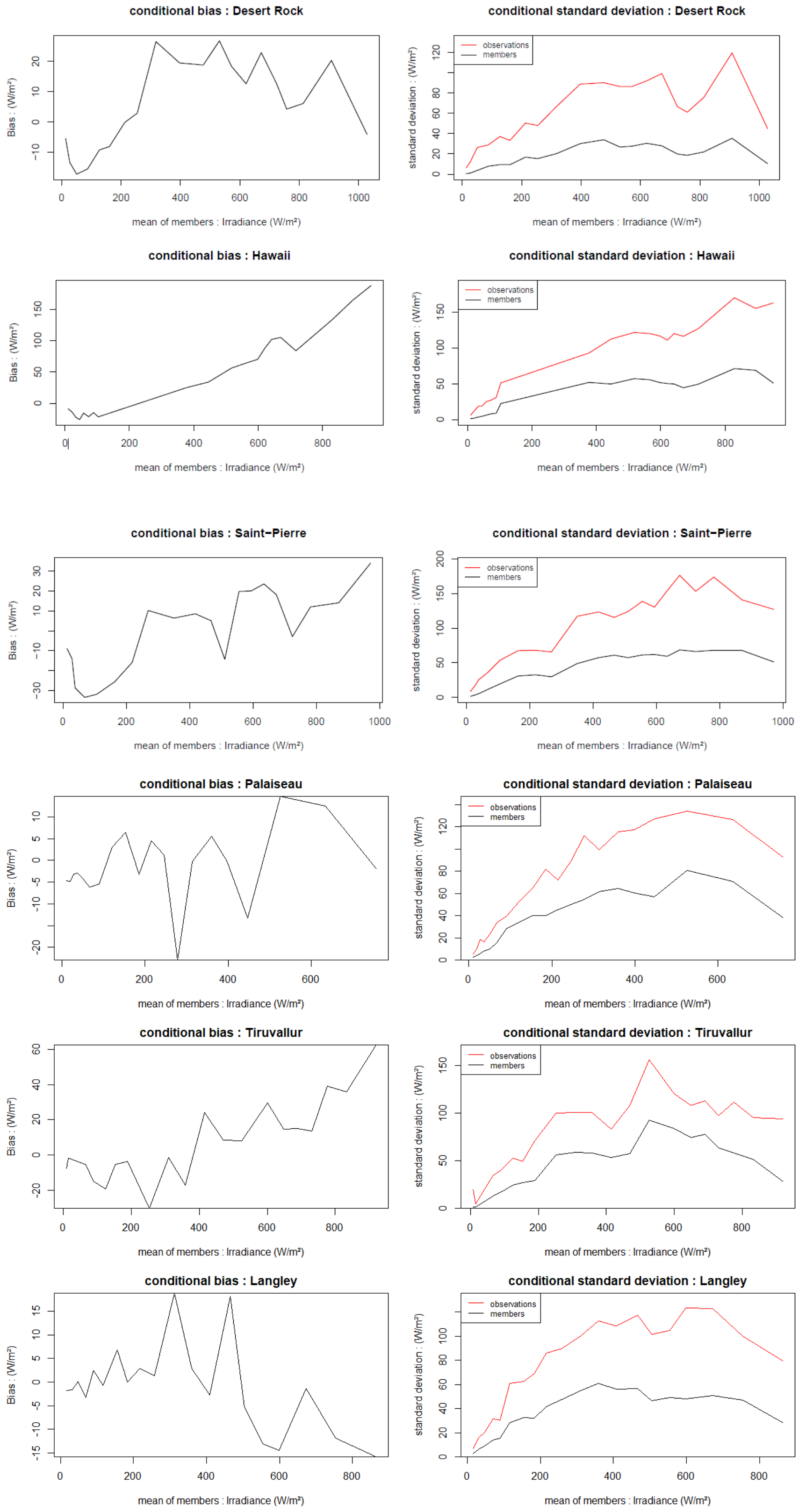
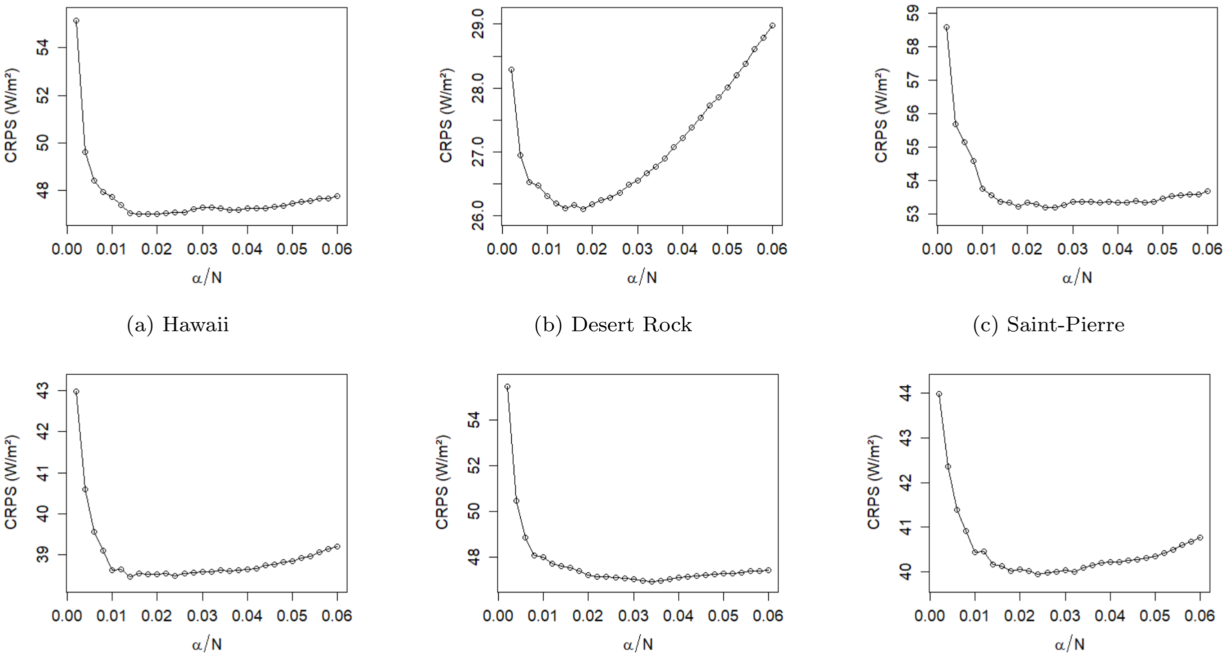


Figure C.25: Bias and standard deviation of EPS members distribution and observations for the ~~three~~ six sites, depending on the level of forecasting

Appendix D. Selection of the optimal α

Figure D.26 presents the results related to the optimal selection of the parameter α. As shown by Figure D.26 , regardless of the site under study, the optimal value corresponds to the minimum of the CRPS calculated on the training evaluation set.



(a) Hawaii

(b) Desert Rock

(c) Saint-Pierre

(d) Palaiseau

(e) Tiruvallur

(f) Langley

Figure D.26: Determination of the α. The optimal value corresponds to the minimum of the CRPS.