A new climatology reference model to benchmark probabilistic solar forecasts

Josselin Le Gal La Salle, Mathieu David, Philippe Lauret

▶ **To cite this version:**

Josselin Le Gal La Salle, Mathieu David, Philippe Lauret. A new climatology reference model to benchmark probabilistic solar forecasts. Solar Energy, 2021, 223, pp.398-414. 10.1016/j.solener.2021.05.037 . hal-03268282

HAL Id: hal-03268282
https://hal.univ-reunion.fr/hal-03268282

Submitted on 13 Jun 2023

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

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

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0038092X21004072
Manuscript_58e1f2aa578b8caa30611c51e2f54224

A new reference model to benchmark probabilistic solar forecasts

Josselin Le Gal La Salle[a], Mathieu David[a], 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*

Abstract

Probabilistic solar forecasting is becoming a major topic in the solar research community as it provides more information about the uncertainty of the forecast compared to deterministic forecasting. However, to facilitate the adoption of probabilistic forecasts within solar forecasting communities (industry and academic), the definition and the use of standardized best practices are a prerequisite. Among others, there is a need for benchmark models that are able to properly assess the performance of new probabilistic forecasting methods. In this work, we propose a new benchmark model called "CSD-CLIM" (for Clear-Sky Dependent Climatology). This reference model is evaluated against two other climatology benchmark models namely the naive climatology and a well-referenced model in the literature, the CH-PeEn (for Complete History Persistence Ensemble). The verification of compliance with a set of properties that a climatology benchmark model must follow demonstrates that the new CSD-CLIM model outperforms the naive climatology and that it can be a viable alternative to the CH-PeEn model. It is shown that the better performance of CSD-CLIM is due to a specific binning of the historical irradiance data based on the clear-sky irradiance values.

Keywords: `Benchmarking`, `Solar irradiance`, `Probabilistic forecasting`, `Climatology`, `Reference model`

Contents

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

© 2021 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

1. Introduction

It is now commonly accepted that solar forecasting is a cost-effective way to increase the share of solar energy in the electrical grid (Pierro et al., 2019). Recently, there has been a growing interest in solar probabilistic forecasting (Van der Meer et al., 2018). Indeed, contrary to deterministic forecasts, probabilistic forecasts of intrinsically highly variable weather predictands like wind or solar bring more value to the grid operator as demonstrated by (Zhu et al., 2002) or (Buizza, 2008).

However, as for deterministic forecasts, a wide adoption of probabilistic forecasts in the solar forecasting communities (industry and academic) requires a set of best practices. For instance, one can cite first the existence of a specific framework for verifying the quality of solar probabilistic forecasts. Raising concerns about the verification of probabilistic forecasts and notably the use of improper scores to measure the performance of the probabilistic methods, Lauret et al. (2019) have recently recommended a set of diagnostic tools and numerical scoring rules like the Continuous Ranked Probability Score (CRPS) to assess the quality of solar probabilistic forecasts.

A second point is related to the use of well-accepted reference models to fairly benchmark any new proposed forecasting methods on preferably standardized datasets (Yang et al., 2020b). In the realm of solar probabilistic forecasts, a reference model called the Persistence ensemble (PeEn) model (Alessandrini et al., 2015) is routinely proposed to benchmark new probabilistic models (David et al., 2016). Unfortunately, Doubleday et al. (2020) and Yang (2019) noted a wide spectrum of implementations of the PeEn model in the literature. Besides, Yang (2019) emphasized the need for universal benchmark models whose implementation must depend only on the data at a particular site. This is why Yang (2019) proposed a universal benchmarking model, the Complete History-Persistence ensemble (CH-PeEn). As stated by Yang (2019), the CH-PeEn model constitutes a consistent baseline model for assessing the skill of a forecasting method. The requirement of such benchmarks models has been also highlighted in Doubleday et al. (2020). In their work, the authors compared ten variants of six reference models, and the class of climatology reference models such as naive (no-skill) classical climatology and the CH-PeEn were implemented.

Similarly to skill scores used in the case of deterministic forecasts (Yang et al., 2020a), skill scores like the CRPS skill score, or "CRPSS" can be used to gauge the performance of a new probabilistic forecasting model against a reference easy-to-implement method. However, as noted by Yang (2019), different implementations of the benchmark model can hamper the interpretability of the skill score. Therefore, the computation of these skill scores should be done using a universal well-accepted benchmark model. Such a practice will promote a fair evaluation of probabilistic forecasting techniques.

Let us stress here the importance of using skills scores. Indeed, a score like CRPS obtained by a forecasting method is not in itself a measure of the skills of the forecast as the score strongly depends on the sky conditions of the considered location. For example, Alessandrini et al. (2015) pointed out that meteorological conditions of a site impact the quality of solar probabilistic forecasts. Hence, the CRPS score must be compared with the CRPS score of the reference model. The latter is expected to reflect the difficulty of forecasting at a particular site or equivalently to quantify the predictability of the solar irradiance at that specific location.

Following the need of easy-to-implement (naive) and universal reference models, it has been shown in (Murphy, 1973) that one component of the CRPS called the uncertainty corresponds to the CRPS of the climatology. The score of this naive climatology 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. Thus, one way to avoid a CRPSS that depends on the implementation of the reference model is to use the uncertainty part of the CRPS as the baseline value. Moreover, it must be noted that, for meteorologists, the baseline model for computing skill scores is usually the climatology - see for instance (Cusack and Arribas, 2008) or (Binter, 2012).

However, while appealing, we will show in this work that the naive climatology is not the best candidate for being a reference model for the particular case of solar irradiance forecasting. Indeed, the raw GHI time series exhibit specific diurnal and seasonal patterns which are not taken into account by the naive climatology model. Consequently, these deterministic patterns increase the CRPS of this benchmark model (denoted hereafter UNC)[1] Hence, it would be desirable to design a benchmark model which does not suffer from this issue.

In this work, we propose to take advantage of the clear-sky irradiance in order to compute a new reference model called the Clear-Sky dependent climatology or "CSD-CLIM". Unlike the CH-PeEn reference model which relies on hour-dependent predictive distributions, the CSD-CLIM model makes use of a binning process of the clear-sky irradiance to compute its CRPS. Additionally, it will be shown that, unlike CH-PeEn, the score of the CSD-CLIM can be directly computed from the historical data at hand without needing to first form the predictive distribution and then calculate the score on the historical dataset.

We will show also that the new model improves on the notion of universal benchmarking. Besides, we will demonstrate that the CRPS of CSD-CLIM and CH-PeEn can reflect the difficulty of forecasting at a particular site.

This paper is structured as follows. Section 2 discusses the required properties that a good benchmark model should exhibit. Section 3 presents the context of the study and in particular, the data and sites used to assess the performance of the different climatology benchmark models. Section 4 details the benchmark models while section 5 verifies the compliance of the reference models with the required properties. A discussion related to the score obtained by the CSD-CLIM model and a detailed comparison of the methodologies pertaining to the CSD-CLIM and CH-PeEn models is conducted in section 6. Finally, section 7 will present our conclusions.

2. Required properties for a good climatology benchmark model

Particular attention must be paid to the selection of a benchmark model and the amount of information used to feed it. Since a benchmark model is mainly used to calculate a skill score of a new forecasting method [2], choosing a benchmark model requires addressing these questions:

1. What information should be given by a skill score ?
2. What score should correspond to the "0" skill score?

This could be subject to discussion, but in our opinion, the purpose of a skill score should be to indicate which part of the information given by the new forecast is naive (i.e. captured by the benchmark model), and to what extent it provides valuable extra-information (which should be credited to the particular skill of the forecast).

[1] Note that UNC is also the uncertainty component of the CRPS.

[2] For a negatively oriented score like the CRPS for instance, the value of the skill score ranges from $-\infty$ for the worst forecast to 1 for a perfect forecast.

Furthermore, the "0" skill score should be defined by the best possible exploitation of all information derived from historical observations (i.e. the climatology). Thus, all the historical data should be exploitable by the benchmark model. Conversely, all extra-information (from meteorological data, satellite observation, etc.) treated by the new forecasting method and the potentially associated better performance should be credited to the merit of the forecast.

From the answer of these questions, we propose in this section to establish a set of required properties that benchmark models should meet. Doubleday et al. (2020) and Yang (2019) have already highlighted four properties that we retain in this study. Table 1 lists these four attributes. We propose also in Table 1 two additional properties that we believe a benchmark model should have. This set of properties is intended to reflect the discussion conducted above about the role of a benchmark model.

Code	Property
P1	The benchmark model should be easy-to-implement
P2	The implementation of the model must depend exclusively on the historical data at hand (and not on any other kind of parameters such as number of past measurements, forecast horizon/lead time)
P3	The score (CRPS) of the model (for a specific location and time resolution of the data) must be unique (or near unique) irrespective of the period or length of the period used to compute the score ("Time-invariance property")
P4	The model should verify the statistical consistency of the naive climatology (i.e. a perfect reliability when compared to new observations)
A1	The quality in terms of CRPS of the benchmark model should be as high as possible
A2	The score of the benchmark model should reflect the difficulty of forecasting at a particular location

Table 1: Desired properties of a good climatology benchmark model. "P" refers to properties already mentioned in the literature. We propose in this study to add the properties denoted by "A".

Consequently, this set of properties implies the following exclusions : P2 excludes the model PeEn ("Persistence Ensemble") developed by (Alessandrini et al., 2015) and used for example in (David et al., 2016; Lauret et al., 2017; Pedro et al., 2018). P2 and P4 exclude raw ensemble forecasts used in (Golestaneh et al., 2016) and (Thorey et al., 2018) and other benchmark models based on simple post-processing of raw ensemble forecasts.

Note that, in this study, we focus on the class of climatology reference models which can be used to benchmark either intra-hour or hourly forecasts according the terminology used by Doubleday et al. (2020).

3. Context of the study

3.1. Data

A selection of 20 sites serves as support for the comparison of the different benchmark models. This choice was made by trying to keep the widest possible spectrum in terms of different sky conditions and locations around the world. The vast majority of data was chosen from BSRN (HTTPS://BSRN.AWI.DE/) collection data in order to minimize the differences in data acquisition and data validation between sites. The remaining data comes from previous works dedicated to state-of-the-art of solar forecasting (David et al. (2016, 2018); Le Gal La Salle et al. (2020)). The complete list of sites is given in Table 2. The observation data considered are global horizontal irradiance (GHI) measurements. As BSRN data is submitted to strict data quality checks, we consider that the quality of data is good. Note that the quality checks of all BSRN data are available in BSRN website. For quality checks of data which is not part of BSRN, please refer to Le Gal La Salle et al. (2020).

	Provider	Data	Longitude	Latitude	Climate	Acronym
Saint-Pierre	PIMENT	2012-2013	55.49	-21.34	Insular tropic	SPI
Hawaii	NREL	2010-2011	-158.08	21.31	Insular tropic	HAW
Desert Rock	BSRN	2012-2013	-116.03	36.62	Desert	DRO
Fort Peck	BSRN	2012-2013	-106.50	48.00	Continent	FPE
Fouillole	LARGE	2010-2011	-61.52	16.22	Insular tropic	FOU
Payerne	BSRN	2017-2018	6.94	46.82	Mountainous	PAY
Palaiseau	BSRN	2016-2017	2.21	48.71	Temperate	PAL
Toravere	BSRN	2017-2018	26.46	58.25	Temperate	TOR
Adelaïde	BSRN	2016-2017	138.51	-35.00	Desert	ADE
Tiruvallur	BSRN	2015-2016	79.97	13.09	Monsoon	TIR
Sioux Falls	BSRN	2017-2018	-96.62	43.73	Continent	SXF
Nauru Islands	BSRN	2011-2012	166.92	-0.52	Insular tropic	NAU
Marshall Islands	BSRN	2014-2015	167.73	8.72	Insular tropic	MAR
Barrow	BSRN	2015-2016	-156.61	71.32	Arctic	BAR
Cocos Islands	BSRN	2017-2018	96.84	-12.19	Insular tropic	COC
Manus Islands	BSRN	2011-2012	147.43	-2.06	Insular tropic	MAN
Tenerife	BSRN	2017-2018	-16.5	28.31	Temperate	TEN
Minamitorishima	BSRN	2017-2018	153.98	24.29	Insular	MIN
Bermuda Islands	BSRN	2011-2012	-64.67	32.27	Insular	BER
Langley	BSRN	2018-2019	-76.39	37.10	Temperate	LAN

Table 2: Characteristics of sites used for the study

3.2. Clear-sky model

The building of the CH-PeEn and CSD-CLIM benchmark models (see section 4) requires a clear-sky model. As demonstrated by Yang (2019) who compared two clear-sky models in the assessment of the performance of the CH-PeEn model, the choice of the clear-sky model is of primary importance (See also (Yang, 2020)).

Among the different clear-sky models that can be found in the literature, one can cite the Bird Model (Bird and Hulstrom, 1981), the Ineichen-Perez model (Ineichen and Perez, 2002) or the McClear model (Lefèvre et al., 2013). In this work, and following the work of Yang (2019), we have selected the McClear clear-sky model. McClear Clear-sky GHI estimates are publicly available at 1-min resolution from the CAMS (Copernicus Atmosphere Monitoring Service) McClear Service (WWW.SODA-PRO.COM).

3.3. Generating the predictive cumalative distrubution function (CDF)

The computation of the CRPS requires the building of the predictive cumulative distribution function (CDF) $\hat{F}_{fcst}$ (see Equation 3). In this section, we describe briefly how these forecast CDFs can be generated either for benchmark models or for Ensemble Prediction System (EPS).

3.3.1. Generation of climatology predictive distributions

For instance, the naive climatology forecast is an empirical (CDF) based on long period of N historical sorted measurements $(Y_1, Y_2, \cdots, Y_N)$. Here, to generate the predictive CDF for the benchmark models, we implement the classical approach (Lauret et al., 2019) that consists in building a piecewise constant function with a jump probability of $\frac{1}{N}$ at each Y_i and null probabilities for events outside the set of historical measurements. The predictive CDF is given by

$$\hat{F}_{fcst}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{x \geq Y_i\}}, \tag{1}$$

where $\mathbb{1}_{\{u\}}$ is the indicator function which has the value of 1 if its argument u is true and 0 otherwise. Note that in the case of the CH-PeEn model, predictive CDFs are built with a training set of historical ordered measurements for each hour of the day.

3.3.2. Generation of the day-ahead LQR probabilistic forecasts

Property A2 states that the CRPS of the benchmark model should be a proxy for judging a priori the quality of a probabilistic forecast. In order to be able to evaluate the proposed benchmark models regarding this property, we generate day-ahead probabilistic forecasts for the different sites listed in Table 2.

The day-ahead GHI ensemble forecast has been provided by the European Centre of Medium-Range Weather Forecasts (ECMWF). This ensemble forecast also called EPS (for ensemble prediction system) is constituted of 51 members : one unperturbed member (control member) and 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 times series recorded on-site are compared with the nearest ECWMF pixel.

For an EPS with M ordered members $(E_1, E_2, \cdots, E_M)$, we use again the classical approach described in Lauret et al. (2019) to build the predictive CDF related to the raw

ECMWF ensemble forecast which reads as :

$$\hat{F}_{fcst}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\{x \geq E_i\}}. \quad (2)$$

However, numerous studies (see for instance Vannitsem et al. (2018)) have shown that CDFs drawn from raw EPS ensemble with the classical construction are statistically unreliable, meaning that the probability assigned to an event is not consistent with the observations (Hamill and Colucci, 1997). Hence, the use of calibration models (i.e. techniques which improve the reliability of raw ensemble forecasts) is a common practice. The interested reader is referred to (Gneiting et al., 2005) or (Le Gal La Salle et al., 2020) for details regarding the implementation of calibration techniques.

In this study, we propose to use a state-of-the-art non-parametric and very flexible calibration method, the Linear Quantile Regression (LQR) technique. The LQR method is depicted at length in (Le Gal La Salle et al., 2020).

3.4. Verification of probabilistic forecasts

In the verification framework proposed by Lauret et al. (2019), the authors recommend the computation of a proper score like the Continuous Ranked Probability Score (CRPS) to evaluate the overall quality of a probabilistic forecast. We will recall here the mathematical definition of the CRPS.

The CRPS measures the distance between the forecast CDF and the CDF associated with the measurement x_{obs} (Hersbach, 2000). The CRPS is defined as

$$CRPS = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} \left[\hat{F}^i_{fcst}(x) - F^i_{x_{obs}}(x) \right]^2 dx. \quad (3)$$

$\hat{F}_{fcst}(x)$ is the predictive CDF and $F_{x_{obs}}(x)$ is the cumulative distribution given by the Heaviside (or step) function $H(x - x_{obs})$, which is zero if $x < x_{obs}$ and one if $x \geq x_{obs}$. The squared difference between the two CDFs is averaged over the N forecast/observation pairs. The CRPS is negatively oriented (smaller values indicate a better forecast). Like the Brier score (see Appendix for the definition of the Brier Score), the CRPS is a proper score and can be decomposed into the three important attributes detailed in Appendix B. The decomposition is as follows

$$CRPS = REL - RES + UNC, \quad (4)$$

where REL, RES and UNC are respectively the reliability part, the resolution part and the uncertainty part of the CRPS.

Reliability is an indication of the statistical consistency between the forecasts and the observations while resolution indicates how far the observations are discriminated from the climatological mean by the forecasts. Finally, the uncertainty term depends on the variability of the observations and will be further developed in section 4.1. As mentioned in the introduction, the uncertainty term corresponds theoretically to the CRPS of the naive climatology. Indeed, if we assume an infinite historical time series, the reliability is perfect (i.e. $REL = 0$) and the resolution is null (i.e. $RES = 0$). More details regarding the calculation of the different components of the CRPS can be found in Appendix B, Appendix C and to Lauret et al. (2019).

Note that, in order to assess the reliability and sharpness properties of the different benchmark models, Doubleday et al. (2020) used visual diagnostic tools like reliability and sharpness diagrams. In this work, we rely on the decomposition of the CRPS in order to obtain a quantitative measure of these two attributes.

Finally, and as mentioned in the introduction, the purpose of the benchmark models is mainly to be used as references in the calculation of skill scores. A skill score is the level of improvement of a forecasting model over the reference model. For example, the CRPS skill score reads as

$$CRPSS = 1 - \frac{CRPS_{model}}{CRPS_{reference}}. \tag{5}$$

4. Climatology benchmark models for solar probabilistic forecasting

In this section, three reference models, the naive climatology (CLIM), the Clearsky-dependent uncertainty (CSD-CLIM) and the Complete-History Persistence Ensemble (CH-PeEn) are presented and their possible pros and cons are discussed.

4.1. The naive climatology (CLIM)

The climatology refers to the ensemble of all the observed values of a weather variable over a long period of time. The predictive CDF created from the aggregation of all these past observations forms the climatological predictive distribution. We denote hereafter the corresponding model as the naive climatology model (CLIM). The CRPS of the CLIM model can be computed either by using Equation 3 or by computing the uncertainty part of the CRPS which reads as (Todter and Ahrens, 2012)

$$UNC = \int_0^{GHI_{MAX}} UNC_{BS}(x)dx. \tag{6}$$

$UNC_{BS}(x)$ is the uncertainty relative to the Brier Score (BS) (see Appendix B for the decomposition of the Brier Score) for a fixed level of irradiance x and GHI_{MAX} is the maximum possible value of irradiance (also called climatological bound) proposed by (Yang, 2019) et (Long and Shi, 2008).

Let Y be the measurement of the predictand (here the GHI). For a fixed level x of GHI, *$UNC_{BS}(x)$* is defined as

$$UNC_{BS}(x) = \overline{o}(x)\bigg(1 - \overline{o}(x)\bigg), \tag{7}$$

where $\overline{o}$ is the frequency with which Y is lower or equal to x i.e.

$$\overline{o}(x) = \frac{\mathbb{1}_{\{Y \leq x\}}}{N}, \tag{8}$$

where N is the number of historical observations and $\mathbb{1}_{\{u\}}$ the indicator function.

Note that Equation 6 allows a graphical representation of the UNC which will be extensively used in this study. For each level of GHI x, a point equal to $UNC_{BS}(x)$ is plotted, and UNC is given by the area under the curve created by this set of points. Such a representation is given in Figure 1.

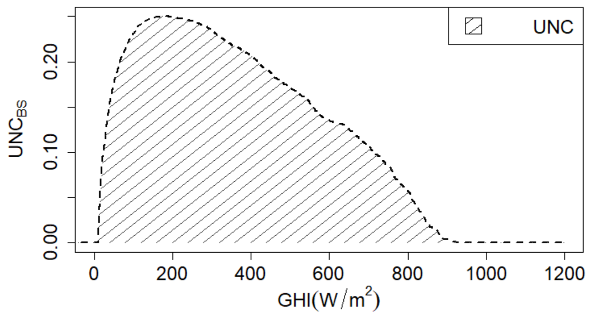


Figure 1: Uncertainty of the Brier Score in relation with the level of GHI for the site of Payerne. The uncertainty part of the CRPS (UNC) is given by the shaded area.

4.2. The clearsky-dependent Climatology (CSD-CLIM)

In the meteorological community, the naive climatology depicted above is used as a reference model for the calculation of skill scores. However, it does not account for the strong time dependence experienced by a variable such as GHI (daily and seasonnal patterns).

It is well-known that the irradiance can be decomposed as

$$GHI = GHI_{CS} * k_t^*, \tag{9}$$

where GHI_{CS} stands for the clear-sky irradiance and represents the irradiance if no cloud cover is observed, and is given by a chosen clear-sky model (see section 3.2). The clear-sky irradiance is fundamentally time-dependent and follows strong temporal patterns.

k_t^* is the clear-sky index and ranges theoretically between 0 to 1[3] and represents the share of GHI_{CS} which is lost due to cloud cover. Equation 9 shows that forecasting k_t^* and forecasting irradiance are equivalent tasks. Indeed, GHI_{CS} is deterministically and fully determined by the clear-sky model. However, this is not taken into account by the naive climatology, which only reflects the variability of the GHI.

For example, a site which always experiences clear sky conditions (i.e $k_t^* = 1$) will exhibit a high uncertainty score UNC, even though the forecast is not difficult (knowing that $k_t^* = 1$). Also, for various cases (e.g. early morning, late evening), the clear-sky model alone, which

[3]In practice, cloud enhancement events (i.e. multi-reflections of the sun beams by the clouds) can produce over-irradiance with clear-sky indices superior to 1.

is not the uncertain part of the forecast, permits to determine that GHI is limited to low-irradiance values. These considerations lead to the idea that k_t^* and the clear-sky model should play a key role in the building of a benchmark model.

This is why we propose here a new benchmark model that exploits the periodic variations of the clear-sky model. In order to get rid of this strong temporal pattern, we propose to bin the climatology according the clear-sky irradiance values and to call this baseline model the clearsky-dependent climatology ("CSD-CLIM"). For each bin, CSD-CLIM computes the CDF in a similar manner as the climatology does, using only the historical data belonging to the bin. Considering the close relationship between this new model and the naive climatology, we propose to call the score of this benchmark model the "clearsky-dependent uncertainty" or " *CSD-UNC*", just as the score of the naive climatology is called uncertainty (UNC) in the literature (see for instance Todter and Ahrens (2012)).

Thus, the score (i.e. the CRPS) of this new benchmark model can be calculated using Equation 3, or equivalently by computing

$$CSD\text{-}UNC = \sum_{i}^{Nb} f_i * UNC^i. \tag{10}$$

For each bin i of clear-sky irradiance values (see Table 3), we calculate UNC^i as defined by Equation 6 and the frequency f_i represents the relative frequency of each bin in the clear-sky model. The main goal of such a model is to discriminate the situations that the clear-sky model alone permits to separate. An example of the binning process for Desert Rock is shown in Table 3.

In this study, we chose a number of $Nb = 30$ bins. This choice could be questioned and is discussed in Appendix A.

bin i (W/m^2)	0-40	40-80	80-120	120-160	160-200	200-240
relative frequency f_i	0.06	0.07	0.06	0.05	0.04	0.04
bin i (W/m^2)	240-280	280-320	320-360	360-400	400-440	440-480
relative frequency f_i	0.04	0.04	0.02	0.03	0.02	0.02
bin i (W/m^2)	480-520	520-560	560-600	600-640	640-680	680-720
relative frequency f_i	0.02	0.05	0.05	0.04	0.04	0.05
bin i (W/m^2)	720-760	760-800	800-840	840-880	880-920	920-960
relative frequency f_i	0.06	0.04	0.04	0.02	0.02	0.02
bin i (W/m^2)	960-1000	1000-1040	1040-1080	1080-1120	1120-1160	1160-1200
relative frequency f_i	0.03	0.02	0.01	0.00	0.00	0.00

Table 3: Binning process for Desert Rock. The relative frequency f_i is the number of clear-sky irradiance values in the bin out of the total number of irradiances values.

Let us stress here that there is no need to form a predictive CDF and compute its CRPS score with Equation 3. Instead, Equation 10 together with Equation 6 fully determine the score *CSD-UNC* of the proposed new reference model CSD-CLIM. Hence, based on the

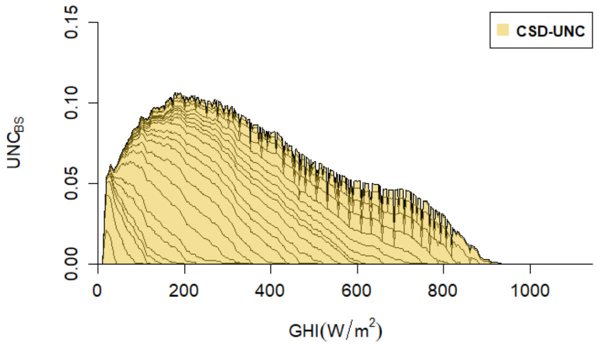
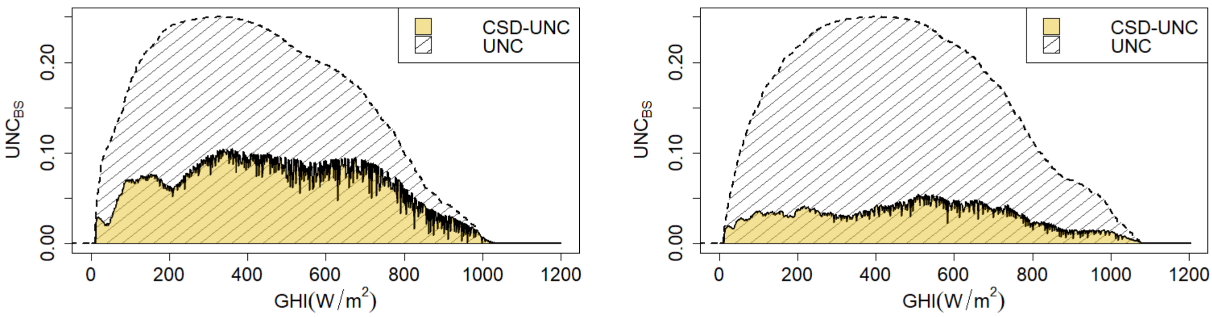


Figure 2: Construction of *CSD-UNC* for Payerne. The sum of the 30 colored areas gives *CSD-UNC*

historical data at hand (property P2 verified), we can argue that the CSD-CLIM model meets the "easy-to-implement" property P1.

Graphically, we propose to represent this *CSD-UNC* by stacking all the UNC^i scaled by their frequencies. As an illustration, Figure 2 shows the stacked 30 UNC^i. The sum of these areas amounts to the *CSD-UNC* score. It becomes then possible to graphically compare *CSD-UNC* and the classical *UNC* defined in section 4.1. Such a comparison for two sites (Marshall Island and Desert Rock) is done in Figure 3.



(a) Marshall Island

(b) Desert Rock

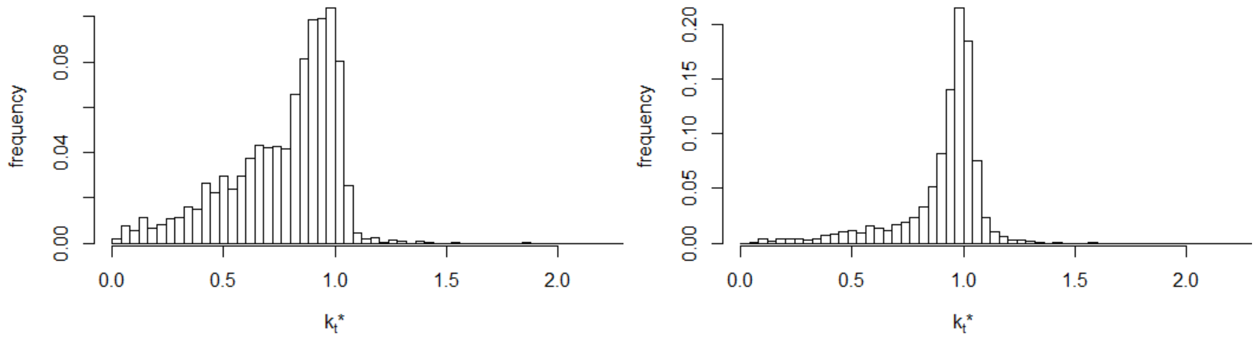
Figure 3: Comparison between *UNC* and *CSD-UNC*.

As shown by Figure 3, a substantial difference can exist between *UNC* and *CSD-UNC* (and by extension between the two reference models CLIM and CSD-CLIM). While for Marshall Island, *CSD-UNC* represents approximately 50% of *UNC*, the reduction is more drastic in the case of Desert Rock. This discrepancy can be explained with the histograms of k_t^* presented in Figure 4 which reveals that k_t^* is both lower and much more variable in Marshall Island than in Desert Rock. This has two consequences:

1. *UNC* is more important in Desert Rock, mainly because k_t^* is in average higher leading

to a distribution of higher values of GHI (see Figure 3)

2. The forecasting task, which essentially consists in guessing the most probable value of k_t^*, is much more difficult in Marshall Island.



(a) Marshall Island

(b) Desert Rock

Figure 4: Histograms of k_t^* for Marshall Island and Desert Rock

4.3. The Complete-History Persistence Ensemble (CH-PeEn)

By pointing out the deficiencies of the benchmark model widely used by the solar forecasting community namely the Persistence ensemble PeEn (Alessandrini et al. (2015), David et al. (2016)), Yang (2019) proposed a new reference model i.e. the Complete-History Persistence Ensemble (CH-PeEn). Unlike the PeEn model whose implementation depends heavily on the number of previous measurements used to build the predictive CDF, the proposed CH-PeEn baseline model uses the entire set of past measurements to form predictive distributions conditioned by each hour of the day. As discussed by Yang (2019), such a benchmark model like CH-PeEn will have a near unique CRPS and therefore will ease the interpretation of skill scores.

Regarding more precisely the implementation of the CH-PeEn model, based on the entire history of available data, the empirical CDF of the clear-sky indices k_t^* for each hour is formed. To compute the CRPS (see Equation 3) of the CH-PeEn model, the predictive GHI distribution is obtained by multiplying the empirical set of k_t^* of a given hour by the clear-sky irradiance value at this specific forecast hour. As a set of k_t^* is independently created for each hour, the resulting predictive CDF is strongly dependent on the local hour.

It must be stressed also that, unlike *CSD-UNC* whose computation does not require the construction of predictive GHI distributions, the derivation of CH-PeEn implies the generation of forecasts and the assessment of the CRPS of these forecasts.

5. Results

In this section, we verify the compliance of the different benchmark models with the properties listed in Table 1. It must be noted that, by construction, the 3 climatology benchmark models meet property P1. The goal of this section is then to find whether the models follow P2 and P3, and to evaluate the performance of the benchmark models in the light of properties P4, A1 and A2.

5.1. Compliance with property P2

As claimed by P2 in Table 1, the implementation of a good benchmark model should require a minimum number of parameters, for both simplicity of calculation and universality of results. The naive climatology is the only model which, by construction, indisputably fully respects this criteria. On the contrary, the other models presented here use some additional parameters.

CH-PeEn needs the time of the day to bin the data and a clear-sky model to compute the clear-sky indices k_t^*. The time of the day is not strictly speaking an additional parameter, because it does not depend on a discutable model. The time resolution of the bin could influence the results (note that this is also true for the naive climatology), but should be governed by the available data, and thus should depend only on the data at hand, as stated in P2. The choice of the clear-sky model could be more problematic, because different models could provide different results. However, it seems reasonable to accept this compromise given the benefits of using a clear sky model.

CSD-CLIM uses also a clear-sky model to bin the data. The same argument used to justify this usage for CH-PeEn is also valid for CSD-CLIM. Besides, *CSD-UNC* could also depend on the chosen number of bins N_b. We justify in Appendix D that the number of bins does not have a strong impact on the final result, as long as the number is chosen reasonably large.

Finally, apart from the above details of implementation, we can state that the three climatology benchmark models follow property P2.

5.2. Compliance with property P3

Property P3 states that all benchmark models must be time-invariant i.e. that their resulting CRPS (for a specific location and time resolution of the data) must be unique or near unique. The verification of the time invariance of the CSD-CLIM model is detailed in Appendix D. As demonstrated in it, we can conclude that the CSD-CLIM produces a near unique score regardless of the period or the length of the historical data used to compute it. The CH-PeEn is also time-invariant, as demonstrated by Yang (2019). Furthermore, it is well-known that the naive climatology is time-invariant, as soon as the amount of data considered for its computation is sufficiently large.

5.3. Compliance with property P4

Property P4 emphasizes the importance of reliability of the benchmark model. We recall that a climatology benchmark model should possess the same statistical consistency as for the naive climatology and therefore should exhibit a reliability component as close as possible to zero. In addition, while respecting the statistical consistency property, any other benchmark model should beat the naive CLIM model in terms of resolution.

As mentioned above, contrary to Doubleday et al. (2020) who used visual diagnostic tools like reliability and sharpness diagrams in order to assess the two important attributes of a forecasting scheme i.e reliability and resolution, we prefer here to rely on the quantitative decomposition of the CRPS. Table 4 details the decomposition of the CRPS into reliability and resolution obtained by the 3 reference models.

	Reliability (W/m^2)									
site	SPI	HAW	DRO	FPE	FOU	PAY	PAL	TOR	ADE	TIR
CSD-CLIM	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CH-PeEn	7.9	8.0	6.0	6.3	5.1	10.0	6.6	8.8	7.0	5.3
CLIM	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0
site	SXF	NAU	MAR	BAR	COC	MAN	TEN	MIN	BER	LAN
CSD-CLIM	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CH-PeEn	5.3	4.3	5.0	3.7	5.8	4.8	6.0	4.2	6.0	5.5
CLIM	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1
	Resolution (W/m^2)									
site	SPI	HAW	DRO	FPE	FOU	PAY	PAL	TOR	ADE	TIR
CSD-CLIM	113.2	107.1	140.4	89.1	63.9	85.0	72.0	66.7	104.0	110.2
CH-PeEn	119.3	109.4	145.0	92.8	67.7	88.3	74.3	70.4	109.0	113.1
CLIM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
site	SXF	NAU	MAR	BAR	COC	MAN	TEN	MIN	BER	LAN
CSD-CLIM	74.9	119.7	95.2	52.5	105.3	83.1	164.4	126.5	94.7	89.6
CH-PeEn	77.8	122.0	97.9	55.3	109.2	86.5	170.5	130.0	98.5	92.8
CLIM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 4: Decomposition of CRPS of the 3 tested benchmark models for all sites.

As expected, the reliability and resolution components of the naive climatology (CLIM) are zero or near zero. It appears also that while retaining the statistical consistency feature, the CSD-CLIM model improves on the resolution of the basic CLIM model. Conversely, it can be noted that the CH-PeEn exhbits a slighty higher resolution than the CSD-CLIM model but at the expense of a degradation in reliability.

5.4. Compliance with property A1

Additional property A1 states that the benchmark models should obtain the best possible quality using only historical data. The overall quality of the reference forecasts is measured here by the CRPS. Table 5 gives the CRPS of the 3 benchmark models for all the sites listed in Table 2.

Model	SPI	HAW	DRO	FPE	FOU
LQR forecast	51.36	46.82	25.58	39.37	77.89
CLIM	172.7	157.8	175.0	137.8	140.1
CSD-CLIM	59.5	50.7	34.6	48.7	76.2
CH-PeEn	61.2	56.2	35.6	51.2	77.4
Model	PAY	PAL	TOR	ADE	TIR
LQR forecast	40.61	36.14	30.95	41.40	42.49
CLIM	143.4	126.4	116.5	162.4	161.0
CSD-CLIM	58.3	53.5	49.8	48.4	50.8
CH-PeEn	64.8	58.6	54.6	60.4	52.8
Model	SXF	NAU	MAR	BAR	COC
LQR forecast	39.65	54.42	61.12	29.62	48.40
CLIM	137.4	175.7	167.3	88.8	166.8
CSD-CLIM	62.5	56.0	72.1	36.3	61.5
CH-PeEn	65.0	58.0	74.4	37.1	63.3
Model	MAN	TEN	MIN	BER	LAN
LQR forecast	68.25	40.87	33.09	44.16	35.16
CLIM	162.1	199.6	170.0	152.4	154.4
CSD-CLIM	79.0	35.2	43.5	57.7	64.8
CH-PeEn	80.1	35.2	44.0	59.7	67.3

Table 5: CRPS of *LQR* forecast (grey) and CRPS of the 3 benchmark models for all sites

As shown by Table 5, for any given considered site, the CSD-CLIM model exhibits the highest overall quality in terms of CRPS. The previous decomposition of the CRPS of the CSD-CLIM model shows that the overall better performance of CSD-CLIM originates from its high reliability. This finding strengthens the assumption that knowledge on clear-sky irradiance decreases the uncertainty[4] associated with a forecast.

5.5. *Compliance with property A2*

As mentioned in section 3.3.2, in order to verify the compliance of the benchmark models with the additional property A2, we generate day-ahead LQR calibrated forecasts. The assumption of this study is that the score of a benchmark model should be related to the quality of the LQR forecasts (measured here by its CRPS). Figure 5 plots the results of the three benchmark models (x axis) versus the CRPS of the LQR forecasts (y axis). Table 5 also gives the CRPS of the LQR forecasts.

[4]The word "uncertainty" is not used here to refer to the uncertainty term of the decomposition of the CRPS, but to the expected level of variability of the predictand, which a forecast model has to deal with.

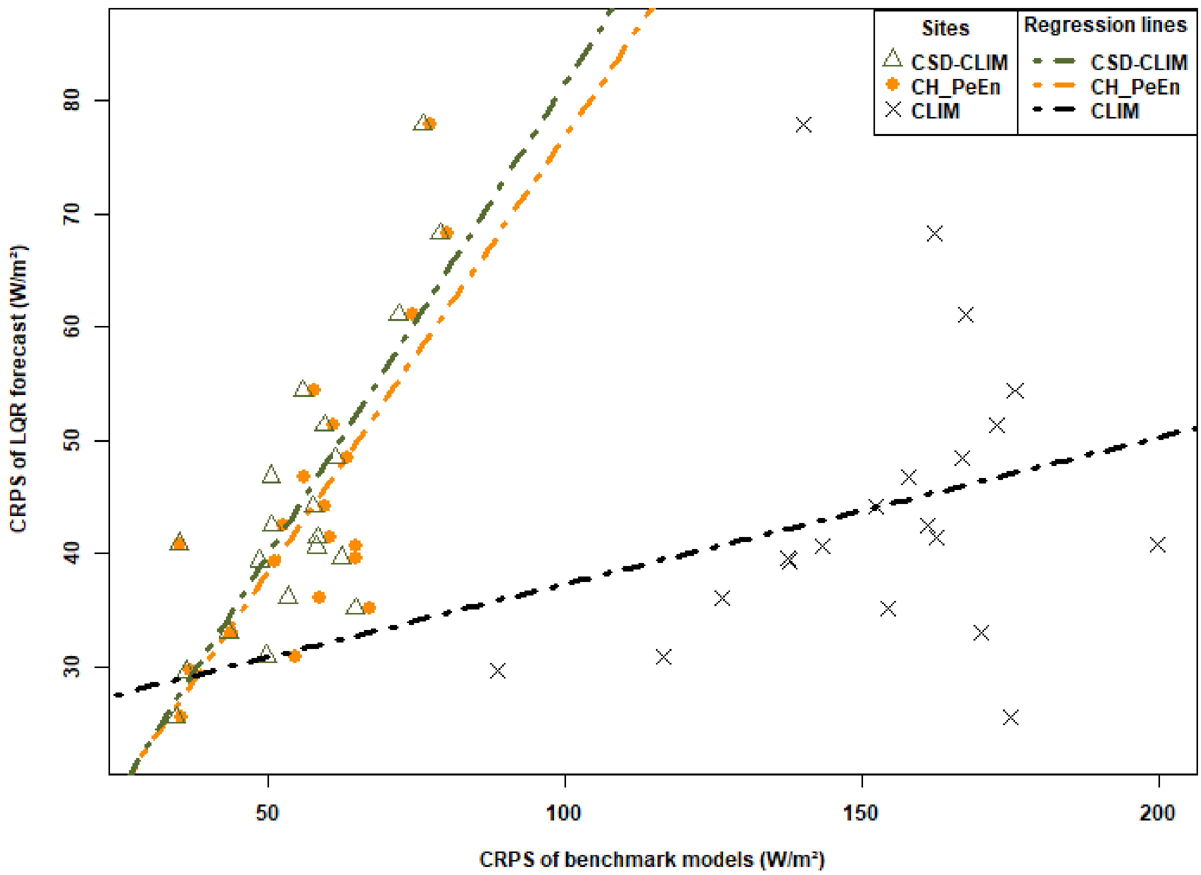


Figure 5: CRPS of the LQR forecasts vs CRPS of the benchmark models computed for each site of Table 2.

Figure 5 clearly shows that, unlike the score of the naive CLIM, the scores of CSD-CLIM and CH-PeEn can be good proxies for judging a priori the quality of a forecast obtained at a particular site. In other words, just like the RMSE score obtained by the clear-sky persistence reference model in case of deterministic forecasts, the score of these two reference models reflects the difficulty of forecasting at a particular location.

To proceed further, we built a linear regression for each considered benchmark model and extracted the coefficient of determination R^2. We found respectively a R^2 coefficient of 0.06, 0.56 and 0.63 for the CLIM, CH-PeEn and CSD-CLIM. Furthermore, it must be noted that a similar ranking has been established for other types of calibrated forecasting models described in Le Gal La Salle et al. (2020). Put differently, whatever the forecasting model, the best correlation is always obtained by CSD-CLIM. As a conclusion, we can state that CH-PeEn and CSD-CLIM outperform the naive climatology regarding property A2.

5.6. Overview

Finally, Table 6 gives an overview of the performance of each benchmark model in the light of properties P4, A1 and A2. Let us recall that all the climatology benchmark models discussed here meet the required rules P1, P2 and P3.

Model	P4 Statistical consistency	A1 Overall quality of the model	A2 Proxy indicator for forecast difficulty
CLIM			
CH-PeEn			
CSD-CLIM			

Best Model 2nd best Model Worst Model

Table 6: Overall qualitative assessment of the benchmark models.

The qualitative results of Table 6 suggest that CSD-CLIM should be preferred as a reference model. It appears to lead to the best trade-off between all the properties required by a climatology benchmark model (see section 2).

6. Discussion

6.1. Discussion on CSD-UNC, the CPRS obtained by CSD-CLIM

An in-depth study was conducted in order to understand what drives the differences in the CRPS of the CSD-CLIM (i.e *CSD-UNC*) obtained at the different sites. However, in this section, we restrict the analysis on two specific sites namely Desert Rock (low *CSD-UNC*) and Fouillole (high *CSD-UNC*). For these two sites, in order to understand what drives the differences in *CSD-UNC*, the UNC^i is plotted for the 30 bins of clear-sky irradiances in Figure 6.

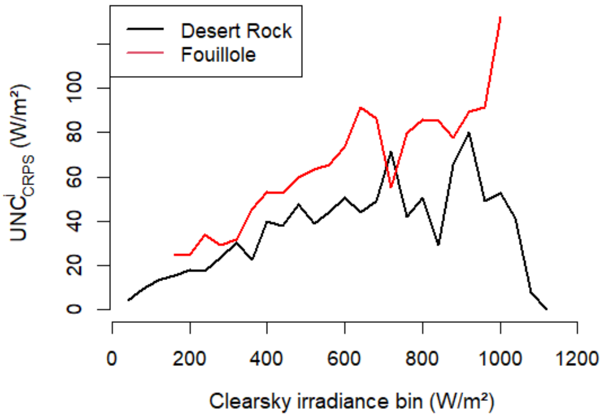


Figure 6: UNC^i in relation with each bin i of clear-sky irradiance for Desert Rock and Fouillole

The values of UNC^i generally increases with the level of clear-sky irradiance. Thus, the bins corresponding to high clear-sky irradiances values are responsible for the major part of *CSD-UNC*. Besides, the uncertainty is highest in Fouillole for almost every bin. A closer look on a bin corresponding to high clear-sky irradiances is presented in Figure 7 and shows from where come these differences.

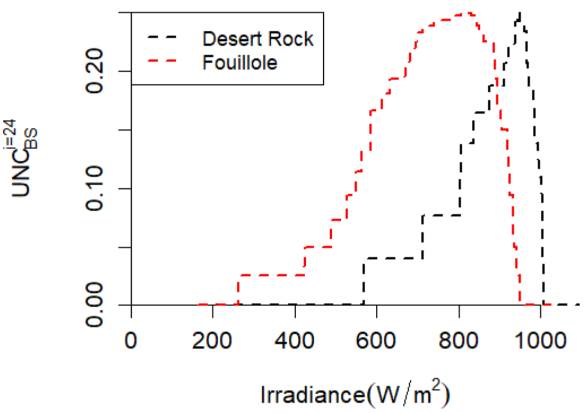
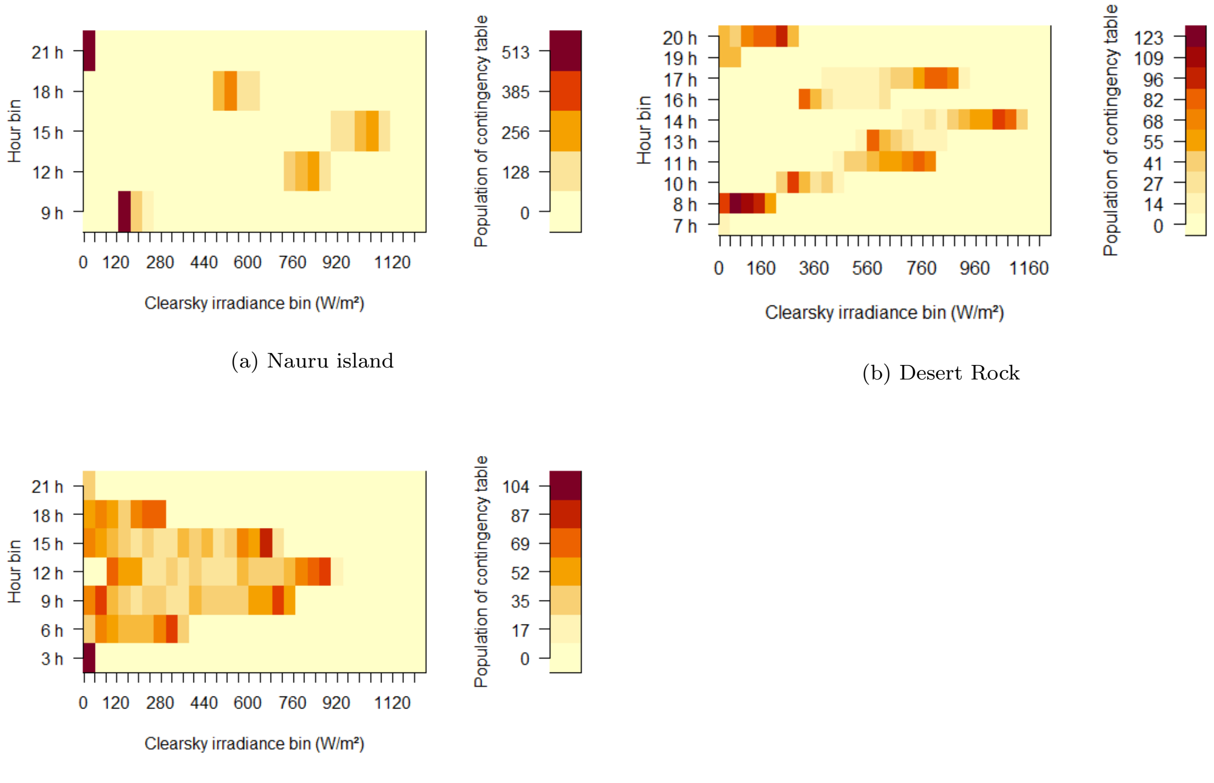


Figure 7: $UNC^{i=24}$ in Desert Rock and Fouillole for clear-sky irradiances between 900 and 940 W/m^2 (24th bin)

For high clear-sky irradiances, even low GHI measurements can occur in Fouillole, which is not the case at Desert Rock. This enlarges the area under $UNC^{i=24}_{BS}$ for Fouillole, leading to a higher $UNC^{i=24}$, and consequently to a higher *CSD-UNC*.

6.2. *Comparison of the binning approaches used by CH-PeEn and CSD-CLIM*

In terms of CRPS, the results obtained with the CH-PeEn turned out to be quite comparable with those of CSD-CLIM. This is not surprising since the general idea behind these two models is very close : giving a time-of-the-day dependent image of the uncertainty. The main difference between the two models is related to their approach to binning. The CH-PeEn model groups together all observations made at the exact same hour whereas the CSD-CLIM model proposes to bin the GHI data according to the clear-sky irradiance value. We propose here to use contingency tables in order to better highlight the relative difference in the binning methodology used by each model. Figure 8 shows such contingency tables for 3 specific sites namely Nauru, Desert Rock and Toravere. Note that the numbers of the contingency table are translated to a color scale to ease readability.



(a) Nauru island

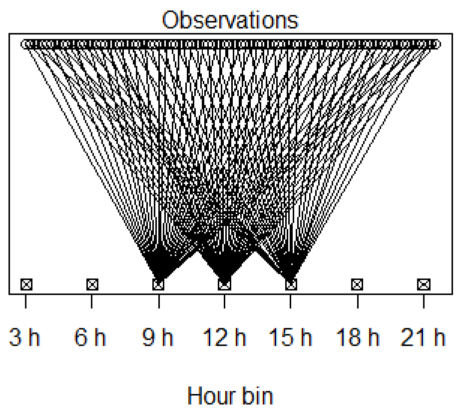
(b) Desert Rock

(c) Toravere

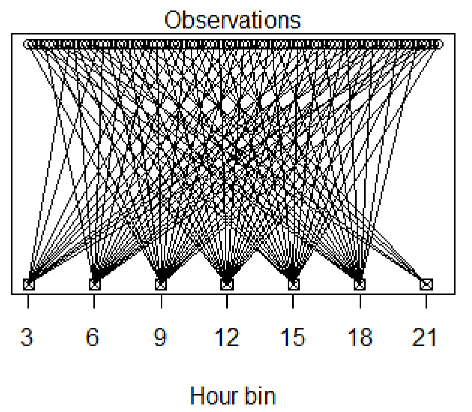
Figure 8: Contingency tables related to the two binning processes for three sites

As shown by Figure 8, the binning process appears very different between the two approaches. Moreover, this difference varies between locations. For instance, in Nauru island, the binning process is more or less equivalent. Indeed, there are no clearsky-irradiance bins of CSD-CLIM which corresponds to several hour bins of CH-PeEn. Conversely, this is not the case for in Desert Rock or Toravere. A possible explanation of these discrepancies may come from the latitude of the considered site and more precisely from the seasonal patterns experienced by a site. As an illustration, the sites of Nauru, Desert Rock and Toravere have respectively the following latitudes : -0.52°(Nauru island), 36.62°(Desert Rock) and 58.25°(Toravere). It can be argued that the further the site is from the equator, the more prominent the seasonal effect. It must be stressed here that this seasonal effect is ignored by CH-PeEn model. In other words, the further the site is from the equator, the more different the two binning processes are.

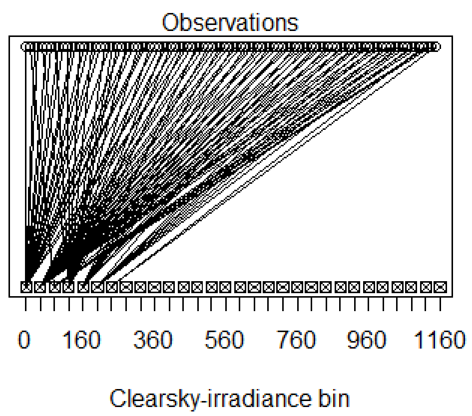
In practice, the difference between the two approaches varies also according to the seasons of the year. An example of this significant difference is illustrated for the site of Toravere in Figure 9.



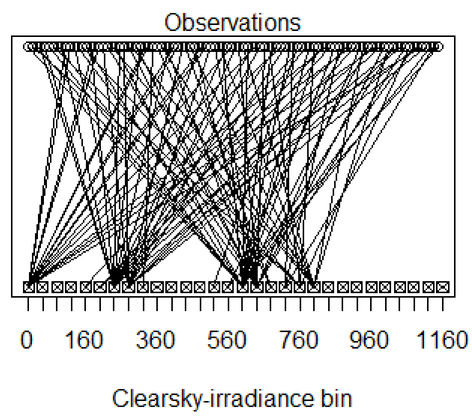
(a) CH-PeEn : binning process in Toravere for winter days



(b) CH-PeEn : binning process in Toravere for summer days



(c) CSD-CLIM : binning process in Toravere for winter days

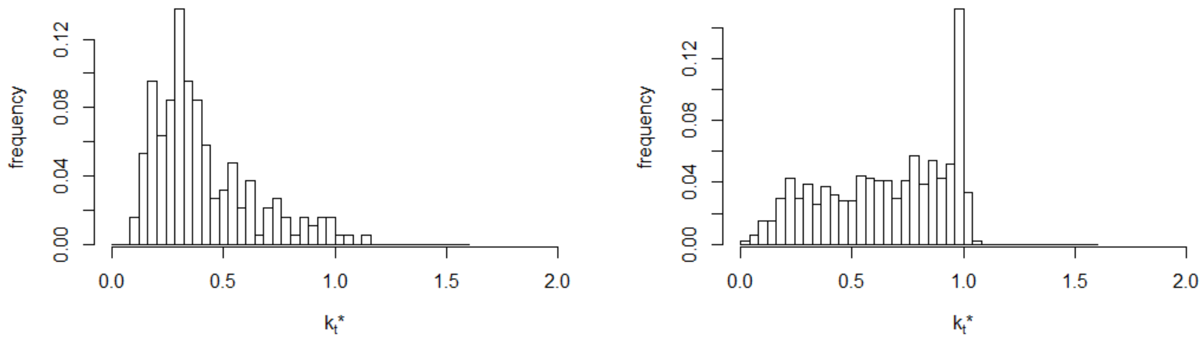


(d) CSD-CLIM : binning process in Toravere for summer days

Figure 9: Differences in the binning process between CH-PeEn and CSD-CLIM models

It is obvious from Figure 9 that the implementation of the binning process is very different between the two models. As shown by Figures 9a and 9b, the winter and summer seasons are treated the same way by CH-PeEn, except that more hour bins are involved during summer days. For CSD-CLIM, in winter (see Figure 9c), all observations are grouped into low clear-sky irradiance bins while they are more equally apportioned in summer (see Figure 9d).

In fact, CH-PeEn relies on a very strong implicit assumption i.e. for a fixed hour, the distribution of the clear-sky indices should be the same at each time of the year. This assumption does not hold completely true in various cases. Not surprisingly, in locations where the seasonal pattern is strong, the differences in the predictive CDF (used to compute the CRPS) related to the winter and summer seasons are non-negligible. An illustration of such a situation is presented below in Figure 10. The clear-sky indices of Toravere of a fixed hour (09h-12h UTC) are distributed according to the level of the clear-sky irradiance (low clear-sky irradiances ,corresponding approximately to winter months, in Figure 10a and high clear-sky irradiances, corresponding approximately to summer months, in Figure 10b).



(a) Months from October to March (clear-sky irradiances inferior to 250 W/m^2)

(b) Months from March to October (clear-sky irradiances superior to 250 W/m^2)

Figure 10: Clear-sky index distributions at Toravere, averaged on the 3-hour window[09h-12h] UTC.

Figure 10 shows a significant difference between the two distributions. This difference cannot be taken into account by CH-PeEn since it groups the GHI data acccording the hour of the day. Thus, the weak point of CH-PeEn is to aggregate together cases that a clear-sky model could easily discriminate. This also explains the difference in the CRPS results between CH-PeEn and CSD-CLIM. The mixture of situations that are statistically different made by CH-PeEn increases its CRPS score. Indeed, it is known that the local hour is not directly correlated with the seasonal and daily cycles of the sun irradiance. Consequently, the variability of the k_t^* bins used for the CH-PeEn includes the variability due to the solar declination, the time equation or the level of Aerosol Optical Depths (AODs). Conversely, the binning made by CSD-CLIM is finer as it is governed by the specification of the number of bins used by the binning process (see Appendix A). Put differently, CH-PeEn only takes profit from the daily periodicity of the climatology, whereas the results of this study tend to show that other periodicities (like the seasonality of the sun path) are non-negligible.

7. Conclusions

In this work, a new reference model is proposed to benchmark solar irradiance probabilistic forecasts. This new model called CSD-CLIM (for Clear-Sky Dependent climatology) is part of the class of climatology benchmark models. CSD-CLIM is evaluated against two existing climatology reference models. The first one is the naive climatology model while the second one is a recommended model in the solar forecasting community namely the CH-PeEn (for Complete History Persistence Ensemble) proposed by (Yang, 2019). After having defined a set of properties that a benchmark model should have, we have shown that CSD-CLIM, similarly to the naive climatology and CH-PeEn,

- is easy-to-implement,
- has an implementation that depends only on the historical data at hand,
- has a performance which is time invariant.

Besides, it was also demonstrated that, unlike the naive climatology, CH-PeEn and CSD-CLIM are able to reflect the difficulty of forecasting at a particular location.

More importantly, it was shown that CSD-CLIM achieves the best trade-off between the two most important attributes of a probabilistic forecast namely reliability and resolution. In particular, CSD-CLIM can be qualified as more statistically consistent than CH-PeEn. As such, in terms of overall performance, CSD-CLIM slightly outperforms CH-PeEn. This improved performance is due to a specific binning of the historical irradiance data based on the clear-sky irradiance values.

Finally, and as a conclusion, we can argue that the CSD-CLIM model can be a viable alternative to the CH-PeEn model.

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. doi:10.1016/j.apenergy.2015.08.011.

Alfred-Wegener-Institute, . World radiation monitoring center-baseline surface radiation network. https://bsrn.awi.de/. Accessed: 2021-02-01.

Binter, R., 2012. Applied Probabilistic Forecasting. Ph.D. thesis. Department of Statistics of the London School of Economics.

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

Brier, G., 1950. Verification of forecasts in terms of probability. Monthly Weather Review 78.

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

Copernicus Atmosphere Monitoring Service, . Cams McClear clear-sky irradiation service, version 3.1. http://www.soda-pro.com/fr/web-services/radiation/cams-mcclear. Accessed: 2021-02-01.

Cusack, S., Arribas, A., 2008. Assessing the usefulness of probabilistic forecasts. Monthly Weather Review 136, 1492–1504. doi:10.1175/2007MWR2160.1.

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. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169207018300384, doi:10.1016/j.ijforecast.2018.02.003.

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. URL: http://linkinghub.elsevier.com/retrieve/pii/S0038092X16300172, doi:10.1016/j.solener.2016.03.064.

Doubleday, K., Van Scyoc Hernandez, V., Hodge, B., 2020. Benchmark probabilistic solar forecasts: Characteristics and recommendations. Solar Energy 206, 52–67. doi:10.1016/j.solener.2020.05.051.

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. doi:10.1175/MWR2904.1.

Golestaneh, F., Pinson, P., Gooi, H., 2016. Very short-term nonparametric probabilistic forecasting of renewable energy generation— with application to solar energy. IEEE Transactions on Power Systems 31, 3850–3863. doi:10.1109/TPWRS.2015.2502423.

Hamill, T.M., Colucci, S., 1997. Verification of eta–rsm short-range ensemble forecasts. Monthly Weather Review 125, 1312–1327. doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting 15, 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.

Ineichen, P., Perez, R., 2002. A new airmass independent formulation for the Linke turbidity coefficient. Solar Energy 73, 151–157. URL: https://linkinghub.elsevier.com/retrieve/pii/S0038092X02000452, doi:10.1016/S0038-092X(02)00045-2.

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

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

Le Gal La Salle, J., Badosa, J., David, M., Pinson, P., Lauret, P., 2020. Added-value of ensemble prediction system on the quality of solar irradiance probabilistic forecasts. Renewable Energy .

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J., Morcrette, J., 2013. Mcclear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos. Meas. Tech. , 17.

Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. The Open Atmospheric Science Journal 2, 23–37. doi:10.2174/1874282300802010023.

Van der Meer, D., Widén, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renewable and Sustainable Energy Reviews 81, 1484 – 1512. doi:https://doi.org/10.1016/j.rser.2017.05.212.

Murphy, A.H., 1973. A new vector partition of the probability score. Journal of Applied Meteorology 12, 595–600.

Pedro, H., Coimbra, C., David, M., Lauret, P., 2018. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renewable Energy 123, 191–203. doi:10.1016/j.renene.2018.02.006.

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

Thorey, J., Chaussin, C., Mallet, V., 2018. Ensemble forecast of photovoltaic power with online crps learning. International Journal of Forecasting 34, 762–773. doi:10.1016/j.ijforecast.2018.05.007.

Todter, J., Ahrens, B., 2012. Generalization of the ignorance score: Continuous ranked version and its

decomposition. Monthly Weather Review 140, 2005–2017. doi:10.1175/MWR-D-11-00266.1.
Vannitsem, S., Wilks, D., Messner, J., 2018. Statistical Postprocessing of Ensemble Forecasts. Elsevier.
Yang, D., 2019. A universal benchmarking method for probabilistic solar irradiance forecasting. Solar Energy 184, 410–416. doi:10.1016/j.solener.2019.04.018.
Yang, D., 2020. Choice of clear-sky model in solar forecasting. Journal of Renewable and Sustainable Energy 12, 026101. URL: https://doi.org/10.1063/5.0003495, doi:https://doi.org/10.1063/5.0003495.
Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H.G., Blaga, R., Boland, J., Bright, J.M., Coimbra, C.F., David, M., Frimane, A., Gueymard, C.A., Hong, T., Kay, M.J., Killinger, S., Kleissl, J., Lauret, P., Lorenz, E., van der Meer, D., Paulescu, M., Perez, R., Perpinan-Lamigueiro, O., Peters, I.M., Reikard, G., Renne, D., Saint-Drenan, Y.M., Shuai, Y., Urraca, R., Verbois, H., Vignola, F., Voyant, C., Zhang, J., 2020a. Verification of deterministic solar forecasts. Solar Energy 210, 20 – 37. URL: http://www.sciencedirect.com/science/article/pii/S0038092X20303947, doi:https://doi.org/10.1016/j.solener.2020.04.019. special Issue on Grid Integration.
Yang, D., van der Meer, D., Munkhammar, J., 2020b. Probabilistic solar forecasting benchmarks on a standardized dataset at folsom, california. Solar Energy 206, 628 – 639. URL: http://www.sciencedirect.com/science/article/pii/S0038092X20305090, doi:https://doi.org/10.1016/j.solener.2020.05.020.
Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K., 2002. The economic value of ensemble-based weather forecasts. Bulletin of the American Meteorological Society 83, 73–83.

Acknowledgements

The authors acknowledge the financial support given by Region Reunion and European Regional development Fund (FEDER) under the POE-FEDER 2014-2020 PhD scholarship DIRED/20171387.

Appendix A. Sensitivity Analysis related to the number of bins used by the CSD-CLIM binning process

In this section, an analysis of the sensitivity of *CSD-UNC* related to Nb is conducted. A number $Nb = 30$ has been chosen in this study. However, this choice is arbitrary. It could have a strong impact on the final result and could be questioned. Note that a choice of $Nb = 1$ makes *CSD-UNC* equal to *UNC*. The impact of Nb on *CSD-UNC* is presented in Figure A.11

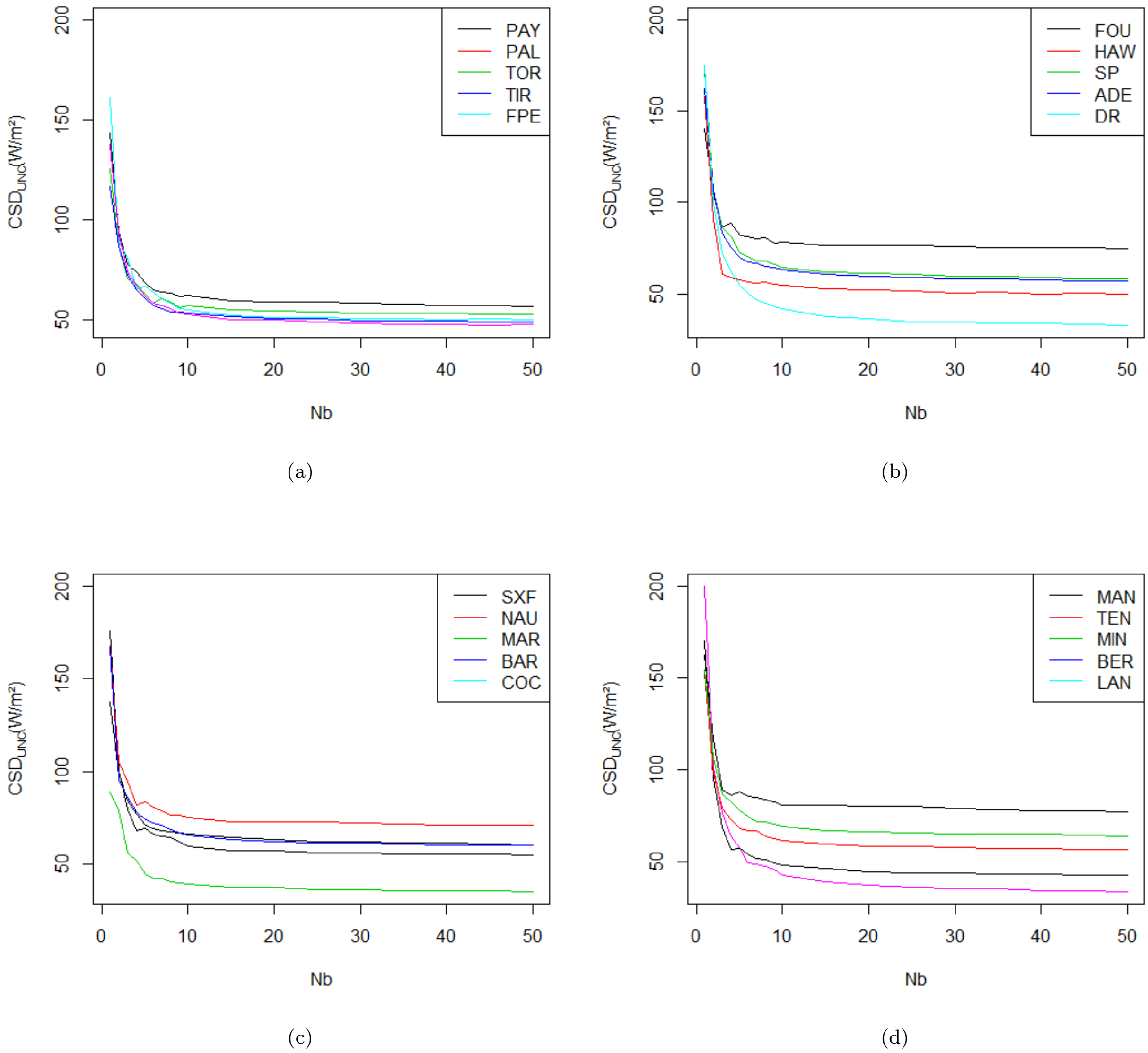


Figure A.11: Impact of Nb on *CSD-UNC*

Two very different regimes are distinguishable. For all sites, *CSD-UNC* decreases dramatically when Nb varies from 1 to 20. From a number $Nb = 20$, *CSD-UNC* is stable and the choice of Nb is no longer of great importance. Thus, a choice of $Nb > 20$ should be preferred. Our choice of $Nb = 30$ meets this requirement. Note that the number of regime switching (here $Nb = 20$) is not absolute and could depend on the size of the data.

Appendix B. Brier Score

The Brier Score (BS) is a probabilistic score used for the evaluation of binary forecasts (i.e. forecast for an event that fully realizes or not). Its mathematical definition is

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{f}_i - o_i)^2, \tag{B.1}$$

where o is the observation (0 if the event does not realize and 1 if it realizes), $\hat{f}$ is the probability forecast (that can take any value between 0 and 1) and N is the number of forecast occurences (Brier, 1950).

The Brier Score can be decomposed into the three main attributes of a forecast namely reliability (REL_{BS}), resolution (RES_{BS}) and uncertainty (UNC_{BS}). This decomposition reads as

$$BS = REL_{BS} - RES_{BS} + UNC_{BS}. \tag{B.2}$$

The reliability measures the difference between the probability forecasts and the observations and is given by

$$REL_{BS} = \frac{1}{N} \sum_{j=1}^{N_k} p_j (\hat{f}_j - \bar{o}_j)^2, \tag{B.3}$$

where $(\hat{f}_j, j = 1, ..., N_k)$ denotes the ensemble of all the different probabilities provided by the forecast, p_j is the number of occurrences of $\hat{f}_j$ over the test period, $\bar{o}_j$ is the mean observation when the forecast is equal to $\hat{f}_j$

The resolution measures to what extent the forecast discriminates the observations from the climatological mean $\bar{o}$. It is given by

$$RES_{BS} = \frac{1}{N} \sum_{j=1}^{N_k} p_j (\bar{o} - \bar{o}_j)^2. \tag{B.4}$$

The uncertainty term depends on the variability of the observations and is defined by

$$UNC_{BS} = \bar{o}(1 - \bar{o}). \tag{B.5}$$

In the case of a continuous variable like GHI, the Brier score can be used to evaluate the probability that GHI exceeds a threshold x.

Appendix C. CRPS as the integral of the Brier score

In the general case of a continuous variable, the CRPS is also the integral of the Brier Score over all thresholds x, as demonstrated by Todter and Ahrens (2012).

$$CRPS = \int_{-\infty}^{+\infty} BS(x)dx, \tag{C.1}$$

and its decomposition comes as :

$$REL = \int_{-\infty}^{+\infty} REL_{BS}(x)dx, \tag{C.2}$$

$$RES = \int_{-\infty}^{+\infty} RES_{BS}(x)dx, \tag{C.3}$$

$$UNC = \int_{-\infty}^{+\infty} UNC_{BS}(x)dx. \tag{C.4}$$

Appendix D. Time-invariance of CSD-CLIM

Since the score *CSD-UNC* of CSD-CLIM is a climatological indicator theoretically based on all historical data, its stability is necessarily achieved when the length of the input data is sufficiently large. Nonetheless, in this section, we investigate the dependency of *CSD-UNC* for different cases of input data.

Four sites of the study i.e. Payerne, Sioux Falls, Tenerife and Bermuda island which experience different sky conditions (see Table 2) have been selected. The *CSD-UNC* was calculated for 3 different periods of 3 years and for 7 historical datasets with different length (from 1 to 7 years). The periods and the lengths of the different datasets are listed respectively in Table D.7 and Table D.8

	Period		
Site	period 1	period 2	period 3
Payerne	2011-2013	2014-2016	2017-2019
Sioux Falls	2010-2012	2013-2015	2016-2018
Tenerife	2012-2014	2015-2017	2018-2020
Bermuda Islands	2004-2006	2007-2009	2010-2012

Table D.7: Periods used for time stability assessment

	Data length						
Site	1	2	3	4	5	6	7
Payerne	2011	2011-2012	2011-2013	2011-2014	2011-2015	2011-2016	2011-2017
Sioux Falls	2010	2010-2011	2010-2012	2010-2013	2010-2014	2010-2015	2010-2016
Tenerife	2012	2012-2013	2012-2014	2012-2015	2012-2016	2012-2017	2012-2018
Bermuda Islands	2004	2004-2005	2004-2006	2004-2007	2004-2008	2004-2009	2004-2010

Table D.8: Data lengths used for time stability assessment.

The resulting CRPS are given respectively in Table D.9 and Table D.10

Site	*CSD-UNC* (W/m^2)		
	period 1	period 2	period 3
Payerne	59.3	60.5	57.9
Sioux Falls	60.1	61.0	61.2
Tenerife	35.1	37.6	35.9
Bermuda Islands	59.0	57.8	56.2

Table D.9: Sensitivity of *CSD-UNC* on data period

Site	*CSD-UNC* (W/m^2)						
	1	2	3	4	5	6	7
Payerne	55.6	57.6	59.3	59.3	59.3	60.0	59.8
Sioux Falls	60.4	61.3	60.1	61.5	61.6	61.4	60.8
Tenerife	38.0	35.6	35.1	36.4	36.2	36.5	36.7
Bermuda Islands	56.9	58.0	59.0	59.5	59.6	58.6	58.6

Table D.10: Sensitivity of *CSD-UNC* on data length

As shown by these tables, *CSD-UNC* is not strongly dependent on the chosen period or length of the dataset used to calculate it. In this work, the data granularity was 3h. However, it must be stressed that the stability in the CRPS results can be improved provided that the time resolution of the data increases. We recall that in a practical case, if the dependency on the input data is found strong, the recommendation should be to extend the length of the input data in order to get closer to the climatological mean.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights (for review)

- A new reference model to benchmark probabilistic solar forecasts called “CSD-CLIM” is introduced
- CSD-CLIM is easy-to-implement, there is no need to form probability distributions to compute its CRPS
- In the light of a baseline for benchmark models, CSD-CLIM is compared with two reference models: the naive climatology and the CH-PeEn
- CSD-CLIM meets all prerequisite properties of a good benchmark model
- CSD-CLIM achieves the best trade-off between reliability and resolution, and can be a viable alternative to existing models