Solar Energy 206 (2020) 52–67

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Benchmark probabilistic solar forecasts: Characteristics and recommendations

Kate Doubleday*, Vanessa Van Scyoc Hernandez, Bri-Mathias Hodge

Power Systems Engineering Center, National Renewable Energy Laboratory, United States
Department of Electrical, Computer, and Energy Engineering, and Renewable and Sustainable Energy Institute, University of Colorado Boulder, United States

ARTICLE INFO

Keywords:
Solar power
Irradiance
Solar forecasts
Probabilistic forecasts
Benchmarking

ABSTRACT

We illustrate and compare commonly used benchmark, or reference, methods for probabilistic solar forecasting that researchers use to measure the performance of their proposed techniques. A thorough review of the literature indicates wide variation in the benchmarks implemented in probabilistic solar forecast studies. To promote consistent and sensible methodological comparisons, we implement and compare ten variants from six common benchmark classes at two temporal scales: intra-hourly forecasts and hourly resolution forecasts. Using open-source Surface Radiation Budget Network (SURFRAD) data from 2018, these benchmark methods are compared using proper probabilistic metrics and common diagnostic tools. Practical implementation issues, such as the impact of missing data and applicability for operational forecasting, are also discussed. We make recommendations for practitioners on the appropriate selection of benchmark methods to properly showcase state-of-the-art improvements in forecast reliability and sharpness. All code and open-source data are available on Github for reproducibility and for other researchers to apply the same benchmark methods to their own data.

1. Introduction

1.1. Background

As power systems worldwide experience ever-increasing penetrations of variable and uncertain renewable generation resources, such as wind and solar photovoltaics (PV), renewable energy forecasting is gaining increasing attention and finding new applications (Kroposki et al., March 2017). Historical efforts focused on deterministic or point forecasts, but interest is shifting toward probabilistic forecasts that fully capture future uncertainty (van der Meer et al., 2018; Hong et al., 2016). Probabilistic formulations can achieve cost and/or reliability benefits over their deterministic equivalents (Appino et al., 2018). Therefore, probabilistic forecasts are valuable for both power system operators and market participants (Bessa et al., 2017). Applications of probabilistic renewable energy forecasts have been proposed in market bidding strategies (Li and Park, 2018), adaptive reserve algorithms (Fahiman et al., 2019), and robust and/or stochastic unit commitment and economic dispatch models (Bukhsh et al., 2016; Li et al., 2018).

As interest in probabilistic forecasting applications has increased, research into advanced probabilistic forecasting methods has expanded as well. Although wind and load forecasts have received significant research attention in the past, solar irradiance/power forecasting has been a developing field during the past few years (Hong et al., 2016). As recently as 2013, in a comparison of state-of-the-art solar and wind forecast methods with 18 entrants, none provided probabilistic solar forecasts (Sperati et al., 2015). In early attempts at probabilistic solar forecasting, efforts focused on assessing confidence intervals around a deterministic forecast, such as the 5–95% interval (Mathiesen et al., 2013; Lorenz et al., 2009; Boland and Soubdhan, 2015; Chu et al., 2015; Almeida et al., 2015; Scolari et al., 2016; Torregrossa et al., 2016). As the field has expanded, attention has turned to predicting a full probability distribution (Alessandrini et al., 2015; Pedro et al., 2018). The format of information in these forecasts is distinct from point forecasts, and it must be assessed and validated appropriately.

The recent abundance in probabilistic solar forecasting literature has given rise to concerns about proper method verification and comparisons, given inconsistent practices in forecast validation. A common inconsistency is the application of improper or inappropriate metrics (van der Meer et al., 2018). Another is the lack of an appropriate benchmark method against which to measure improvements (Bracale et al., 2013; Dong et al., 2013; Abuella and Chowdhury, 2015; Liu et al., 2016; Nagy et al., 2016). To address these concerns, Lauret et al. (2019) recently proposed a verification framework for probabilistic solar forecasting focusing on proper metrics and visual diagnostic tools. To complement that work, this article delves into benchmark, or reference,

* Corresponding author.
E-mail address: kado4165@colorado.edu (K. Doubleday).

https://doi.org/10.1016/j.solener.2020.05.051
Received 4 March 2020; Received in revised form 15 May 2020; Accepted 18 May 2020
0038-092X/ Published by Elsevier Ltd on behalf of International Solar Energy Society.

methods for probabilistic solar forecasting—baseline methods that all researchers can use to measure the performance of their proposed techniques.

Benchmark forecasts can serve two key purposes. The first is simply as a reference for comparison, analogous to a yardstick. For this purpose, the benchmark should be consistent, accessible, and easily reproducible, though it does not necessarily need be considered a "good" forecast. The second purpose of a benchmark is to provide a target for new methods to outperform, analogous to a point on the yardstick. For this purpose, the benchmark should be close to the state of the art, but accessibility and reproducibility are still important. Common benchmark methods—including numerical weather prediction (NWP) ensembles, climatology, and persistence ensembles—will be discussed in detail. As will be shown, however, the range of implementations is wide, even within the same general benchmark methodology, resulting in fundamentally different benchmark forecasts.

The objective of this article is to promote consistent and sensible methodological comparisons in the probabilistic solar forecasting community by summarizing current practices, illustrating key differences among benchmark methods, discussing considerations for missing data, and making recommendations for researchers. Following a detailed literature review of benchmark probabilistic solar forecasts, ten common benchmark variants are implemented at two temporal scales. Benchmark forecasts are generated for the entire year 2018 using publicly available data from the seven Surface Radiation Budget Network (SURFRAD) facilities, which are located in diverse climates throughout the United States (Augustine et al., 2005; NOAA Earth System Research Laboratory, n.d.). Along with the SURFRAD data, the `R` code used to generate these forecasts is made open source with this paper for reproducibility and for future researchers to apply the same benchmark methods to their own data (R Core Team, 2017).

This section concludes with an introduction of the relevant timescales and terminology for solar forecasting. Section 2 presents a detailed literature review of probabilistic benchmarks seen in the solar forecasting field. Section 3 introduces the case study data used to compare common benchmark methods, focusing on the seven SURFRAD sites. For six of the benchmark classes found in the literature review, Section 4 describes the methodology and considerations for implementing these benchmarks at two temporal scales. Section 5 introduces metrics and tools for assessing the quality of a probabilistic forecast, including key attributes such as reliability, sharpness, and resolution. Section 6 compares the performance of ten benchmark variants during the entire year 2018 for each SURFRAD site. Finally, Section 7 summarizes recommendations and concludes.

1.2. Forecast terminology

When discussing any type of forecast, we are commonly discussing a forecast run, which is a series of forecasts spaced through time. The temporal parameters that describe a forecast run are illustrated in Fig. 1 (Yang, 2019a). Forecast run k is issued at a time, t_k, in advance of its first forecast valid time by a lead time $\mathcal{L}$. The series of forecasts is equally spaced with resolution $\mathcal{R}$ during a period of time, the forecast horizon $\mathcal{H}$. Therefore, the series of forecasts is valid for times $t = t_k + \mathcal{L},\ t_k + \mathcal{L} + \mathcal{R},\ \ldots, t_k + \mathcal{L} + (\mathcal{H} - 1)\mathcal{R}$. As new information becomes available, forecast runs are updated at an update rate, $\mathcal{U}$, which is the period between the issue times of sequential forecast runs.

In this article, we consider the general categories of intra-hourly forecasts, which have forecast horizons up to a few hours, and hourly-resolution forecasts, which have forecast horizons up to a few days. These temporal scales are typical for power systems operations. Longer term seasonal or annual forecasts are beyond the scope of this paper. For example, an intra-hourly forecast might have a 5-min resolution with a 5-min lead time over a 1-h horizon with a 15-min update rate; it might be applied in an hourly economic dispatch model updated every 15 min. In contrast, an hourly forecast might have hourly resolution with a 12-h lead time over a 48-h horizon with a 24-h update rate for use in a day-ahead unit commitment model.

Different forecasting methods are applied at these different temporal scales. Hourly-resolution forecasts typically employ NWP models, which are physics-based models run by weather agencies with time horizons of a week or 10 days (van der Meer et al., 2018). These models are highly useful for forecasting developing weather patterns, but they are very computationally intensive and have coarse spatial and temporal scales. Many have only 3-, 6-, or 12-h update rates, which make them unsuitable for intra-hourly forecasting. One of the highest temporal resolution NWP models is the National Oceanic and Atmospheric Administration's (NOAA's) deterministic High-Resolution Rapid Refresh (HRRR), which is updated hourly with 15-min resolution (NOAA National Weather Service, n.d). As a result, many intra-hourly forecasting techniques make use of machine learning or time-series methods that require only historical observations as inputs rather than exogenous weather data (Pedro et al., 2018; David et al., 2016; Munkhammar et al., 2019a).

2. Literature review

To assess current practices relevant to probabilistic solar forecast benchmarking, 42 recent journal articles and conference papers were reviewed. Of these, 8 did not contain a benchmark method per se; rather, they compared variants of their own proposed methods against each other (Bracale et al., 2013; Dong et al., 2013; Abuella and Chowdhury, 2015; Almeida et al., 2015; Huang and Perry, 2016; Liu et al., 2016; Nagy et al., 2016; Lotfi et al., 2020). Including benchmark forecasts allows the reader a point of comparison to assess the relative merit of the proposed methods and provides that yardstick for comparing salient forecast features, such as sharpness and reliability (Section 5). Omitting such benchmarks obscures the value of the new work.

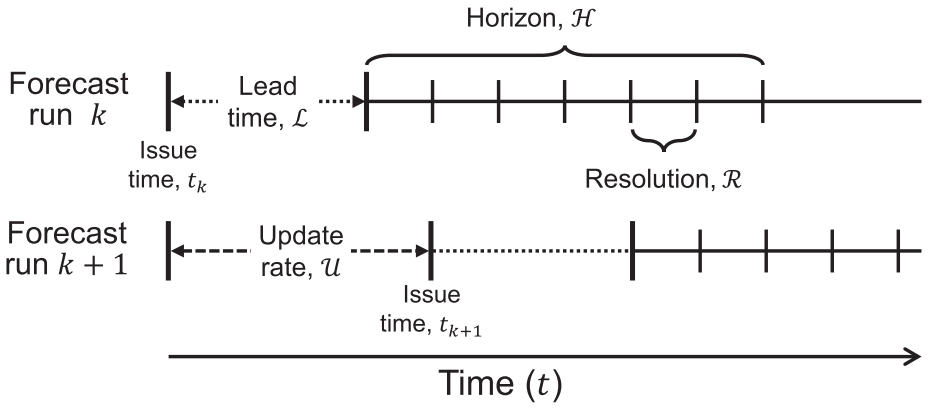

Fig. 1. Illustration of temporal attributes of a forecast run.

Table 1
Summary of Forecast Benchmarks from the Literature.

Reference	Forecast Variable	$\mathcal{R}$	$\mathcal{L} + \mathcal{H}$	Benchmark Method(s)	Training Data Details
Pinson et al. (2008)	Wind speed/power	1 h	48 h	Climatology	All available measurements
				Raw NWP ensemble	75-member NWP ensemble
Thorarinsdottir and Gneiting (2010)	Wind speed	Daily max	48 h	Climatology	All measurements from a ±15-d window around forecast date in 5 yr training period
				Raw NWP ensemble	8-member NWP ensemble
Sloughter et al. (2010)	Wind speed	1 h	48 h	Climatology	All available measurements
				Raw NWP ensemble	8-member NWP ensemble
Iversen et al. (2014)	GHI	1 h	24 h	Climatology	All available measurements
				Climatology variant 1: CH-PeEn	All measurements from same hour of day
				Climatology variant 2: PeEn	All measurements from same month and hour of day
Aryaputera et al. (2016)	Accumulated GHI	6 h	6 h	Climatology with kernel density estimation	All available measurements
				NWP ensemble with bias correction	23-, 26-, and 50-member NWP ensembles
Thorey et al. (2018)	PV power	30 min	6 d	Climatology	All measurements from a ±1-mth window around the forecast date
				Raw ensemble	50- and 34- member NWP ensembles, transformed to power
Golestaneh et al. (2016b)	PV power	1 min	10 min or 1 h	Climatology	All available measurements
				Gaussian error distribution	Power persistence errors from the past hour
				HLA (Wan et al., 2014a)	
				BELM (Wan et al., 2014b)	
Möller et al. (2013)	Wind speed, max and min temperature, precipitation, pressure	Daily	48 h	Raw NWP ensemble	8-member NWP ensemble
Lauret et al. (2019)	GHI	1 h	6 h	Quantile random forests (Grantham et al., 2016; Lauret et al., 2017)	5 previous measurements, solar zenith angle, and hour angle
	Accumulated GHI	3 h	24 h	Raw NWP ensemble	50-member NWP ensemble
Alessandrini et al. (2015)	PV power	1 h	72 h	PeEn	Last 20 measurements at same hour of day
Sperati et al. (2016)	PV power	3 h	72 h	PeEn	Last 51 measurements at same time of day
David et al. (2016)	GHI	10 min	1 h	PeEn	Last 10 estimates of CSI
		1 h	6 h		
Lauret et al. (2017)	GHI	1 h	6 h	PeEn	Last 10 GHI measurements
Pedro et al. (2018)	GHI, DNI	5 min	30 min	PeEn	All measurements from previous 2 h
El-Baz et al. (2018)	PV power	1 min	1 min	PeEn	Last 20 observations from same hour
Munkhammar et al. (2019a)	CSI	1 min	5 min	PeEn	Last 9 CSIs
				Quantile regression	Unclear
				Markov-chain mixture (MCM)	Various, including previous 20 or 50 days of CSI estimates, or every other day
Panamtash et al. (2020)	PV power	15 min	15 min	PeEn	Last 20 CSI estimates from same time of day
				CH-PeEn	All CSI estimates from same time of day
Chu and Coimbra (2017)	DNI	1 min	20 min	PeEn	DNI observations from past hour
				Gaussian distribution	k-nearest neighbors
Ni et al. (2017)	PV power	5 min	5 min	Gaussian distribution of PeEn	10 most recent measurements
Mathiesen et al. (2013)	GHI	1 h	24 h	Empirical error distribution	Errors from times with similar CSI and atmospheric flow
Lorenz et al. (2009)	GHI/PV power	1 h	72 h	Gaussian error distribution	Errors from times with similar CSI and solar zenith angle
Boland and Soubdhan (2015)	GHI	1 h	1 h	Gaussian error distribution	ARCH model of variance
Chu et al. (2015)	DNI	1 min	20 min	Gaussian error distribution	Smart persistence errors from the past hour
Torregrossa et al. (2016)	GHI	250 ms	750 ms	Gaussian error distribution	Holt-Winters or persistence forecast errors
					GARCH(1,1) model of variance
				Empirical error distribution	Bootstrap of Holt-Winters forecast errors
Scolari et al. (2016)	GHI	0.5 s to 5 min	0.5 s to 5 min	Gaussian error distribution	Errors from ARIMA deterministic forecast
				DIP (Torregrossa et al., 2016)	Persistence deterministic forecast
Grantham et al. (2016)	GHI	1 h	1 h	Gaussian error distribution	All point forecasting errors from 8 yr training period
Yang (2019b)	GHI	15 min	6.25 h	CH-PeEn	All CSIs from same hour of day
Davò et al. (2016)	Accumulated GHI	Daily	1 d	Analog ensemble (Alessandrini et al., 2015)	11 member NWP ensemble and 3 yrs of historical forecasts and measurements

(*continued on next page*)

Table 1 (*continued*)

Reference	Forecast Variable	$\mathcal{R}$	$\mathcal{L} + \mathcal{H}$	Benchmark Method(s)	Training Data Details
Cervone et al. (2017)	PV power	1 h	72 h	Analog ensemble (Alessandrini et al., 2015)	Deterministic NWP forecast and 1 yr of historical forecasts and measurements
Verbois et al. (2018)	GHI	1 h	24 h	Analog ensemble (Alessandrini et al., 2015)	Deterministic NWP forecast and 2 yrs of historical forecasts and measurements

The remaining literature includes a variety of benchmark methods, from which four major classes emerge. Table 1 summarizes the benchmarks used in these references, including their application's forecast resolution, maximum look-ahead time (notated as $\mathcal{L} + \mathcal{H}$ because of the ambiguity between lead time and horizon in many of these papers), the general class of the benchmark method, and any specifics about the training data chosen to calculate the benchmark. Most of these references focus on solar applications: common forecast variables are global horizontal irradiance (GHI), direct normal irradiance (DNI), clear-sky index (CSI—the ratio of GHI to estimated clear-sky GHI), accumulated GHI (the sum of GHI over the forecast window/resolution), and PV power output. A few references forecast related renewable energy variables, such as wind speed.

Note that a few references in Table 1 contain only information about the fundamental methods that we call benchmarks, but most references propose a novel method and compare that method relative to a benchmark. For papers with both novel and benchmark methods, the table summarizes the benchmark only.

The four major benchmark classes are briefly introduced here; Section 4 goes into greater detail on their implementations. The first common class is the climatology benchmark, which is a staple of related meteorology fields, including wind forecasting (Pinson et al., 2008; Thorarinsdottir and Gneiting, 2010; Sloughter et al., 2010). A basic climatology forecast is an empirical cumulative distribution function (CDF) based on historical measurements over a long period of time; it is the most naive forecast one can generate, given the statistical properties of the forecast variable and no knowledge of upcoming conditions. Though less common in the solar forecasting literature thus far, it has been applied for both hourly (Iversen et al., 2014; Aryaputera et al., 2016; Thorey et al., 2018) and intra-hourly (Golestaneh et al., 2016b) solar applications.

Another standard benchmark class from meteorology is the raw NWP ensemble (Pinson et al., 2008; Thorarinsdottir and Gneiting, 2010; Sloughter et al., 2010; Möller et al., 2013). Although each NWP simulation gives one deterministic forecast of future conditions, a probabilistic perspective can be gained from an ensemble of NWP forecasts, collected from a variety of NWP models or by perturbing a model's initial conditions (van der Meer et al., 2018; Leutbecher and Palmer, 2008). These ensembles are often post-processed to address bias and underdispersion, which is the tendency of the ensemble to underestimate uncertainty in the forecast (Leutbecher and Palmer, 2008). Therefore, a natural benchmark for post-processing techniques is an empirical CDF of the "raw," unprocessed ensemble. Because of the temporal restrictions of NWP modeling, this benchmark has been applied only to hourly rather than intra-hourly solar forecasts (Aryaputera et al., 2016; Thorey et al., 2018; Lauret et al., 2019).

The next, very common class of benchmarks is the persistence ensemble (PeEn) (Sperati et al., 2016; David et al., 2016; Lauret et al., 2017; Pedro et al., 2018; El-Baz et al., 2018; Munkhammar et al., 2019a; Chu and Coimbra, 2017; Ni et al., 2017; Panamtash et al., 2020), often attributed to Alessandrini et al. (2015). PeEn's are intended to capture weather patterns by collecting recent observations or CSIs into an empirical CDF. For intra-hourly forecasts, the common practice is to collect observations from the last few hours, (David et al., 2016; Pedro et al., 2018; El-Baz et al., 2018), whereas for hourly forecasts, researchers often use observations at the same time of day from the last several days (Alessandrini et al., 2015; Sperati et al., 2016).

The fourth class is the Gaussian error distribution, which is particularly popular for intra-hourly or continuously-updated, "rolling" forecasts ($\mathcal{H} = \mathcal{R}$). These probabilistic forecasts are generally extensions of a deterministic forecast, in which the deterministic forecast is dressed in a distribution based on historical errors between the point forecasts and the observations. This distribution can be a simple empirical CDF (Mathiesen et al., 2013; Torregrossa et al., 2016). It is much more common, however, to fit a Gaussian distribution to the errors

Table 2
Temporal Configurations of the Intra-Hourly and Hourly Resolution Case Study Forecast Runs.

	Resolution $\mathcal{R}$	Lead Time $\mathcal{L}$	Horizon $\mathcal{H}$	Update Rate $\mathcal{U}$
Intra-hourly	5 min	5 min	1 h	1 h
Hourly	1 h	1 h	6 h	6 h

(Lorenz et al., 2009; Boland and Soubdhan, 2015; Chu et al., 2015; Grantham et al., 2016; Golestaneh et al., 2016b; Scolari et al., 2016; Torregrossa et al., 2016; Chu and Coimbra, 2017), even though solar and other renewable energy deterministic errors do not typically follow a Gaussian distribution (Bludszuweit et al., 2008; Zhang et al., 2015; Golestaneh et al., 2016b; Chu and Coimbra, 2017). Within this class, researchers have proposed a wide variety of screening methods to select the most relevant historical errors: similarity of atmospheric condition, including clear-sky index, solar zenith angle, and/or wind direction (Lorenz et al., 2009; Mathiesen et al., 2013); most recent errors (Chu et al., 2015; Golestaneh et al., 2016b); nearest neighbor errors (Chu and Coimbra, 2017); or all errors in the training data set (Grantham et al., 2016).

Although the majority of the benchmarks can be categorized into these four general areas, there is also blurring among them and significant internal variation in their implementations. Particularly for hourly forecasts in which training data are accumulated during previous days, PeEn variants can blend into climatology variants (Iversen et al., 2014; Sperati et al., 2016). Iversen et al. (2014) illustrates this well in Table 1: the authors apply three climatology variants, with increasing down-selection of historical data, until the third variant uses only observations from the same month and hour of day—very similar to the PeEn methods in Alessandrini et al. (2015) and Sperati et al. (2016). The second, intermediate variant in Iversen et al. (2014) uses all historical measurements at the same hour of day, a benchmark that was codified in Yang (2019b) as the complete history persistence ensemble (CH-PeEn). The CH-PeEn hybrid has the statistical consistency of climatology, but it follows the diurnal solar trend like a PeEn forecast. This hybrid, also used by Panamtash et al. (2020), is the fifth basic benchmark class detailed in Section 4.

Finally, a Markov-chain mixture (MCM) model was recently proposed as an intra-hourly probabilistic benchmark in Munkhammar et al. (2019a) as well as Munkhammar et al. (2019b). This model uses a transition matrix based on historical CSIs to model changes through time over a forecast run. MCM was shown to outperform a PeEn at short time scales (Munkhammar et al., 2019a), and this is the sixth and final method illustrated below.

Note that many of these papers use two or more benchmark forecasts, reflective of the two purposes of a benchmark. For example, some studies use climatology as a yardstick and a near state-of-the-art, raw NWP ensemble as the point on the yardstick. Similarly, a handful of papers implement methods from recently published work as state-of-the-art benchmarks to outperform (Golestaneh et al., 2016b; Scolari et al., 2016; Lauret et al., 2019), including a series of studies building on the National Center for Atmospheric Research's analog ensemble approach (Alessandrini et al., 2015; Davò et al., 2016; Cervone et al., 2017; Verbois et al., 2018). In the remaining sections, the key features of the four major forecast classes, plus CH-PeEn and MCM, will be further distinguished and illustrated, but remember that more than one benchmark is often useful to frame the contributions of proposed improvements. Lastly, it is important to note that these benchmark methods produce probabilistic forecasts in the form of CDFs at each forecast time, rather than scenarios or trajectories over time, as in Golestaneh et al. (2016a), Woodruff et al. (2018), Sun et al. (2020). Benchmarking trajectories requires its own set of methods and is beyond the scope of this article.

3. Case study data

The remainder of this paper illustrates six of the benchmark classes using a case study of commonly available data. Because of the wider availability of solar irradiance than power data, we focus primarily on irradiance forecasting, though these methods can be adapted to PV power forecasting as well (e.g., with an upper limit at the power plant's AC power rating). Three data sources are used: irradiance observations, CSI estimates, and NWP ensemble forecasts. All data sets are retrieved for the entire year 2018 to enable a long-term comparison of benchmark methods.

For the irradiance observations, 1-min resolution measurements are retrieved from NOAA's seven SURFRAD sites, located in diverse climates throughout the United States (www.esrl.noaa.gov/gmd/grad/surfrad). For each of the seven locations, clear-sky irradiance estimates are available at 1-min resolution from the CAMS McClear Service (Copernicus Atmosphere Monitoring Service (CAMS), 2019; Lefèvre et al., 2013). Note that these clear-sky estimates are not operational and are only available 2 days after the fact; to operationalize the benchmarks that use a CSI would require these historical clear-sky "observations" plus an operational deterministic clear-sky forecast. Here, the CAMS McClear values are used as both. Finally, the European Centre for Medium-Range Weather Forecasts' (ECMWF) 51-member ensemble is selected as the case study NWP ensemble (www.ecmwf.int). The ECMWF ensemble is available at an hourly resolution and updated four times per day.

Based on the resolutions of these three data sets, benchmark forecasts are compared at two temporal configurations, summarized in Table 2. First, an intra-hourly forecast is implemented with 5-min resolution, 5-min lead time, 1-h horizon, and 1-h update rate. This schedule is based on a likely economic dispatch schedule and the high resolution of the SURFRAD observations. Second, an hourly-resolution, intra-day forecast is implemented with 1-h resolution, 1-h lead time, 6-h horizon, and 6-h update rate. This schedule is selected based on the ECMWF resolution, always using the most recent forecast (issued every 6 h). In practical use cases, an hourly resolution forecast might have a 24- to 48-h horizon for applications such as the day-ahead unit commitment. For this exercise, however, forecast runs are configured with equal horizon and update rate so that each observation is used to validate a single forecast rather than multiple forecasts submitted at different issue times. The same methods illustrated here for the intra-day forecasts are directly applicable to day-ahead forecasts with longer horizons.

Based on these two configurations, the three data sets are preprocessed, as detailed in the appendices. The `R` scripts used to execute the preprocessing are made available online in the `kdayday/solarbenchmarks` Github repository, as are the open-source SURFRAD and CAMS McClear data. Although the ECMWF data cannot be directly shared, the scripts used to access and preprocess the ECMWF data are also provided for those users with ECMWF permissions.

4. Benchmark forecast methods

Based on the literature review in Section 2, six probabilistic benchmark types were selected for comparison using the case study data: climatology, a Ch-PeEn, a PeEn, a raw NWP ensemble, a Gaussian error distribution, and an MCM model. In this section, we show example implementations of each class for both hourly forecasts and intra-hourly forecasts, where applicable. For each class, we also discuss practical considerations based on data availability. Forecasts are generated for all times when the sun is up (see Appendix); for all methods, and the CDF for times when the sun is down is a step function at 0 W/m^2.

Before delving into each forecast type, we will discuss a method that is common to five of them: the empirical CDF. Our objective is to generate a full CDF rather than a confidence interval or selected

quantiles. For this purpose, the empirical CDF is widely applied because it is ostensibly the simplest approach to generate a complete CDF from a set of discrete data. For a set of n data points $S(t) = \{X_1(t),...,X_n(t)\}$, the empirical CDF $\hat{P}$ of the forecast variable x is defined by:

$$\hat{P}\left(x, t\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i(t) \leqslant x), \tag{1}$$

where $\mathbb{1}$ is the indicator function. That is, each point in the data set is considered equally likely, resulting in a "stepped" CDF with a jump at each discrete value.

Although the empirical CDF appears straightforward, there are myriad permutations on its implementation. For example, the `quantile` function in the `R stats` package offers nine variations on the empirical CDF. The implementation can change the interpolation among the discrete members and might result in very different tail behavior. For instance, linearly interpolating between the members rather than applying discrete steps is a simple variant that smooths the forecast distribution (`quantile` function `type = 4`), resulting in the binned probability forecast proposed in Anderson (1996). For a non-negative parameter such as solar irradiance, reasonably assuming the lower tail of the empirical CDF begins at 0 W/m^2 rather than the lowest NWP ensemble member could result in very different tail behavior. Further discussion about three common implementations is also given in Lauret et al. (2019), including whether and how to assign probability outside the bounds of $S(t)$, i.e., below the lowest data point and above the highest data point.

Whatever choice is made, authors should clearly state which implementation has been applied. For the case study shown here, we obey the strict definition in (1). The classic stepped empirical CDF has minimum and maximum values delimited by the minimum and maximum values in the forecast data set (`quantile` function `type = 1`). Except for the Gaussian error distribution, the other five reference classes use this empirical CDF; the key differences are in the selection of the input data points (i.e., the X_is).

4.1. Climatology

The climatology forecast is the empirical CDF of measurements during a long period of time. That is, the set of input data S in (1) is a long set of observations, so the distribution $\hat{P}(x, t)$ is essentially static (no dependence on t) or changes very slowly as new data accumulates. The implementation of a climatology forecast is the same irrespective of forecast lead time, update rate, and horizon. Hourly average observations could be used for an hourly resolution forecast, 5-min averages for a 5-min resolution forecast, etc. In addition to this simplicity, missing data handling is trivial for climatology—missing values are simply left out of the training set.

To generate the data set S, a long historical data set (e.g., during many decades) is typically seen as ideal, though it assumes that the climate is static. In solar forecasting applications, two potential issues arise. First, some sites might have only very recent history available, given the newness of the hardware installation, e.g., pyranometer, PV power meter. In the solar forecasting field, the SURFRAD sites used in this case study are some of the best examples of a long historical data set with observations extending to between 1994 and 2003 (Augustine et al., 2005; NOAA Earth System Research Laboratory, n.d.). Second, there are indicators that the static climate assumption does not hold and that climate change is impacting available renewable resources (Craig et al., 2018; Craig et al., 2019). In this regard, using only recent observations (e.g., last year or few years) might be preferable to capture recent trends and extremes.

In our example implementations, the hourly and intra-hourly forecast training sets S consist of all available (nonmissing), sunup hourly average or 5-min average GHI observations, respectively, from 2018. That is, we generate and validate in-sample forecasts useful for hindcasting validation only. To provide an operational forecast, the training data set must contain only observations available prior to the issue time. Purely from the perspective of methods comparisons, using the in-sample validation data to create climatological hindcasts has the advantage of representing a perfectly reliable forecast. Using a sufficiently long (> year) training data set, however, should show similar though not quite as exact reliability, so this could be a less important consideration for practitioners (Yang, 2019b).

The classic characteristics of a climatology forecast are evident in Fig. 2, which shows 3 days of time-series forecast distributions and observations during the spring for the SURFRAD site in Boulder, CO. The hourly example consists of 12 consecutive forecast runs with 6-h horizons each; the intra-hourly example shows 72 consecutive forecast runs with 1-h horizons each. As illustrated, the climatology forecast lacks a key attribute: forecast resolution, or the ability to generate case-dependent forecasts (Lauret et al., 2019). This definition of resolution is distinct from the temporal resolution ($\mathcal{R}$) of a forecast run. All daylight times are forecasted identically regardless of knowledge about the present conditions. This forecast has long-term reliability (Section 6), but it lacks both sharpness and the ability to represent solar irradiance/power's dependency on sun position and weather.

4.2. Complete-history persistence ensemble

The CH-PeEn forecast is a variant on climatology that captures solar's diurnal trend through a static daily cycle. These case study CH-PeEn implementations are based on the method in Yang (2019b). First, the GHI values in the historical data set are translated to CSI using the CAMS McClear CSI estimate. Then, CSIs are binned into 24 sets by hour of day for both the hourly and intra-hourly resolutions. To generate a

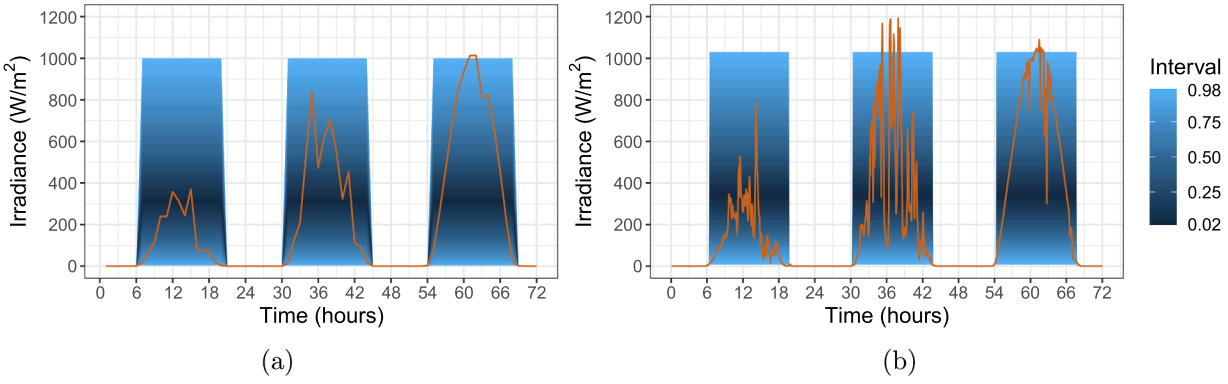

Fig. 2. Climatology forecast samples for 3 spring days in Boulder for the (a) hourly forecast and (b) intra-hourly forecast. The fan plots show the prediction intervals of the probabilistic forecasts, from the 1% to 99% central intervals. The orange line shows the observed hourly or 5-min average irradiance, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

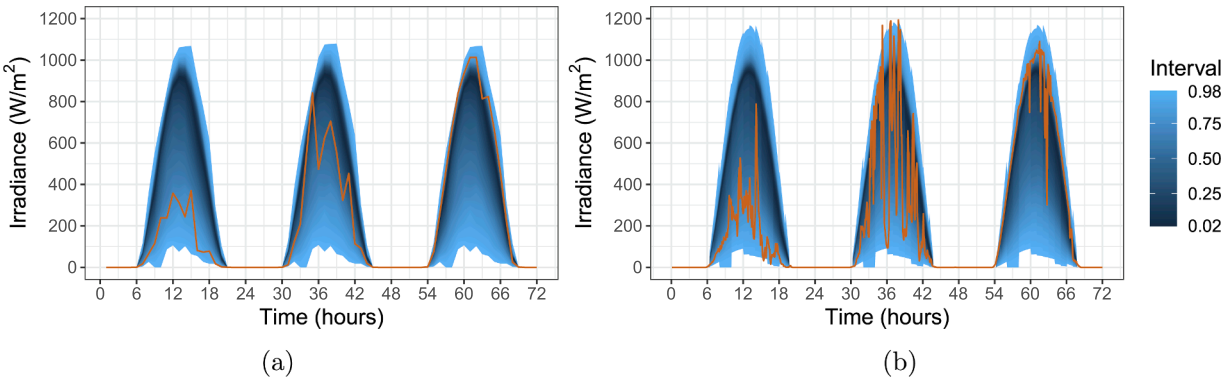

Fig. 3. CH-PeEn forecast samples for 3 spring days in Boulder for the (a) hourly forecast and (b) intra-hourly forecast.

set of GHI values to define an empirical CDF, the data set $S(t)$ is defined as the product of the CSI estimate at time t and the set of CSIs from the same hour of day. For the intra-hourly forecast with 5-min resolution, all 12 forecasts that fall in the same hour of day receive the same set of clear-sky indices.

Typical time-series forecasts are shown in Fig. 3. The envelope of the forecast follows the sun's diurnal trajectory, but the daily forecast shape is repeated, accounting only for the slow change in clear-sky irradiance. This is a baseline that accounts for current sun position and aerosols but not clouds and other weather impacts. As with the climatology forecast, the case study hindcasting comparison uses the in-sample 2018 data to generate the data set $S(t)$. For hindcasting purposes, Yang (2019b) shows the calendar year of training data selected (in-sample vs. out-of-sample) to generate a CH-PeEn forecast has minor impact on the forecast characteristics. For operational, out-of-sample forecasts, this CH-PeEn implementation would require both a different year of training data and updates to an operational clear-sky irradiance estimate.

4.3. Persistence ensemble

Compared to the previous benchmarks, the PeEn benchmark attempts to achieve basic forecast resolution (Lauret et al., 2019) by accounting for some combination of sun position and weather. Unlike the static climatology forecast or the CH-PeEn's simple dependence on clear-sky irradiance, the PeEn benchmark attempts to forecast each time uniquely based on recent data. Each probabilistic forecast is an empirical CDF where the set $S(t)$ in (1) comprises a much smaller subset of recent data, assuming that those conditions will persist into the future. Typical PeEn methods for intra-hourly and hourly forecasts in the literature result in quite distinct forecast characteristics, so we consider each timescale separately.

A key consideration for a reasonable PeEn benchmark is the selection of which variable should be persisted at a given temporal scale. Based on the available measurements, the options include CSI, GHI/DHI, or PV power. For intra-hourly forecasts, the set $S(t)$ commonly comprises the most recent n GHI/DNI/power observations from the past hour or two, for which any of these available variables should be reasonable choices (Chu and Coimbra, 2017; El-Baz et al., 2018). The exceptions are in the shoulder hours when the sun rises and sets quickly, which justifies the use of CSI instead. More fundamentally, the first hour after sunrise presents a complication of how a PeEn should be defined when no data are available to persist yet. This is usually simply ignored in the literature.

For the intra-hourly PeEn implemented in this case study, the $n = 24$ CSI estimates from the previous 2 h are persisted, inspired by David et al. (2016) and Pedro and Coimbra (2015). To address the issue of forecasting early in the day when CSI estimates are unavailable, the first forecast run issued each day is a deterministic clear-sky GHI forecast, i.e., the probabilistic forecast is a step function at the forecasted clear-sky GHI. The second forecast run uses a PeEn of the last hour's $n = 12$ observed CSIs, and by the third forecast run, a full 24-member PeEn is available.

In contrast, hourly PeEns typically persist observations at the same hour of the day from the last n days rather than using intra-day information (Alessandrini et al., 2015; Sperati et al., 2016). If looking at the same hour of the day during an intra-seasonal period (e.g., 20 days), persisting GHI is reasonable; however, there are alternative implementations where CSI is more appropriate, if available. Persisting GHI across more than an hour within the same day as in Lauret et al. (2017) should be avoided because of known changes in solar position; in this case, persisting CSI would be a more realistic alternative. In the case study implementation, the hourly PeEn comprises the previous $n = 20$ available GHI measurements at the same hour of the day, following the commonly referenced method in Alessandrini et al. (2015).

Examples of the two PeEn implementations are shown in Fig. 4. The intra-hourly forecast is highly dependent on recent conditions, whereas the hourly forecast is broader, capturing a range of conditions experienced during previous days. The hourly resolution forecast is more

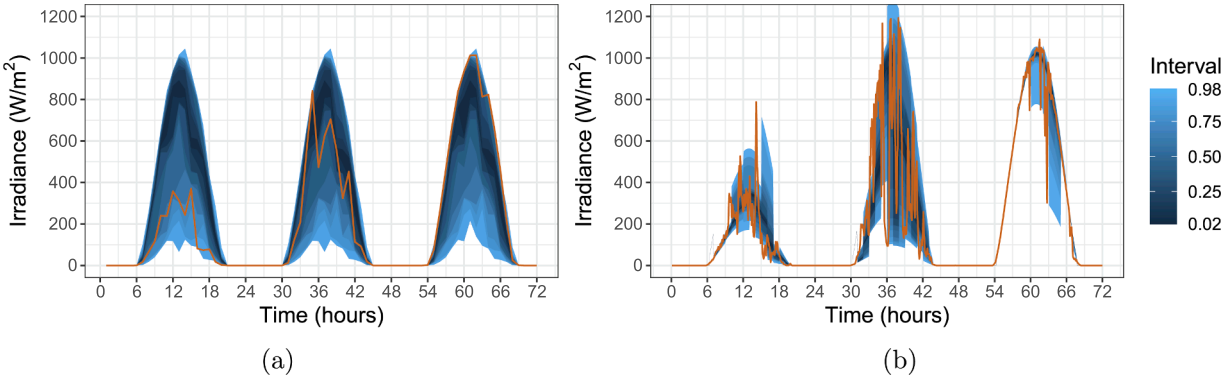

Fig. 4. PeEn forecast samples for 3 spring days in Boulder for the (a) hourly forecast and (b) intra-hourly forecast.

granular than the CH-PeEn and shows slow evolution as the PeEn changes from one day to the next. The intra-hourly forecast, in contrast, changes very quickly. Because the CSI PeEn is maintained over each hourly forecast run, outliers are persisted through the next set of forecasts, resulting in erratic jumps in the forecast.

Finally, a practical implementation of a PeEn must also consider impacts of missing data. There is a choice between selecting only up to n members, even if fewer than n members are available in the typical training period, versus always selecting n members, even if that requires extending the training period to find sufficient available data. For these case study implementations, an equal length training period is enforced, even if that results in a PeEn size $< n$. This selection can be customized based on the application, and the PeEn might simply be unsuitable for data sets with significant missing data.

4.4. NWP raw ensemble

NWP ensemble benchmarks are very straightforward to implement: the training set $S(t)$ comprises the NWP ensemble members valid at time t. Each member is weighted equally in the empirical CDF. The case study implementation uses the ECMWF control forecast plus the 50 members of its perturbed ensemble forecast to produce a 51-member ensemble, available at hourly resolution. This ensemble is suitable for hourly forecasting, but because of the NWP computation time, few, if any, NWP ensembles are currently available for intra-hour forecasting. NOAA's HRRR NWP model is one of the only that provides 15-min resolution forecasts, but its ensemble version (HRRR ensemble) is still experimental and at hourly resolution. Therefore, the NWP raw ensemble benchmark is implemented only for the hourly forecast.

The example forecast in Fig. 5 shows typical characteristics: the NWP ensemble is very sharp in comparison to the previous hourly benchmarks and captures both upcoming weather and the diurnal trend. The ensemble members are often too clustered, however, and the observed irradiance falls outside the ensemble—this is a classic NWP ensemble underdispersion, which many post-processing methods seek to address (Leutbecher and Palmer, 2008). Finally, note that NWP models forecast conditions over a grid rather than a single point, so care should be taken during validation to ensure consistency between the grid forecast and the point validation. For the case study, ECMWF's recommended practices for comparing NWP models to pyranometer measurements were followed and are available in the supplementary `R` code.

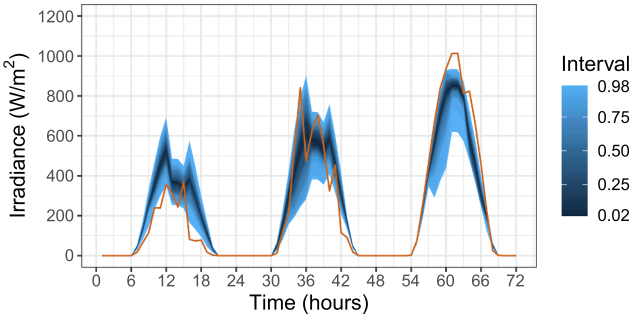

Fig. 5. Hourly ECMWF ensemble forecast samples for 3 spring days in Boulder.

4.5. Gaussian error distribution

The final benchmark, the Gaussian error distribution, is distinct from the others. Rather than generating an empirical CDF as in (1), a doubly truncated Gaussian forecast of irradiance x is issued at time t_k valid for time t according to:

$$P_\Phi\left(x,\, t_k,\, t\right) = \frac{\Phi\left(\frac{x-\mu(t)}{\sigma(t_k)}\right) - \Phi\left(\frac{0-\mu(t)}{\sigma(t_k)}\right)}{\Phi\left(\frac{I_{CS}(t)-\mu(t)}{\sigma(t_k)}\right) - \Phi\left(\frac{0-\mu(t)}{\sigma(t_k)}\right)}, \tag{2}$$

where $I_{CS}(t)$ is the clear-sky irradiance, Φ is the CDF of the standard normal distribution, and μ and σ are the mean and standard deviation of the forecast, respectively. The doubly truncated Gaussian distribution ensures that no probability density is allocated less than 0 W/m^5 or more than the clear-sky irradiance, given the feasible bounds for solar irradiance. Given the uncertainty in true clear-sky irradiance, the lower bound is the more important to enforce, yet untruncated Gaussian distributions are still applied in the literature.

Applying this model requires selecting μ and σ. The mean, μ, centers the forecast, so it is typically a deterministic forecast valid at time t. The standard deviation, σ, determines the forecast uncertainty, and it is typically fit to historical data at the forecast issue time, t_k, and is valid for the entire forecast run. As noted, the literature contains a range of variants on the idea of a Gaussian error distribution, but there is not a unified approach to fitting μ and σ (Lorenz et al., 2009; Mathiesen et al., 2013; Chu et al., 2015; Golestaneh et al., 2016b; Chu and Coimbra, 2017; Grantham et al., 2016). In this case study, implementations are selected to be easily reproducible and require minimal decision-making.

For the intra-hourly forecast, a smart persistence forecast is used for the forecast mean, μ. Smart persistence, or persistence of cloudiness, assumes that the last available CSI estimate will be persisted over the forecast run, persisting weather conditions but accounting for the diurnal trend (Zhang et al., 2015). The standard deviation, σ, is calculated from the past 2 h of smart persistence errors, inspired by Chu and Coimbra (2017) and analogous to the training period used for the intra-hour PeEn. As with the PeEn, the first 1-h forecast run of each day is a deterministic clear-sky GHI forecast. The second hour is the smart persistence forecast based on the last CSI, dressed in a Gaussian distribution based on the first (up to) 12 accumulated residuals—times when the sun is down during the training hour are skipped. Assuming an operational CSI estimate, this approach is readily applicable for operational forecasting, with the same sensitivity to missing data as the PeEn described in Section 4.3.

Rather than relying on smart persistence, the hourly implementation uses the ECMWF control forecast as the forecast mean, μ. All available residuals from the same hour of the day in the 2018 data set are used to calculate the standard deviation, σ, which echoes the CH-PeEn compromise between simply using all available errors (Grantham et al., 2016) and developing a more advanced model based on sun position (Mathiesen et al., 2013; Lorenz et al., 2009). Like the Ch-PeEn forecast, the long residual data set makes this forecast resilient to missing data, but the selection of the relevant residuals would need to be updated to operationalize the forecast.

Resulting forecasts are shown in Fig. 6. The hourly forecast shows how the Gaussian approach is much smoother than the stepped empirical CDFs produced by other approaches. Although the standard deviation at, for example, 12 p.m., is the same from one day to the next, the forecast shape follows the updated information from the ECMWF control forecast. Like the PeEn, the intra-hourly forecast can be extremely sharp during clear skies (hours 54–60), but it shows that the erratic behavior in the observations is reflected in the next forecast run.

4.6. Markov-chain mixture model

In contrast with the Gaussian error distribution's parametric approach, the MCM method is an intra-hourly nonparametric method intended to replicate state changes over the steps in a forecast run. The implementation used here is based on the approach introduced in Munkhammar et al. (2019b) and finalized in Munkhammar et al. (2019a). First, a time-series of training CSIs are binned in N bins evenly

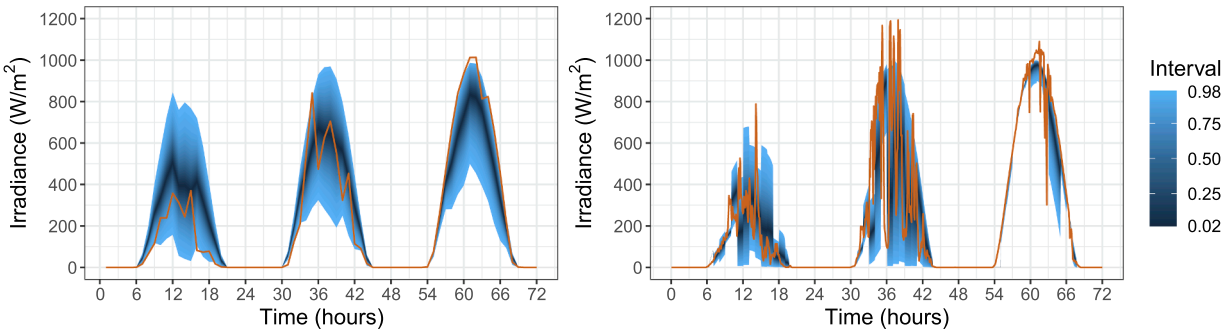

Fig. 6. Gaussian error distribution forecast samples for 3 spring days in Boulder. The hourly forecast (a) is fit to the deterministic ECMWF ensemble, i.e., the control member of the 51-member ensemble. The intra-hourly forecast (b) is fit to a smart persistence deterministic forecast.

divided on $[a, b]$ from the minimum to the maximum value in the training set. Second, an $N \times N$ transition matrix M is estimated from the transitions among the bins in the training time-series. The bin $i \in [1, ...,N]$ of the most recent estimate of CSI at the issue of the forecast run, $CSI\,(t_k)$, is determined; the piece-wise uniform distribution forecast of CSI for step D in the forecast run then corresponds to the i^{th} row of the D^{th} multiple of the transition matrix: $M^D = M{\cdot}M{\cdot}...{\cdot}M$. Like the PeEn and Gaussian error distributions, $CSI\,(t_k)$ is assumed to be 1 for the first forecast run of each day when no data is available yet. The distribution will evolve as D is stepped over the course of a forecast run. To achieve a forecast in terms of uniform quantiles rather than uniform CSI bins, the model is sampled 1000 times for use in the empirical CDF, $\widehat{P}$.

For this implementation, $N = 100$ bins are used, and time-series of training data is selected on a rolling basis using the last 20 days of CSIs, when the sun is up. This approach is similar to training procedure (B) in Munkhammar et al. (2019a) and was selected for its similarity to other 20-day approaches, such as the 20-day hourly-resolution PeEn. While this is an intra-hourly method, the longer training window was chosen because the shorter 2-h training windows for the intra-hourly PeEn and Gaussian error distribution approaches would not yield enough transitions to populate the matrix M. Munkhammar et al. (2019a) note that with a high number of bins, a bin can have zero probability of transition if there were no transitions into and out of the bin in the training set. If the test observation $CSI\,(t_k)$ falls within that bin, the model cannot forecast forward. Using the 20-day training set, this issue was not observed in this case study data. In instances where it does occur, Munkhammar et al. (2019a) suggest forecasting a uniform distribution from $[a, b]$ to transition to the next time step. Repeated occurrences might make this model less resilient to missing data, requiring a lower number of bins or a longer training horizon.

A second potential issue identified in Munkhammar et al. (2019a), which was not observed in their case study, was observed here. The MCM model cannot account for test data that falls outside the range of the training set (i.e., below a or above b). To account for these instances in the case study, the MCM model is extended to account for instances when the observation $CSI\,(t_k)$ falls outside the boundaries $[a, b]$ by assigning it the closest boundary value, a or b, based on the recommendation in Munkhammar et al. (2019a).

The characteristics of the intra-hourly MCM forecast are shown in Fig. 7. The probabilistic forecasts are much broader than either the intra-hourly PeEn or Gaussian error distribution approaches, even during clear skies. It is also noted that the upper tail of the forecast can often extend far beyond the likely clear-sky irradiance. The MCM case study in Munkhammar et al. (2019a) restricted its training and testing data to the 2 h around noon to avoid low solar angles, while we test a broader range of times and solar angles. There are outlier CSIs in the current data set, with CSIs > 2 observed in $< 1\%$ of the 5-min average values; these outlier CSIs, however, can range as high as 72. The impact

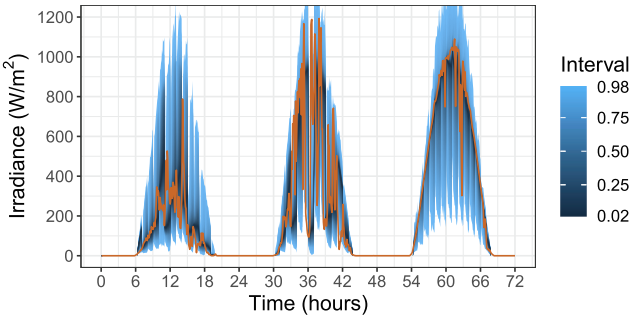

Fig. 7. Intra-hourly MCM forecast samples for 3 spring days in Boulder.

of these outliers can be seen in the very high upper quantiles of the MCM distributions in Fig. 7. In a practical study, a few options might be considered: retaining outliers if the proposed method has more advanced outlier handling or low solar angle modeling or restricting the evaluation period if the model is intended for midday forecasting, as in Munkhammar et al. (2019a).

Finally, it should be noted that while MCM model is only shown here for intra-hourly forecasting, there is no inherent reason why it cannot be applied to hourly-resolution forecasting. However, the lack of exogenous data inputs, such as NWP models, will likely deteriorate its quality over longer horizons, similar to other time-series and machine learning methods. For that reason, we restrict our view to intra-hourly forecasting, which is also the focus of its proposed applications in Munkhammar et al. (2019a) and Munkhammar et al. (2019b).

5. Probabilistic forecast evaluation

In addition to describing the time-series characteristics of the benchmark methods, their aggregate long-term performances will be measured using proper probabilistic metrics and diagnostic techniques. Three salient probabilistic forecast characteristics have been described in the meteorology literature: reliability/calibration, resolution, and sharpness (Gneiting et al., 2007; Gneiting and Raftery, 2007; Lauret et al., 2019). Reliability, or calibration, is the statistical consistency of a probabilistic forecast. That is, in a sufficiently long data set, the nominal coverage rate should equal the observed coverage rate, e.g., the 20% confidence level covers 20% of the observations. A calibrated forecast avoids systemic bias. Reliability/calibration can be assessed visually using a reliability diagram, which plots the observed coverage vs. the nominal coverage to observe deviation from the ideal (Pinson et al., 2007; Lauret et al., 2019).

Forecast resolution is the ability of the method to produce case-dependent forecasts; a static climatology forecast provides a counter-example with zero resolution; however, forecast resolution cannot be measured independently and is usually inferred from other analyses.

Sharpness is a measure of how concentrated the probabilistic information is without considering the resulting observation—a forecast with narrow prediction intervals is sharper than one with broad prediction intervals. Sharpness on its own is not a measure of forecast quality, however, because a sharp forecast can be wildly unreliable. Therefore, Gneiting et al. (2007) states that the objective of a probabilistic forecast is to maximize forecast sharpness, subject to calibration. Sharpness is intuitive to measure via the average interval width, $\bar{\delta}$, of a central $(1-\rho)\times 100\%$interval of interest during an evaluation period T (Pinson et al., 2007):

$$\bar{\delta} = \frac{1}{T}\sum_{t=1}^{T} P^{-1}\left(1-\frac{\rho}{2}, t\right) - P^{-1}\left(\frac{\rho}{2}, t\right), \tag{3}$$

where P is the forecast method's CDF (e.g., $\hat{P}$ or P_Φ). A sharpness diagram that plots the average widths of the 10%, 20%, ···, 90% central intervals can be used to visually assess sharpness (Lauret et al., 2019).

In addition to individual assessments of these characteristics, the continuous ranked probability score (CRPS) has been widely applied to measure all three in one proper metric. CRPS is defined as the squared difference between the forecasted and observed CDFs, where the observed CDF is simply a step function at the observation, y (Hersbach, 2000):

$$\mathrm{CRPS}(t) = \int_{-\infty}^{\infty} [P(x,t) - \mathbb{1}\{x \geqslant y(t)\}]^2 dx. \tag{4}$$

P is the CDF of the forecast variable, x, and $\mathbb{1}$ is the indicator function representing the observed CDF. CRPS is negatively oriented (i.e., lower is better) and presented in the units of the forecasted variable. A forecast that is narrow (sharp) and close to the observed value (resolved) will achieve the best CRPS. During an evaluation period T, an average CRPS, $\overline{\mathrm{CRPS}}$, can be calculated as:

$$\overline{\mathrm{CRPS}} = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty} [P(x,t) - \mathbb{1}\{x \geqslant y(t)\}]^2 dx. \tag{5}$$

Given that this single metric describes multiple desired forecast traits, methods have been proposed to decompose $\overline{\mathrm{CRPS}}$ into its reliability, resolution, and uncertainty components, where uncertainty is a function of the variability in the observations only. Hersbach (2000) proposed a $\overline{\mathrm{CRPS}}$ decomposition suitable for raw ensemble forecasts, which Candille and Talagrand (2005) generalized for any continuous distribution function. Although these decompositions are widely cited as theoretically feasible, only the simpler decomposition in Hersbach (2000) for ensemble forecasts is commonly implemented. The authors note that the decomposition in Candille and Talagrand (2005) is non-standardized and subject to user decision-making, severely limiting its practical implementation; in fact, the authors opted to disregard this decomposition in subsequent work (Candille and Talagrand, 2008). Given the need to compare multiple probabilistic methods and not only ensemble forecasts, we opt to rely on the visual diagnostics of reliability and sharpness diagrams and turn instead to an investigation of forecast tail behavior.

Rather than decomposing $\overline{\mathrm{CRPS}}$ into reliability and resolution components, it can also be decomposed by quantile to illustrate strengths and weaknesses in different regions of the forecast distribution. In the literature, $\overline{\mathrm{CRPS}}$ is often described as the integral of the Brier score (BS) over all thresholds in the dimension of the forecast variable (Hersbach, 2000): $\int_{-\infty}^{\infty}\mathrm{BS}(x,t)dx$; however, each threshold in x does not have a clear interpretation in solar forecasting, where a 200-$\mathrm{W/m^2}$ threshold has a much different meaning at sunrise than at noon. Instead, $\overline{\mathrm{CRPS}}$ can alternatively be decomposed as the integral of the quantile score (QS) over all quantiles (Laio and Tamea, 2007; Gneiting and Ranjan, 2011):

$$\overline{\mathrm{CRPS}} = \int_0^1 \frac{1}{T}\sum_{t=1}^{T}\mathrm{QS}_\xi(P^{-1}(\xi,t), y(t))d\xi, \tag{6}$$

where the QS at the level $\xi \in (0,1)$ is:

$$\mathrm{QS}_\xi = 2(\mathbb{1}\{y(t) \leqslant P^{-1}(\xi,t)\} - \xi)(P^{-1}(\xi,t) - y(t)). \tag{7}$$

Without the constant scaling factor of 2, QS is also known as the pinball loss (Steinwart and Christmann, 2011). Through this decomposition, diurnal trends are naturally accounted for in the time-varying forecast CDFs, P. Additionally, different regions of the distribution can be weighted more heavily to illustrate strengths and weaknesses that can be obscured by $\overline{\mathrm{CRPS}}$ (Lauret et al., 2019; Gneiting and Ranjan, 2011). For example, the lower tail of the distribution might be of particular interest to power system operators because times when solar power is unexpectedly low are more likely to impact system reliability. Gneiting and Ranjan (2011) proposes weighted quantile scores of the form $\mathrm{wQS}_\xi = w(\xi)\mathrm{QS}_\xi$, which can be substituted into (6) to calculate a weighted average CRPS, $\overline{\mathrm{wCRPS}}$, which preferentially scores selected areas of the distribution. Two quantile-weighting functions are applied as defined by Gneiting and Ranjan (2011): a left tail (w_l) and right tail (w_r) weighting function, for each level $\xi \in (0,1)$:

$$w(\xi) = \begin{cases} w_l(\xi) = (1-\xi)^2, & \text{if left-tail weighted} \\ w_r(\xi) = \xi^2, & \text{if right-tail weighted} \end{cases} \tag{8}$$

In the case study results shown next, the benchmark forecasts are compared using these diagnostics: unweighted and weighted $\overline{\mathrm{CRPS}}$ values to compare aggregate performance as well as reliability and sharpness diagrams to illustrate different characteristics. For these benchmarks to be useful in practice, however, the proposed methods should be compared to a selected benchmark to illustrate improvement. Authors can use a skill score, such as the CRPS skill score, to quantify improvement over the benchmark (Lauret et al., 2019):

$$\mathrm{CRPSS} = 1 - \frac{\overline{\mathrm{CRPS}}_{proposed}}{\overline{\mathrm{CRPS}}_{benchmark}}. \tag{9}$$

Using a proper score such as $\overline{\mathrm{CRPS}}$, a forecast with negative skill score is worse than the benchmark, a skill score of 0 is on par with the benchmark, and a skill score of 1 is ideal.

6. Comparative results of benchmark methods

Using the entire year 2018 data set, each of the ten benchmark methods summarized in Table 3 (5 hourly, 5 intra-hourly) were assessed for the 7 SURFRAD sites using the diagnostic techniques described in Section 5. The annual aggregate results further demonstrate the main features of each benchmark method illustrated for the 3 days in Figs. 2–7. For example, Fig. 8 shows the average widths of the central 10%–90% intervals for the hourly forecasts for the SURFRAD site in Boulder, CO. As expected, climatology is consistently the least sharp and has the broadest intervals, whereas the NWP ensemble—known for having clustered ensemble members—is consistently the sharpest. Adding the Gaussian distribution around one ECMWF member produces a significantly broader forecast than the ensemble approach. Interestingly, the PeEn and Ch-PeEn show similar sharpness, showing that for this site, a 20-day subset can have about as much spread as the entire year.

As stated, however, sharpness in the absence of reliability does not indicate a high-quality forecast. To investigate statistical reliability/calibration, Fig. 9 shows the reliability of the 1st to 99th percentiles by comparing the nominal percentile to the proportion of observations that fell below the corresponding quantile. Though not sharp, climatology and its relative, the CH-PeEn, show perfect reliability—this is expected by definition, given that the forecast is made from the sample set. The 20-day PeEn shows reasonable reliability as well as a stepped characteristic because of the relatively small set of discrete points used to define its empirical CDF. The ECMWF ensemble, in contrast, shows clear reliability deficiencies. Although it is the sharpest forecast, it is underdispersed: the lower quantiles are too high (observed far more

Table 3
Summary of Case Study Benchmark Implementations.

Benchmark Class	CDF	Temporal Scale	Training Data Selection
Climatology	$\hat{P}$	Hourly	2018 1-h average GHI
		Intra-hourly	2018 5-min average GHI
CH-PeEn	$\hat{P}$	Hourly	2018 CSIs from same hour-of-day + hourly clear-sky GHI
		Intra-hourly	2018 CSIs from same hour-of-day + 5-min clear-sky GHI
PeEn	$\hat{P}$	Hourly	GHI from same hour-of-day from previous 20 d
		Intra-hourly	CSI from previous 2 h + clear-sky GHI forecast
ECMWF Ensemble	$\hat{P}$	Hourly	51-member ECMWF ensemble
		Intra-hourly	—
Gaussian Error Distribution	P_Φ	Hourly	ECMWF control forecast + errors from same hour-of-day in 2018
		Intra-hourly	Smart persistence forecast + errors from previous 2 h
MCM	$\hat{P}$	Hourly	—
		Intra-hourly	1000 samples from MCM model trained on previous 20 days of CSIs

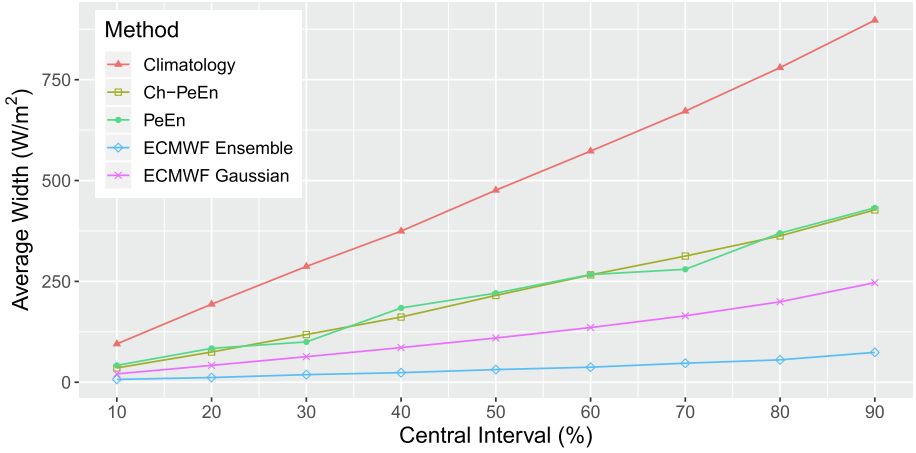

Fig. 8. Average width of 10%–90% central intervals for five benchmark hourly resolution forecasts for the Boulder SURFRAD site in 2018.

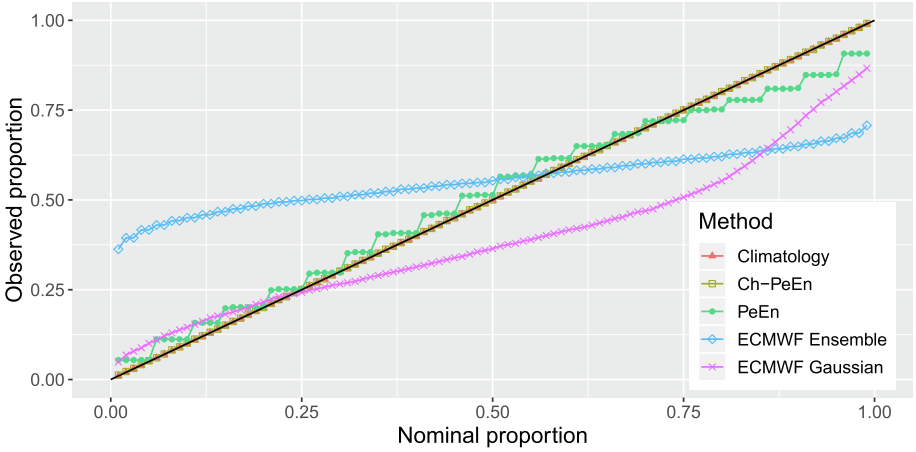

Fig. 9. Reliability diagram of the 1st to 99th percentiles for five benchmark hourly resolution forecasts for the Boulder SURFRAD site in 2018.

frequently than expected), and the upper quantiles are too low (not observed frequently enough), and it shows that the 50th percentile might be slightly biased too low. Dressing the ECMWF control member in a Gaussian distribution produces a more reliable lower tail, but the forecast is regularly biased too high.

When turning to the intra-hourly temporal resolution, Figs. 10 and 11 show that the reliability and sharpness characteristics of the climatology and CH-PeEn methods are the same. Two of the other methods—PeEn using the last 2 h of data and the smart persistence Gaussian error distribution—are much sharper. The MCM model, trained on the last 20 days of data, is not quite as sharp, particularly at the outer (70–90%) central intervals. In contrast with the 20-day PeEn for hourly resolution forecasts, the intra-hourly version is not nearly as reliable and tends to be underdispersed—that is, it underestimates the spread of uncertainty. The smart persistence Gaussian error distribution shows even more bias and regularly underestimates the observed irradiance. The MCM model, in contrast, shows much better calibration than either the PeEn or Gaussian error distribution, while being slightly

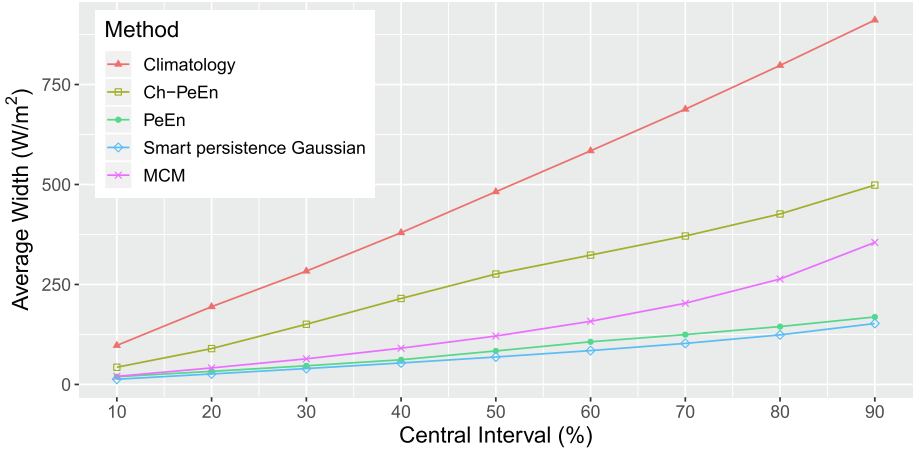

Fig. 10. Average width of 10%–90% central intervals for five benchmark intra-hourly forecasts for Boulder SURFRAD site in 2018.

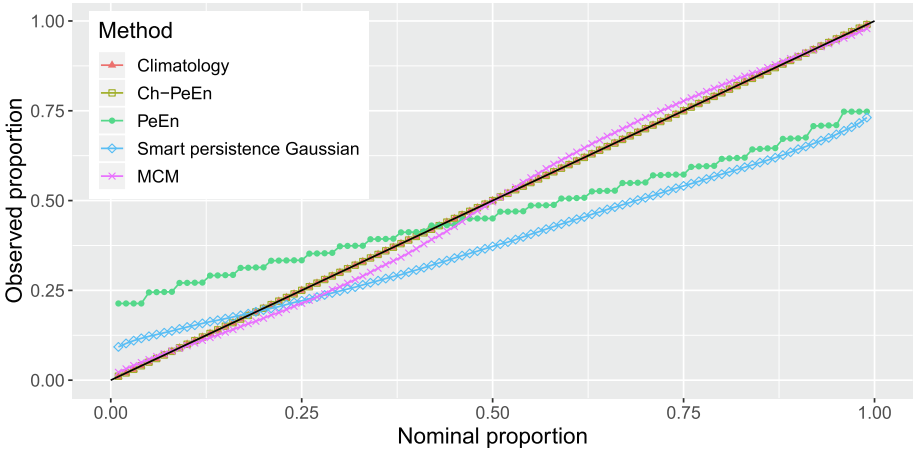

Fig. 11. Reliability diagram for five benchmark intra-hourly forecasts for Boulder SURFRAD site in 2018.

overdispersed in the middle of the distribution.

The $\overline{\text{CRPS}}$—both unweighted and weighted—measures the overall performance of each of these methods for each SURFRAD site, reported in Table 4 for the hourly resolution forecasts and Table 5 for the intra-hourly forecasts. Though not achieving perfect reliability, the ECMWF ensemble and Gaussian error distribution of the ECMWF control member achieve the best (lowest) $\overline{\text{CRPS}}$ for the hourly resolution forecasts, reflecting the benefits of their forecast resolution and sharpness. The one exception is for the Desert Rock site in Nevada, which achieves the best $\overline{\text{CRPS}}$ using the CH-PeEn forecast. This site is an outlier because of its very dry and clear weather, for which this simple forecast capturing the diurnal trend appears competitive.

The ECMWF ensemble and Gaussian error distribution also achieve the best left-tail and right-tail weighted $\overline{\text{CRPS}}$, with the NWP ensemble performing slightly better for the right tail and the Gaussian error distribution performing slightly better for the left. These trends can be seen in more detail in Fig. 12, which shows the quantile score decompositions both with and without the tail-weighting functions. In Fig. 12(a), the ECMWF ensemble's quantile scores are skewed, with the right tail significantly outperforming the left tail because of the ensemble's tendency to underestimate the range of low outcomes. The Gaussian error distribution scores in Fig. 12(b) are much more symmetrical, with the right tail only slightly outperforming the left tail.

For the intra-hourly forecasts, the MCM model achieves the best

Table 4
Average unweighted and weighted CRPS [W/m^2] over 2018 for hourly-resolution forecast methods: climatology (CLI), CH-PeEn (CH-P), PeEn, ECMWF ensemble (NWP), and ECMWF control Gaussian error distribution (GAU). The best scores are in bold.

	Unweighted $\overline{\text{CRPS}}(w=1)$					Left-weighted $\overline{\text{CRPS}}(w=w_l)$					Right-weighted $\overline{\text{CRPS}}(w=w_r)$				
	CLI	CH-P	PeEn	NWP	GAU	CLI	CH-P	PeEn	NWP	GAU	CLI	CH-P	PeEn	NWP	GAU
Bondville, IL	153	78.1	84.8	**50.8**	52.7	41.2	26.5	27.7	16.7	**16.6**	50.7	20.3	23.3	**15.5**	15.6
Boulder, CO	163	75.7	85.0	64.6	**64.2**	44.8	26.4	29.2	23.9	**21.2**	53.1	19.2	22.2	**17.4**	18.0
Desert Rock, NV	177	**37.7**	47.0	39.2	42.5	51.6	15.0	17.5	**11.0**	13.9	54.7	**8.5**	11.7	14.1	12.1
Fort Peck, MT	146	64.8	70.1	**48.0**	49.9	39.2	22.5	23.7	16.6	**16.0**	48.8	16.5	18.7	**13.9**	14.6
Goodwin Creek, MS	163	82.3	87.8	**56.4**	58.3	44.1	28.4	29.1	**18.3**	18.5	53.5	21.0	23.7	17.4	**17.0**
Penn State, PA	140	83.4	88.0	57.4	**55.1**	35.9	25.5	26.7	19.1	**16.9**	48.4	24.2	26.0	17.1	**16.8**
Sioux Falls, SD	145	74.3	83.5	**49.7**	50.6	38.6	24.9	27.2	17.3	**16.4**	48.9	19.6	22.9	**14.1**	14.5

Table 5
Average unweighted and weighted CRPS [W/m^2] over 2018 for intra-hourly forecast methods: climatology (CLI), CH-PeEn (CH-P), PeEn, smart persistence Gaussian error distribution (GAU), and MCM. The best scores are in bold.

	Unweighted $\overline{\text{CRPS}}(w = 1)$					Left-weighted $\overline{\text{CRPS}}(w = w_l)$					Right-weighted $\overline{\text{CRPS}}(w = w_r)$				
	CLI	CH-P	PeEn	GAU	MCM	CLI	CH-P	PeEn	GAU	MCM	CLI	CH-P	PeEn	GAU	MCM
Bondville, IL	157	92.1	52.8	52.8	**48.7**	42.4	30.5	16.9	**15.7**	15.8	52.1	24.5	15.7	17.2	**14.1**
Boulder, CO	166	91.3	61.6	56.7	**51.6**	45.8	31.3	19.9	16.8	**16.6**	54.1	23.7	18.2	18.7	**15.2**
Desert Rock, NV	173	47.3	35.2	36.2	**29.4**	50.6	18.5	11.7	10.7	**10.2**	53.0	11.0	10.1	12.0	**8.1**
Fort Peck, MT	149	77.0	46.3	46.1	**39.8**	40.1	26.2	14.6	13.3	**12.9**	49.5	20.1	14.0	15.7	**11.7**
Goodwin Creek, MS	168	98.4	59.7	57.9	**52.5**	45.6	33.0	19.1	17.2	**17.1**	55.1	25.9	17.7	18.9	**15.0**
Penn State, PA	146	98.1	60.0	56.4	**53.0**	37.4	29.3	18.4	16.5	**16.4**	50.6	29.1	18.6	18.7	**16.0**
Sioux Falls, SD	149	86.8	47.8	44.0	**41.0**	39.7	28.6	15.0	**12.9**	13.1	49.8	23.4	14.6	14.6	**12.2**

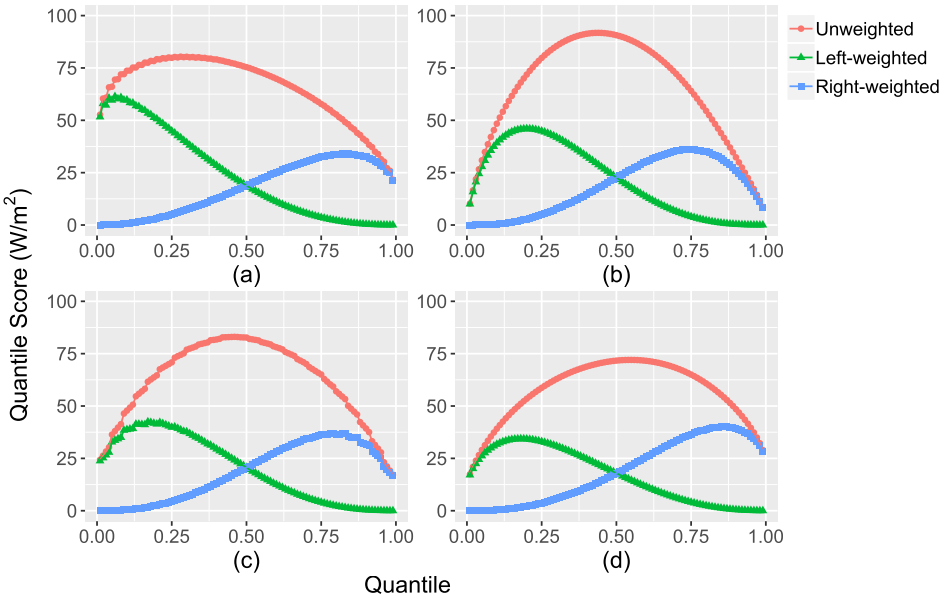

Fig. 12. Quantile score decompositions of the unweighted and weighted $\overline{\text{CRPS}}$ for selected methods for the Boulder, CO SURFRAD site. Top row shows hourly (a) ECMWF ensemble and (b) ECMWF Gaussian error distribution methods. Bottom row shows intra-hourly (c) PeEn and (d) smart persistence Gaussian error distribution methods.

scores in almost all cases, including for its highly reliable tails. Following MCM, the PeEn and smart persistence Gaussian error distribution are competitive with each other, with the Gaussian error distribution outperforming PeEn for the majority of sites. In all cases, the CH-PeEn achieves a distant fourth place, with climatology achieving the worst $\overline{\text{CRPS}}$. In this instance, the MCM model's high reliability and intermediate sharpness outperforms the sharper but unreliable PeEn and Gaussian approaches. Putting aside the MCM model, the Gaussian error distribution consistently performs better for the left tail-weighted $\overline{\text{CRPS}}$, while the PeEn consistently performs better for the right tail, echoing the trends shown in Fig. 11. The quantile score decompositions for these two methods for the Boulder site are shown in Fig. 12c) and – both are relatively symmetrical, but slightly skewed, resulting in the tail performance also shown in the weighted $\overline{\text{CRPS}}$ scores.

In addition to the selected results shown in Figs. 2–11 for the Boulder site, the open-source `kdayday/solarbenchmarks R` repository on Github generates more than 200 figures showing the results for all sites, including probability integral transform (PIT) histograms, quantile score decompositions, and sharpness and reliability diagrams. Additionally, the `solarbenchmarks` code generates data files with the 1st to 99th forecast quantiles for each benchmark and site, for use by other researchers.

7. Recommendations and conclusions

This paper reviewed and implemented ten probabilistic solar forecast variants from six benchmark classes relevant for both hourly and intra-hourly forecasting. While the hourly resolution forecasts were illustrated here with an intra-day horizon, it is important to note that the hourly-resolution methods are equally applicable to day-ahead forecasts with longer and potentially overlapping horizons. The benchmarks range from reliable but low resolution—such as climatology and CH-PeEn—to the very sharp but less reliable NWP ensemble and Gaussian error distribution approaches. None is an ideal forecast in every respect; however, the strengths of each can be used to bound the space where method improvements can be made.

To that end, we recommend that future researchers use at least two benchmarks to compare their proposed methods: a highly reliable (though potentially naive) yardstick benchmark and a highly resolved or state-of-the art point-on-the-yardstick benchmark. Although meteorology has classically relied on climatology for the former, we recommend CH-PeEn as in Yang (2019b). CH-PeEn, which is essentially climatology tailored to solar forecasting's diurnal trend, has the same firm reliability as climatology but gains some basic forecast resolution. Unlike PeEn, a long historical data set makes CH-PeEn resilient to missing data issues. CH-PeEn also improves upon the coarse reliability of a PeEn, and it can be applied similarly for both intra-hourly and

hourly forecasts. Any decent operational forecast should be able to beat this benchmark in terms of resolution and CRPS, though likely with some degradation in reliability.

The second benchmark should describe a less naive benchmark, ideally closer to the state of the art. For an hourly forecast, a raw NWP ensemble is a natural choice. NWP models are widely used in operation and therefore give a useful point of comparison. Typically, raw NWP ensembles are very underdispersed, resulting in a highly sharp though unreliable forecast. Together, CH-PeEn and the raw NWP ensemble can bound the desired region of forecast characteristics: on one extreme is a benchmark with high reliability/low sharpness, and on the other is a benchmark with high sharpness/low reliability. Comparing the reliability diagrams, sharpness diagrams, and CRPS skill scores of a proposed method to these benchmarks can help position it within that region.

For intra-hourly forecasting, there is not an obvious choice to use as this second benchmark. Intra-hourly probabilistic forecasts are still rare in operation. The proposed methods in the literature rely on a variety of statistical and machine learning approaches, so there is less of a clear state of the art than with hourly resolution forecasting. Of the five methods reviewed here, however, the MCM model clearly outperformed the other four in terms of $\overline{\text{CRPS}}$, with the PeEn and Gaussian error distribution providing alternative options. While it can experience convergence failure due to zero probability transitions or outlier test data, reasonable workarounds are identified in Munkhammar et al. (2019a).

In addition to the two generic benchmarks that allow easier comparison among articles, authors may wish to include more advanced benchmarks to contrast among methods within the article. While the benchmark methods provided here are developed to be reproducible and widely applicable, they do not make use of exogenous information such as wind speed and temperature. For novel methods that use exogenous information, a fairer comparison will also include benchmarks with the same amount of information as inputs. For example, a paper could include the generic MCM approach, as well as a variant that downselects training data, conditional on the values of exogenous variables, like the atmospheric flow approach in Mathiesen et al. (2013).

Finally, we also recommend that other authors include a data handling section or appendix to describe small but important details, such as missing data handling, boundary conditions applied to the forecast distribution, and the programmatic implementation. For example, a truncated vs. an untruncated Gaussian error distribution can result in very different tail behavior—or even a nonsensical forecast. In this paper, implementation details are described throughout, and data preprocessing is documented in the appendices, which follow.

Declaration of Competing Interest

None.

Acknowledgments

The authors thank Will Holmgren and Cliff Hansen for the insightful discussion that inspired this article. Atmospheric ensemble (ENS 15-day) forecast data provided courtesy of ECMWF. Contains modified Copernicus Atmosphere Monitoring Service Information 2018; neither the European Commission nor ECMWF is responsible for any use that may be made of the information this work contains.

This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 33505 and by the National Science Foundation under Grant No. HRD 1619673. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Appendix A. SURFRAD and CAMS McClear Data Preprocessing

Based on the temporal resolutions of the intra-hourly and hourly forecasts, the SURFRAD and McClear data are averaged from 1-min values to 5-min or 1-h values, respectively. Only time periods when average SURFRAD solar zenith angle indicates the sun is at least 5° above the horizon (i.e., zenith angle is $\leqslant 85°$) are used, both for training and validation data. A NetCDF file is produced for each site with the preprocessed average GHI observations, clear-sky GHI estimates, and logical indicators of whether the sun is up for each temporal resolution during the year 2018. A similar file is produced with data from the last 20 days of 2017 for use in the hourly PeEn. All the associated files are provided online the `kdayday/solarbenchmarks` Github repository, including the raw SURFRAD and CAMS McClear data files, the `R` preprocessing script, and the final NetCDF data files.

Appendix B. ECMWF Forecast Preprocessing

Temporal and spatial preprocessing of the ECMWF NWP ensemble members are described here. Although the ECMWF data are not open source, researchers are often given permission to access historical data for research purposes. Unlike the SURFRAD data, the ECMWF data used in this case study cannot be included with this paper; however, the accompanying files include both a batch script used to retrieve data from ECMWF's Meteorological Archival and Retrieval System (MARS) and an `R` script that executes the preprocessing steps summarized as follows. Other researchers with permission to access MARS can replicate these steps to gather the same set of input data used here.

Fifty-one historical forecasts (50 members of the perturbed ensemble forecast + 1 control forecast) were retrieved from MARS for the entire year 2018. Forecast runs are issued at midnight, 6 a.m., noon, and 6 p.m. ($\mathcal{U} = 6$ h) with hourly resolution; the first six forecasts from each run are retained, so that every time step is forecasted using the most recent run. ECMWF reports surface irradiance through its surface solar radiation downward (SSRD) (J/m^2) parameter, which is accumulated irradiance during the modeling period (https://apps.ecmwf.int/codes/grib/param-db?id=169). Average hourly irradiance (W/m^2) is assessed as the difference between subsequent values, divided by the modeling period. The data were retrieved over a latitude/longitude grid of $0.2° \times 0.2°$. The forecast values at the four closest grid points were spatially interpolated using the coordinates of the SURFRAD sites to generate a forecast for each location. Like the SURFRAD and CAMS McClear data, the preprocessed ECMWF data are saved to NetCDF files for each site for use in the `benchmark_forecast_comparison.R` script available in the `kdayday/solarbenchmarks` Github repository.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.solener.2020.05.051.

References

Abuella, M., Chowdhury, B., 2015. Solar power probabilistic forecasting by using multiple linear regression analysis. SoutheastCon 2015, 1–5.

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

Almeida, M.P., Perpiñán, O., Narvarte, L., 2015. PV power forecast using a nonparametric PV model. Sol. Energy 115, 354–368.

Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J. Clim. 9 (7), 1518–1530.

Appino, R.R., Ángel González Ordiano, J., Mikut, R., Faulwasser, T., Hagenmeyer, V., 2018. On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages. Appl. Energy 210, 1207–1218.

Aryaputera, A.W., Verbois, H., Walsh, W.M., 2016. Probabilistic accumulated irradiance forecast for Singapore using ensemble techniques. In: Proceedings of the IEEE 43rd Photovoltaic Specialists Conference (PVSC), pp. 1113–1118.

Augustine, J.A., Hodges, G.B., Cornwall, C.R., Michalsky, J.J., Medina, C.I., 2005. An update on SURFRAD–the GCOS surface radiation budget network for the continental United States. J. Atmos. Ocean. Technol. 22 (10), 1460–1472.

Bessa, R.J., Möhrlen, C., Fundel, V., Siefert, M., Browell, J., Haglund El Gaidi, S., Hodge, B.-M., Cali, U., Kariniotakis, G., 2017. Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry. Energies 10 (9).

Bludszuweit, H., Domínguez-Navarro, J.A., Llombart, A., 2008. Statistical analysis of wind power forecast error. IEEE Trans. Power Syst. 23 (3), 983–991.

Boland, J., Soubdhan, T., 2015. Spatial-temporal forecasting of solar radiation. Renew. Energy 75, 607–616.

Bracale, A., Caramia, P., Carpinelli, G., Di Fazio, A.R., Ferruzzi, G., 2013. A Bayesian method for short-term probabilistic forecasting of photovoltaic generation in smart grid operation and control. Energies 6 (2), 733–747.

Bukhsh, W.A., Zhang, C., Pinson, P., 2016. An integrated multiperiod OPF model with demand response and renewable generation uncertainty. IEEE Trans. Smart Grid 7 (3), 1495–1503.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Quart. J. Roy. Meteorol. Soc. 131 (609), 2131–2150.

Candille, G., Talagrand, O., 2008. Impact of observational error on the validation of ensemble prediction systems. Quart. J. Roy. Meteorol. Soc. 134 (633), 959–971.

Cervone, G., Clemente-Harding, L., Alessandrini, S., Monache, L.D., 2017. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. Renew. Energy 108, 274–286.

Chu, Y., Coimbra, C.F., 2017. Short-term probabilistic forecasts for direct normal irradiance. Renew. Energy 101, 526–536.

Chu, Y., Li, M., Pedro, H.T., Coimbra, C.F., 2015. Real-time prediction intervals for intra-hour DNI forecasts. Renew. Energy 83, 234–244.

Copernicus Atmosphere Monitoring Service (CAMS), n.d., CAMS McClear Clear-Sky Irradiation Service, version 3.1. <http://www.soda-pro.com/web-services/radiation/cams-mcclear>, Accessed: 2019-11-11.

Craig, M.T., Carreño, I.L., Rossol, M., Hodge, B.-M., Brancucci, C., 2019. Effects on power system operations of potential changes in wind and solar generation potential under climate change. Environ. Res. Lett. 14 (3), 034014.

Craig, M.T., Cohen, S., Macknick, J., Draxl, C., Guerra, O.J., Sengupta, M., Haupt, S.E., Hodge, B.-M., Brancucci, C., 2018. A review of the potential impacts of climate change on bulk power system planning and operations in the United States. Renew. Sustain. Energy Rev. 98, 255–267.

David, M., Ramahatana, F., Trombe, P., Lauret, P., 2016. Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. Sol. Energy 133, 55–72.

Davò, F., Alessandrini, S., Sperati, S., Monache, L.D., Airoldi, D., Vespucci, M.T., 2016. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. Sol. Energy 134, 327–338.

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2013. Short-term solar irradiance forecasting using exponential smoothing state space model. Energy 55, 1104–1113.

El-Baz, W., Tzscheutschler, P., Wagner, U., 2018. Day-ahead probabilistic PV generation forecast for buildings energy management systems. Sol. Energy 171, 478–490.

Fahiman, F., Disano, S., Erfani, S.M., Mancarella, P., Leckie, C., 2019. Data-driven dynamic probabilistic reserve sizing based on dynamic Bayesian belief networks. IEEE Trans. Power Syst. 34 (3), 2281–2291.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. Roy. Stat. Soc.: Series B (Stat. Methodol.) 69 (2), 243–268.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. J. Bus. Econ. Stat. 29 (3), 411–422.

Golestaneh, F., Gooi, H.B., Pinson, P., 2016a. Generation and evaluation of space–time trajectories of photovoltaic power. Appl. Energy 176, 80–91.

Golestaneh, F., Pinson, P., Gooi, H.B., 2016b. Very short-term nonparametric probabilistic forecasting of renewable energy generation–with application to solar energy. IEEE Trans. Power Syst. 31 (5), 3850–3863.

Grantham, A., Gel, Y.R., Boland, J., 2016. Nonparametric short-term probabilistic forecasting for solar radiation. Sol. Energy 133, 465–475.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weath. Forecast. 15 (5), 559–570.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. Int. J. Forecast. 32 (3), 896–913.

Huang, J., Perry, M., 2016. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. Int. J. Forecast. 32 (3), 1081–1086.

Iversen, E.B., Morales, J.M., Møller, J.K., Madsen, H., 2014. Probabilistic forecasts of solar irradiance using stochastic differential equations. Environmetrics 25 (3), 152–164.

Kroposki, B., Johnson, B., Zhang, Y., Gevorgian, V., Denholm, P., Hodge, B.-M., Hannegan, B., 2017. Achieving a 100% renewable grid: Operating electric power systems with extremely high levels of variable renewable energy. IEEE Power Energ. Mag. 15 (2), 61–73.

Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci 11, 1267–1277.

Lauret, P., David, M., Pedro, H.T.C., 2017. Probabilistic solar forecasting using quantile regression models. Energies 10 (10), 1591.

Lauret, P., David, M., Pinson, P., 2019. Verification of solar irradiance probabilistic forecasts. Sol. Energy 194, 254–271.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J.W., Morcrette, J.-J., 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos. Meas. Tech. 6 (9), 2403–2418.

Leutbecher, M., Palmer, T., 2008. Ensemble forecasting. J. Comput. Phys. 227 (7), 3515–3539.

Li, P., Yu, D., Yang, M., Wang, J., 2018. Flexible look-ahead dispatch realized by robust optimization considering CVaR of wind power. IEEE Trans. Power Syst. 33 (5), 5330–5340.

Li, S., Park, C.S., 2018. Wind power bidding strategy in the short-term electricity market. Energy Econ. 75, 336–344.

Liu, Y., Shimada, S., Yoshino, J., Kobayashi, T., Miwa, Y., Furuta, K., 2016. Ensemble forecasting of solar irradiance by applying a mesoscale meteorological model. Sol. Energy 136, 597–605.

Lorenz, E., Hurka, J., Heinemann, D., Beyer, H.G., 2009. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2 (1), 2–10.

Lotfi, M., Javadi, M., Osório, G.J., Monteiro, C., Catalão, J.P.S., 2020. A novel ensemble algorithm for solar power forecasting based on kernel density estimation. Energies 13 (1), 216.

Mathiesen, P., Brown, J.M., Kleissl, J., 2013. Geostrophic wind dependent probabilistic irradiance forecasts for coastal California. IEEE Trans. Sustain. Energy 4 (2), 510–518.

Möller, A., Lenkoski, A., Thorarinsdottir, T.L., 2013. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. Quart. J. Roy. Meteorol. Soc. 139 (673), 982–991.

Munkhammar, J., van der Meer, D., Widén, J., 2019a. Probabilistic forecasting of high-resolution clear-sky index time-series using a Markov-chain mixture distribution model. Sol. Energy 184, 688–695.

Munkhammar, J., van der Meer, D., Widén, J., 2019b. Probabilistic forecasting of the clear-sky index using Markov-chain mixture distribution and copula models. In: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC). pp. 2428–2433.

Nagy, G.I., Barta, G., Kazi, S., Borbély, G., Simon, G., 2016. GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. Int. J. Forecast. 32 (3), 1087–1093.

Ni, Q., Zhuang, S., Sheng, H., Kang, G., Xiao, J., 2017. An ensemble prediction intervals approach for short-term PV power forecasting. Sol. Energy 155, 1072–1083.

NOAA Earth System Research Laboratory, n.d., SURFRAD (Surface Radiation Budget) Network. <https://www.esrl.noaa.gov/gmd/grad/surfrad/index.html>.

NOAA National Weather Service, n.d., Environmental Modeling Center. <https://www.emc.ncep.noaa.gov/>.

Panamtash, H., Zhou, Q., Hong, T., Qu, Z., Davis, K.O., 2020. A copula-based Bayesian method for probabilistic solar power forecasting. Sol. Energy 196, 336–345.

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

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

Pinson, P., Madsen, H., 2008. Ensemble-based probabilistic forecasting at Horns Rev. Wind Energy 12, 137–155.

Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G.N., 2007. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 10 (6), 497–516.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

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

Sloughter, J.M., Gneiting, T., Raftery, A.E., 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. J. Am. Stat. Assoc. 105 (489), 25–35.

Sperati, S., Alessandrini, S., Monache, L.D., 2016. An application of the ECMWF ensemble prediction system for short-term solar power forecasting. Sol. Energy 133, 437–450.

Sperati, S., Alessandrini, S., Pinson, P., Kariniotakis, G., 2015. The "Weather Intelligence for Renewable Energies benchmarking exercise on short-term forecasting of wind and solar power generation. Energies 8 (9), 9594–9619.

Steinwart, I., Christmann, A., 2011. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17 (1), 211–225.

Sun, M., Feng, C., Zhang, J., 2020. Probabilistic solar power forecasting based on weather scenario generation. Appl. Energy 266, 114823.

Thorarinsdottir, T.L., Gneiting, T., 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. J. Roy. Stat. Soc.: Ser. A (Stat. Soc.) 173 (2), 371–388.

Thorey, J., Chaussin, C., Mallet, V., 2018. Ensemble forecast of photovoltaic power with online CRPS learning. Int. J. Forecast. 34 (4), 762–773.

Torregrossa, D., Boudec], J.-Y.L., Paolone, M., 2016. Model-free computation of ultra-short-term prediction intervals of solar irradiance. Sol. Energy 124, 57–67.

van der Meer, D., Widén, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew. Sustain. Energy Rev. 81, 1484–1512.

Verbois, H., Rusydi, A., Thiery, A., 2018. Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting. Sol. Energy 173, 313–327.
Wan, C., Xu, Z., Pinson, P., Dong, Z.Y., Wong, K.P., 2014a. Optimal prediction intervals of wind power generation. IEEE Trans. Power Syst. 29 (3), 1166–1174.
Wan, C., Xu, Z., Pinson, P., Dong, Z.Y., Wong, K.P., 2014b. Probabilistic forecasting of wind power generation using extreme learning machine. IEEE Trans. Power Syst. 29 (3), 1033–1044.
Woodruff, D.L., Deride, J., Staid, A., Watson, J.-P., Slevogt, G., Silva-Monroy, C., 2018. Constructing probabilistic scenarios for wide-area solar power generation. Sol. Energy 160, 153–167.
Yang, D., 2019a. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). J. Renew. Sustain. Energy 11 (2), 022701.
Yang, D., 2019b. A universal benchmarking method for probabilistic solar irradiance forecasting. Sol. Energy 184, 410–416.
Zhang, J., Florita, A., Hodge, B.-M., Lu, S., Hamann, H.F., Banunarayanan, V., Brockway, A.M., 2015. A suite of metrics for assessing the performance of solar power forecasting. Sol. Energy 111, 157–175.