Verification of deterministic solar forecasts

Dazhi Yang[a,*], Dennis van der Meer[b], Jan Kleissl[c], Bri-Mathias Hodge[d], Jamie M. Bright[e], Jie Zhang[f], Cyril Voyant[g], Yves-Marie Saint-Drenan[h], Merlinde J. Kay[i], Sven Killinger[j], Gordon Reikard[k], Philippe Lauret[g], Mathieu David[g], Elke Lorenz[j], Robert Blaga[l], Marius Paulescu[l], Viorel Badescu[m], Christian A. Gueymard[n], Carlos F.M. Coimbra[c], John Boland[o], Javier Antonanzas[p], Ruben Urraca[p], Oscar Perpiñán-Lamigueiro[q], Fernando Antonanzas-Torres[r], Hadrien Verbois[h], Loïc Vallance[h], Philippe Blanc[h], Dave Renné[s], Hans-Georg Beyer[t]

[a] *Solar Energy Research Institute of Singapore, National University of Singapore, Singapore*
[b] *Department of Engineering Sciences, Uppsala University, Uppsala, Sweden*
[c] *Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA*
[d] *Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, CO, USA*
[e] *Fenner School of Environment and Society, Australian National University, Canberra, Australia*
[f] *Department of Mechanical Engineering, University of Texas at Dallas, Richardson, TX, USA*
[g] *PIMENT Laboratory, University of La Reunion, Reunion, France*
[h] *MINES ParisTech, PSL Research University, Sophia Antipolis, France*
[i] *School of Photovoltaic and Renewable Energy Engineering, University of South Wales, Sydney, NSW, Australia*
[j] *Fraunhofer Institute for Solar Energy Systems ISE, Freiburg, Germany*
[k] *Statistics Department, U.S. Cellular, Chicago, IL, USA*
[l] *Faculty of Physics, West University of Timisoara, Timisoara, Romania*
[m] *Candida Oancea Institute, Polytechnic University of Bucharest, Bucharest, Romania*
[n] *Solar Consulting Services, Colebrook, NH, USA*
[o] *Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA, Australia*
[p] *Department of Mechanical Engineering, University of La Rioja, Logroño, Spain*
[q] *Escuela Técnica Superior de Ingeniería y Diseño Industrial, Universidad Politécnica de Madrid, Madrid, Spain*
[r] *Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile*
[s] *Dave Renné Renewables, Boulder, CO, USA*
[t] *University of the Faroe Islands, Thorshavn, Faroe Islands*

Abstract

The field of energy forecasting has attracted many researchers from different fields (e.g., meteorology, data sciences, mechanical or electrical engineering) over the past decade. Solar forecasting is a fast-growing subdomain of energy forecasting. Despite several previous attempts, the methods and measures used for verification of deterministic (also known as single-valued or point) solar forecasts are still far from being standardized, making forecast diagnosis and comparison difficult.

To diagnose and compare forecasts, the well-established Murphy–Winkler framework for distribution oriented forecast verification is introduced to the solar forecasting community. This framework examines aspects of forecast quality, such as reliability, resolution, association, or discrimination, and diagnoses the joint distribution of forecasts and observations, which contains all time-independent information relevant to verification. To verify forecasts, one can use any graphical display or mathematical/statistical measure to provide insights and summarize the aspects of forecast quality. The majority of graphical methods and accuracy measures known to solar forecasters are specific methods under this general framework.

Additionally, measuring the skillfulness of forecasters is also of general interest. The use of the root mean square error (RMSE) skill score based on clear-sky adjusted persistence is recommended. By standardizing the accuracy measure (i.e., RMSE) and reference forecasting method (i.e., clear-sky adjusted persistence), RMSE skill score allows—with appropriate caveats—comparison of forecasts made using different models, across different locations and time periods.

Keywords: Solar forecasting, Measure-oriented forecast verification, Distribution-oriented forecast verification, Skill score, Clear-sky adjusted persistence

1. Introduction

Climate change intensified by the on-going anthropogenic greenhouse gas emissions poses a broad threat to humanity (Mora et al., 2018).[1] To limit global warming, the mitigation pathways would require substantial emissions reductions over the next few decades (IPCC, 2014). The utilization of solar energy and other forms of renewable and clean energy therefore must step up to fulfill humanity's ever-increasing energy demand. The power grids, which transmit and distribute electricity to end users, are being monitored and controlled by system operators at all times to ensure reliable power delivery. Considering that solar energy is inherently variable, till utility-scale energy storage becomes economically viable globally, operational excellence of the power grids can benefit from accurate solar forecasts. Consequently, solar forecasting is now considered of high value (Martinez-Anido et al., 2016; Huang and Thatcher, 2017).

Surface shortwave radiation fluctuates as a function of cloud cover, aerosols, and other weather variables. Forecasts are used in switching energy sources, planning backup, calculating reserves, and constantly trading power on the electricity market. The horizons covered by modern solar forecasting range from a few seconds to a few days. Over the last decade, the literature has bloomed. A wide spectrum of methods, either physics-based (e.g., sky or shadow imagery, remote sensing, or numerical weather prediction), data-driven (e.g., time series, spatio-temporal statistics, or machine learning), or and a combination or hybrid of both, has been proposed (see Blaga et al., 2019; Yang et al., 2018; van der Meer et al., 2018; Voyant et al., 2017; Antonanzas et al., 2016; Ren et al., 2015; Inman et al., 2013, for reviews). Furthermore, the existing studies span a range of time intervals and locations, with different weather conditions. Because of these differences, the field would benefit from having a general verification framework for forecast diagnosis and standardization of accuracy measures or metrics [2] for forecast comparison.

This article has three missions. The first is to introduce the distribution-oriented forecast verification framework to the solar forecasting community. The idea of using distributions—in particular the joint distribution of forecasts and observations—originates in the work of Murphy and Winkler (1987). A joint distribution contains all time-independent information relevant to verification. As such, it offers a more detailed view than the traditional measure-oriented approach in terms of forecast diagnosis. The second mission is to recommend an accuracy measure that should be universally reported in solar forecasting studies—the root mean square error (RMSE) skill score based on clear-sky adjusted persistence. The third mission is to look into a series of practical issues in terms of forecast verification, so that the recommended procedures can be adopted worldwide.

While there will always be trouble in gaining universal consensus within the community on the appropriate measures and methods, the authors of this work represent a broad range of active researchers in the solar forecasting community. The authors wish the forecast verification procedure herein discussed can allow for greater interpretability of results, and direct—"apples to apples"—comparisons of techniques. We anticipate positive feedback from the rest of the community.

The organization of this study is as follows. The forecast verification problem and the perceived difficulties are elaborated in Section 2. Distribution-oriented forecast verification is discussed and exemplified in Section 3. The recommended accuracy measure is justified in Section 4, alongside with some discussions on practical concerns. Section 5 concludes with a series of recommendations.

*Corresponding author. Tel.: +65 9159 0888.
Email address: yangdazhi.nus@gmail.com (Dazhi Yang)

[1] The Intergovernmental Panel on Climate Change (IPCC) is 95 percent certain that humans are the main cause of current global warming (IPCC, 2014).

[2] "Measures" and "metrics" are distinct concepts in *measure theory*. A measure μ on a set X is a mapping $\mu : \mathcal{A} \to [0, \infty]$ defined on a σ-algebra $\mathcal{A}$ that satisfies non-negativity, null empty set, and σ-additivity, that is $\mu(A) \geq 0 \; \forall \; A \in \mathcal{A}$, $\mu(\emptyset) = 0$, and $\mu(\sqcup_{j\in\mathbb{N}} A_j) = \sum_{j\in\mathbb{N}} \mu(A_j)$, where symbol $\sqcup$ denotes disjoint union (Schilling, 2017). On the other hand, a metric is a distance measure $d : X \times X \to [0, \infty]$ that satisfies definiteness, symmetry, and triangle inequality, that is $d(x, y) = 0$ iff $x = y$, $d(x, y) = d(y, x)$, and $d(x, y) \leq d(x, z) + d(z, y)$, $\forall \; x, y, z \in X$. (Schilling, 2017). Nonetheless, moving out from measure theory, the two terms are often used interchangeably, e.g., "accuracy measure" and "error metrics" use the words "measure" and "metric" in their everyday sense. To most forecasters, especially forecast practitioners, they both refer to functions of forecast errors, such as MBE, MAE, or RMSE.

2. Problem description

Solar forecasting is a term applied to any form of estimating the solar energy resource ahead of time. With a fast-growing global portfolio of installed solar energy technology, the need for solar forecasting to facilitate improved operations orchestration and market compatibility is paramount. A rapidly expanding scientific community in the subdomain of energy forecasting have contributed numerous methodologies and approaches towards solar forecasting (Hong et al., 2016). A major goal of all forecasters is accuracy. The variability in solar irradiance intrinsically governs predictability (Pedro and Coimbra, 2015). Therefore, it is particularly interesting to compare forecasts generated by different models, using data from different locations,[3] or data from different time periods (Yang, 2019).

Current methods of solar forecast verification are mostly limited to using measures as indicators of goodness of forecasts. In other words, solar forecasters compare models based on some error metrics, and thus draw conclusions. Under this type of verification procedure, any conclusion is ambiguous in at least two ways: (1) it is unclear what the forecast objective is, and (2) it is unclear how the model of interest performs against other models that are not included in the study. These problems are described in the following two sub-sections, respectively.

2.1. What is a good forecast?

The word "objective" refers to goals given to a forecaster prior to verification. It is natural to think of the objective as "small RMSE," "high skill score," or "high economic value." Nonetheless, these trivial objectives are often ill-motivated, conflicting, and inconsistent. Consequently, one may end up collecting a large, and possibly redundant, set of error metrics (as exemplified by the work of Zhang et al., 2015, in which a suite of 17 metrics was assembled after a lengthy stakeholder process involving members from both the meteorological and power systems communities). In other cases, new metrics are proposed to meet a specific objective (as exemplified by the work of Vallance et al., 2017, in which the ability to follow the ramps in irradiance transients is gauged by two new metrics).

Assembling or introducing new members to a pool of error metrics is meaningful to the field of solar forecasting. By presenting a wide spectrum of error metrics, forecasters are able to choose freely the metrics that can "best" elaborate the strengths of their proposals. There are many studies that propose, contrast, and recommend error metrics to forecasters (e.g., Vallance et al., 2017; Zhang et al., 2015; Hoff et al., 2013; Beyer et al., 2009). However, despite the well-argued discussions, these works can rarely change another forecaster's sentiment towards some specific metrics: for each argument that favors a metric, one may find a counter-argument against it (see Chai and Draxler, 2014; Willmott and Matsuura, 2005).[4] Furthermore, since there are countless publications that discuss and conclude that one metric is better than the other, it is not difficult at all to cite those articles to support any choice the author wishes to make (Chai and Draxler, 2014). The obvious consequence is a field with diverse, subjective, manipulative, and incoherent usage of error metrics. Nonetheless, this is not unique to the emerging field of solar forecasting. Historically, the lack of unified forecast verification procedure has been discussed countless times by many experts from other relatively mature fields (e.g., Murphy and Winkler, 1987; Armstrong, 2001; Fildes et al., 2008), but nothing seems to have changed (Gneiting, 2011).

At this stage, it is essential to ask the question: "what is a good forecast?" It is known, *a priori*, that different metrics favor different forecasts. To put this issue forward, a simulation study is presented. Suppose the hourly clear-sky index, which is the ratio between the global horizontal irradiance (GHI) and clear sky GHI, in Desert Rock (DRA), Nevada (36.624°N, 116.019°W) follows:

$$k_t^* = 1 - z_t^2, \tag{1}$$

[3] Verification of forecasts, particularly those made by a numerical weather prediction (NWP) model, can be carried out spatially (see Gilleland et al., 2010). Spatial averaging or spatial scale has a strong impact on forecast accuracy (Lorenz et al., 2016). However, in this work, we constrain the discussions to forecast verification at single locations.

[4] During the initial stage of this study, the original idea was to propose a specific suite of metrics to the community. However, soon it became obvious that it is literally impossible to make everyone agree. No consensus can be established on things such as whether we should favor MAE or RSME, whether normalized metrics should be used, or whether we should normalize the RMSE using mean, maximum, 1000 W/m^2, or square root of second moment.

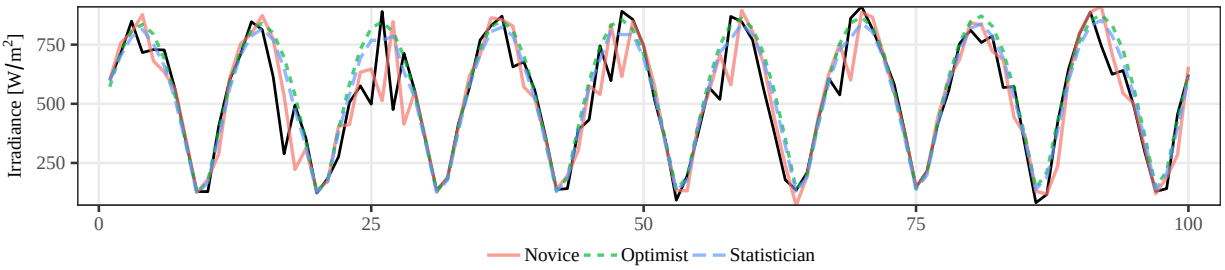

Figure 1: A window of 100 simulated global horizontal irradiance data points (with zenith angle < 85°). Forecasts generated by three forecasters over the same window are overlaid. The *novice* uses 1-step-ahead clear-sky persistence, the *optimist* always uses 0.95 times of clear-sky irradiance as forecasts, and the *statistician* uses the true conditional mean as forecasts.

k_t^* denotes the clear-sky index at time t, and $z_t \sim \mathcal{N}(0, \sigma_t^2)$ follows GARCH(1,1) model, where GARCH is the abbreviation of generalized autoregressive conditional heteroskedasticity, and parameterized as:

$$\sigma_t^2 = 0.15 z_{t-1}^2 + 0.3\sigma_{t-1}^2 + 0.07. \tag{2}$$

With initial values $z_0 = 0$ and $\sigma_0^2 = 0.01$, the GHI time series is simulated for 1000 daylight hours (herein defined to be data points with a zenith angle < 85°), using the actual clear-sky GHI calculated for DRA in spring (Mar, Apr, May). Fig. 1 shows the first 100 data points in the simulated GHI time series. Based on the simulated time series, three forecasters are asked to generate forecasts. The *novice* has no skill to offer, and thus issues 1-step-ahead persistence forecasts on k^*, i.e., $\hat{k}_t^* = \hat{k}_{t-1}^*$. The *optimist* knows it is sunny in Nevada, and always uses $k_t^* = 0.95$. The *statistician* has knowledge about the inherent model, and thus issues the true conditional mean, i.e., $\hat{k}_t^* = 1 - \sigma_t^2$. A length-100 window of their forecasts is overlaid in Fig. 1. To measure the forecast accuracy, mean bias error (MBE), mean absolute error (MAE), and RMSE are used.[5] The results are tabulated in Table 1. The results are inconclusive, because each forecaster is the best in terms of a particular error metric.

Table 1: MBE, MAE, and RMSE, in W/m^2, of the three forecasters in the simulation study. Column-wise best results are in bold.

Forecaster	MBE	MAE	RMSE
Novice	**0.60**	68.83	115.79
Optimist	25.22	**50.16**	91.74
Statistician	1.38	54.77	**86.96**

The result of the above simulation study contradicts the common belief of knowing the inherent (physical or statistical) process is the sole key to making good forecasts. This contradiction is attributed to how we define the goodness of forecasts. To most solar forecasters, "good forecast" is equivalent to "small error." However, the pitfalls of this definition become apparent whenever contradicting rankings of models materialize. In order to resolve such conflicts, we seek solutions from the field of meteorology, where forecast verification is well-studied.

Murphy (1993) outlined three types of goodness that jointly define a good forecast:

I. *consistency*—correspondence between forecasts and judgements;

II. *quality*—correspondence between forecasts and observations; and

III. *value*—incremental benefits of forecasts to users.

[5] By surveying 1000 recent forecasting papers, Yang et al. (2018) found that there are about 20 commonly used metrics in solar forecasting, with MBE, MAE, and RMSE being the most popular ones. They are hence used in the simulation study.

2.1.1. Consistency

The type I goodness,[6] consistency, is quite an abstract concept—a forecast is consistent if it corresponds with the forecaster's best judgement. Murphy (1993) argued that such a judgement must contain an element of uncertainty, because the forecaster's knowledge on the forecasting task is necessarily incomplete. In probabilistic forecasting, consistency can be ensured by adopting *strictly proper scoring rules* (see Gneiting and Raftery, 2007), with that forecasters are rewarded with the best scores if and only if their forecasts correspond with their judgement (Murphy and Winkler, 1971). The Brier score and continuous ranked probability score (CRPS) frequently used in probabilistic solar forecasting are both strictly proper (van der Meer et al., 2018).

On the other hand, in deterministic forecasting, one has to translate his probabilistic judgement through a statistical functional,[7] $T(F)$, which summarizes the forecast distribution, F. While the reader is referred to Gneiting (2011) for the formal definition, informally, the scoring function S is consistent if $\mathbb{E}[S(f, x)] \leq \mathbb{E}[S(r, x)]$, for all $f \in T(F)$, where f is an evaluation of the functional, r is any forecast, and x is a future observation. This definition implies that S is consistent if and only if any $f \in T(F)$ is an optimal forecast under S. For example, if the mean value of a forecaster's judgmental probability distribution is of interest, then RMSE is a consistent accuracy measure, because RMSE is minimized by forecasting the mean of the predictive distribution. In the above simulation study, the *statistician* provided the optimal forecasts under RMSE. The *optimist*, although winning the competition with respect to MAE, did not provide optimal forecasts under MAE since he did not issue forecasts according to the median values (MAE is minimized by forecasting the median of the predictive distribution).

The underlying assumption of using consistency as a measure of goodness of forecasts is that the forecaster receives a *directive* in the form of a statistical functional to convert his probabilistic judgment to a deterministic forecast. For instance, the directive could be "forecast the mean of your probabilistic judgment." Only then, a scoring rule can be identified as consistent if it is optimized by the chosen directive. However, Jolliffe (2008) noted that the definition for consistency is circular, namely, a forecaster could also start by choosing a scoring rule. Once the forecasts are made by optimizing the scoring rule, the consistent directive naturally follows.

Consistency implies an important guideline in choosing accuracy measure during forecasting. For most statistical and machine learning models, the model parameter or weights are estimated or fitted according to some cost function. In this regard, a consistent error measure should be used during verification. For instance, ordinary least squares regression minimizes the sum of squared errors, hence, RMSE is an appropriate metric to report. This guideline is also applicable to statistical ensemble forecasting (also known as forecast combination), where forecasts from different component models are weighed. If the weights are optimized through MAE, then MAE should be used for verification.

2.1.2. Quality

The type II goodness, quality, is a familiar concept to all solar forecasters, as it refers to the correspondence between forecasts and observations. For example, MAE and RMSE are both measures that assess the overall accuracy of forecasts. Accuracy is an aspect of forecast quality. It can be interpreted through measures, which are quantitative. Besides accuracy, other aspects of forecast quality known to solar forecasters, such as bias, association, skill, or uncertainty, can be assessed through MBE, correlation, skill score, or variance. In forecast verification, the traditional way of comparing measures, may it be positively oriented (the larger, the better, such as skill score), negatively oriented (the smaller, the better, such as RMSE), or center oriented (the closer to a center value, the better, such as MBE), is known as the *measure-oriented approach*.

As mentioned earlier, one disadvantage of using the measure-oriented approach is the subjectivity in choosing measures. Since choosing which measures to report is essentially a decision that is internal to a forecaster, it is by default unknown to anyone who is observing the forecast verification procedure from an external view point. In academia, forecasters are authors, whereas observers are editors, reviewers, and readers. If the *optimist* in the simulation study only reports MAE in his article, the observers will not be able to fully realize the underlying pitfalls of his forecasts, but to accept his proposal based on available information. While the simulation study might over-

[6] This terminology follows Murphy (1993) and shall not to be confused with any other definitions.

[7] Any function of a probabilistic distribution is called a statistical functional. Examples of functionals include mean, median, or variance. Generally, it is written as $T(F)$, where F is a distribution.

simplify the state-of-the-art solar forecasting scenarios, it is thought that when the forecasting model gets complex, it would be even more difficult to interpret the forecast results through a few measures.

That said, if two forecasting methods yield the same MBE, RMSE, or skill score, are they equally good? The obvious answer is "no." Measures only provide an *overall* assessment of forecast quality. Since error metrics are often computed based on a collection of samples (e.g., rolling hourly forecasts made over a year), this gives infinite ways to result in the same error-metric value. The reader is referred to Fig. 1 in Vallance et al. (2017) for an example on how two sets of drastically different forecasts can lead to the same RMSE. One solution frequently being used in the solar forecasting literature is to report the regime-dependent error metrics, i.e., to differentiate the errors by classes of prevailing situations. For instance, one can report errors for overcast-, clear-, and all-sky conditions, separately. Alternatively, one can also report errors for different times of day, different times of year, or different day types. However, the dimensionality of forecast verification scales with the number of classes, e.g., an RMSE table will become three, if three sky conditions are analyzed separately, or ten, if ten day types are defined. The error contingency table often gets out of control quickly. What has just been discussed is known as *forecast diagnosis*, which generally means the procedure to understand the *composition* of the overall quality.[8]

Since both assessment and diagnosis of forecast quality are driven by the information embedded in (true out-of-sample) forecast–observation pairs, it is useful to define the total amount of information available to a forecaster during verification. By defining the entirety of information, a forecaster is no longer limited by the set of summary statistics. Stated differently, if the time order of forecast–observation pairs is not of interest, the joint distribution of forecast and observation can be used to study the skillfulness of the forecasts, since it contains *all* time-independent information relevant to forecast verification. This *distribution-oriented approach* to forecast verification was proposed in 1987 by Allan Murphy, together with Robert Winkler (Murphy and Winkler, 1987). It has gained high popularity in the field of meteorological forecasting, but is less known by solar forecasters.

This particular framework needs to be discussed because it provides an alternative view to the traditional measure-oriented approach. It offers high flexibility in terms of accessing the information. In fact, the majority of graphical methods (e.g., Taylor diagram, target diagram, or error heat map) and accuracy measures (e.g., MBE, RMSE, or Kolmogorov–Smirnov test integral) known to solar forecasters are specific methods under this general framework. More importantly, the Murphy–Winkler framework is augmented by Bayes' theorem, in that the joint distribution can be written equivalently as the product of marginal and conditional distributions, making the embedded information more accessible. Last but not least, the distribution-oriented approach establishes communication between forecast quality and accuracy measure. Aside from those aspects of forecast quality mentioned earlier, other aspects such as reliability, resolution, or discrimination can easily be defined and quantified. Whereas more details on the Murphy–Winkler framework are provided in Section 3, with a case study, it is noted that the framework is essential to understanding the goodness of forecasts.

2.1.3. Value

The type III goodness, value, relates to the benefits realized, or cost incurred, by individuals or organizations who use the forecasts during their decision making. Murphy (1993) pointed out that the forecasts by themselves possess no value, as they only acquire value through influencing the decisions made by their users. Most often, the value of solar forecasts is translated into and measured in monetary units, e.g., by reducing the RMSE of the forecasts by x W/m^2, the owner of a photovoltaic (PV) system with energy storage gains an additional \$$y$ per year through optimizing the feed-in strategy of the system. For example, Law et al. (2016) discussed the benefits of improvements in irradiance forecasting for a concentrated solar thermal (CST) power plant in this context. An alternative view was given by Antonanzas et al. (2017) where they compared the profit from different forecasting methods with respect to that from a perfect forecast.

Naturally, such benefits or cost would depend on the characteristics of a particular decision-making problem. Thus, the type III goodness is not under the control of forecasters, but is determined and appreciated by decision makers. Furthermore, this goodness of forecast is non-transferable by default. That is, one cannot simply scale the value realized by others, using the characteristics of the problem at hand. Because of the different courses of action and payoff structures available to different decision makers, there is little reason to assume an *ex post* value would

[8] A related issue is how to decompose these overall quality metrics. Some options are will be discussed in a later section.

apply in an *ex ante* study. A good forecasting strategy that creates high value to some users might not be appreciated by others.

That said, it is believed that for a fixed and well-defined decision-making problem, the mapping between types I and II goodness and type III goodness is *monotone*. In other words, higher forecast quality corresponds to higher value. This gives a forecaster no motivation to issue any suboptimal forecast, and a forecaster can do no better than providing the optimal forecasts (or forecasts to the best of his capability).

To give a perspective on "well-defined decision-making problem," the case of the Australian National Electricity Market (NEM) is considered. In NEM, conventional generators submit bid stacks every five minutes to the Australian Energy Market Operator, that runs a linear program to see how far up the stacks they have to proceed to meet their forecast net load (regional load forecast minus forecast of domestic and commercial PV generation). If the conventional generators miss their promised amount by more than a given tolerance, either above or below, they are penalized. There is presently a dramatic expansion in solar farm construction and these installations are price takers, i.e., not involved in making the spot market price. Hence, the solar plants would not be fined for poor forecasts, but can be curtailed when necessary. Under this regime, the decision-making problem might not be well-defined, as the plant owners could always use the highest possible power generation as forecasts, since there is no monetary penalty on over-forecast. To make fair play in a new regime, the solar plants could be penalized accordingly, if they do not meet their forecasts. Given the new payoff structure, the cost of over-forecasts is equivalent to the cost of running spinning reserve to fulfill the difference between the forecast and the generated solar energy, and the cost of under-forecasts is the lost revenue that would have been generated if that extra potential energy were sold at the prevailing spot price at the time. In this case, if the two costs are the same, then the decision-making problem is well-defined, and the forecasters should submit their optimized forecasts truthfully.

2.2. The skill score

The three types of goodness defined by Murphy (1993) provide a clear objective during forecast verification—while maintaining consistency, one should aim at maximizing quality. However, having a well-defined objective only helps a forecaster to diagnose and thus make conclusions based on his own forecasting experiment. As the literature expands rapidly, it is unrealistic to expand the scope of the experiment simultaneously, that is, to include all previously proposed methods as benchmarks. The obvious reason is that the data (information) available to one forecaster might not be available to others, similarly, not all types of information available at one location or time is available at other places or times. Thus, to ensure the field progresses, the community is forced to make comparison among different research works, based on the reported measures of forecast quality.

2.2.1. A false sense of comparability

The variability of solar irradiance depends on geographical location and timescale. Even if the same forecast-generating strategy is employed, the hourly forecasts made for a location with predominant clear-sky conditions will have a significantly smaller RMSE as compared to 10-min forecasts made for a site with a tropical climate, where cloud formation is rapid and difficult to predict. Hence, if one wishes to compare forecast skills, some form of scaling (normalization) is needed for scale-dependent errors, such as MAE or RMSE. In a recent review paper, Blaga et al. (2019) used the normalized RMSE (nRMSE) as a basis of such comparison.

The particular form of normalization considered in Blaga et al. (2019) is through mean of observations, i.e., nRMSE is computed by dividing RMSE with the mean of observations. Whereas the final conclusion—nRMSE reported in the solar forecasting literature is getting smaller over the years—is factual, the methodology (directly comparing nRMSE) used by the authors can be misleading. "Researchers are getting better at forecasting solar" is *a priori* knowledge, and it is easy to find evidence supporting that.[9] However, nRMSE gives a false sense of comparability, which cannot be used to justify one forecaster has better skill than the other.

[9]It is possible to invoke the law of large numbers (LLN) to justify the conclusion. LLN states that the sample average converges in probability to the expectation. If one treats the reported nRMSE as a random variable, by comparing the large-sample mean of nRMSE reported in earlier 2010 to that reported in the recent years, it is possible to make a conclusion that the expectation of nRMSE has reduced. The underlying assumption is that the nRMSEs reported are independently produced and sufficiently span the multi-dimensional sample space (e.g., climate condition, weather condition, timescale, forecast method). Notwithstanding, the same conclusion can be drawn using RMSE, MAE, nMAE, or any other error measure. This does not mean that all of these measures can be used to directly compare the skill of forecasters.

As mentioned earlier, forecast error is tightly linked to predictability, which is essentially related to variability and uncertainty. Equivalently, we can say that conditions with higher variability and uncertainty are harder to forecast, and thus one should expect larger errors. In forecasting, variability and uncertainty are often quantified by step change and variance, respectively. Since the ultimate aim is to have a measure that quantifies forecast skill, its dependency on variability and uncertainty has to be minimized if not removed completely. It is now clear that mean-normalized nRMSE cannot be used to compare forecast skills, because the mean is related to neither variability nor uncertainty. The same arguments can be applied to range-normalized nRMSE, max-normalized nRMSE, capacity-normalized nRMSE, and the various versions of nMAE.

2.2.2. *On the propagation of normalized accuracy measures in solar forecasting*

Normalized accuracy measures are popular in solar forecasting (Blaga et al., 2019). This contrasts the field of meteorology, where normalized accuracy measures are rarely used. For instance, in the 267-page book "Forecast Verification: A Practioner's Guide in Atmospheric Science" by Jolliffe and Stephenson (2012), there is not a single sentence that discusses normalized accuracy measures. Similarly, no trace of normalized accuracy measures can be found in Hyndman and Koehler (2006), in which the accuracy measures are discussed in the context of general-purpose univariate time series forecasting. Hence, a probable explanation on why normalized accuracy measures is so popular in solar forecasting is given next.

The notion of normalization develops naturally when the forecast quantities are at different scales. On this point, the class of accuracy measures based on percentage errors, such as the mean absolute percentage error (MAPE), needs to be discussed. Measures based on percentage errors are not quite feasible in solar forecasting since the irradiance and PV-generated power is near zero during early mornings and late afternoons, or when the clouds move in. Although the early morning and late afternoon cases can be trimmed with a zenith-angle filter before verification, a few missed forecasts on large irradiance swings during mid-day are enough to result in a very large MAPE. Therefore, to allow the forecast errors to be interpreted as a (not-so-crazy) percentage, the normalization is taken out of the summation, i.e., normalization is performed after aggregation.

Normalized accuracy measures are used in wind forecasting, a more developed sub-domain of energy forecasting. In an effort to standardize metric usage in wind forecasting, Madsen et al. (2005) noted that the purpose of using normalized accuracy measures is to produce results independent of wind farm sizes. In addition, the authors recommended normalization by the installed capacity or mean observation. At that time (2005), published studies on solar forecasting were almost non-existent. When the field of solar forecasting started to bloom in the early 2010s, such normalization became default for solar forecasters, to whom wind forecasting was the most relevant literature to follow.

A compelling reason why normalized accuracy metrics are frequently used in solar (or wind) forecasting is that the typical end users (or "stakeholders") are typically electric engineers, business analysts or financial experts. These professionals are very familiar with percents, but not with a solar radiation unit such as W/m^2 or a power unit such as MW. From this standpoint, the use of normalization is essentially dictated by the necessity for the end users to understand and correctly use the forecast results. Nonetheless, grid operators almost never compare their forecasts across seasons, time scales, and let alone to forecasts of other grid operators. Therefore, in that context, the choice of normalized accuracy metrics is for convenience and internal communications, since a percentage metric is more accessible to a non-technical audience (decision-makers) than a MW metric. Since for grid operators the normalizing quantity, i.e. the denominator, is either constant (peak load) or similar (average load), the normalization does not affect the ranking of forecast accuracy. The grid operator preference for normalized accuracy metrics in their "small world" should not be construed as a motivation to adopt normalized error metrics in the academic community as we should strive for global inter-comparability of forecast quality.

This section was intended to motivate why normalized accuracy has become prevalent in the academic community. But as discussed in Section 4.2, we believe that normalized accuracy measures are inferior to the skill score as they do not appropriate consider variability.

2.2.3. *Problems with the skill score*

Since normalized accuracy measures cannot be used to compare forecasts made at different locations and timescales, an alternative has to be sought. In modern solar forecasting, the first attempt to address the problem of comparability was made by Carlos Coimbra and his former student Ricardo Marquez in several conference papers around 2011 (e.g.,

Marquez and Coimbra, 2011), a journal version (Marquez and Coimbra, 2013), and a book chapter published in 2013 (Coimbra et al., 2013). In those documents, a well-known concept in meteorological forecasting called *skill score* was introduced to the young field of solar forecasting. In the field of meteorology, the skill score, s, can be defined based on some measure of accuracy A, namely,

$$s = \frac{A_f - A_r}{A_p - A_r}, \tag{3}$$

where A_f, A_p, and A_r are the accuracy of the forecasts of interest, accuracy of the perfect forecasts, and accuracy of the reference forecasts, respectively (Murphy, 1988). For instance, s based on RMSE is

$$s = 1 - \frac{\mathrm{RMSE}(f, x)}{\mathrm{RMSE}(r, x)}, \tag{4}$$

where f, r, and x are forecasts of interest, reference forecasts, and observations, respectively.[10] For N samples,

$$\mathrm{RMSE}(f, x) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(f_t - x_t)^2}, \tag{5}$$

$$\mathrm{RMSE}(r, x) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(r_t - x_t)^2}. \tag{6}$$

Skill score s is often written in percentage, representing the percentage improvements in accuracy of the forecasts over the reference forecasts. If $s > 0$, the forecasts of interest have a smaller RMSE than that of the reference forecasts, otherwise, $s \leq 0$ indicates that the model of interest fails to outperform the reference forecasts. There are, however, two problems with using s to compare forecasts: (1) the choice of accuracy measure can be arbitrary, and (2) the choice of the reference forecasting model can be arbitrary.

The first problem can be understood with a simple example. The computation of s requires a measure of forecast accuracy, A, which is based on a scoring function. Depending on the choice of A, s can be dramatically different. For instance, suppose $\mathrm{RMSE}(r, x) = 200\ \mathrm{W/m^2}$ and $\mathrm{RMSE}(f, x) = 100\ \mathrm{W/m^2}$, then $s_{\mathrm{RMSE}} = 0.5$. However, when the mean square error (MSE) is used, $s_{\mathrm{MSE}} = 1 - (1 - s_{\mathrm{RMSE}})^2$ is boosted to 0.75. Whereas the conversion between s_{RMSE} and s_{MSE} is straightforward, s calculated based on other metrics, such as MAE, would be different, and cannot be inferred from s_{RMSE} or s_{MSE}. Hence, there is no obvious solution to this but to consider a consensus-based approach—at the moment, RMSE is the most common form of A in the literature (Blaga et al., 2019; Yang et al., 2018), and thus should be used in skill score computation (In a later section, this choice will be discussed further). Hereafter, the symbol s only denotes s_{RMSE}, unless otherwise stated.

One remedy to the second problem is to use a universally-accepted naïve reference model, so that s can be used—with appropriate caveats—to compare the accuracy of forecasts made across different locations or time periods. Skill score is built upon the notion that the "no skill" reference forecasts should sufficiently represent the difficulty of the forecasting scenario. In business forecasting, random walk is often used as the reference, and the relative performance of the model of interest is gauged using the Theil's U statistic, a concept similar to skill score (Makridakis et al., 2008). In meteorology, the so-called "climatology" is often used as the naïve reference (Jolliffe and Stephenson, 2012).[11]

[10] It should be noted that Eq (4) assumes the RMSE of the perfect forecast accuracy, A_p, is 0. However, in almost all statistical forecasting frameworks, the models would assume some unpredictable white noise, i.e., non-zero RMSE even if a model perfectly describes the data-generating process. Hence, the assumption here is that the $A_p \ll A_r$, so that it can be neglected.

[11] There are different types of climatology. Murphy (1988) considered single-valued internal climatology, multiple-valued internal climatology, single-valued external climatology, and multiple-valued external climatology, each gives a different skill score expression. The definitions of "internal" and "external" are based on whether the reference forecasts are derived from experimental periods or historical periods. Given one year of irradiance observations, the daytime sample mean, $\mathbb{E}(x)$, could be considered as the single-valued internal climatology forecasts. Under this definition, the MSE of climatology is simply the sample variance, $\mathbb{V}(x) = \mathbb{E}([x - \mathbb{E}(x)]^2)$. It might be however argued that the single-valued climatology would be inappropriate when the forecasts of interest are to be evaluated over a time period longer than a month or a season (Murphy, 1988).

In deterministic solar forecasting, the most popular naïve reference model is the *clear-sky adjusted persistence*, or simply, clear-sky persistence. Clear-sky persistence is conceptually no different from the seasonal naïve method described in Makridakis et al. (2008). More precisely, the clear-sky irradiance can effectively describe both seasonal cycles (a yearly cycle and a daily cycle) in an irradiance time series. Details of the clear-sky persistence baseline, its alternatives, as well as other important implementation aspects are addressed at length in Section 4 and Appendix A. For now, we note that the 1-step-ahead clear-sky persistence forecast for time t is $r_t = x_{t-1} \cdot c_t/c_{t-1}$, where c is the clear-sky expectation.[12]

2.3. The Coimbra skill score

Skill score is not limited to verification of deterministic forecasts of continuous random variable. It is also used in verification of deterministic forecasts of binary events (e.g, Gilbert skill score or Doolittle skill score), multi-category events (e.g., Gandin and Murphy score), and probabilistic forecast verification (e.g., Brier skill score or CRPS skill score). While the reader is referred to Jolliffe and Stephenson (2012) for more details on skill score, the version proposed by Marquez and Coimbra (2011) needs to be discussed. In deterministic solar forecasting, their version is the only notable alternative to the skill score defined in Eq. (3), we denote it as s_{Coimbra}.

Marquez and Coimbra (2011) proposed their skill score based on the concept of "uncertainty" (U) and "variability" (V):

$$s_{\text{Coimbra}} = 1 - \frac{U}{V}, \tag{7}$$

where

$$U = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{f_t - x_t}{c_t}\right)^2}, \tag{8}$$

$$V = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{x_t}{c_t} - \frac{x_{t-1}}{c_{t-1}}\right)^2}, \tag{9}$$

and f, x, and c are forecast, observation, and clear-sky expectation, respectively. The authors claimed that the ratio between U and V can be approximated by the ratio of the RMSE of the model of interest and the RMSE of clear-sky persistence, i.e., $s_{\text{Coimbra}} \approx 1 - \text{RMSE}(f, x)/\text{RMSE}(r, x) = s$. However, no detailed theoretical support was given in the different versions of the proposal (Marquez and Coimbra, 2011, 2013; Coimbra et al., 2013). Instead, the approximation was demonstrated empirically, through the results of several time series models.

By comparing s_{Coimbra} to s defined in Eq. (4), it is clear that the two skill scores generally should not be used interchangeably. In fact, s is the RMSE skill score of irradiance forecasts, whereas s_{Coimbra} is the RMSE skill score of clear-sky index forecasts. Stated differently, the two scores verify different forecast quantities—the former verifies irradiance forecasts, and the latter verifies clear-sky index forecasts. Appendix B shows that in order for the two scores to be approximately equal, both the clear-sky persistence and the model of interest need to have homoscedastic error, that is, the variances of the forecast errors are independent of the magnitude of clear-sky irradiance. This is, however, unlikely owing to the bell-shaped transient of irradiance. Hence, they should not be treated as alternatives.

2.4. Section summary

This section has discussed and put forward quite a few new concepts to deterministic solar forecast verification. In general, there are two approaches for forecast verification, namely, the measure-oriented approach and the distribution-oriented approach. The goodness of forecasts contains three elements: (1) consistency, (2) quality, and (3) value. Particularly interesting is that the distribution-oriented approach could help forecasters to assess the quality of their

[12] The word "expectation" refers to the fact that true clear-sky irradiance is always unknown, and only the estimates from a so-called "clear-sky model" are available. One should not confuse the current usage of the word with the usage in "statistical expectation."

forecasts in a systematic way, by relating various aspects of forecast quality to different accuracy measures, so that they can be interpreted. This will be discussed further in Section 3.

Whereas the distribution-oriented approach is primarily recommended for forecast diagnosis, i.e., to be used within a forecasting case study, the skill score is recommended for cross-work forecast comparison. It is important to note that the measure-oriented and distribution-oriented approaches are complementary, not substitutive. Skill score is computed based on the accuracy measure of a reference model that can sufficiently describe the difficulty (variability and uncertainty) of a forecast situation. It gauges the overall skillfulness of a forecaster. The RMSE skill score computed based on the clear-sky persistence model and its implementation should be mandated and standardized, see Section 4.

3. Distribution-oriented approach for forecast verification

The distribution-oriented forecast verification framework is quite general. Before one starts to wonder what joint distribution[13] has to do with forecast verification, as a forecaster, most likely he or she has already used this framework. It is common to use a forecast–observation scatter plot to check forecast quality. One may draw some conclusions based on whether the point cloud is centered on the identity line, or how dispersed the scatter points are. In other cases, a forecaster may wish to check how the scatter points are distributed along the x-axis, or whether the spread of forecasts vary for different observation ranges. In fact, most forecast accuracy quantification—visually or through accuracy measures—are just summaries of the joint distributions, or equivalently, the marginal and conditional distributions. The relationship between joint, marginal, and conditional distributions of two random variables can be expressed using Bayes' theorem. When these variables are the forecast and the observations, the same relationship is referred to as Murphy–Winkler factorization in meteorology.

Murphy–Winkler factorizations are:

$$p(f, x) = p(x|f)p(f), \tag{10}$$

$$p(f, x) = p(f|x)p(x), \tag{11}$$

where p denotes distribution, f and x represent forecasts and observations, respectively. Eq. (10) is called the *calibration–refinement factorization*, whereas Eq. (11) is called the *likelihood–base rate factorization*. The naming convention is quite intuitive. For example, the $p(x|f)$ term in Eq. (10) describes the spread of the observations, given a particular forecast. For a good correspondence, the forecast is said to be *calibrated* or *reliable*. Mathematically, the forecasts are perfectly calibrated if $\mathbb{E}(x|f) = \int xp(x|f)df = f$. The reader is referred to Table 2 for an interpretation of other conditional and marginal distributions, and Murphy (1997) for a list of aspects of forecast quality.

Verifying the above conditional and marginal distributions is equivalent to verifying the joint distribution. For instance, given two sets of forecasts, f_1 and f_2, by comparing $p(x|f_1)$ and $p(x|f_2)$, one can conclude whether one set of forecasts is more reliable than the other, see Moskaitis (2008); Murphy et al. (1989) for case studies. Whereas linking the forecast distributions to aspects of forecast quality provides forecasters with insights regarding their forecasts, it would be easier to interpret if the different aspects of forecast quality can be quantified using measures. For instance, consider the well-known bias–variance decomposition of MSE:

$$\begin{aligned}
\mathrm{MSE} &= \iint (f - x)^2 p(f, x) df dx \\
&= \mathbb{E}[(f - x)^2] \\
&= \mathbb{V}(f - x) + [\mathbb{E}(f) - \mathbb{E}(x)]^2 \\
&= \overbrace{\mathbb{V}(f) + \mathbb{V}(x)}^{\text{marginal dist.}} - \overbrace{2\mathsf{cov}(f, x)}^{\text{association}} + \overbrace{[\mathbb{E}(f) - \mathbb{E}(x)]^2}^{\text{unconditional bias}},
\end{aligned} \tag{12}$$

[13] Formally, we call a function $p(x, y)$ the joint distribution of random variables X and Y if $p(x, y) \geq 0$, $\forall(x, y)$; $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y)dxdy = 1$; and for any set $\mathcal{A} \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}[(X, Y) \in \mathcal{A}] = \iint_{\mathcal{A}} p(x, y)dxdy$, $\mathbb{P}[(X, Y) \in \mathcal{A}]$ denotes the probability of (X, Y) in set $\mathcal{A}$ (Wasserman, 2013).

Table 2: Definition, interpretation, and quantification of Murphy–Winkler factorizations (Jolliffe and Stephenson, 2012; Murphy and Winkler, 1987).

distribution	definition	interpretation	(some specific methods of) quantification
$p(x\|f)$	(related to) *calibration*	A set of deterministic forecasts is perfectly calibrated if $\mathbb{E}(x\|f) = \int xp(x\|f)df = f$.	(1) *calibration*, *reliability*, or *type 1 conditional bias*: $\mathbb{E}_f[f - \mathbb{E}(x\|f)]^2$; (2) *resolution*: $\mathbb{E}_f[\mathbb{E}(x\|f) - \mathbb{E}(x)]^2$.
$p(f)$	(related to) *refinement*, or *sharpness*	Refinement or sharpness is an aspect that usually applies only to probabilistic forecasts (Murphy, 1997). In deterministic forecast verification, if a forecaster produces the same forecast all the time, it is said to be completely unrefined. However, the complete refinement is difficult to define for deterministic forecasting (Murphy et al., 1989), but $p(f)$ has to be equal to $p(x)$ for perfect forecasts.	(1) *Kolmogorov–Smirnov test statistic*: $\max\|F(f) - F(x)\|$; (2) *earth mover's distance* or *first Wasserstein distance*: the area between the two ECDFs. (The formal definition is technical and thus omitted.)
$p(f\|x)$	*likelihood*	If $p(f\|x)$ is zero for all values x but one, the forecast is perfectly discriminatory. If $p(f\|x)$ is the same for all values of x, the forecast is not at all discriminatory.	(1) *discrimination 1*, or *type 2 conditional bias*: $\mathbb{E}_x[x - \mathbb{E}(f\|x)]^2$; (2) *discrimination 2*, or simply *discrimination*: $\mathbb{E}_x[\mathbb{E}(f\|x) - \mathbb{E}(f)]^2$.
$p(x)$	*uncertainty*, or *base rate*	If $p(x)$ is a fairly peaked distribution, the scenario has relatively small uncertainty (and thus easier to forecast) as compared to a scenario where $p(x)$ is fairly uniform.	(1) *variance*: $\mathbb{V}(x)$; (2) *kurtosis*: $\frac{\mathbb{E}[(x - \mathbb{E}(x))^4]}{\left(\mathbb{E}[(x - \mathbb{E}(x))^2]\right)^2}$.

where the overhead braces show the representation of each term. In this decomposition, $\mathbb{V}(f)$ and $\mathbb{V}(x)$ are variances of forecasts and observations, respectively. Their values can be used as a proxy for measuring the similarity between $p(f)$ and $p(x)$. If the forecasts were perfect, the two marginal distributions would be exactly the same, so would the variances.[14] Similarly, the $\mathtt{cov}(f, x)$ term can be written as correlation, namely, $\sqrt{\mathbb{V}(f)\mathbb{V}(x)} \cdot \mathtt{cor}(f, x)$, which denotes the *association* between forecasts and observations. Lastly, the $[\mathbb{E}(f) - \mathbb{E}(x)]^2$ term represents the squared unconditional bias, i.e., MBE^2. This example illustrates the complementarity between the measure-oriented approach (e.g., verification using MBE, correlation, or variance of the forecasts) and the distribution-oriented approach (analyzing the joint, conditional, and marginal distributions of forecasts and observations), see Appendix C for further information.

Besides the bias–variance decomposition, MSE can also be decomposed following the calibration–refinement and likelihood–base rate factorizations:

$$\mathrm{MSE} = \mathbb{V}(x) + \overbrace{\mathbb{E}_f[f - \mathbb{E}(x|f)]^2}^{\text{type 1 conditional bias}} - \overbrace{\mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2}^{\text{resolution}}, \tag{13}$$

$$\mathrm{MSE} = \mathbb{V}(f) + \overbrace{\mathbb{E}_x[x - \mathbb{E}(f|x)]^2}^{\text{type 2 conditional bias}} - \overbrace{\mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2}^{\text{discrimination}}. \tag{14}$$

The derivation of these decompositions are shown in Moskaitis (2008). As indicated in the equation, different terms in the decomposed forms explain different aspects of forecast quality.

Type 1 conditional bias, $\mathbb{E}_f[f - \mathbb{E}(x|f)]^2$, indicates the degree of correspondence between the mean observation given a particular forecast, i.e., calibration or reliability. Recall perfect calibration is when $\mathbb{E}(x|f) = f$, so the smaller this term the better. *Resolution* accounts for the difference between conditional and unconditional mean observation, which is reflected by $\mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2$. If $\mathbb{E}(x|f) = \mathbb{E}(x)$, it means the data samples have no resolution. It is desired to have the generated forecasts to be followed by different observations (so that the forecasts are meaningful), this term should be maximized, which is also reflected by the negative sign in front of the term. *Type 2 conditional bias*, $\mathbb{E}_x[x - \mathbb{E}(f|x)]^2$, indicates the degree of correspondence between the mean forecast given a particular observation and the observation. Naturally, this term should be as small as possible. Lastly, *discrimination* denotes the difference between conditional and unconditional mean forecast, i.e., $\mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2$, which indicates how forecasts are differentiated for different observation values. This terms needs to be maximized.

The numerical evaluation of these decomposed factors can be difficult. When Murphy and Winkler (1987) proposed these decompositions, a binary x was used, which greatly simplifies the computation. In Moskaitis (2008), the evaluation was performed by discretizing the continuous random variable—tropical cyclone intensity—into bins.

[14] However, having identical variances does not imply identical distributions; and having identical distributions does not imply the forecasts are perfect.

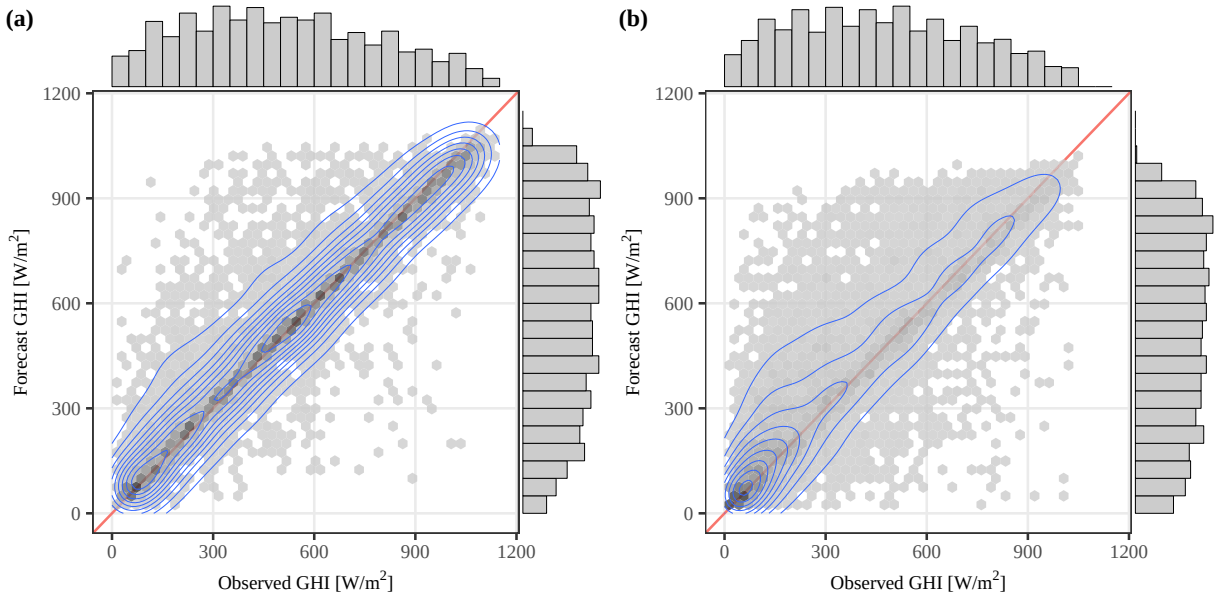

Figure 2: Joint and marginal distributions of 24-h-ahead hourly NAM forecasts and SURFRAD observations at (a) Desert Rock, Nevada (36.624°N, 116.019°W), and (b) Penn. State Univ., Pennsylvania (40.720°N, 77.931°W), from 2015 to 2016. The contour lines show the 2d kernel densities.

Recently, Yang and Perez (2019) used kernel conditional density estimation (KCDE) to estimate the conditional expectations, namely, $\mathbb{E}(x|f)$ and $\mathbb{E}(f|x)$, which removes the dependency on binning. The code for the KCDE-based approach is available in the supplementary material of that paper.

In contrast to numerical evaluation of Eqs. (13) and (14), visual inspection is more straightforward and enables a forecaster to appreciate the properties of the forecasts in great detail. In general, visualizing the error distribution is a powerful way of communicating the performance of a model. In line with the Murphy–Winkler factorizations, an x–f scatter plot displays the joint distribution between observations and forecasts, allows for visualizing the marginal distributions as well as specific conditional distributions.

To exemplify the forecast verification procedure discussed in this section, a case study is presented. Fig. 2 shows the joint and marginal distributions of 24-h-ahead hourly forecasts of global horizontal irradiance (GHI) produced by North American Mesoscale (NAM) forecast system against the observations collected by the Surface Radiation Budget Network (SURFRAD), at two locations with distinct climate over a period of two years. Whereas the joint distribution at the Desert Rock (DRA) station has approximately equal probabilities on both sides of the diagonal, the forecasts at the Penn. State Univ. (PSU) station over-predict GHI, i.e., higher probability is concentrated above the diagonal, where $f > x$. A closer examination of the 2d kernel density contours reveals that the NAM forecasts at DRA drift slightly below the identity line for high-irradiance conditions. For mid- and low-irradiance conditions at DRA, the forecasts are slightly above the identity line. This observation warrants an irradiance-condition-based post-processing treatment. Similar observations could be made for forecasts at PSU.

The histograms shown in Fig. 2 denote marginal distributions, $p(f)$ (on the right) and $p(x)$ (on the top). Since the shape of the histograms depends largely on bin width, different choices may affect the forecaster's judgement differently. In this regard, overlaying the empirical cumulative distribution functions (ECDFs) of f and x could be useful at times. Fig. 3 demonstrates such plots using the same data. Visually, the ECDFs of forecasts and observations at DRA align better than those at PSU. At PSU, f is stochastically greater than x (the ECDF of f lies below and hence to the right of that for x). Formally, the Kolmogorov–Smirnov (KS) test computes the statistic $D_n = \max|F_n(f) - F_n(x)|$, i.e., the maximum absolute distance between the ECDFs of forecasts and observations. In the present case, KS tests conducted at the two stations both reject the null hypothesis—two distributions are equal—at a significance level of 0.05.

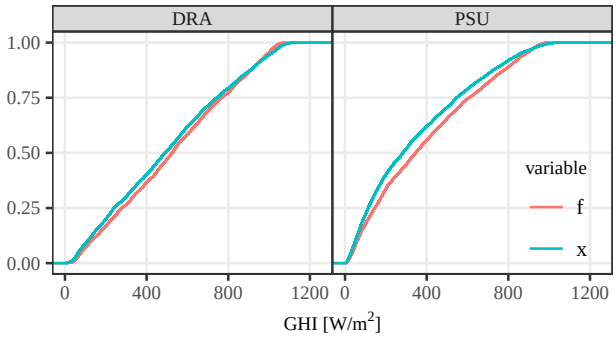

Figure 3: Marginal distributions of forecasts and observations described in Fig. 2.

Table 3: Bias–variance decomposition (see Eq. 12) and Murphy–Winkler factorization (see, Eqs. 13 and 14) of 1–24-h-ahead NAM (f) against SURFRAD GHI (x), at DRA and PSU stations over 2015–2016. For interpretability, all metrics are written in squared form, so that the bases have the unit of W/m^2, except for correlation ρ, which is dimensionless.

	MSE	$\mathbb{V}(x)$	$\mathbb{V}(f)$	$\rho(f,x)$	$[\mathbb{E}(f)-\mathbb{E}(x)]^2$	$\mathbb{E}_f[f-\hat{\mathbb{E}}(x\|f)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x\|f)-\mathbb{E}(x)]^2$	$\mathbb{E}_x[x_g-\hat{\mathbb{E}}(f\|x)]^2$	$\mathbb{E}_x[\hat{\mathbb{E}}(f\|x)-\mathbb{E}(f)]^2$
DRA	108.22^2	297.97^2	288.75^2	0.94	22.65^2	28.00^2	279.00^2	38.82^2	270.46^2
PSU	154.99^2	269.89^2	275.40^2	0.85	41.67^2	64.36^2	230.01^2	61.17^2	235.62^2

As compared to joint and marginal distributions, the visualization of conditional distribution is more challenging. Whereas some authors plot the individual quantiles or use box plots to represent the distributions, ridgeline plots are employed here, see Fig. 4. In this plot, $p(x|f)$ and $p(f|x)$ at both stations are represented using overlapping lines, which create the impression of a mountain range. Fig. 4 (a) and (c) reveal $p(x|f)$ is mostly centered on the forecast value, i.e., $\mathbb{E}(x|f)$ is close to f, indicating small type 1 conditional bias in NAM forecasts. On the other hand, type 2 conditional bias is found to be significant for high values of x, see $p(f|x)$ for $x = 1050$ W/m^2 in Fig. 4 (b) and (d).

The distribution-oriented verification technique is often complemented with *summary measures* of the different aspects of forecast quality. Table 3 shows the quantification of these aspects using the bias–variance decomposition and Murphy–Winkler factorization, as stated in Eqs. (12), (13), and (14). Note that the decomposed terms listed in the table do not add up exactly to MSE, due to the uncertainty introduced during KCDE. Such discrepancy is however small, and thus does not affect our analysis. In terms of correlation, a higher $\rho = 0.94$ is observed at DRA as compared to 0.85 at PSU, indicating a better association between forecasts and observations at DRA. The square of unconditional bias, $[\mathbb{E}(f)-\mathbb{E}(x)]^2$, is also significantly smaller at DRA, agreeing with the earlier observation made using the joint distribution plots. Since smaller type 1 conditional bias, $\mathbb{E}_f[f-\hat{\mathbb{E}}(x|f)]^2$, means higher calibration—the forecasts at DRA are more reliable than those at PSU. Similarly, smaller type 2 conditional bias, higher resolution, and higher discrimination observed at DRA all lead to the conclusion that the NAM forecasts at DRA have better forecast quality than those at PSU.

Based on the case study above, it is evident that the distribution-oriented forecast verification is useful in assisting forecasters to make informed decision based on forecast quality. In other cases, the same methodology can be applied to compare forecasts made using different methods, which provides more information than using MSE values alone. This verification procedure leads to more meaningful conclusions than statements such as "the MSE at location A is smaller than that at location B, and thus the forecasts at location A are better". That said, should one wish to examine the relative accuracy gain from the reference method, the quantification of aspects of forecast quality can be carried out with the reference forecasts, as exemplified in Table 4. Comparing the two tables, one thing certain is that in the case of NAM, the NWP-based model dominates persistence in all aspects except for the unconditional bias. Nonetheless, such unconditional bias could be easily trimmed with regression-based post-processing, and thus does not affect one's confidence in opting for the NAM model.

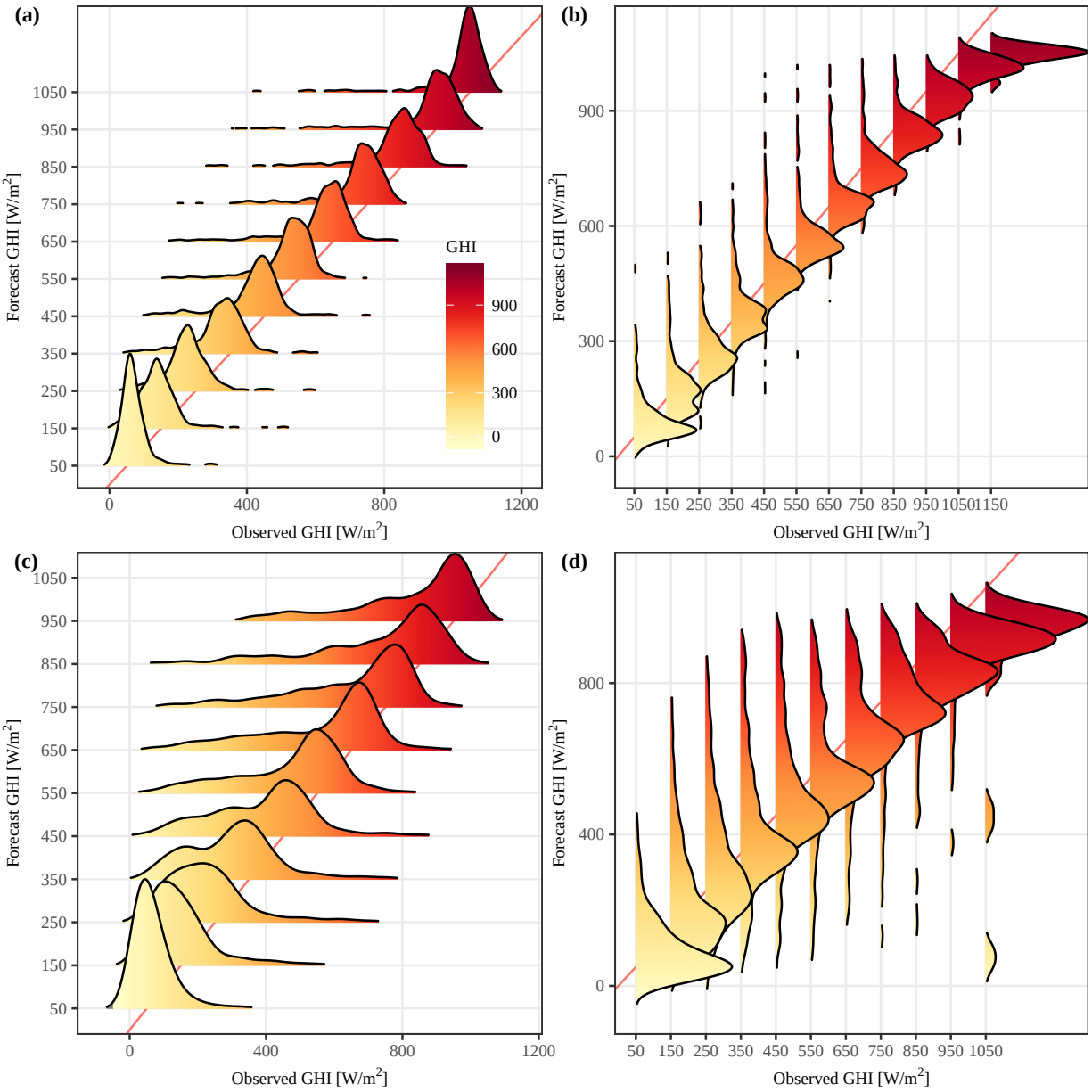

Figure 4: Conditional distributions of 24-h-ahead hourly NAM forecasts and SURFRAD observations. $p(x|f)$ are shown in (a) and (c) for Desert Rock, Nevada (36.624°N, 116.019°W) and Penn. State Univ., Pennsylvania (40.720°N, 77.931°W), respectively. $p(f|x)$ are shown in (b) and (d) for the two stations, respectively.

4. Recommendations and practical concerns

The traditional measure-oriented approach is complemented by the distribution-oriented approach, which can reflect different aspects of forecast quality and help forecasters to diagnose their forecasts. However, at the end of the day, a "one-number summary" of forecasts is still highly desirable, especially when scientists bring their forecasts to non-technical personnel, e.g., the sales team, politicians, or the general public. Therefore, in this section, some recommendations and practical concerns regarding the use of skill score are discussed.

Table 4: Same as Table 3, but tabulating the forecast quality of 1–24-h-ahead persistence forecasts.

	MSE	$\mathbb{V}(x)$	$\mathbb{V}(f)$	$\rho(f,x)$	$[\mathbb{E}(f)-\mathbb{E}(x)]^2$	$\mathbb{E}_f[f-\hat{\mathbb{E}}(x\|f)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x\|f)-\mathbb{E}(x)]^2$	$\mathbb{E}_x[x_g-\hat{\mathbb{E}}(f\|x)]^2$	$\mathbb{E}_x[\hat{\mathbb{E}}(f\|x)-\mathbb{E}(f)]^2$
DRA	151.83^2	297.97^2	297.39^2	0.87	0.52^2	42.43^2	259.75^2	42.89^2	259.10^2
PSU	220.97^2	269.89^2	269.75^2	0.66	1.14^2	94.62^2	181.54^2	95.14^2	181.50^2

4.1. MBE, MAE, or RMSE?

Skill score belongs to the class of relative measures (Hyndman and Koehler, 2006).[15] In that, a scale-dependent measure is needed for its computation. Since MBE, MAE, and RMSE are the most popular metrics at the moment (Yang et al., 2018), one of them will be recommended for skill score computation.

MBE is defined as $\mathbb{E}(f-x)$, or $\mathbb{E}(f)-\mathbb{E}(x)$. This is different from how bias is described in some statistics literature, namely, $\mathbb{E}(x-f)$, which originates from how a predictive model is constructed (see Makridakis et al., 2008).[16] Defining the MBE to be "forecast minus observation" is more natural for solar forecasting, since an over-prediction (forecasts are, on average, higher than observations) corresponds to a positive MBE, and an under-prediction corresponds to a negative MBE. MBE describes unconditional bias, and most statistical forecasting methods have MBE $\to 0$. State-of-the-art operational solar forecasts would have some form of bias correction implemented, e.g., model output statistics (MOS). Therefore, having small MBE is more of a baseline requirement, rather than a credit-worthy feature among state-of-the-art forecasts. Furthermore, the MBE for the reference forecasts, clear-sky persistence, has an expectation of zero, and thus makes the skill score undefined. To that end, MBE is unsuitable for skill score computation.

The main difference between MAE, defined as $\mathbb{E}(|f-x|)$, and RMSE, defined as $\sqrt{\mathbb{E}[(f-x)^2]}$, is that the latter penalizes large errors while the former gives the same weight to all errors. Since large errors are particularly concerning for grid integration of solar power (e.g., a loss of load becomes more likely), RMSE is more suitable when a set of forecasts contain several large errors, which is usually the case for solar forecasts. On this point, the percentage improvement in RMSE, in the form of s, might attract more interests than the MAE skill score.

The second reason for using the RMSE skill score is related to the distribution-oriented forecast verification. The Murphy–Winkler MSE factorization has been recommended for forecast diagnosis in the previous section. As a result, RMSE values become readily available after the various aspects of forecast quality are quantified. It must be highlighted that in the field of meteorology, and many other fields such as statistics, researchers generally do not distinguish the use of words "RMSE" and "MSE" in their writing (Jolliffe and Stephenson, 2012), since RMSE and MSE differ only by a square root. Nonetheless, they should not be mixed up during skill score computation, recall the example given in Section 2.2.3.

Lastly, the popularity of RMSE is higher than that of MAE, not only in solar forecasting, but in other forecasting domains as well. Gneiting (2011) found that the usage of RMSE in four related domains, namely, forecasting, statistics, econometrics, and meteorology, dominates as compared to MAE. Whereas the precise reasons are unknown, it is hypothesized that consistency might be one of the main reasons, since there are more models minimizing MSE than minimizing MAE. In additional, squares are more amenable than absolute values in many mathematical operation (Chai and Draxler, 2014). Hence, for this and the above reasons, RMSE skill score is recommended in deterministic solar forecasting.

4.2. Normalized versus raw measures

Although skill scores computed using the normalized and raw measures are identical (as long as the normalization value is computed based on observations), it is of interest to discuss the various issues related to normalization. There are four main ways to normalize the RMSE in irradiance forecasting, namely, by mean, by maximum,[17] by 1000 W/m^2,[18] or by the square root of second moment, i.e., $\sqrt{\mathbb{E}(x^2)}$.[19] On the other hand, normalization by the installed capacity is the most common approach in PV power forecasting.

[15] One should distinguish relative measure from measure of relative error. The former performs division after the primary measure is computed, i.e., $\mathbb{E}[S(f,x)]/\mathbb{E}[S(r,x)]$, whereas the latter performs averaging on relative errors, i.e., $\mathbb{E}[S(f,x)/S(r,x)]$, where S is a scoring function.

[16] For a regression model, $y = g(x)+e$, where y is the response, x is the regressor, and bias e is in fact $\mathbb{E}[y-g(x)]$, or observation minus prediction.

[17] Using the maximum is equivalent to using the range, i.e., maximum minus minimum.

[18] This is the standard test condition for global irradiance. Such a normalization practice in irradiance forecasting corresponds to normalizing the error with rated power in PV power forecasting

[19] This normalization resembles the Theil's U statistic used in finance (Brown and Rozeff, 1978).

Some of the above choices of normalization have been discussed by Hoff et al. (2013). Nonetheless, it is rather meaningless to pursue which normalization strategy is the best. As discussed in Section 2, these normalized errors have little if not no relevance in comparing forecasts made at different locations and timescales—the normalization values represent neither variability nor uncertainty. For instance, it is common to see MAPE of (hourly) transmission-level load forecasts reaching 2% or less. Does this mean that the load forecasters are a lot better in doing forecasting than solar forecasters, or is it simply that the transmission-level load profile is very predictable? When a forecaster who lives in Sahara Desert (for whatever reasons) reports a 5% mean-normalized nRMSE for his PV plant, other forecasters who live in Southeast Asia should really think twice before making that percentage a benchmark for their forecasts.

Aside from the intrinsic differences in climates and forecast timescales, other implementation issues (see below) can also distort the interpretation of these normalized error metrics. It seems that the only advantage of using normalized error metrics is the convenience of quoting it as a percentage, e.g., "the forecast error is 10%." Although many authors of this article agree to abandon the use of normalized measures, they are too deeply rooted in solar forecasting. In this regard, we call for the moderate use of normalized error metrics, and tabulating the values used for normalization is necessary.

4.3. Some seemingly trivial implementation issues

Forecast verification must be performed based on experimental data, i.e., out-of-sample forecasts and corresponding observations. There are several seemingly trivial implementation issues such as data trimming, normalization, counting nighttime hours, and data aggregation, that can strongly affect the verification results.

Data trimming refers to quality control (QC) applied to experimental data prior to forecasting and verification. Owing to factors such as measurement uncertainty or irradiance modeling error, experimental data often contain spurious values. There is no universally-accepted QC procedure (Gueymard and Ruiz-Arias, 2016), but recommended QC for surface radiation measurements (Long and Shi, 2008), PV power output (Killinger et al., 2017), and satellite-based products (Urraca et al., 2017) is available. Partly attributed to the advent of statistical and machine-learning software, details of implementation become more opaque to forecasters. Therefore, forecasters should check the output of each step during forecasting, responsibly, to prevent spurious data from entering the final verification stage. To ensure forecast verification is performed with reasonably trimmed data, visual inspection as outlined in the previous section is necessary. It is however noted that data trimming based on forecast error is *not* recommended. One should not remove a forecast–observation pair just because it produces a large error, instead, the cause behind it should be investigated.

There is also no well-accepted answers to questions "which normalization value should be used" and "whether the nighttime hours should be included during validation." For instance, normalizing the RMSE using the maximum observation would give a smaller value than normalizing with mean observation. To address such ambiguity, and allow the transformation from a normalized to an scale-dependent error metric, it is necessary to report the normalizing value alongside the accuracy measures. Similarly, inclusion of nighttime hours would make the RMSE smaller, since the forecasts—0 W/m^2—are perfect during those hours. Hence, nighttime data should always be excluded from verification. This could be ensured by using a zenith angle filter of $< 85°$. Additionally, the filter eliminates some modeling and measurement artifacts under low-sun conditions.

In state-of-the-art solar forecasting, it is common to have more than one data source involved (e.g., ground measurements, satellite-derived irradiance, reanalysis data, NWP output, or PV output). Even in the simple case of verifying NWP forecasts, one needs to compare the NWP forecasts to ground-based measurements. The issue of data aggregation naturally comes into play, since the ground-based measurements are usually at a higher temporal resolution (e.g., 1 min) than the NWP output (e.g., hourly).

There are three schemes of averaging to aggregate a high-temporal-resolution time series to a low-temporal-resolution one, namely, floor, ceiling, and round. A floor aggregation means that the data within a time interval are aggregated to the earliest time stamp in that interval, e.g., 1-min data points between 1:00 to 2:00 are aggregated and stamped as 1:00. Similarly, the ceiling-aggregation scheme stamps the aggregated data with the last time stamp of an interval, and the round-aggregation scheme collapses the data to the center time stamp of an interval.

In this regard, inappropriate data aggregation creates temporal misalignment between different datasets, hence amplifies forecast errors unnecessarily (see Fig. 1 in Yang, 2018). To select the correct data-aggregation scheme, it is

necessary to understand how each dataset is produced. In other words, one must always read the data documentation. For instance, satellite-derived irradiance and some NWP outputs have a "snapshot" nature, and the round-aggregation schemes is appropriate. In the case of reanalysis, the data often represent the condition over the past hour, and the ceiling-aggregation schemes is appropriate. Finally, it is important to re-emphasize that here we are concerned with deterministic forecasts only. Other strategies and metrics are needed when dealing with probabilistic forecasts and prediction intervals (see, e.g., Chu et al., 2015).

4.4. Implementing the RMSE skill score

The skill score, s, is perhaps the one number that interests solar forecasters the most, the version stated in Eq. (4) is strongly recommended, i.e., $s = 1 - \text{RMSE}(f, x)/\text{RMSE}(r, x)$, where r is the clear-sky persistence forecasts. In a day-ahead scenario, one often assumes the diurnal transient of clear-sky irradiance does not change over the forecast horizon. In other words, clear-sky persistence forecasts for day d are the observations from the most recent day that has complete records. More formally,

$$r_{t+h} = x_{t+h-\lceil h/m \rceil \cdot m - \lceil l/m \rceil \cdot m}, \tag{15}$$

where t is the start of the operating day, h is forecast horizon, l is forecast submission lead time, m is frequency of data in a day (number of observations per day), and $\lceil \cdot \rceil$ is the ceiling operator. Time parameters t, h, l and m have a unit equals to the temporal resolution of observations, e.g., 15 min, 30 min, or hourly. For example, the California Independent System Operator (CAISO) requires the 24-h-ahead hourly forecasts to be submitted 18.5 h prior to the operating day. In this case, $h \in \{1, 2, \cdots, 24\}$, $l = 18.5$, $m = 24$, and Eq. (15) reduces to $r_{t+h} = x_{t+h-48}$. If the observations are half-hourly, and a forecaster is interested in forecasts out to 36 h with no lead time, i.e., $h \in \{1, 2, \cdots, 72\}$, $l = 0$, $m = 48$, Eq. (15) becomes $r_{t+h} = x_{t+h-48}$ for $h \in \{1, 2, \cdots, 48\}$ and $r_{t+h} = x_{t+h-96}$ for $h \in \{49, 50, \cdots, 72\}$. In the literature, these operational forecasting time parameters are rarely considered (Yang et al., 2017). In this regard, the forecast verification results reported in academic literature do not represent the "true" accuracy of a model—without lead time, the errors are smaller than the actual errors one should anticipate.

Another way of implementing day-ahead clear-sky persistence was used by Perez et al. (2013). In their version, the forecast is given by:

$$r_{t+h} = c_{t+h} \cdot \frac{1}{m} \sum_{h'=1}^{m} \frac{x_{t+h'-\lceil h/m \rceil \cdot m - \lceil l/m \rceil \cdot m}}{c_{t+h'-\lceil h/m \rceil \cdot m - \lceil l/m \rceil \cdot m}}. \tag{16}$$

Stated in words, it means that the daily clear-sky index from the last available day is projected for all forecast horizons. Eq. (16) assumes the daily clear-sky index is averaged after the divisions. Alternatively, one can also interpret the daily clear-sky index as the ratio of averaged irradiance and clear-sky irradiance. That is,

$$r_{t+h} = c_{t+h} \cdot \frac{\sum_{h'=1}^{m} x_{t+h'-\lceil h/m \rceil \cdot m - \lceil l/m \rceil \cdot m}}{\sum_{h'=1}^{m} c_{t+h'-\lceil h/m \rceil \cdot m - \lceil l/m \rceil \cdot m}}. \tag{17}$$

It is generally unclear which version produces better results. Thus, a case study using hourly GHI data from 7 SURFRAD stations over 2015–2016 is conducted. The McClear model (Lefèvre et al., 2013) is used to compute the clear-sky GHI. The CAISO day-ahead forecast submission requirement is used, so Eqs. (16) and (17) reduce to $r_{t+h} = (1/24) \cdot c_{t+h} \cdot \sum_{h'=1}^{24} (x_{t+h'-48}/c_{t+h'-48})$ and $r_{t+h} = c_{t+h} \cdot \sum_{h'=1}^{24} x_{t+h'-48} / \sum_{h'=1}^{24} c_{t+h'-48}$, respectively. Table 5 depicts the results. It is found that Eq. (17) is the most accurate in terms of RMSE,[20] it is therefore recommended for future works on deterministic solar forecast verification.

To make reference forecasts for an intra-day or intra-hour scenario, the procedure is straightforward. Using the earlier notations, we have

$$r_{t+h} = c_{t+h} \cdot \frac{x_{t-l}}{c_{t-l}}, \tag{18}$$

[20] It is noted that only RMSE is of interest here, since the clear-sky persistence is used for RMSE skill score computation.

Table 5: RMSEs, in W/m^2, of day-ahead clear-sky persistence using hourly data from 7 SURFRAD stations over 2015–2016. Three versions as shown in Eqs. (15)–(17) are compared. The row below the header shows station-specific $\mathbb{E}(x)$. The McClear model (Lefèvre et al., 2013) is used to compute the clear-sky GHI.

	BON	DRA	FPK	GWN	PSU	SXF	TBL
	338.00	495.26	350.44	388.69	331.33	361.53	405.23
Eq. (15)	229.55	157.06	192.19	238.48	232.95	224.13	224.54
Eq. (16)	203.93	138.20	171.11	212.10	206.65	195.81	198.76
Eq. (17)	202.24	137.46	169.13	208.46	203.98	192.38	198.13

i.e., the reference forecast from the clear-sky persistence model for time $t+h$ is the clear-sky index observation at time $t-l$ adjusted to the clear-sky irradiance at $t+h$. For example, CAISO's real-time market requires hour-ahead forecasts to be submitted 75 min prior to the operating hour at 15-min resolution out to 5 h. In this case, assuming observations have a resolution of 15-min, then $h = \{1, 2, \cdots, 20\}$ and $l = 5$. In Eq. (18), a same clear-sky index observation is used for all h.

There are many clear-sky models available. To that end, several extensive reviews on clear-sky models were published recently (Antonanzas-Torres et al., 2019; Ruiz-Arias and Gueymard, 2018). One particular issue is that most of the high-performance clear-sky models require several inputs, such as total column ozone amount, precipitable water, or aerosol single-scattering albedo, which have to be sourced from remote-sensing or reanalysis databases; this limits the worldwide uptake of these models. On the other hand, simple models that require only a few input variables usually have limited performance. Therefore, the McClear model (Lefèvre et al., 2013) might be the best choice for solar forecasters. Being a physical model based on radiative transfer, the performance of McClear is among the best. Better still, McClear is available as a web service[21] for all locations in the world, from 2004-01-01 up to two days ago.[22] The R package "camsRad" is also freely available, and offers access to McClear through an API.

5. Conclusion

The increasing amount of solar forecasting research calls for harmonization of forecast verification measures and methods among researchers. This paper has discussed a wide spectrum of issues relevant to verification of deterministic solar forecasts. The final recommendation is listed as follows.

- The distribution-oriented approach to forecast verification can be used for forecast diagnosis. Since the joint distribution contains all time-independent information relevant to verification, it is more general than the traditional measure-oriented approach. It is recommended to use the distribution-oriented approach to visualize and quantify forecast quality.
- Bias–variance factorization and Murphy–Winkler factorization link various qualitative aspects of the skillfulness of the forecasts, such as uncertainty, reliability, resolution, association, or discrimination, to quantitative measures.
- Small MBE is a prerequisite of all solar forecasts and therefore not a critical metric to judge forecast quality.
- When the normalized errors are used, it is necessary to also tabulate the normalization values.
- Generally, forecasters are encouraged to use any meaningful measure to gauge their forecasts. However, if a chosen measure is inconsistent with the given forecast directive, it is inappropriate.
- Implementation issues are important for the final interpretation of forecast accuracies. Nighttime data should be excluded from forecast verification. Special care is needed during data trimming and aggregation.

[21] http://www.soda-pro.com/web-services/radiation/cams-mcclear

[22] This two-day lag makes it unsuitable for operational forecast verification. Nonetheless, during forecast verification, one is only interested in analyzing the long-term behavior of different models, this lag does not concern us.

- We strongly recommend using the RMSE skill score based on clear-sky persistence in all solar forecasting works. Skill score denotes the relative improvement of a model of interest from clear-sky persistence, and it can be used to compare forecasts produced in different works.

- As forecasting workflow is getting more and more complex, it is advised to perform sanity checks throughout the course of producing forecasts. To ensure the worldwide uptake of any proposed forecasting model, source code and data should be made available whenever possible. Without reproducibility, it would be cumbersome, if not impossible, to verify the reported forecast accuracies.

Appendix A. On choice of naïve reference

As discussed in Section 2, the clear-sky persistence model should be used as the naïve reference for deterministic solar forecasting. In the simplest case, a persistence model uses the most recent observation available as forecasts; this is the definition of persistence. Such forecast is sometimes referred to as "random walk" (i.e., $r_t = r_{t-1} + e_t$), or "no change" forecast. In the case of solar irradiance forecasting, raw persistence forecasts should be the most recent observed irradiance. Notwithstanding, given the bell-shaped diurnal transient of irradiance due to the apparent movement of the Sun, it is important to take this cycle into consideration. Makridakis et al. (2008) noted in their book that seasonally adjusted persistence can frequently do much better than the raw persistence. So the question is "how do we adjust for the seasonality?"

Besides using the clear-sky irradiance, one can use the extraterrestrial irradiance (the irradiance at the top of the atmosphere) for adjustment. The ratio between surface and extraterrestrial GHI is known as clearness index, k. In other words, the persistence is performed on clearness index, namely, $r_{t+h} = x_t \cdot k_{t+h}/k_t$. Both clear-sky persistence and clearness persistence adopt a multiplicative seasonality modeling approach. A particular problem with multiplicative seasonality is that during sunrise and sunset (small solar elevation angle), clear-sky index can become quite large, owing to the measurement uncertainty and the inaccuracy in the clear-sky models, and thus the forecast errors at those times can be large. To exclude those undesirable forecasts that severely distort the error metrics, solar forecasters usually apply a zenith angle filter, e.g., zenith angle $< 85°$, before error computation. Alternatively, one can opt an additive seasonality modeling, e.g., $r_{t+h} = x_t - c_t + c_{t+h}$. However, the remainder series (i.e., $x_t - c_t$) in this case is still heteroscedastic.

One can also use a "cloudiness index" where the reference forecast is referred to as "smart persistence" by Inman et al. (2015). This reference model includes the effects of air mass, aerosols, turbidity, i.e., every major atmospheric effect but clouds. Because the timescale for turbidity variations is much larger than the timescale for cloud-cover variations, the cloudiness index provides the best possible reference for short-time forecasts. Any skill calculated over the cloudiness-index persistence measures the ability to capture cloud cover changes over short periods of time, and thus the qualifier "smart". Note that skills reported over smart persistence are necessarily lower than skills reported over clear-sky persistence since it is virtually impossible for a forecast to improve over smart persistence for cloudless skies since the smart persistence reference model includes all effects of diurnal variability (air mass or the cosine of the solar zenith angle corrections) plus the combined effect of water vapor path and aerosols. These types of smart persistence reference forecasts are typically updated sub-hourly (Inman et al., 2015; Reno and Hansen, 2016).

The literature on treatment of seasonal and multi-seasonal (e.g., diurnal and yearly cycles in solar irradiance) time series is rich. Chapter 3 of Makridakis et al. (2008) provides a detailed account for various statistical techniques for time series decomposition. In the solar forecasting literature, various techniques such as Fourier series, exponential smoothing, STL decomposition, additive clear-sky decomposition have also been extensively explored (e.g., Dong et al., 2013; Yang et al., 2015; Voyant and Notton, 2018). However, as compared to the clear-sky persistence, these methods usually require more steps, which may be a reason for their limited uptake. Since the main goal of seasonally adjusted persistence is to construct a better reference model than raw persistence, clear-sky persistence makes a good trade-off between implementation difficulty and baseline accuracy.

Appendix B. On Coimbra's skill score $s_{Coimbra}$ and s in meteorology

Consider the square of Coimbra's definition for U in Eq. (8):

$$U^2 = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{f_t - x_t}{c_t}\right)^2, \tag{B.1}$$

where f, x, and c are forecasts, observations, and the clear-sky irradiance expectations, respectively. If we assume the clear-sky irradiance has m discrete states, namely, $c^{(1)}, \cdots, c^{(m)}$, the summation can be rewritten as:

$$\begin{aligned}
U^2 =& \frac{1}{N}\left\{\sum_{t\in\mathcal{N}_1}\left(\frac{f_t - x_t}{c^{(1)}}\right)^2 + \sum_{t\in\mathcal{N}_2}\left(\frac{f_t - x_t}{c^{(2)}}\right)^2 + \cdots + \sum_{t\in\mathcal{N}_m}\left(\frac{f_t - x_t}{c^{(m)}}\right)^2\right\} \\
=& \frac{1}{N}\left\{\frac{1}{\left(c^{(1)}\right)^2}\sum_{t\in\mathcal{N}_1}(f_t - x_t)^2 + \frac{1}{\left(c^{(2)}\right)^2}\sum_{t\in\mathcal{N}_2}(f_t - x_t)^2 + \cdots + \frac{1}{\left(c^{(m)}\right)^2}\sum_{t\in\mathcal{N}_m}(f_t - x_t)^2\right\} \\
=& \frac{1}{N}\left\{\frac{|\mathcal{N}_1|}{\left(c^{(1)}\right)^2}\mathbb{E}\left[(f-x)^2\middle|c = c^{(1)}\right] + \frac{|\mathcal{N}_2|}{\left(c^{(2)}\right)^2}\mathbb{E}\left[(f-x)^2\middle|c = c^{(2)}\right] + \cdots + \frac{|\mathcal{N}_m|}{\left(c^{(m)}\right)^2}\mathbb{E}\left[(f-x)^2\middle|c = c^{(m)}\right]\right\},
\end{aligned} \tag{B.2}$$

where $\mathcal{N}_1, \cdots, \mathcal{N}_m$ are the sets of data index that satisfy events $c_t = c^{(1)}, \cdots, c_t = c^{(m)}$, respectively. The notation $|\mathcal{N}_1|$ denotes the cardinality of $\mathcal{N}_1$. If the expected squared error of forecast f is same for all c, i.e.,

$$\mathbb{E}\left[(f-x)^2\middle|c = c^{(1)}\right] = \mathbb{E}\left[(f-x)^2\middle|c = c^{(2)}\right] = \cdots = \mathbb{E}\left[(f-x)^2\middle|c = c^{(m)}\right] = \mathbb{E}\left[(f-x)^2\right], \tag{B.3}$$

then

$$U^2 = \frac{1}{N}\left\{\frac{|\mathcal{N}_1|}{\left(c^{(1)}\right)^2} + \frac{|\mathcal{N}_2|}{\left(c^{(2)}\right)^2} + \cdots \frac{|\mathcal{N}_m|}{\left(c^{(m)}\right)^2}\right\}\mathbb{E}\left[(f-x)^2\right]. \tag{B.4}$$

Similarly, the square of Coimbra's definition for V in Eq. (9):

$$\begin{aligned}
V^2 =& \frac{1}{N}\sum_{t=1}^{N}\left(\frac{x_{t-1}}{c_{t-1}} - \frac{x_t}{c_t}\right)^2 \\
=& \frac{1}{N}\sum_{t=1}^{N}\left(\frac{r_t - x_t}{c_t}\right)^2 \\
=& \frac{1}{N}\left\{\frac{|\mathcal{N}_1|}{\left(c^{(1)}\right)^2} + \frac{|\mathcal{N}_2|}{\left(c^{(2)}\right)^2} + \cdots \frac{|\mathcal{N}_m|}{\left(c^{(m)}\right)^2}\right\}\mathbb{E}[(r-x)^2],
\end{aligned} \tag{B.5}$$

if the squared reference forecast error, $(r-x)^2$, is independent of the clear-sky irradiance. Hence, the skill score

definition by Marquez and Coimbra (2013):

$$
\begin{aligned}
s_{\text{Coimbra}} =& 1 - \frac{U}{V} \\
=& 1 - \frac{\sqrt{\frac{1}{N}\left\{\frac{|\mathcal{N}_1|}{\left(c^{(1)}\right)^2} + \frac{|\mathcal{N}_2|}{\left(c^{(2)}\right)^2} + \cdots \frac{|\mathcal{N}_m|}{\left(c^{(m)}\right)^2}\right\}}\sqrt{\mathbb{E}\left[(f-x)^2\right]}}{\sqrt{\frac{1}{N}\left\{\frac{|\mathcal{N}_1|}{\left(c^{(1)}\right)^2} + \frac{|\mathcal{N}_2|}{\left(c^{(2)}\right)^2} + \cdots \frac{|\mathcal{N}_m|}{\left(c^{(m)}\right)^2}\right\}}\sqrt{\mathbb{E}\left[(r-x)^2\right]}} \\
=& 1 - \frac{\sqrt{\mathbb{E}\left[(f-x)^2\right]}}{\sqrt{\mathbb{E}\left[(r-x)^2\right]}} \\
=& 1 - \frac{\text{RMSE}(f,x)}{\text{RMSE}(r,x)} = s,
\end{aligned}
\tag{B.6}
$$

if both the squared forecast error and squared clear-sky persistence forecast error are independent of the clear-sky irradiance.

Appendix C. Links between measure-oriented and distribution-oriented forecast verification approaches

The *rule of the lazy statistician* states (Wasserman, 2013): Let $y = g(x)$, then

$$
\mathbb{E}(y) = \mathbb{E}[g(x)] = \int g(x)dF(x) = \int g(x)p(x)dx. \tag{C.1}
$$

The two-variable case is handled in a similar way: Let $z = g(x, y)$, then

$$
\mathbb{E}(z) = \mathbb{E}[g(x,y)] = \int g(x,y)dF(x,y) = \iint g(x,y)p(x,y)dxdy. \tag{C.2}
$$

This rule links the joint distribution to a large collection of error metrics. For example, MBE, MAE, and RMSE can be written as:

$$
\text{MBE} = \mathbb{E}[(f-x)] = \iint (f-x)p(f,x)dfdx, \tag{C.3}
$$

$$
\text{MAE} = \mathbb{E}(|f-x|) = \iint |f-x|p(f,x)dfdx, \tag{C.4}
$$

$$
\text{RMSE} = \sqrt{\mathbb{E}[(f-x)^2]} = \left[\iint (f-x)^2 p(f,x)dfdx\right]^{\frac{1}{2}}. \tag{C.5}
$$

Similarly, it is possible to express nMBE, nMAE, nRMSE, maximum absolute error, mean average percentage error, etc., in this form. Hence, it is clear that all of these metrics are just different ways to summarize the joint distribution.

In the report by Beyer et al. (2009), four metrics based on the Kolmogrov–Smirnov test were proposed, namely, the Kolmogrov–Smirnov test integral (KSI), the OVER index, KSE (linear combination of KSI and OVER), and RIO (sum of KSD and RMSE, divided by 2). For instance, KSI calculates approximately the area between the ECDFs of f and x, OVER calculates the area of those instances between the two ECDFs that exceed the critical value at 99% level of confidence. While the descriptions of these metrics can be confusing, and the calculation can be ambiguous (the number of intervals during the trapezoidal integration needs to be defined), they are in fact solar engineers' early attempts to summarize the (differences between) marginal distributions of f and x.

Lastly, Section 3 demonstrates several ways to summarize the conditional distributions. In these cases, the summaries are performed together on marginal distributions. More specifically, type 1 conditional bias and resolution are summaries of p(x|f) and p(f), whereas type 2 conditional bias and discrimination are summaries of p(f|x) and p(x) (Murphy, 1997).

To that end, the open questions "are there better ways to summarize these distributions," "which summaries allow cross-work forecast comparison," "how to analyze the summaries graphically," etc., jointly motivate future research on verification of deterministic solar forecasts.

References

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F.M., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Solar Energy 136, 78 – 111. URL: http://www.sciencedirect.com/science/article/pii/S0038092X1630250X, doi:https://doi.org/10.1016/j.solener.2016.06.069.

Antonanzas, J., Pozo-Vázquez, D., Fernandez-Jimenez, L., de Pison, F.M., 2017. The value of day-ahead forecasting for photovoltaics in the Spanish electricity market. Solar Energy 158, 140 – 146. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17308307, doi:https://doi.org/10.1016/j.solener.2017.09.043.

Antonanzas-Torres, F., Urraca, R., Polo, J., Perpiñán-Lamigueiro, O., Escobar, R., 2019. Clear sky solar irradiance models: A review of seventy models. Renewable and Sustainable Energy Reviews 107, 374 – 387. URL: http://www.sciencedirect.com/science/article/pii/S1364032119301261, doi:https://doi.org/10.1016/j.rser.2019.02.032.

Armstrong, J.S., 2001. Evaluating forecasting methods, in: Principles of forecasting. Springer, pp. 443 – 472.

Beyer, H.G., Polo Martinez, J., Suri, M., Torres, J.L., Lorenz, E., Müller, S.C., Hoyer-Klick, C., Ineichen, P., 2009. Benchmarking of Radiation Products. Technical Report 038665. Mesor Report D.1.1.3.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M., Badescu, V., 2019. A current perspective on the accuracy of incoming solar energy forecasting. Progress in Energy and Combustion Science 70, 119 – 144. URL: http://www.sciencedirect.com/science/article/pii/S0360128518300303, doi:https://doi.org/10.1016/j.pecs.2018.10.003.

Brown, L.D., Rozeff, M.S., 1978. The superiority of analyst forecasts as measures of expectations: Evidence from earnings. The Journal of Finance 33, 1 – 16. URL: http://www.jstor.org/stable/2326346, doi:https://doi.org/10.2307/2326346.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development 7, 1247 – 1250. URL: https://www.geosci-model-dev.net/7/1247/2014/, doi:https://doi.org/10.5194/gmd-7-1247-2014.

Chu, Y., Li, M., Pedro, H.T.C., Coimbra, C.F.M., 2015. Real-time prediction intervals for intra-hour DNI forecasts. Renewable Energy 83, 234 – 244. URL: https://www.sciencedirect.com/science/article/pii/S096014811500302X, doi:https://doi.org/10.1016/j.renene.2015.04.022.

Coimbra, C.F.M., Kleissl, J., Marquez, R., 2013. Chapter 8 - Overview of solar-forecasting methods and a metric for accuracy evaluation, in: Kleissl, J. (Ed.), Solar Energy Forecasting and Resource Assessment. Academic Press, Boston, pp. 171 – 194. URL: http://www.sciencedirect.com/science/article/pii/B9780123971777000085, doi:https://doi.org/10.1016/B978-0-12-397177-7.00008-5.

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. URL: http://www.sciencedirect.com/science/article/pii/S0360544213003381, doi:https://doi.org/10.1016/j.energy.2013.04.027.

Fildes, R., Nikolopoulos, K., Crone, S.F., Syntetos, A.A., 2008. Forecasting and operational research: a review. Journal of the Operational Research Society 59, 1150 – 1172. doi:https://doi.org/10.1057/palgrave.jors.2602597.

Gilleland, E., Ahijevych, D.A., Brown, B.G., Ebert, E.E., 2010. Verifying forecasts spatially. Bulletin of the American Meteorological Society 91, 1365 – 1376. doi:https://doi.org/10.1175/2010BAMS2819.1.

Gneiting, T., 2011. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746 – 762. URL: http://www.jstor.org/stable/41416407, doi:https://doi.org/10.2307/41416407.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359 – 378. doi:https://doi.org/10.1198/016214506000001437.

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Solar Energy 128, 1 – 30. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15005435, doi:https://doi.org/10.1016/j.solener.2015.10.010. special issue: Progress in Solar Energy.

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

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. International Journal of Forecasting 32, 896 – 913. URL: http://www.sciencedirect.com/science/article/pii/S0169207016000133, doi:https://doi.org/10.1016/j.ijforecast.2016.02.001.

Huang, J., Thatcher, M., 2017. Assessing the value of simulated regional weather variability in solar forecasting using numerical weather prediction. Solar Energy 144, 529 – 539. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17300774, doi:https://doi.org/10.1016/j.solener.2017.01.058.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679 – 688. URL: http://www.sciencedirect.com/science/article/pii/S0169207006000239, doi:https://doi.org/10.1016/j.ijforecast.2006.03.001.

Inman, R.H., Edson, J.G., Coimbra, C.F.M., 2015. Impact of local broadband turbidity estimation on forecasting of clear sky direct normal irradiance. Solar Energy 117, 125 – 138. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15002200, doi:https://doi.org/10.1016/j.solener.2015.04.032.

Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Progress in Energy and

Combustion Science 39, 535 – 576. URL: http://www.sciencedirect.com/science/article/pii/S0360128513000294, doi:https://doi.org/10.1016/j.pecs.2013.06.002.
IPCC, 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland.
Jolliffe, I.T., 2008. The impenetrable hedge: a note on propriety, equitability and consistency. Meteorological Applications 15, 25 – 29. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.60, doi:https://doi.org/10.1002/met.60.
Jolliffe, I.T., Stephenson, D.B., 2012. Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons.
Killinger, S., Engerer, N., Müller, B., 2017. QCPV: A quality control algorithm for distributed photovoltaic array power output. Solar Energy 143, 120 – 131.
Law, E.W., Kay, M., Taylor, R.A., 2016. Calculating the financial value of a concentrated solar thermal plant operated using direct normal irradiance forecasts. Solar Energy 125, 267 – 281. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15007045, doi:https://doi.org/10.1016/j.solener.2015.12.031.
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. Atmospheric Measurement Techniques 6, 2403 – 2418. URL: https://www.atmos-meas-tech.net/6/2403/2013/, doi:http://dx.doi.org/10.5194/amt-6-2403-2013.
Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. The Open Atmospheric Science Journal 2, 23 – 37.
Lorenz, E., Kühnert, J., Heinemann, D., Nielsen, K.P., Remund, J., Müller, S.C., 2016. Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. Progress in Photovoltaics: Research and Applications 24, 1626 – 1640. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2799, doi:10.1002/pip.2799.
Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H.A., Nielsen, T.S., 2005. Standardizing the performance evaluation of short-term wind power prediction models. Wind Engineering 29, 475 – 489. doi:https://doi.org/10.1260/030952405776234599.
Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting methods and applications. John wiley & sons.
Marquez, R., Coimbra, C.F.M., 2011. A novel metric for evaluation of solar forecasting models, in: ASME 2011 5th International Conference on Energy Sustainability, ASME. pp. 1459 – 1467. doi:https://doi.org/10.1115/ES2011-54519.
Marquez, R., Coimbra, C.F.M., 2013. Proposed metric for evaluation of solar forecasting models. Journal of solar energy engineering 135, 011016. doi:https://doi.org/10.1115/1.4007496.
Martinez-Anido, C.B., Botor, B., Florita, A.R., Draxl, C., Lu, S., Hamann, H.F., Hodge, B.M., 2016. The value of day-ahead solar power forecasting improvement. Solar Energy 129, 192 – 203. URL: http://www.sciencedirect.com/science/article/pii/S0038092X16000736, doi:https://doi.org/10.1016/j.solener.2016.01.049.
van der Meer, D., Widén, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renewable and Sustainable Energy Reviews 81, 1484 – 1512. URL: http://www.sciencedirect.com/science/article/pii/S1364032117308523, doi:https://doi.org/10.1016/j.rser.2017.05.212.
Mora, C., Spirandelli, D., Franklin, E.C., Lynham, J., Kantar, M.B., Miles, W., Smith, C.Z., Freel, K., Moy, J., Louis, L.V., Barba, E.W., Bettinger, K., Frazier, A.G., Colburn IX, J.F., Hanasaki, N., Hawkins, E., Hirabayashi, Y., Knorr, W., Little, C.M., Emanuel, K., Sheffield, J., Patz, J.A., Hunter, C.L., 2018. Broad threat to humanity from cumulative climate hazards intensified by greenhouse gas emissions. Nature Climate Change 8, 1062 – 1071. doi:https://doi.org/10.1038/s41558-018-0315-6.
Moskaitis, J.R., 2008. A case study of deterministic forecast verification: Tropical cyclone intensity. Weather and Forecasting 23, 1195 – 1220. doi:https://doi.org/10.1175/2008WAF2222133.1.
Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review 116, 2417 – 2424. doi:https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.
Murphy, A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8, 281 – 293. doi:https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.
Murphy, A.H., 1997. Forecast verification, in: Katz, R.W., Murphy, A.H. (Eds.), Economic Value of Weather and Climate Forecasts. Cambridge University Press, pp. 19 – 74. doi:https://doi.org/10.1017/CBO9780511608278.003.
Murphy, A.H., Brown, B.G., Chen, Y.S., 1989. Diagnostic verification of temperature forecasts. Weather and Forecasting 4, 485–501. doi:https://doi.org/10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2.
Murphy, A.H., Winkler, R.L., 1971. forecasters and probability forecasts: some current problems,. Bulletin of the American Meteorological Society 52, 239 – 248. doi:https://doi.org/10.1175/1520-0477(1971)052<0239:FAPFSC>2.0.CO;2.
Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Monthly Weather Review 115, 1330 – 1338. doi:https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.
Pedro, H.T.C., Coimbra, C.F.M., 2015. Short-term irradiance forecastability for various solar micro-climates. Solar Energy 122, 587 – 602. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15005162, doi:https://doi.org/10.1016/j.solener.2015.09.031.
Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G.V., Hemker, K., Heinemann, D., Remund, J., Müller, S.C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J.A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L.M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Solar Energy 94, 305 – 326. URL: http://www.sciencedirect.com/science/article/pii/S0038092X13001886, doi:https://doi.org/10.1016/j.solener.2013.05.005.
Ren, Y., Suganthan, P., Srikanth, N., 2015. Ensemble methods for wind and solar power forecasting—ĂŤA state-of-the-art review. Renewable and Sustainable Energy Reviews 50, 82 – 91. URL: http://www.sciencedirect.com/science/article/pii/S1364032115003512, doi:https://doi.org/10.1016/j.rser.2015.04.081.
Reno, M.J., Hansen, C.W., 2016. Global horizontal irradiance clear sky models: Implementation and analysis. Renewable Energy 90, 520 – 531. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15002200, doi:https://doi.org/10.1016/j.renene.2015.12.031.

Ruiz-Arias, J.A., Gueymard, C.A., 2018. Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface. Solar Energy 168, 10 – 29. URL: http://www.sciencedirect.com/science/article/pii/S0038092X18301257, doi:https://doi.org/10.1016/j.solener.2018.02.008. advances in Solar Resource Assessment and Forecasting.

Schilling, R.L., 2017. Measures, integrals and martingales. Cambridge University Press.

Urraca, R., Gracia-Amillo, A.M., Huld, T., de Pison, F.J.M., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017. Quality control of global solar radiation data with satellite-based products. Solar Energy 158, 49 – 62.

Vallance, L., Charbonnier, B., Paul, N., Dubost, S., Blanc, P., 2017. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. Solar Energy 150, 408 – 422. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17303687, doi:https://doi.org/10.1016/j.solener.2017.04.064.

Voyant, C., Notton, G., 2018. Solar irradiation nowcasting by stochastic persistence: A new parsimonious, simple and efficient forecasting tool. Renewable and Sustainable Energy Reviews 92, 343 – 352. URL: http://www.sciencedirect.com/science/article/pii/S1364032118303344, doi:https://doi.org/10.1016/j.rser.2018.04.116.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. Renewable Energy 105, 569 – 582. URL: http://www.sciencedirect.com/science/article/pii/S0960148116311648, doi:https://doi.org/10.1016/j.renene.2016.12.095.

Wasserman, L., 2013. All of statistics: a concise course in statistical inference. Springer Science & Business Media.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30, 79 – 82. URL: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/, doi:https://doi.org/10.3354/cr030079.

Yang, D., 2018. A correct validation of the National Solar Radiation Data Base (NSRDB). Renewable and Sustainable Energy Reviews 97, 152 – 155. URL: http://www.sciencedirect.com/science/article/pii/S1364032118306087, doi:https://doi.org/10.1016/j.rser.2018.08.023.

Yang, D., 2019. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). Journal of Renewable and Sustainable Energy 11, 022701. doi:https://doi.org/10.1063/1.5087462.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T.C., Coimbra, C.F.M., 2018. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar Energy 168, 60–101. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17310022, doi:https://doi.org/10.1016/j.solener.2017.11.023. Advances in Solar Resource Assessment and Forecasting.

Yang, D., Perez, R., 2019. Can we gauge forecasts using satellite-derived solar irradiance? Journal of Renewable and Sustainable Energy 11, 023704. doi:https://doi.org/10.1063/1.5087588.

Yang, D., Quan, H., Disfani, V.R., Rodríŋguez-Gallegos, C.D., 2017. Reconciling solar forecasts: Temporal hierarchy. Solar Energy 158, 332 – 346. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17308423, doi:https://doi.org/10.1016/j.solener.2017.09.055.

Yang, D., Sharma, V., Ye, Z., Lim, L.I., Zhao, L., Aryaputera, A.W., 2015. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. Energy 81, 111 – 119. URL: http://www.sciencedirect.com/science/article/pii/S0360544214013528, doi:https://doi.org/10.1016/j.energy.2014.11.082.

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. Solar Energy 111, 157 – 175. URL: http://www.sciencedirect.com/science/article/pii/S0038092X14005027, doi:https://doi.org/10.1016/j.solener.2014.10.016.