Solar Energy 252 (2023) 72–80

Contents lists available at ScienceDirect

Solar Energy

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

Probabilistic solar forecasting: Benchmarks, post-processing, verification

Tilmann Gneiting [a,b,*], Sebastian Lerch [c,a], Benedikt Schulz [b]

[a] Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
[b] Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Englerstr. 2, 76131 Karlsruhe, Germany
[c] Chair of Statistical Methods and Econometrics, Karlsruhe Institute of Technology (KIT), Bluecherstr. 17, 76185 Karlsruhe, Germany

ARTICLE INFO

Keywords:
Clear-sky index
Empirical copula
Isotonic distributional regression
Neural network
Reliability diagram
Post-processing

ABSTRACT

Probabilistic solar forecasts may take the form of predictive probability distributions, ensembles, quantiles, or interval forecasts. State-of-the-art approaches build on input from numerical weather prediction (NWP) models and post-processing with statistical and machine learning methods. We propose a probabilistic benchmark based on a deterministic forecast of clear-sky irradiance, introduce new methods for post-processing that merge statistical techniques with modern neural networks, discuss methods for spatio-temporal scenario forecasts, and illustrate the assessment of predictive ability via proper scoring rules and calibration checks. We expect future solar forecasting efforts to be increasingly probabilistic, and encourage continuing close interaction with operational weather prediction, where innovations based on sophisticated neural networks supplement and challenge traditional approaches.

1. The case for probabilistic solar forecasting

The science of solar resource assessment and forecasting is flourishing, with "hundreds, if not thousands of review papers" having been published in these areas (Yang et al., 2022c, p. 1240) and further growth being anticipated (Hong et al., 2020; Sweeney et al., 2020; Yang et al., 2022a). Arguably, the most critical recent development is a transition from single-valued deterministic to probabilistic forecasts (Gneiting and Katzfuss, 2014; van der Meer et al., 2018; Haupt et al., 2019; Yang, 2019a). Probabilistic forecasts can be issued in the form of probability distributions, ensembles, quantiles, or prediction intervals (Lauret et al., 2019; Hong et al., 2020) that allow for uncertainty quantification and provide crucial input to stochastic programming problems, where optimal strategies for decision makers in the face of uncertainty are sought (Appino et al., 2018; Beykirch et al., 2022), both at individual sites and in the context of spatio-temporal trajectory forecasts over multiple locations and time periods (van der Meer et al., 2020).

A broad consensus has developed that solar forecasts – comprising both solar irradiance and solar power – ought to rely on the output from physics-based numerical weather prediction (NWP) models in concert with post-processing using techniques of statistics and machine learning. Yang et al. (2022c) distinguish five major aspects of solar forecasting, namely, forecasting methodology, post-processing, irradiance-to-power conversion, verification, and materialization of values. We touch on essentially all of these facets, with emphasis on areas where the solar forecasting community might benefit from recent advances in statistical theory and methodology.

Over the past few years, solar forecasting has undertaken major steps towards reproducible science (Stodden et al., 2016; Yang, 2019a), and we applaud the development of forecast contests (Hong et al., 2016) and benchmark datasets (Yang, 2018; Yang et al., 2020b; Wang et al., 2022). In this article, we draw on data and code from Yang et al. (2022b) to illustrate concepts in forecast generation, post-processing, and the assessment of predictive ability. Specifically, we consider day-ahead forecasts of hourly irradiance at SURFRAD stations in the continental United States (Augustine et al., 2005), based on the operational deterministic high-resolution NWP model run by the European Centre for Medium-Range Weather Forecasts (ECMWF). Yang et al. (2022b) apply the Analogue Ensemble (AnEn) method of Alessandrini et al. (2015) to yield a post-processed 11-member ensemble forecast. Generally, an ensemble forecast with m members corresponds to a probability distribution that assigns mass $1/m$ to each of the member values, and the member values can be interpreted as quantiles at level $1/(m+1), \ldots, m/(m+1)$, respectively. The range of the ensemble value yields an equal-tailed prediction interval with nominal coverage $(m-1)/(m+1)$, as illustrated in Fig. 1. We return to the benchmark setting from Yang et al. (2022b) throughout the paper.

The discussion in the remainder of the article is methodological in character, and applies to forecasts of both solar irradiance and solar power, and both at intra-hour, intra-day, and day-ahead horizons. In Section 2 we review recent advances in the development

* Corresponding author at: Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany.
E-mail address: tilmann.gneiting@h-its.org (T. Gneiting).

https://doi.org/10.1016/j.solener.2022.12.054
Received 14 November 2022; Received in revised form 17 December 2022; Accepted 31 December 2022
Available online 1 February 2023

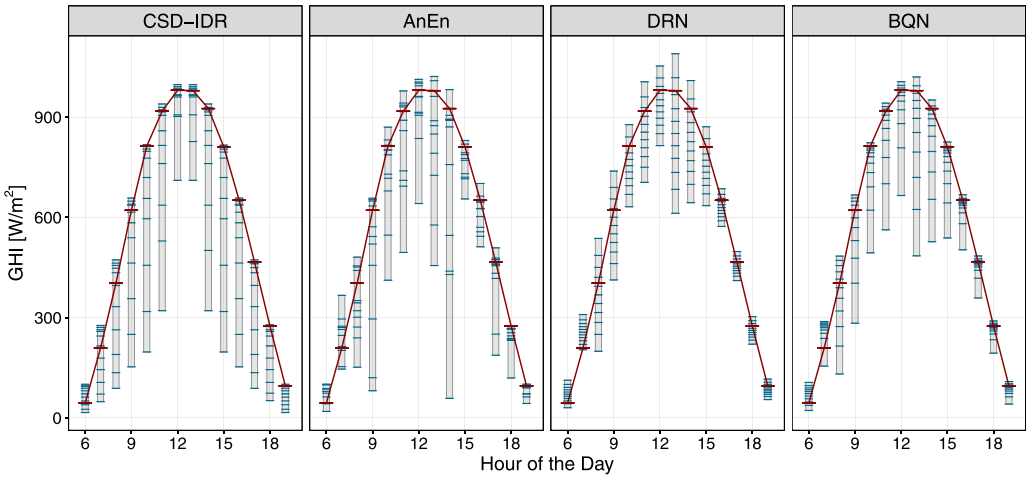

Fig. 1. Probabilistic forecasts of hourly solar irradiance on 10 July 2020 at station Bondville (BON) by the CSD-IDR benchmark of Section 2, the 11-member Analogue Ensemble (AnEn) implemented by Yang et al. (2022b), and the Distributional Regression Network (DRN) and Bernstein Quantile Network (BQN) techniques of Section 3, in W/m^2. The forecasts are restricted to the hours of the day within the benchmark setting of Yang et al. (2022b). The blue bars show predictive quantiles at level $1/12, \ldots, 11/12$, so their range forms a prediction interval with nominal coverage 10/12 or 83.33%. The red bars represent the observed value of Global Horizontal Irradiance (GHI). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of reference forecasts and propose the use of the Isotonic Distributional Regression (IDR) technique of Henzi et al. (2021) to generate probabilistic benchmark forecasts from a deterministic forecast of clear-sky irradiance. In Section 3 we adapt the distributional regression network (DRN) and Bernstein quantile network (BQN) techniques, which have been developed for weather prediction (Rasp and Lerch, 2018; Bremnes, 2020; Schulz and Lerch, 2022), to forecasts of solar irradiance. Section 4 reviews recent progress in the evaluation of probabilistic forecasts (Dimitriadis et al., 2021; Gneiting et al., 2023) and illustrates the use of proper scoring rules and reliability diagrams on a comparison of the mentioned probabilistic forecast techniques. The paper closes with a discussion in Section 5, where we summarize comments and suggestions and make predictions about the future of probabilistic solar forecasting. For implementation details we refer to the respective original literature and the accompanying replication material for the paper (Schulz, 2022).

2. Reference forecasts

Regardless of application field, progress in forecasting techniques needs to be demonstrated relative to reference methods. However, as Hong et al. (2020, p. 382) note, "many papers avoid direct comparisons with classic, established, and state-of-the-art models. Some even skip comparisons with naive models". We appreciate recent advances in solar forecasting, where several authors have studied benchmark forecasts (Yang, 2019a,b,c; Doubleday et al., 2020; Le Gal La Salle et al., 2021; Yang and van der Meer, 2021). In particular, Yang (2019c) proposed the Complete History Persistence Ensemble (CH-PeEn) as a probabilistic benchmark forecast that is based on climatology, while following the diurnal solar cycle — a technique similar in spirit to the Extended Probabilistic Climatology (EPC) approach of Walz et al. (2021) that follows the seasonal cycle of quantitative precipitation. Recently, Le Gal La Salle et al. (2021) proposed Clear-Sky Dependent Climatology (CSD-CLIM) as an alternative probabilistic benchmark. In contrast to CH-PeEn, which stratifies by the time of the day, CSD-CLIM stratifies by clear-sky irradiance values. Essentially, the method uses historical data to bin observed irradiance (or power) values by clear-sky irradiance, and takes the respective empirical distributions as probabilistic forecasts.

Perhaps serendipitously, recent advances in statistical methodology permit the implementation of a very similar technique that enjoys the same desirable properties, while avoiding any binning and guaranteeing optimality on training data. Specifically, Henzi et al. (2021) introduced Isotonic Distributional Regression (IDR), a nonparametric technique that yields simple and flexible probabilistic forecasts based on training data of deterministic predictor variables (e.g., clear-sky irradiance and/or smart persistence) and associated outcomes (irradiance or power). The technique is illustrated in Fig. 2, where we use clear-sky irradiance as the sole predictor variable and GHI observations as the outcome. Described informally, IDR operates under the sole assumption of isotonicity, namely, that higher values of the predictors generate probabilistic forecasts that are stochastically larger, in the sense that the graphs of the respective cumulative distribution functions (CDFs) are nested from left to right. In the univariate case with a single predictor variable, isotonicity refers to the linear order on the real line. In the case of multiple predictor variables, a partial order, such as the componentwise order, needs to be employed.

The constraint of isotonicity regularizes the computational solution and is perfectly suited to solar forecasting; for example, we expect higher values of clear-sky irradiance to yield probabilistic forecasts that are stochastically larger. Subject to the constraint, IDR generates probabilistic forecasts that on the historical data are simultaneously optimal in terms of the continuous ranked probability score (CRPS), the Brier score, and the pinball loss (Section 4, Eqs. (2), (3), and (5)) and many other loss functions. When based on clear-sky irradiance and applied to solar data, we refer to the resulting probabilistic reference forecast, which is a discrete distribution concentrated on the outcomes in the training set, as the Clear-Sky Dependent IDR (CSD-IDR) forecast, as illustrated in Figs. 1 and 2. While here we use the clear-sky irradiance provided with the benchmark data from Yang et al. (2022a), other types of clear-sky models can be used and might be more appropriate choices for real-time forecasting (Yang, 2020a).

As the CSD-IDR technique uses historical data only, it does not require any implementation choices or parameter tuning, and has in-sample guarantees of outperforming CSD-CLIM, hence it satisfies the desirable properties of probabilistic benchmark forecasts listed in Table 1 of Le Gal La Salle et al. (2021). While we provide an initial evaluation of CSD-IDR in Section 4 in this article, we encourage follow-up work that includes direct comparisons to the CH-PeEn and CSD-CLIM benchmarks, ideally with extensions to multivariate probabilistic forecast distributions (van der Meer, 2021).

3. Post-processing

As noted, state-of-the-art solar forecasting relies on output from either a single numerical weather prediction (NWP) model or – often,

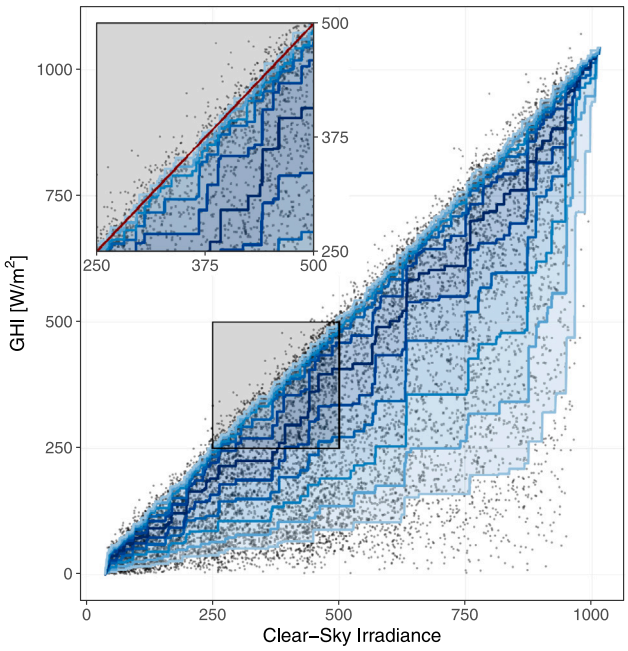

Fig. 2. Illustration of the Isotonic Distributional Regression (IDR) approach on training data from 2017–2019 at station Bondville (BON). The IDR forecast distributions for Global Horizontal Irradiance (GHI) as a function of clear-sky irradiance are represented by predictive quantiles at level $1/12, \ldots, 11/12$, respectively. While the quantile functions are non-crossing, they cluster at high levels, and thus we show a close-up view at upper left. The black dots represent clear-sky irradiance and GHI observations from the training archive. The close-up view also shows the diagonal.

but not necessarily – from ensembles consisting of multiple NWP model runs (Gneiting and Raftery, 2005; Mathiesen and Kleissl, 2011; Bauer et al., 2015; Sperati et al., 2016; Zhang et al., 2022). The generation of the NWP ensemble can be tailored to solar applications in various ways, including but not limited to perturbations of initial conditions, stochastic perturbations, multi physics and multi model approaches (Kim et al., 2022). In this context, the term post-processing refers to the conversion and improvement of raw output from NWP models to skillful solar forecasts, by using statistical and machine learning techniques. While developments of this type were pioneered in weather prediction (Vannitsem et al., 2018, 2021), they are now commonly applied in solar forecasting (Pinson and Messner, 2018; Yang and van der Meer, 2021).

Yang and van der Meer (2021) propose a typology of deterministic-to-deterministic (D2D), probabilistic-to-deterministic (P2D), deterministic-to-probabilistic (D2P), and probabilistic-to-probabilistic (P2P) post-processing. We focus on D2P post-processing, due to our emphasis on probabilistic forecasts, and because a probabilistic forecast can readily be converted to a deterministic one, by extracting the desired summary, such as the mean or a quantile, of the distribution (Gneiting, 2011). The AnEn approach of Alessandrini et al. (2015) is of this type; in a nutshell, it picks forecasts cases that are similar to the one at hand from a historic database, and uses the collection of the respective outcomes as an ensemble forecast.

Many if not most facets of P2P are analogous to those for D2P post-processing. In particular, the commonly used P2P post-processing techniques Bayesian Model Averaging (BMA: Raftery et al., 2005; Doubleday et al., 2021) and Ensemble Model Output Statistics (EMOS: Gneiting et al., 2005; Yagli et al., 2020; Yang, 2020b,c; Schulz et al., 2021) nest D2P post-processing as special cases. The EMOS approach, which is also known as nonhomogeneous regression, assumes that, conditional on raw predictor variables x, the outcome of interest follows a distribution from a parametric family with parameter vector $\theta = g(x)$, where the function g is specified in parametric (e.g., linear) form and commonly referred to as link function. The choice of the parametric families for the forecast distribution and the link function depend on the outcome, with the simplest case arising under the normal family (Gneiting et al., 2005). Recently, post-processing approaches based on neural networks have gained considerable popularity (Wang et al., 2019; Nielsen et al., 2020; Yagli et al., 2022). Here we introduce the Distributional Regression Network (DRN: Rasp and Lerch, 2018) and Bernstein Quantile Network (BQN: Bremnes, 2020; Schulz and Lerch, 2022) approaches and adapt them from meteorological settings, as in the original references, to solar forecasting.

The DRN approach of Rasp and Lerch (2018) builds on, and extends, the framework of EMOS. In lieu of the typically linear, fixed form link function that expresses the EMOS parameter vector in terms of input predictors, the DRN technique uses modern neural networks to learn flexible, nonlinear relations between predictor variables and parameter vectors. Owing to the increased flexibility, the vector of predictor variables may now include NWP forecasts for a range of weather quantities well beyond irradiance. The output of the neural network thus consists of the parameters of the forecast distribution, based on the predictors at hand. The BQN technique introduced by Bremnes (2020) is a fully nonparametric approach, where quantile functions are expressed in terms of basis polynomials. The BQN forecast distribution is defined by the basis coefficients, and the neural network is trained to link the predictors to these coefficients, based on training data. For both DRN and BQN, we adopt and adapt recent implementations by Schulz and Lerch (2022), which differ from the original proposals in technical detail, with the DRN forecast taking the form of a truncated normal distribution, as documented in Schulz (2022). The DRN approach has also been adapted to solar forecasting by Baran and Baran (2022), who employ censored normal distributions and build exclusively on predictor variables that derive from NWP ensemble forecasts of irradiation. Fig. 1 provides an illustration in the benchmark setting of Yang et al. (2022b), with the DRN and BQN forecast distributions being more concentrated than the CSD-IDR and AnEn forecasts. The truncated normal DRN forecast distributions are unimodal and (essentially) symmetric, whereas the nonparametric CSD-IDR and BQN distributions attain flexible shapes and tend to be skewed.

The discussed post-processing methods cover univariate settings only, where a solar variable at a single location and a single lead time is considered. In recent work, van der Meer et al. (2020) study probabilistic forecasts of spatio-temporal trajectories that cover multiple locations and/or lead times. Not surprisingly, the handling of spatial and/or temporal dependencies and interaction poses challenges, and van der Meer et al. (2020, p. 12) conclude that the empirical copula approach, which learns multivariate dependence structures from historical spatio-temporal data, "is a favorable candidate for clear-sky index space–time trajectory generation" from a collection of probabilistic solar forecasts that have been post-processed in a univariate fashion. Their findings echo experiences in hydrometeorological forecasting, where a reordering technique called Schaake shuffle (Clark et al., 2004) has been used for this purpose to much success. As Schefzik et al. (2013) note, the use of the Schaake shuffle is equivalent to the empirical copula approach. Schefzik et al. (2013) furthermore studied an Ensemble Copula Coupling (ECC) approach that derives the empirical copula from the ensemble forecast at hand, as opposed to historical outcomes. For a recent comparison from a meteorological perspective see Lerch et al. (2020). In solar applications, the usage of empirical copula techniques remains underexplored, with the notable exception of work by Alessandrini and McCandless (2020), and we encourage follow-up studies.

4. Forecast evaluation

Across application domains, improvements in forecast methods hinge on our ability to compare competing forecasts, and to diagnose their strengths and weaknesses. The science of forecast verification

Table 1
Brief description of the CSD-IDR, AnEn, DRN, and BQN methods for probabilistic forecasts of hourly solar irradiance, including acronym, full name, dependence (or not) on numerical weather prediction (NWP) model output, and key reference.

Acronym	Name	NWP	Reference
CSD-IDR	Clear-Sky Dependent Isotonic Distributional Regression	No	Henzi et al. (2021)
AnEn	Analogue Ensemble	Yes	Alessandrini et al. (2015)
DRN	Distributional Regression Network	Yes	Schulz and Lerch (2022)
BQN	Bernstein Quantile Network	Yes	Bremnes (2020)

Table 2
Mean CRPS for day-ahead probabilistic forecasts of hourly solar irradiance with the CSD-IDR, AnEn, DRN, and BQN methods at SURFRAD stations in the benchmark setting of Yang et al. (2022a), in W/m^2.

Method	BON	DRA	FPK	GWN	PSU	SXF	TBL
CSD-IDR	82.9	37.9	62.4	85.6	84.6	74.8	74.2
AnEn	56.2	31.5	49.2	59.9	59.5	54.4	61.3
DRN	52.5	28.9	45.5	55.2	55.9	52.5	57.5
BQN	50.8	28.3	44.2	53.9	55.1	50.4	55.6

serves to address these needs, and recent reviews in the solar community cover both deterministic and probabilistic forecasts (Lauret et al., 2019; Yang et al., 2020a). In solar forecasting, probabilistic forecasts are issued in the form of probability distributions, ensembles, quantiles, or prediction intervals (Hong et al., 2020), and we tend to these settings now. A general principle is that probabilistic forecasts ought to maximize sharpness – that is, be as focused and informative as possible – subject to calibration, where the term calibration refers to the statistical compatibility between the forecasts and the outcomes (Gneiting and Raftery, 2007). The terms calibration and reliability are used interchangeably.

Proper scoring rules are omnibus performance measures for comparative forecast evaluation (Gneiting and Raftery, 2007; Jordan et al., 2019). Slightly informally, a scoring rule S is a function that assigns a score or penalty $\mathrm{S}(F, y)$ based on a probabilistic forecast F and the respective outcome y, with smaller values being preferred. A scoring rule is proper if it is designed such that forecasters minimize the expected score or penalty if they issue forecasts that follow their true beliefs about the uncertain outcome, in a well defined technical sense (Gneiting and Raftery, 2007) that induces the closely related notion of a consistent scoring function for deterministic forecasts (Gneiting, 2011). Scores are then averaged across the forecast cases $i = 1, \ldots, n$ in the test set, to yield a mean score of the form

$$\bar{\mathrm{S}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(F_i, y_i), \tag{1}$$

and the method with the lowest mean score $\bar{\mathrm{S}}$ is preferred. Often, skill scores are employed, which equal one minus the ratio between the mean score for the method at hand and the mean score for a reference forecast. Conditional on the reference forecast, this is simply an affine transformation to the skill scale, where positive values indicate performance better than the reference forecast, and negative values performance inferior to the reference forecast. Calibration can be checked diagnostically using Probability Integral Transform (PIT) histograms and reliability diagrams, and sharpness is typically diagnosed via the mean width of an equal-tailed prediction interval.

In what follows, we illustrate these concepts on the benchmark data from Yang et al. (2022b) and compare CSD-IDR, AnEN, DRN, and BQN one day-ahead forecasts of Global Horizontal Irradiance (GHI) at the SURFRAD (Augustine et al., 2005) stations Bondville (BON), Desert Rock (DRA), Fort Peck (FPK), Goodwin Creek (GWN), Penn State (PSU), Sioux Falls (SXF), and Table Mountain (TBL) in the continental United States. Table 1 provides a succinct summary of the forecasting methods that we assess, based on the descriptions in earlier sections and the original references. While the CSD-IDR reference does not depend on NWP models, the AnEn, DRN, and BQN methods leverage output from the high resolution (HRES) model operated by the ECMWF, by involving the solar zenith angle and ECMWF HRES forecasts of 2-m temperature, surface pressure, relative humidity and GHI (via the clear-sky index). We adopt the benchmark setting of Yang et al. (2022a) in all detail; in particular, we restrict the probabilistic forecasts and their evaluation to (essentially) daylight hours, and we aggregate across the considered hours. Data from the years 2017–2019 are used for training, whereas the evaluation period comprises calendar year 2020.

4.1. Predictive probability distributions

In the most general and most powerful setting, probabilistic forecasts take the form of a fully specified probability distribution, F. This distribution F could be parametric, such as the truncated normal distribution in the case of the DRN forecast, a nonparametric distribution specified by a quantile function as for the BQN method, an ensemble forecast as in the case of the AnEn technique, or a nonparametric discrete distribution as for the CSD-IDR benchmark. Intermediate and mixed types of forecast distributions can be employed as well.

In this setting, Lauret et al. (2019) recommend the use of the CRPS, defined by

$$\mathrm{S}(F, y) = \int_{-\infty}^{\infty} (F(t) - \mathbb{1}\{y \le t\})^2 \, \mathrm{d}t, \tag{2}$$

where the probabilistic forecast F is interpreted as a CDF, and we support the recommendation, due to the desirable properties of the CRPS listed in their Table 4. In particular, the CRPS is proper, it reduces to the mean absolute error (MAE) for a deterministic forecast, and it is reported in the same unit as the outcome. In Table 2 we return to the benchmark setting of Yang et al. (2022a) and report the mean CRPS in the unit of Watts per square meter (W/m^2). At all seven stations, there is a clear ordering: the BQN forecast has the lowest (best) mean CRPS, followed by the DRN, AnEn, and CSD-IDR forecasts. At station Bondville (BON), the AnEn, DRN, and BQN forecasts display CRPS skill scores relative to the CSD-IDR benchmark of 0.32, 0.37, and 0.39, respectively.

The PIT is simply the value that the CDF of the probabilistic forecast attains at the outcome. For a calibrated probabilistic forecast, the PIT has a uniform distribution, and deviations from uniformity can be interpreted diagnostically (Gneiting et al., 2007). In Fig. 3 we show histograms of PIT values at station Bondville in 12 equi-spaced bins, a choice that accommodates the 11-member AnEn ensemble forecast, where the bins represent the rank of the outcome relative to the 11 ensemble member values. We note that, while the histograms for the CSD-IDR, AnEn, and BQN forecasts are nearly uniform, there are major deviations for the DRN method, due to its truncated normal assumption and lack of flexibility in the shape of the predictive distributions.

4.2. Probability forecasts

At any given threshold t, a probabilistic forecast in the form of a CDF reduces to a probability forecast $p = F(t)$ for the binary event of the outcome being less than or equal to t. Probability forecasts are often evaluated using the Brier score,

$$\mathrm{S}(p, y) = p^2 \, \mathbb{1}\{y \le t\} + (1 - p)^2 \, \mathbb{1}\{y > t\}, \tag{3}$$

which is proper. As can be seen from Eq. (2), the CRPS of a probabilistic forecast F equals the integral over the Brier score for the induced probability forecasts. In Table 3 we compare the CSD-IDR, AnEn, DRN, and BQN forecasts in terms of the mean Brier score for probability forecasts at the threshold of 250 W/m^2.

To assess the calibration or reliability of a probability forecast, one typically uses reliability diagrams, where the observed nonexceedance probability in the test set is plotted versus the forecast probability. For a reliable forecast, the graph of this function lies on or near the

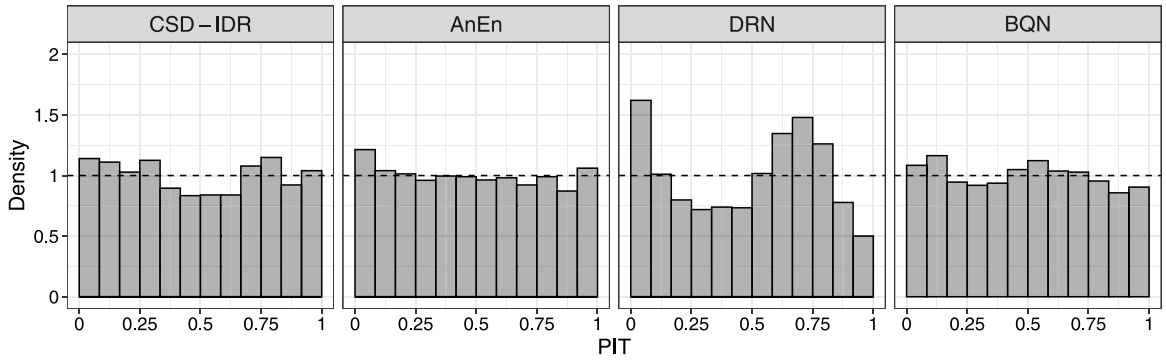

Fig. 3. PIT histograms for day-ahead probabilistic forecasts of hourly solar irradiance with the CSD-IDR, AnEn, DRN, and BQN methods at station Bondville (BON).

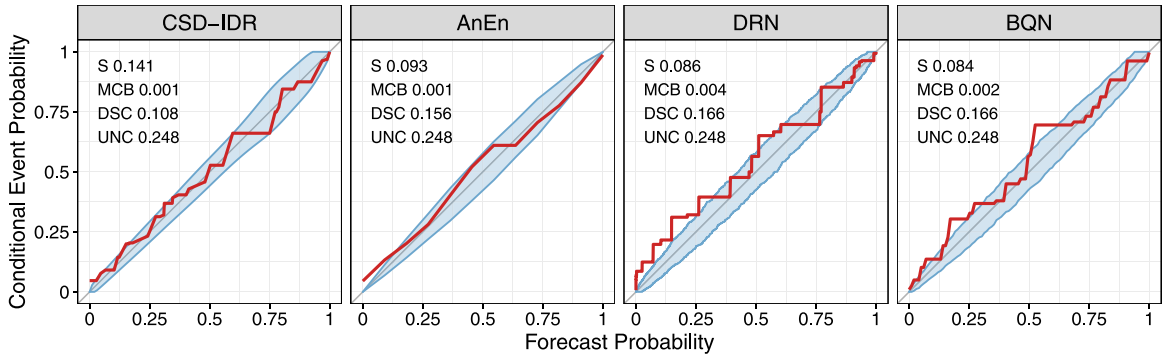

Fig. 4. CORP reliability diagrams for day-ahead probability forecasts of the binary event of solar irradiance being less than or equal to 250 W/m^2, as induced by the CSD-IDR, AnEn, DRN, and BQN forecast CDFs at station Bondville (BON) station. The blue regions show 90% consistency bands under the hypothesis of reliable probability forecasts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Mean Brier score for day-ahead probability forecasts of hourly solar irradiance being less than or equal to 250 W/m^2, as induced by the CSD-IDR, AnEn, DRN, and BQN forecast CDFs at SURFRAD stations in the benchmark setting of Yang et al. (2022a).

Method	BON	DRA	FPK	GWN	PSU	SXF	TBL
CSD-IDR	0.141	0.040	0.107	0.133	0.142	0.124	0.104
AnEn	0.093	0.031	0.086	0.087	0.088	0.085	0.083
DRN	0.086	0.030	0.078	0.082	0.083	0.080	0.079
BQN	0.084	0.028	0.077	0.079	0.083	0.078	0.077

diagonal. However, the usual binning-and-counting approach, where forecast probabilities are binned, and the conditional nonexceedance probability of the respective outcomes is plotted vs. the midpoint of the bin, has serious drawbacks, such as instabilities under the choice of bins. For these reasons, Dimitriadis et al. (2021) introduce a new type of reliability diagram, which they call CORP, for Consistent, Optimal, Reproducible, and Pool-adjacent-violators (PAV) algorithm based, as shown in Fig. 4, along with consistency bands that show variability under the hypothesis of a reliable probability forecast. The CORP approach uses nonparametric isotonic regression and the PAV algorithm (de Leeuw et al., 2009) to generate nondecreasing empirical reliability curves that are optimal in mathematically well-defined ways.

The CORP approach of Dimitriadis et al. (2021) also generates a new type of score decomposition that expresses the mean Brier score,

$$\bar{S} = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC} \tag{4}$$

in terms of miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components, as given in the top-left corner of the CORP reliability diagrams in Fig. 4. The UNC component is simply the mean Brier score for a simple reference forecast that equals the unconditional nonexceedance probability in the test set. Thus, the UNC component does not depend on the forecast considered. The MSC component is the difference of the mean Brier score for the method at hand in its original form, and the mean score after recalibration based on the CORP reliability curve. The DSC component is the difference of the mean Brier score for the recalibrated probabilities and the mean score constant reference forecast.

While in principle the decomposition in (4) is classical, the novelty of the CORP approach lies in its judicious use of the PAV algorithm, which guarantees stability and ensures both nondecreasing reliability diagrams and nonnegative score components (Dimitriadis et al., 2021). From Fig. 4 we see, not surprisingly, that the probability forecasts induced by the DRN and BQN techniques have the highest discrimination ability, which is a proxy for sharpness in the binary case, followed by the AnEn and CSD-IDR methods. However, the DRN forecast is the least reliable.

4.3. Quantile forecasts

Forecasts in the form of quantiles enjoy increasing prominence, and can readily be deduced from fully probabilistic forecasts. For the comparative evaluation of quantile forecasts, consistent scoring functions ought to be used, just as proper scoring rules ought to be used for fully probabilistic forecasts, and we refer to Gneiting (2011), Ehm et al. (2016), Yang et al. (2020a, Section 2.1.1), and Yang and van der Meer (2021, Section 3.1.1) for technical discussion. The most widely used consistent scoring function for the quantile at level α is the asymmetric piecewise linear scoring function or pinball loss that assigns a penalty of

$$\mathrm{S}_\alpha(x, y) = (\mathbb{1}\{y \le x\} - \alpha)(x - y) = \begin{cases} (1-\alpha)(x-y), & y \le x, \\ \alpha(y-x), & y \ge x, \end{cases} \tag{5}$$

for a quantile forecast x and outcome y, with an obvious interpretation in the unit of the outcome.

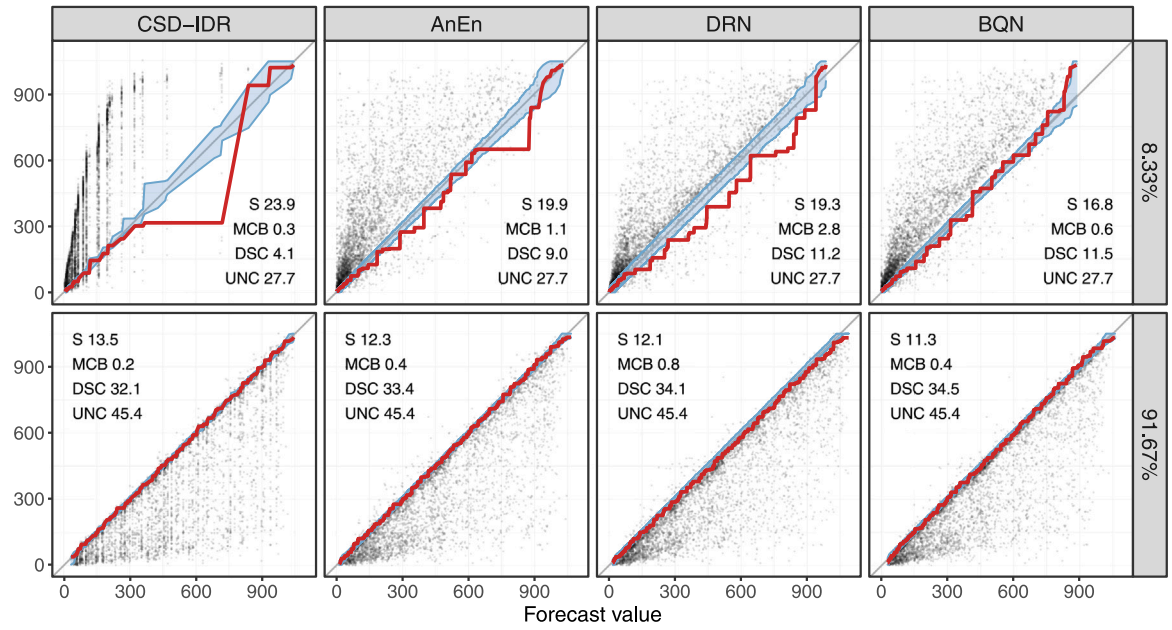

Fig. 5. CORP reliability diagrams for day-ahead quantile forecasts at the 8.33%- and 91.67%-level, as induced by the CSD-IDR, AnEn, DRN, and BQN forecast CDFs at station Bondville (BON). The blue regions show 90% consistency bands under the hypothesis of reliable quantile forecasts. The scatter diagrams show the respective quantile forecasts and outcomes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Mean pinball loss for day-ahead quantile forecasts of hourly solar irradiance at the 1/12 or 8.33%-level, and 11/12 or 91.67%-level, as induced by the CSD-IDR, AnEn, DRN, and BQN methods at SURFRAD stations in the benchmark setting of Yang et al. (2022a), in W/m^2.

Method	BON	DRA	FPK	GWN	PSU	SXF	TBL
8.33%							
CSD-IDR	23.9	21.0	22.1	24.7	22.1	24.1	26.1
AnEn	19.9	15.2	18.2	20.5	18.4	20.7	22.8
DRN	19.3	12.8	17.4	19.7	17.4	21.2	21.2
BQN	16.8	12.7	15.6	17.4	15.7	17.3	19.5
91.67%							
CSD-IDR	13.5	6.1	10.0	13.9	15.6	11.8	11.9
AnEn	12.3	6.6	10.3	13.2	14.8	10.9	12.4
DRN	12.1	6.2	10.5	13.3	14.6	11.8	13.0
BQN	11.3	5.8	9.3	12.1	13.5	10.6	11.2

In Table 4 we compare quantile forecasts derived from the CSD-IDR, AnEn, DRN, and BQN distributions at level $\alpha = 1/12$ and $\alpha = 11/12$, respectively. The levels have been chosen such that they correspond to the smallest and the largest of the 11 members of the AnEn ensemble. Generally, we see the familiar pattern, in that the BQN forecast is superior, followed by the DRN, AnEn, and CSD-IDR methods. Interestingly, at some stations, the CSD-IDR reference outperforms the AnEn and DRN quantile forecasts at the higher level.

As Gneiting et al. (2023) demonstrate, concepts of reliability for quantile forecasts are more subtle than commonly assumed. In particular, unconditional and conditional quantile calibration can be distinguished. For outcomes with continuous distributions, unconditional quantile reliability posits that the fraction of forecast cases in which the outcome is less than or equal to the α-quantile forecast ought to be $(\alpha \times 100)\%$. For example, the diagrams in Lauret et al. (2019, Figure 4) address unconditional quantile calibration in this form. Matters get more complicated when outcomes or forecasts have discrete components, as nonnegligible fractions of data that coincide with the quantile forecast require care and re-interpretation. For details we refer to Section 2.2 of Gneiting et al. (2023).

Conditional quantile calibration is a stronger notion; in a nutshell, the conditional notion posits that, conditional on the quantile forecast attaining a certain value, the distribution of the outcome ought to have said value as its conditional quantile. For diagnostic checks, Gneiting et al. (2023) recommend the use of CORP reliability diagrams for quantiles, as illustrated on the CSD-IDR, AnEn, DRN, and BQN forecasts in Fig. 5. Following the lead of Dimitriadis et al. (2021) and Gneiting and Resin (2021), and now applying to quantiles, rather than probabilities, the CORP approach uses nonparametric isotonic regression and the PAV algorithm to estimate conditional quantiles. The CORP reliability curve shows, conditional on the value of the α-quantile forecast on the horizontal axis, the estimated value of the α-quantile for the outcome distribution. For a forecast that is conditionally calibrated, the respective graph lies at or near the diagonal. Deviations from the diagonal indicate miscalibration.

The CORP approach also generates a decomposition of the type in (4) that now applies to the mean pinball loss. The respective miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) components are shown in the bottom-right resp. top-left corner of the quantile reliability diagrams in Fig. 5. The components are computed and interpreted in ways analogous to those for the Brier score, and they enjoy the same appealing properties. The quantile forecasts induced by the DRN and BQN techniques have the highest discrimination ability, which again is a proxy for sharpness, followed by the AnEn and CSD-IDR methods. An interesting observation applies to the reliability diagram for the CSD-IDR reference forecast for $\alpha = 1/12$. While the reliability curve deviates quite strongly from the diagonal, the miscalibration (MCB) component remains very small, owing to the fact that the CSD-IDR forecast has very few forecast values in the respective range from about 300 to 700 W/m^2. In contrast, the deviations for the DRN forecast occur in regions with very many forecast values, for a much higher MCB component. The BQN quantile forecasts are well calibrated and superior in both the DSC component and the total score.

4.4. Interval forecasts

The most natural and most persuasive way of deriving an interval forecast from a predictive distribution is to consider the equal-tailed,

Table 5
Mean width of equal-tailed interval forecasts of hourly solar irradiance at the 10/12 or 83.33%-level, as induced by the CSD-IDR, AnEn, DRN, and BQN methods at SURFRAD stations in the benchmark setting of Yang et al. (2022a), in W/m^2.

Method	BON	DRA	FPK	GWN	PSU	SXF	TBL
CSD-IDR	387.8	233.0	324.6	414.0	401.6	386.6	398.4
AnEn	271.9	161.2	249.1	290.6	282.5	265.1	312.7
DRN	214.7	106.7	198.0	226.3	234.3	204.4	255.8
BQN	245.6	144.5	218.2	263.4	251.6	239.0	294.3

or central, $(1-\alpha)\times 100\%$ prediction interval, whose lower and upper endpoints are given by the predictive quantile at level $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, respectively. The prediction intervals shown in Fig. 1 use $\alpha = 10/12$.

Issuing an interval forecast of this type is the same as issuing two quantile forecasts. Thus, to derive a proper scoring rule $S_\alpha(l,u;x)$ for interval forecasts, where l and u represent the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantile forecast, a natural approach is to add up the respective pinball losses from (5). After rescaling and reshuffling terms, we obtain the interval score (Gneiting and Raftery, 2007; van der Meer et al., 2018; Yang and van der Meer, 2021),

$$S_\alpha^{\text{int}}(l,u;y) = (u-l) + \frac{2}{\alpha}(l-y)\mathbb{1}\{y<l\} + \frac{2}{\alpha}(y-u)\mathbb{1}\{y>u\}, \tag{6}$$

which is proper and again reported in the unit of the outcome. As Bracher et al. (2021) note, the interval score has three intuitively meaningful, nonnegative components, namely, the width $u-l$ of the interval, which quantifies sharpness, the undershoot penalty term $\frac{2}{\alpha}(l-y)\mathbb{1}(y<l)$ for outcomes y below the lower endpoint l, and an analogous overshoot penalty term $\frac{2}{\alpha}(y-u)\mathbb{1}(y>u)$ for outcomes that exceed the upper endpoint u. Thus, the mean pinball loss decomposes into the respective mean components. Alternatively, the aforementioned CORP decomposition can be applied to each of the pinball losses, and terms can be added up, to yield a decomposition of the form in (4). Either approach is meaningful, and we invite the solar forecasting community to experiment with them, to see which one fits best, to yield the decomposition that Lauret et al. (2019, Table 4) have sought.

To assess calibration or reliability the prediction interval coverage probability (PICP: van der Meer et al., 2018, equation (2.38)) is frequently reported. In the present case of the equal-tailed interval with nominal coverage 10/12, the PICP can be read off the PIT histograms in Fig. 3, namely, as the proportion of the area of the inner 10 bins relative to the total area. Similarly, PICPs at other nominal levels can be read off from PIT histograms. However, nominal PICP alone is insufficient to judge reliability for an equal-tailed interval, as both endpoints could be biased in the same direction — a behavior that can be diagnosed in calibration checks for the endpoints, when viewed as quantile forecasts. The mean width of the prediction interval serves to quantify sharpness, as exemplified in Table 5.

Not surprisingly, the verification results echo those for fully probabilistic, probability, and quantile forecasts. Generally, DRN and BQN outperform AnEn and CSD-IDR interval forecasts. At every station considered, the DRN forecast is sharpest, but also is underdispersed, with a PICP below the nominal level of 83.3%, except at station TBL. The BQN forecast is less sharp, but well calibrated due to the flexible shape of the forecast distribution that is not bound to parametric assumptions such as a truncated normal distribution. Thus, BQN excels in terms of maximizing sharpness subject to calibration and achieves the lowest mean score at every station.

5. Discussion

In recent years, the solar community has embraced advances in the multidisciplinary science of forecasting, by leveraging cumulative progress in numerical weather prediction (NWP), developing post-processing techniques, furthering probabilistic forecasting, and addressing the compelling case for reproducibility, benchmark data, forecast contests, and theoretically principled, both comparative and diagnostic forecast evaluation. Our goal in this article was to draw attention to recent advances in statistical and machine learning methodology that may yield or facilitate further progress. Documented code for the reproduction of our results is available in R (R Core Team, 2022; Schulz, 2022) and can readily be adapted to other settings, for experimentation and comparative evaluation of the proposed CSD-IDR, DRN, and BQN techniques for probabilistic solar forecasting. Furthermore, we encourage the use of proper scoring rules, PIT histograms and CORP reliability diagrams in the assessment of predictive performance.

Not surprisingly, avenues for further research abound. While we have proposed the use of Isotonic Distribution Regression (IDR: Henzi et al., 2021) to generate a probabilistic benchmark technique based on clear-sky irradiance (CSD-IDR), the IDR technique can be applied for post-processing as well. For example, IDR can be applied based on clear-sky irradiance and the irradiance forecast from the NWP model jointly. In the benchmark setting of Yang et al. (2022a), such an approach underperforms the DRN and BQN forecasts, but is on a par with the AnEn approach (Schulz, 2022).

Turning to forecast verification, we have illustrated the recently developed CORP approach that yields improved reliability diagrams, and decomposes the mean Brier score and the mean pinball loss into miscalibration (MCB), discrimination (DSC), and uncertainty components. Similar types of score decompositions for the CRPS pose technical challenges. As Lauret et al. (2019, Appendic C) note, the CRPS can be written as an integral over Brier scores, as in our Eq. (2), and so it can be decomposed into the aforementioned three terms, by integrating over the respective components of the Brier score. However, the CRPS can also be written as an integral over pinball losses (Gneiting and Ranjan, 2011), and so it can also be decomposed by integrating over the respective components for the pinball loss. This raises the question which of the alternatives, amongst other options, ought to be pursued in practice.

Many applied settings call for genuinely multivariate probabilistic forecasts that honor inter-variable, spatial, temporal, and/or spatio-temporal dependence structures, and we have encouraged the use of empirical copula techniques for this purpose. While methods for the verification of multivariate probabilistic forecasts have been developed over the years (Gneiting et al., 2008; Scheuerer and Hamill, 2015; Golestaneh et al., 2016; Thorarinsdottir et al., 2016), progress has been slow and much remains to be done.

Ideally, comparative forecast evaluation ought to be based on a direct assessment of the costs or benefits incurred by the use of competing probabilistic forecast methods (van der Meer et al., 2018, Section 5). Perhaps surprisingly, from a theoretical perspective, an economic assessment of this type is equivalent to the use of proper scoring rules. Stated informally, as shown by Grünwald and Dawid (2004) and reviewed by Brehmer and Gneiting (2020), if one considers the actual cost–loss structure in an applied problem, finds an action that minimizes the expected cost under the probabilistic forecast distribution at hand, computes the actual loss based on said action and the outcome, and averages monetary results over a test set, the approach is equivalent to using the mean score under a proper scoring rule.

We end the paper by encouraging continued close interaction with the community in operational weather prediction, where at this time we are witnessing vigorous development and progress, particularly in innovative uses of sophisticated neural networks that have the potential to supplement or supersede physics-based models, as reviewed by Schultz et al. (2021). While approaches of this type continue to depend on gridded initial conditions from real time data assimilation, as provided by operational weather centers, they promise improved predictive performance at substantially lower cost and generation time than presently used models (Pathak et al., 2022; Keisler, 2022; Bi et al., 2022), and allow for adaptation and tailoring with particular emphasis on solar variables. In all these efforts, in addition to the linkage to weather prediction, a stark emphasis on probabilistic forecasts, comparisons to probabilistic reference techniques, and the use of theoretically principled methods for comparative and diagnostic forecast verification will remain critical to progress in solar forecasting.

Declaration of competing interest

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

Acknowledgments

We thank Dazhi Yang for insightful comments and generous advice on the handling of the benchmark data from Yang et al. (2022b), Daniel Wolffram for help with the illustrations, Alexander Jordan for technical advice, and three anonymous reviewers for insightful and constructive comments. The authors have been working within projects C2 and C5 of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Science Foundation (DFG), Germany. Tilmann Gneiting and Sebastian Lerch are grateful for support by the Klaus Tschira Foundation, Germany.

References

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.

Alessandrini, S., McCandless, T., 2020. The Schaake shuffle technique to combine solar and wind power probabilistic forecasting. Energies 13, 2503.

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

Augustine, R.R., 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. 210, 1207–1218.

Baran, Á., Baran, S., 2022. A two-step machine learning approach to statistical post-processing of weather forecasts for power generation. Preprint, available at https://arxiv.org/abs/2207.07589.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525, 47–55.

Beykirch, M., Janke, T., Steinke, F., 2022. Bidding and scheduling in energy markets: Which probabilistic forecast do we need? In: 17th International Conference on Probabilistic Methods Applied to Power Systems. PMAPS.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2022. Pangu-weather: A 3d high-resolution system for fast and accurate global weather forecast. Preprint, available at https://arxiv.org/abs/2211.02556.

Bracher, J., Ray, E.L., Gneiting, T., Reich, N.G., 2021. Evaluating epidemic forecasts in an interval format. PLoS Comput. Biol. 17, e1008618.

Brehmer, J., Gneiting, T., 2020. Properization: Constructing proper scoring rules via Bayes acts. Ann. Inst. Statist. Math. 72, 659–673.

Bremnes, J.B., 2020. Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. Mon. Weather Rev. 148, 403–414.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004. The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. J. Hydrometeorol. 5, 243–262.

de Leeuw, J., Hornik, K., Mair, P., 2009. Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. J. Stat. Softw. 32, 1–24.

Dimitriadis, T., Gneiting, T., Jordan, A.I., 2021. Stable reliability diagrams for probabilistic classifiers. Proc. Natl. Acad. Sci. 118, e2016191.

Doubleday, K., Jascourt, S., Kleiber, W., Hodge, B.M., 2021. Probabilistic solar power forecasting using Bayesian model averaging. IEEE Trans. Sustain. Energy 12, 325–337.

Doubleday, K., Van Scyoc Hernandez, V., Hodge, B.M., 2020. Benchmark probabilistic solar forecasts: Characteristics and recommendations. Sol. Energy 206, 52–67.

Ehm, W., Gneiting, T., Jordan, A.I., Krüger, F., 2016. Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings (with discussion and rejoinder). J. R. Stat. Soc. Ser. B Stat. Methodol. 78, 505–562.

Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.

Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 243–268.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annu. Rev. Stat. Appl. 1, 125–151.

Gneiting, T., Raftery, A.E., 2005. Weather forecasting with ensemble methods. Science 310, 248–249.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. J. Bus. Econom. Statist. 29, 411–422.

Gneiting, T., Resin, J., 2021. Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. Preprint, available at https://arxiv.org/abs/2108.03210.

Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17, 211–235.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A.I., Lerch, S., Phipps, K., Schienle, M., 2023. Model diagnostics and forecast evaluation for quantiles. Annu. Rev. Stat. Appl. Available at https://doi.org/10.1146/annurev-statistics-032921-020240 (in press).

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

Grünwald, P.D., Dawid, A.P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Ann. Statist. 32, 1367–1433.

Haupt, S.E., Casado, M.G., Davidson, M., Dobschinski, J., Du, P., Lange, M., Miller, T., Mohrlen, C., Motley, A., Pestana, R., 2019. The use of probabilistic forecasts: Applying them in theory and practice. IEEE Power Energy Mag. 17, 46–57.

Henzi, A., Ziegel, J.F., Gneiting, T., 2021. Isotonic distributional regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 83, 963–993.

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, 896–913.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., Zareipour, H., 2020. Energy forecasting: A review and outlook. IEEE Open Access J. Power Energy 7, 376–388.

Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringRules. J. Stat. Softw. 90, 1–37.

Keisler, R., 2022. Forecasting global weather with graph neural networks. Preprint, available at https://arxiv.org/abs/2202.07575.

Kim, J.H., Sengupta, M., Dudhia, J., Yang, J., Alessandrini, S., 2022. The impact of stochastic perturbations in physics variables for predicting surface solar irradiance. Atmosphere 13, 1932.

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

Le Gal La Salle, J., David, M., Lauret, P., 2021. A new climatology reference model to benchmark probabilistic solar forecasts. Sol. Energy 223, 398–414.

Lerch, S., Baran, S., Möller, A., Gross, J., Schefzik, R., Hemri, S., Graeter, M., 2020. Simulation-based comparison of multivariate ensemble post-processing methods. Nonlinear Process. Geophys. 27, 349–371.

Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol. Energy 85, 967–977.

Nielsen, A.H., Iosifidis, A., Karstoft, H., 2020. IrradianceNet: Spatiotemporal deep learning model for satellite-derived solar irradiance short-term forecasting. Sol. Energy 228, 659–669.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., Anandkumar, A., 2022. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. Preprint, available at https://arxiv.org/abs/2202.11214.

Pinson, P., Messner, J.W., 2018. Application of postprocessing for renewable energy. In: Vannitsem, S., Wilks, D.S., Messner, J.W. (Eds.), Statistical Postprocessing of Ensemble Forecasts. Elsevier, pp. 241–266.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.

Rasp, S., Lerch, S., 2018. Neural networks for postprocessing ensemble weather forecasts. Mon. Weather Rev. 146, 3885–3900.

Schefzik, R., Thorarinsdottir, T.L., Gneiting, T., 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. Statist. Sci. 28, 616–640.

Scheuerer, M., Hamill, T.M., 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Mon. Weather Rev. 143, 1321–1334.

Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A., Stadtler, S., 2021. Can deep learning beat numerical weather prediction? Philos. Trans. R. Soc. Ser. A 379, 20200097.

Schulz, B., 2022. Replication code for "probabilistic solar forecasts: Benchmarks, post-processing, verification". Available at http://dx.doi.org/10.5281/zenodo.7436179.

Schulz, B., El Ayari, M., Lerch, S., Baran, S., 2021. Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting. Sol. Energy 220, 1016–1031.

Schulz, B., Lerch, S., 2022. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. Mon. Weather Rev. 150, 235–257.

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

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., Taufer, M., 2016. Enhancing reproducibility for computational methods. Science 354, 1240–1241.

Sweeney, C., Bessa, R.J., Browell, J., Pinson, P., 2020. The future of forecasting for renewable energy. WIREs Energy Environ. 9, e365.

Thorarinsdottir, T.L., Scheuerer, M., Heinz, C., 2016. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. J. Comput. Graph. Statist. 25, 105–122.

van der Meer, D., 2021. A benchmark for multivariate probabilistic solar irradiance forecasts. Sol. Energy 225, 286–296.

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

van der Meer, D.W., Yang, D., Widén, J., Munkhammar, J., 2020. Clear-sky index space-time trajectories from probabilistic solar forecasts: Comparing promising copulas. J. Renew. Sustain. Energy 12, 026102.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenković, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., Ylhaisi, J., 2021. Statistical postprocessing for weather forecasts – Review, challenges and avenues in a big data world. Bull. Am. Meteorol. Soc. 102, E681–E699.

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

Walz, E.M., Maranan, M., van der Linden, R., Fink, A.H., Knippertz, P., 2021. An IMERG-based optimal extended probabilistic climatology (EPC) as a benchmark ensemble forecast for precipitation in the tropics and subtropics. Weather Forecast. 36, 1561–1573.

Wang, H., Lei, Z., Zhang, X., Zhou, B., Peng, J., 2019. A review of deep learning for renewable energy forecasting. Energy Convers. Manage. 198, 111799.

Wang, W., Yang, D., Hong, T., Kleissl, J., 2022. An archived dataset from the ECMWF ensemble prediction system for probabilistic solar power forecasting. Sol. Energy 248, 64–75.

Yagli, G.M., Yang, D., Srinivasan, D., 2020. Ensemble solar forecasting using data-driven models with probabilistic post-processing through GAMLSS. Sol. Energy 208, 612–622.

Yagli, G.M., Yang, D., Srinivasan, D., 2022. Ensemble solar forecasting and post-processing using dropout neural network and information from neighboring satellite pixels. Renew. Sustain. Energy Rev. 155, 111909.

Yang, D., 2018. SolarData: An R package for easy access of publicly available solar datasets. Sol. Energy 171, A3–A12.

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, 022701.

Yang, D., 2019b. Making reference solar forecasts with climatology, persistence, and their optimal convex combination. Sol. Energy 193, 981–985.

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

Yang, D., 2020a. Choice of clear-sky model in solar forecasting. J. Renew. Sustain. Energy 12, 026101.

Yang, D., 2020b. Ensemble model output statistics as a probabilistic site-adaptation tool for satellite-derived and reanalysis solar irradiance. J. Renew. Sustain. Energy 12, 016102.

Yang, D., 2020c. Ensemble model output statistics as a probabilistic site-adaptation tool for solar irradiance: A revisit. J. Renew. Sustain. Energy 12, 036101.

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H.G., Blaga, R., Boland, J., Bright, J.M., Coimbra, C.F.M., David, M., Frimane, A., Gueymard, C.A., Hong, T., Kay, M.J., Killinger, S., Kleisl, J., Lauret, P., Lorenz, E., van der Meer, D., Paulescu, M., Perez, R., Perpinan-Lamigueiro, O., Peters, I.M., Reikard, G., Renné, D., Saint-Drenan, Y.M., Shuai, Y., Urraca, R., Verbois, H., Vignola, F., Voyant, C., Zhang, J., 2020a. Verification of deterministic solar forecasts. Sol. Energy 210, 20–37.

Yang, D., van der Meer, D., 2021. Post-processing in solar forecasting: Ten overarching thinking tools. Renew. Sustain. Energy Rev. 140, 110735.

Yang, D., van der Meer, D., Munkhammar, K., 2020b. Probabilistic solar forecasting benchmarks on a standardized dataset at Folsom, California. Sol. Energy 206, 628–639.

Yang, D., Wang, W., Gueymard, C.A., Hong, T., Kleissl, J., Huang, J., Perez, M.J., Perez, R., Bright, J.M., Xia, X., van der Meer, D., Peters, I.M., 2022a. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. Renew. Sustain. Energy Rev. 161, 112348.

Yang, D., Wang, W., Hong, T., 2022b. A historical weather forecast dataset from the European centre for medium-range weather forecasts (ECMWF) for energy forecasting. Sol. Energy 232, 263–274.

Yang, D., Wang, W., Xia, X., 2022c. A concise overview on solar resource assessment and forecasting. Adv. Atmos. Sci. 39, 1239–1251.

Zhang, G., Yang, D., Galanis, G., Androulakis, E., 2022. Solar forecasting with hourly updated numerical weather prediction. Renew. Sustain. Energy Rev. 154, 111768.