International Journal of Forecasting 39 (2023) 981–991

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Calibration of deterministic NWP forecasts and its impact on verification

Martin János Mayer [a,*], Dazhi Yang [b]

[a] *Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111, Budapest, Hungary*
[b] *School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China*

ARTICLE INFO

Keywords:
Solar forecasting
Verification
Calibration
Correlation coefficient
Numerical weather prediction

ABSTRACT

Deterministic forecasts (as opposed to ensemble or probabilistic forecasts) issued by numerical weather prediction (NWP) models require post-processing. Such corrective procedure can be viewed as a form of calibration. It is well known that, based on different objective functions, e.g., minimizing the mean square error or the mean absolute error, the calibrated forecasts have different impacts on verification. In this regard, this paper investigates how a calibration directive can affect various aspects of forecast quality outlined in the Murphy–Winkler distribution-oriented verification framework. It is argued that the correlation coefficient is the best measure for the potential performance of NWP forecast verification when linear calibration is involved, because (1) it is not affected by the directive of linear calibration, (2) it can be used to compute the skill score of the linearly calibrated forecasts, and (3) it can avoid the potential deficiency of using squared error to rank forecasts. Since no single error metric can fully represent all aspects of forecast quality, forecasters need to understand the trade-offs between different calibration strategies. To echo the increasing need to bridge atmospheric sciences, renewable energy engineering, and power system engineering, as to move toward the grand goal of carbon neutrality, this paper first provides a brief introduction to solar forecasting, and then revolves its discussion around a solar forecasting case study, such that the readers of this journal can gain further understanding on the subject and thus potentially contribute to it.

1. A brief introduction to solar forecasting

Over the past few years, many countries have pledged their aggressive plans of action for moving toward carbon neutrality by mid of this century. Since this goal can only be achieved with rapid and radical changes to the ways in which energy is generated and consumed, the need to increase the penetration of renewable energy has been emphasized in all those plans. Insofar as grid integration of renewable energy is concerned, forecasting always plays a key part. Indeed, modern power grid operation and control require not only future information on electric load and price but also forecasts of renewable energy generation, over various horizons. On this point, the reader is referred to the works by Makarov, Etingov, Ma, Huang, and Subbarao (2011) and Yang, Li, Yagli, and Srinivasan (2021) for overviews on the correspondence between various power system operations (such as unit commitment, economic dispatch, or regulation) and forecast horizons, as advised and followed by the California Independent System Operator and State Grid Corporation of China, respectively.

Owing to the rapid increase in penetration of solar photovoltaic (PV) in the recent decade, there has been an

[*] Corresponding author.
E-mail address: mayer@energia.bme.hu (M.J. Mayer).

https://doi.org/10.1016/j.ijforecast.2022.03.008

exponential growth of research in solar forecasting, which is the youngest subdomain of energy forecasting (Hong et al., 2016). The recent survey by Hong et al. (2020) has further confirmed the above finding after conducting a bibliometric analysis on energy forecasting papers published over the past ten years. Although the number of papers does not necessarily indicate progress or level of maturity, the size of solar forecasting literature has overtaken that of price forecasting and is catching up with that of wind forecasting.

Forecasting solar irradiance, as a means to arrive at solar power forecasts, differs from forecasting in a social setting (Makridakis, Hyndman, & Petropoulos, 2020). Given the fact that solar irradiance is a physical (atmospheric) process that is predominantly affected by moving clouds, many advanced solar forecasting techniques leverage instruments that can provide 2D or 3D views of the sky, such as total sky imagers or radiometers onboard geostationary satellites and polar orbiters (e.g., Kazantzidis et al., 2017; Miller, Rogers, Haynes, Sengupta, & Heidinger, 2018). Besides camera- and satellite-based forecasting, another major class of solar forecasting methods relies on numerical weather prediction (NWP) models (e.g., Perez et al., 2013). It is worth noting that statistical and machine-learning methods have also been widely applied to solar forecasting problems, though not in a stand-alone way, but as post-processing tools to complement those physics-based methods (Yang & van der Meer, 2021).

The body of literature on solar forecasting has grown very large, and it is not possible, at this date, to fully elaborate on all aspects of it in just a few pages. In this regard, anyone who wishes to have an in-depth understanding of solar forecasting is referred to a recent uber review, which is a compendium of review papers on solar forecasting (Yang, Wang, & Xia, 2022), as a starting point. Additionally, the recent review by (Yang, Wang, Gueymard, et al., 2022) discussed, in great detail, how solar forecasting depends on atmospheric science and impacts grid integration. In what follows, we should confine ourselves to a niche problem of solar forecasting, that is, calibration of NWP-based irradiance forecasts and its impact on forecast verification.

2. Post-processing and calibrating irradiance forecasts

As mentioned earlier, the use of dynamical weather models, which can capture the trajectories of weather events in mesoscale more effectively than statistical and machine-learning models, has hitherto been attractive to solar forecasters. Ever since the seminal review by Inman, Pedro, and Coimbra (2013), numerical weather prediction (NWP) has been widely accepted as the most suitable approach to solar forecasting for any forecast horizon longer than four hours (Miller et al., 2018; Nielsen, Iosifidis, & Karstoft, 2021; Yang, Kleissl, Gueymard, Pedro, & Coimbra, 2018).

Historically, solar irradiance was not regarded as a major output variable of NWP—as compared to temperature or wind, the social-economic impact of solar irradiance is much lower. Although the science of NWP, particularly the part pertaining to radiation and cloud microphysics, has certainly improved over the years, the production, maintenance, and disappearance of the cumulus and cirrus clouds are still the most challenging weather processes to forecast, of which the imperfect parameterizations translate to the inaccuracy of irradiance forecasts (Bauer, Thorpe, & Brunet, 2015). On this point, owing to the aforementioned emerging needs for large-scale grid integration of solar power, such as PV or concentrating solar power (CSP), NWP models with parameterizations and assimilation techniques that are specifically chosen to favor solar irradiance forecasting have been developed (Jimenez et al., 2016; Sahu, Yang, & Kleissl, 2018).

Furthermore, from a forecast practitioner's perspective (e.g., the perspective of a PV plant owner, who needs to submit forecasts to a utility operator), running specialized NWP models is rarely a feasible option, because setting up such models requires not only a substantial amount of meteorological knowledge but also expansive hardware. Hence, applying statistical or machine learning post-processing techniques to operational NWP forecasts disseminated by national weather centers and space agencies is a far more common way of generating solar power forecasts. The raw irradiance forecasts from NWP models can be converted to solar power forecasts through the so-called *model chain* (Mayer & Gróf, 2021). Analogously, a model chain is simply the "wind power curve" for solar irradiance.

Constructing a model chain can take many steps, but not all are essential. The granularity of modeling depends on both the solar forecaster's skill and the available data (Mayer, 2021). The major steps include a separation model, which splits the global horizontal irradiance (GHI) to the beam normal irradiance (BNI) and the diffuse horizontal irradiance (DHI); a transposition model, which converts GHI, BNI, and DHI to in-plane irradiance, also known as the global tilted irradiance (GTI); and an irradiance-to-PV model, which is available from both free and commercial software packages, such as pvlib, SAM, or PVsyst. Since the construction of all model chains starts from weather input, e.g., GHI forecasts from NWP models, NWP forecast post-processing is vital to uncertainty propagation toward the end of the model chain. Stated differently, because the uncertainty in the GHI forecasts is very likely to be the largest among the uncertainty components at different stages of the model chain, the final uncertainty on the PV power forecast can be significantly lowered with an improved set of GHI forecasts.

Forecast quality, which is one of the three types of the goodness of a forecast (Murphy, 1993), is a general concept that goes far beyond the common error metrics used in solar forecast verification, such as the mean bias error (MBE) or root mean square error (RMSE). On this point, Murphy and Winkler (1987) introduced several useful aspects of forecast quality, such as *calibration*, *resolution*, or *discrimination*, through two factorizations. More specially, they are: (1) the calibration–refinement factorization:

$$\mathrm{MSE}\left(f, x\right) = \mathbb{V}\left(x\right) + \mathbb{E}_f\left[f - \mathbb{E}\left(x|f\right)\right]^2 - \mathbb{E}_f\left[\mathbb{E}\left(x|f\right) - \mathbb{E}\left(x\right)\right]^2, \tag{1}$$

and (2) the likelihood–base rate factorization:

$$\mathrm{MSE}\,(f,x)=\mathbb{V}\,(f)+\mathbb{E}_x\left[x-\mathbb{E}\,(f|x)\right]^2-\mathbb{E}_x\left[\mathbb{E}\,(f|x)-\mathbb{E}\,(f)\right]^2, \tag{2}$$

where f and x denote forecast and observation, respectively; $\mathbb{E}$ and $\mathbb{V}$ are the expectation and variance operators, respectively; MSE stands for mean squared error. In what follows, Eqs. (1) and (2) are jointly referred to as the Murphy–Winkler factorization and will be discussed more in Section 3.

Although the Murphy–Winkler factorization, or more generally, the Murphy–Winkler distribution-oriented verification framework, has been around for decades, it was not known by the solar forecasting community until a recent publication by Yang and Perez (2019). Due to its clear advantages over the traditional measure-oriented verification, the distribution-oriented verification has been quickly recommended as the standard practice in deterministic solar forecast verification by a group of 33 forecasting experts (Yang et al., 2020); the reader is referred to that review for a detailed comparison between the Murphy–Winkler verification framework and the traditional measure-oriented approaches. Extending from its application in forecasting, the Murphy–Winkler factorization has also been applied to the validation of gridded radiation and aerosol products from geostationary satellites and reanalyses (Yang & Bright, 2020; Yang & Gueymard, 2021). Since the Murphy–Winkler verification framework assesses forecast quality by examining the joint distribution of forecast and observation, which contains all time-independent information relevant to verification, it is more general and thus more flexible than the measure-oriented verification. Hence, in this paper, the discussion on forecast verification revolves around the Murphy–Winkler factorization. More specifically, the effects of using an improved set of deterministic NWP forecasts (as opposed to ensemble NWP forecasts) on various aspects of quality are studied in terms of those aspects of forecast quality as described by the Murphy–Winkler factorization.

Next, we discuss forecast post-processing. The most commonly used class of methods for improving deterministic NWP forecasts is the model output statistics (MOS), which seeks error patterns from historical forecast–observation pairs, and thus leverages such information to perform corrections on the current forecasts. The popularity of MOS in solar forecasting can be largely attributed to Lorenz, Hurka, Heinemann, and Beyer (2009). In that work, the bias, i.e., the difference between a forecast and an observation, is modeled as a polynomial function of forecast clear-sky index and cosine of zenith angle. Although such a simplistic polynomial function has been extended numerous times, e.g., into a kernel conditional density estimation (KCDE) function (Yang, 2019b), the main purpose of MOS is to reduce, if not eliminate, the model-led bias in NWP forecasts. In the language of Murphy and Winkler (1987), this kind of post-processing is referred to as *calibration*, because MOS aims to reduce the bias between a given forecast and the materialized observation, i.e., the $\mathbb{E}_f\left[f-\mathbb{E}\,(x|f)\right]^2$ term in Eq. (1).

Even though MOS is mostly used for bias correction, applying it would also influence the variance of the forecasts. Underdispersed forecasts would fail to cover the physically possible range of irradiance values, which may then lead to insufficient commitment of reserves and thus may threaten the stability of the grid. In this regard, this study is the first to reveal the relationship between the dispersion of solar forecasts and the commonly used performance measures. In this regard, the term *calibration* is used in a broader sense, covering not only the bias reduction but also the adjustment of the forecast variance, for either the lowest mean squared error (MSE), the lowest mean absolute error (MAE), or is equal to the observed variance. These three calibration goals are inherently conflicting because optimizing forecasts based on one metric leads to inferiority in others. As both MSE and MAE depend on the dispersion of the forecasts, and even a simple linear calibration significantly may change their values, using these metrics for comparative verification of forecasts with different variances may lead to misleading or inaccurate conclusions.

The underlying issue here is related to the general notion of *consistency*, which refers to the correspondence between judgment and forecast, and has been known since at least the 1960s (Wilks & Murphy, 1986), and has been reiterated numerous times (e.g. Armstrong, 2001; Gneiting, 2011; Murphy, 1993; Stephan & Martin, 2011). In the solar forecasting literature, however, the concept of consistency was not overtly discussed until very recently (Yang et al., 2020). Before the publication of that review, already numerous attempts were made to post-process solar forecasts with advanced statistical and machine-learning methods (see Yang & van der Meer, 2021, for a review). In virtually all works of that sort, superiority claims were concluded based on one or a few accuracy measures, of which the final presented choice of measures is often made after repeated trial-and-error, as to maximize the apparent goodness of the proposed methods. As argued by Jolliffe (2008), this kind of verification is essentially circular.

In view of the consistency problem outlined above, this paper advocates the use of the correlation coefficient between forecast and observation as a measure of performance of NWP forecasts. As an alternative to the measures for accuracy such as MSE or MAE, the correlation coefficient describes the association between forecast and observation, but not the correspondence between the values of forecast and observation. Since the correlation coefficient is invariant with respect to the linear transformation of the forecast, it is *not* affected by the directive under which the calibration is done, and thus escapes from the circularity of consistency during forecast verification. Furthermore, as shown in the following pages, insofar as linear calibration is concerned, the MSE skill score depends on the correlation, regardless of which calibration directive is issued.

The rest of the paper is organized as follows. The theoretical background and derivation of the main findings are summarized in Section 3, based on both a visual distribution-oriented and a formal measure-oriented approach. The results are empirically demonstrated for

NWP irradiance forecasts in Section 4. The practical and verification-related implications and recommendations are discussed in Section 5, while the conclusions are summarized in Section 6.

3. Theoretical foundations

3.1. Error distribution of solar forecasts

The process of making optimal deterministic (or point) forecasts depends on the scoring function, i.e., the optimality depends on the "forecast directive". Many commonly used error metrics can be minimized by different functionals of the predictive distribution, e.g., MSE is minimized by the mean, and MAE is minimized by the median (Gneiting, 2011). Even though the predictive distribution does not appear directly in point forecasts, considering it in the background facilitates the understanding of solar forecasts' characteristics (Kolassa, 2020). The predictive distribution of a deterministic solar forecast can materialize from the conditional distribution of historical observations and forecasts. The $p(x|\,f)$ conditional distribution represents the predictive distribution in those circumstances when a point forecast f is issued. The MSE or the MAE of a forecast can be minimized by replacing all f point forecast values with the mean or the median of $p(x|\,f)$, respectively.

The ridgeline plots drawn in Fig. 1 show the conditional distributions of the North American Mesoscale (NAM) irradiance forecasts and Surface Radiation Budget Network (SURFRAD) observations for Goodwin Creek (GWN), Mississippi—SURFRAD is a seven-station irradiance monitoring network that is of the world's highest radiometry standard. The $p(x|\,f)$ distributions have a negative skew (i.e., a long tail toward the lower values) at high and a positive skew at low forecast irradiance. This asymmetry is not a special trait of this particular NWP model; rather, this characteristic appears to be general, due to the "two-state" nature of solar irradiance. Stated differently, owing to the clear and cloudy states of the sky, a misidentification of the sky condition can lead to a large bias toward the opposite direction. Observed irradiance varies between zero and the maximum clear-sky irradiance (excluding the transient cloud-enhancement events). Therefore, in the case of high forecast irradiance, there is a much bigger margin for overestimation than for underestimation, while the opposite is true for low forecast values. Thus, these large errors resulting from erroneous sky-state identification contribute to the asymmetry of the conditional distributions.

The MSE of the forecasts can be reduced by corrections that bring the mean of $p(x|\,f)$ closer to the $f = x$ identity line. In the Murphy–Winkler factorization, this distance determines the $\mathbb{E}_f\,[f - \mathbb{E}\,(x|f)]^2$ term, i.e., the type-1 conditional bias, which reflects the calibration of the forecasts. A lower value of this term indicates a better calibration. The ridgeline plot of $p(x|\,f)$ for the MSE-optimized NAM forecasts is shown in Fig. 1(c) (the MSE-optimized forecast is created by a linear transformation, as described in the following subsection). However, adjusting $\mathbb{E}\,(x|f)$ to the identity line would shift a larger portion of the distribution below (above) the identity line at high (low) irradiance forecasts, due to the aforementioned skewness of the distributions, which would then lead to under dispersed forecasts (note that the values above 900 W/m^2 are practically eliminated from the forecast). Moreover, as shown in Fig. 1(b) and (d), the underdispersion moves the mean of the $p(f|x)$ distributions further away from the identity line, which increases the $\mathbb{E}_x\,[x - \mathbb{E}\,(f|x)]^2$ term, also known as the type-2 conditional bias, and decreases the $\mathbb{E}_x\,[\mathbb{E}\,(f|x) - \mathbb{E}\,(f)]^2$ term, which is the discrimination of the forecasts. Fig. 1 demonstrates this effect for only one location, but similar patterns can also be observed at all SURFRAD stations, as shown in the figures included in the supplementary materials. Moreover, as long as the conditional distributions are not symmetrical, these findings would likely apply to forecasts issued by any solar forecasting model—the conditional distributions can only be symmetrical if the states of the sky are perfectly identified. (Recall that the asymmetry originates from the misidentification of a clear-sky condition as a cloudy one or vice versa.)

On the other hand, the forecasts can also be optimized for the MAE, by shifting the median of $p(x|\,f)$ closer to the identity line. Fig. 1(e) and (f) show the ridgeline plots for the MAE-optimized forecasts. The MAE-optimized forecasts are also slightly under dispersed, but they have a lower type-2 conditional bias and higher discrimination as compared to the MSE-optimized ones. The higher type-1 conditional bias indicates a worse calibration, but this is due to the simple fact that these forecasts are not calibrated for the mean but the median. The MSE-optimized, mean-calibrated forecasts can be interpreted as the average of the expected solar irradiance. In contrast, in the case of MAE-optimized, median-calibrated forecasts, there is a 50%–50% chance that the actual irradiance will be lower or higher than the forecast value, respectively. These two calibration philosophies result in significantly different forecasts, and the superiority of one set of calibrated forecasts over another depends on the directive under which the verification is conducted.

3.2. Effect of a linear calibration on the forecast performance metrics

The qualitative findings derived from the analysis of the distributions can be supported by quantitative results using the well-known bias–variance decomposition of the MSE:

$$\mathrm{MSE}\,(f,x) = \mathbb{V}\,(f) + \mathbb{V}\,(x) - 2\rho\,(f,x)\,\sqrt{\mathbb{V}\,(f)\,\mathbb{V}\,(x)} + \mathrm{MBE}^2\,(f,x)\,, \tag{3}$$

where $\mathbb{V}\,(.)$ stands for the variance, ρ is the correlation coefficient, and MBE is the mean bias error. This equation can be re-arranged, as proposed by Murphy (1995):

$$\mathrm{MSE}\,(f,x) = \left[1 - \rho^2(f,x)\right]\mathbb{V}\,(x) + [\sigma\,(f) - \rho\,(f,x)\,\sigma\,(x)]^2 + \mathrm{MBE}^2\,(f,x)\,, \tag{4}$$

where $\sigma(.)$ is the standard deviation. If MBE and ρ are kept unchanged, MSE is minimal when $\mathbb{V}\,(f) = \rho^2(f,x)\mathbb{V}\,(x)$,

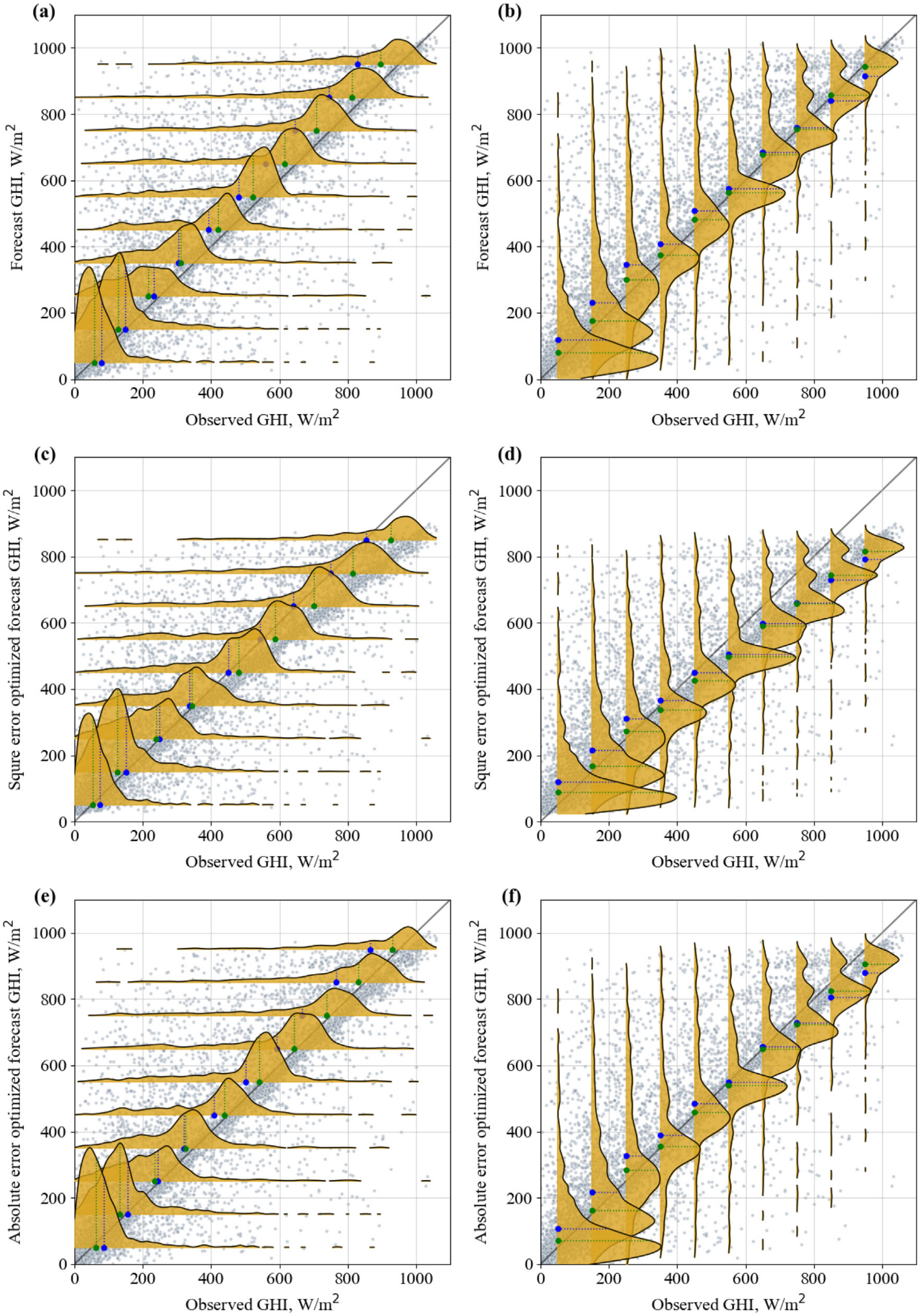

Fig. 1. Conditional distributions of 24-h-ahead hourly NAM forecasts and SURFRAD observations for Goodwin Creek, Mississippi. (a), (c) and (e) show $p(x| f)$ and (b), (d) and (f) show $p(f| x)$ for the raw, MSE-optimized, and MAE-optimized forecasts, respectively. Blue and green dots represent the mean and the median of the distributions, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

i.e., the minimization of MSE leads to the inevitable underdispersion of the forecast. The underdispersion depends on the association (i.e., correlation coefficient) between the forecasts and observations. The higher association there is, the less under dispersed the forecasts are. This result has also been reported by Vannitsem and Hagedorn (2011) who found that post-processing via MOS, as to reduce the MSE, leads to under dispersed forecasts for longer lead times, i.e., the corrected forecasts converge to the climatological mean as the correlation between the forecasts and observations decreases.

Graphically, a well-known tool for the joint visualization of the $\sigma(f)$, $\rho(f,x)$ and $\mathrm{RMSE}(f,x)$, based on the analogy of formulas for the bias–variance decomposition of Eq. (3) and the law of cosines, is the Taylor diagram (Taylor, 2001). The underdispersion of the MSE-optimized forecasts can also be seen from the Taylor diagram: the tangent point of the correlation coefficient and RMSE isolines shift towards decreasing standard deviation as the correlation coefficient decreases. However, Taylor diagrams have not gained wide uptake in the solar forecast verification literature, though it is a familiar concept in the statistical forecasting community.

The simplest way to calibrate a forecast is through a linear transformation:

$$f' = af + b \tag{5}$$

where a and b are the scale and offset parameters, respectively. As long as $a > 0$, this linear transformation does not change the correlation coefficient, i.e., $\rho\left(f', x\right) = \rho(f, x)$.

The bias and variance of the calibrated forecast are $\mathrm{MBE}\left(f', x\right) = a\mathbb{E}(f) + b - \mathbb{E}(x)$ and $\mathbb{V}\left(f'\right) = a^2\mathbb{V}(f)$, respectively. The scale parameter can be used to set the variance to the desired value; then an offset can be chosen such that it eliminates the bias (i.e., $\mathrm{MBE}\left(f', x\right) = 0$). In general, the variance of the forecast can be set to any arbitrary value, which can be described as a proportion of the observed variance with an F variance ratio. The variance ratio, as introduced by Mayer and Gróf (2021) in the context of solar forecasting, provides a one-number summary of the dispersion of forecasts. Two special cases are $F = \rho^2$, which corresponds to the calibration directive of having the lowest squared error, and $F = 1$, which occurs when the forecast and observed variances are the same. The calculation of parameters a and b and the resulting MSE are summarized in Table 1. The expression for the offset parameter is the same across all cases, i.e., $\mathbb{E}(x) - a\mathbb{E}(f)$, but since the expression contains in itself the scale parameter, the offset takes a different numerical value in each case. It also can be seen from the table that parameters a and b can be determined from the $\mathbb{V}(x)$ variance and $\mathbb{E}(x)$ mean of the observations, which are not known exactly in advance but can be effectively estimated from historical data.

The MSE of forecasts depends only on the correlation coefficient between the forecast and observation, the chosen variance ratio, and the variance of observation, which is independent of the forecast. Furthermore, the ratio of the optimized MSE to the MSE of the variance-corrected forecasts is $[1 + \rho(f, x)]/2$. For example, for a set of forecasts with a 0.85 correlation coefficient, the MSE-optimized forecasts have a 7.5% lower MSE as compared to the variance-corrected ones, but it only captures 72.25% of the observed variability. These numbers are the same for any set of forecast–observation pairs that have $\rho = 0.85$, which indicates that the often-overlooked correlation coefficient is a unique indicator of forecast performance. However, since the correlation coefficient does not measure the correspondence between the *values* of forecast and observation, it does not gauge accuracy or skill (Murphy, 1995).

The RMSE skill score is the most commonly used metric to gauge the overall skillfulness of a solar forecaster (Yang et al., 2020). The RMSE skill score is calculated as $s = 1 - \mathrm{RMSE}\left(f', x\right)/\mathrm{RMSE}_{\mathrm{ref}}$, and in our particular case, it expresses the RMSE improvement of the calibrated NWP forecasts over the reference forecasts. The standard of reference can be either persistence, climatology, or the optimal convex combination of climatology and persistence (CLIPER) as suggested by Yang (2019a). Table 2 lists the skill score expressions of the differently calibrated forecasts over the three different standards of reference.

The skill scores only depend on the correlation coefficient, the variance ratio of choice, and $\gamma(h)$, the lag-h autocorrelation of the observations. Since the variance ratio under an MSE-optimized or variance-corrected calibration is fixed, the *potentially* best skill of the calibrated forecasts depends *only* on the correlation coefficient. This result again agrees with the finding of Murphy and Epstein (1989) that the correlation coefficient is "a measure of potential rather than actual skill".

The four terms of the calibration–refinement and likelihood–base rate factorizations are influenced by the linear calibration in the following ways:

- Type-1 conditional bias $\mathbb{E}_f\left[f - \mathbb{E}(x|f)\right]^2$ (lower is better): changes to the same extent as the MSE.[1]
- Resolution $\mathbb{E}_f\left[\mathbb{E}(x|f) - \mathbb{E}(x)\right]^2$ (higher is better): constant, not affected by the linear transformation.
- Type-2 conditional bias $\mathbb{E}_x\left[x - \mathbb{E}(f|x)\right]^2$ (lower is better): has a minimum for overdispersed forecasts with $F = 1/\rho^2$, and its value is higher for the MSE-optimized than for the variance-corrected forecasts.
- Discrimination $\mathbb{E}_x\left[\mathbb{E}(f|x) - \mathbb{E}(f)\right]^2$ (higher is better): changes proportionally to the forecast variance, and its value is lower for the MSE-optimized than for the variance-corrected forecasts.

Eq. (5) can also be used to optimize forecasts based on the lowest-MAE directive. In this case, there is no explicit formula for a and b. Still, they could be effectively calculated by any general-purpose optimization routine, e.g., the *scipy.minimize* function in Python or the *optim* function in R. As MAE optimization is unrelated to squared error, there is no theoretical relationship to knowing the variance and other metrics of an MAE-optimized forecast without actually calculating it. However, as long as the

[1] As the resolution term is constant, all changes in the MSE are due to the change of the type-1 conditional bias, so if the MSE is reduced by X, then type-1 conditional bias is also reduced by X.

Table 1
Scale and offset parameters required for the calibration of different variance ratios and the formula for the resulting mean square error (MSE)

	MSE-optimized ($F = \rho^2$)	General (F variance ratio)	Variance-corrected ($F = 1$)
a	$\rho\left(f, x\right)\sqrt{\mathbb{V}\left(x\right)/\mathbb{V}\left(f\right)}$	$\sqrt{F\mathbb{V}\left(x\right)/\mathbb{V}\left(f\right)}$	$\sqrt{\mathbb{V}\left(x\right)/\mathbb{V}\left(f\right)}$
b		$\mathbb{E}(x) - a\mathbb{E}(f)$	
MSE	$\left[1 - \rho\left(f, x\right)^2\right]\mathbb{V}\left(x\right)$	$\left[1 + F - 2\rho\left(f, x\right)\sqrt{F}\right]\mathbb{V}\left(x\right)$	$2\left[1 - \rho\left(f, x\right)\right]\mathbb{V}\left(x\right)$

Table 2
Skill scores of forecasts calibrated for zero bias and a chosen variance ratio, including the optimal underdispersion for the lowest MSE.

Ref.	Reference RMSE	Skill score		
		Min. MSE (best skill)	Equal variances	General formula
Clim.	$\text{RMSE}_c = \sigma\left(x\right)$	$s_c = 1 - \sqrt{1 - \rho\left(f, x\right)^2}$	$s_c = 1 - \sqrt{2\left[1 - \rho\left(f, x\right)\right]}$	$s_c = 1 - \sqrt{1 + F - 2\rho\left(f, x\right)\sqrt{F}}$
Pers.	$\text{RMSE}_p = \sqrt{2\left[1 - \gamma\left(h\right)\right]}\sigma\left(x\right)$	$s_p = 1 - \sqrt{\frac{1-\rho(f,x)^2}{2[1-\gamma(h)]}}$	$s_p = 1 - \sqrt{\frac{1-\rho(f,x)}{1-\gamma(h)}}$	$s_p = 1 - \sqrt{\frac{1+F-2\rho(f,x)\sqrt{F}}{2[1-\gamma(h)]}}$
CLIPER	$\text{RMSE}_{cp} = \sqrt{1 - \gamma\left(h\right)^2}\sigma\left(x\right)$	$s_{cp} = 1 - \sqrt{\frac{1-\rho(f,x)^2}{1-\gamma(h)^2}}$	$s_{cp} = 1 - \sqrt{\frac{2[1-\rho(f,x)]}{1-\gamma(h)^2}}$	$s_{cp} = 1 - \sqrt{\frac{1+F-2\rho(f,x)\sqrt{F}}{1-\gamma(h)^2}}$

Table 3
Performance metrics for raw, MSE-optimized, MAE-optimized, and variance-corrected NAM forecasts for the Goodwin Creek, Mississippi SURFRAD station.

		Raw	MSE opt.	MAE opt.	Var. corr.
Scale parameter	a	1	0.842	0.969	0.974
Offset parameter	b	0	21.4	−6.8	−38.3
Correlation coefficient	ρ	0.865			
Variance ratio	F	105.4%	74.8%	98.8%	100.0%
Mean absolute error	MAE	93.1	96.4	89.7	95.0
Mean bias error	MBE	49.9	0.0	28.9	0.0
Mean square error	MSE	157.4^2	142.1^2	149.5^2	147.1^2
Clim. skill score	s_c	13.6%	22.0%	17.9%	19.2%
Pers. skill score	s_p	25.6%	32.8%	29.3%	30.4%
CLIPER skill score	s_{cp}	8.6%	17.5%	13.2%	14.6%
Observed variance	$\mathbb{V}(x)$	282.9^2	282.9^2	282.9^2	282.9^2
Forecast variance	$\mathbb{V}(f)$	290.4^2	244.6^2	281.3^2	282.9^2
Type-1 cond. bias	$\mathbb{E}_f\ [f - \mathbb{E}(x\|\ f)]^2$	69.6^2	14.1^2	49.2^2	41.4^2
Resolution	$\mathbb{E}_f\ [\mathbb{E}(x\|\ f) - \mathbb{E}(x)]^2$	244.8^2	244.6^2	244.8^2	244.8^2
Type-2 cond. bias	$\mathbb{E}_x\ [x - \mathbb{E}(f\ \|\ x)]^2$	62.4^2	73.4^2	52.7^2	42.9^2
Discrimination	$\mathbb{E}_x\ [\mathbb{E}(f\ \|\ x) - \mathbb{E}(f)]^2$	251.6^2	212.0^2	243.7^2	245.1^2

$p(x|\,f)$ conditional distributions follow the general asymmetric pattern discussed in Section 3.1 and the medians are more towards the extremes than the means, the MAE-optimized forecasts are less under dispersed than the MSE-optimized ones.

4. Application for the NAM irradiance forecast

The theoretical differences of various calibration methods are demonstrated and quantified for the NAM forecasts issued for 2015–2016 at seven SURFRAD stations. Table 3 summarizes all performance indicators commonly used in solar forecast verification for the GWN station, for the raw, MSE-optimized, MAE-optimized, and variance-corrected forecasts. The standard deviations, correlation coefficients, and RMSEs of these four sets of forecasts are shown in a Taylor diagram in Fig. 2. It is noted that although with different numerical values, the same tendencies can be observed at the six other SURFRAD stations. Therefore, the results for the other stations are not presented in the main text of the paper but can be found in supplementary materials.

The MSE-optimized forecasts are much under dispersed, as they only capture <75% of the observed variance. They have the lowest type-1 but the highest type-2 condition bias and the worst discrimination. However, due to their lowest RMSE, they seem to be considerably more skillful than the other forecasts. This is not surprising, because RMSE changes monotonically with MSE, i.e., minimizing MSE is equivalent to minimizing RMSE.

The variance-corrected forecasts lag behind the row-wise best values in the traditional error metrics like MSE and MAE, but they are well balanced in terms of the various aspects of quality under the Murphy–Winkler factorization: the type-1 conditional bias and type-2 conditional bias are almost the same, which is also true for their resolution and discrimination.

The MAE-optimized forecasts are only slightly under dispersed, and as a trade-off, they retain some of the raw forecasts' positive bias. The type-1 conditional bias is high, which indicates that these forecasts are not well

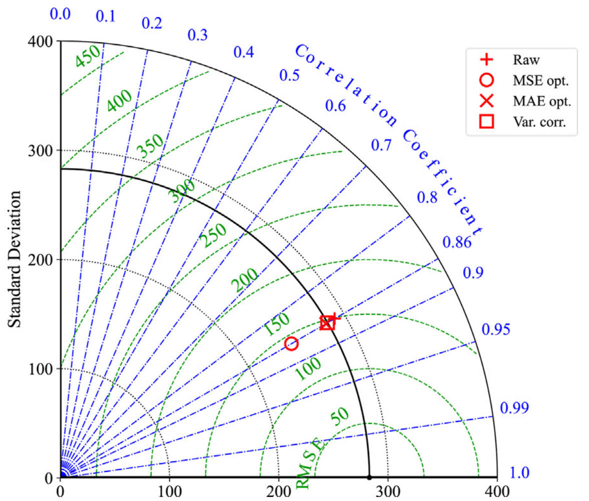

Fig. 2. Taylor-diagram for the raw, MSE-optimized, MAE-optimized, and variance-corrected NAM forecasts for Goodwin Creek, Mississippi. The solid black line indicates the standard deviation of the observations.

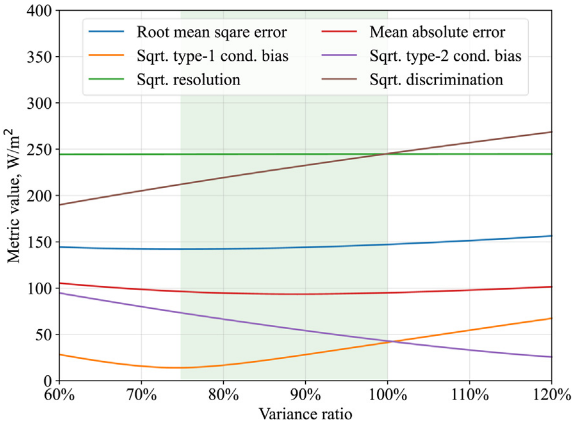

Fig. 3. Performance indicators of the bias-corrected NAM forecasts calibrated for different variance ratios for Goodwin Creek, Mississippi. The green area indicates the $\rho^2 < F < 1$ domain, which is the most important in practical calibrations, and the "sqrt" abbreviation stands for the square root of the given metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

calibrated for the mean. However, as long the MAE-optimized forecasts are deliberately calibrated for the median instead of the mean, it makes no sense to measure their performance using the $\mathbb{E}_f\left[f - \mathbb{E}\left(x|f\right)\right]^2$ term. In the Taylor diagram, the MAE-optimized forecasts are the closest to the variance-corrected forecasts.

Fig. 3 shows the effect of the dispersion of bias-corrected forecasts on six performance indicators. The green zone indicates the $\rho^2 < F < 1$ domain, i.e., a variance ratio ranging from that of the MSE-optimized to that of the variance-corrected forecasts. This is the range in which the variance ratio is recommended to be set during calibration. For more under dispersed forecasts (i.e., $F < \rho^2$), their performance under most metrics starts to deteriorate, hence, it makes no sense to calibrate for any variance ratio lower than the square of the correlation coefficient. Overdispersed forecasts (i.e., $F > 1$) are only recommended if the variance of the MAE-optimized forecasts falls into this domain—the only such example among the NAM forecasts for the seven examined locations is at Sioux Falls (SFX), South Dakota, with an F = 101.5%. An even higher overdispersion may still seem beneficial in terms of type-2 conditional bias and discrimination, but the increase of the absolute and squared errors overrides these apparent benefits. Technically, the discrimination is maximized by forecasts with an infinite variance; thus, it is generally not recommended to calibrate forecasts for the maximum discrimination. In the green zone, different metrics are clearly conflicting. In this regard, there is no theoretically best variance ratio, and the choice depends on the directive under which the forecasts should be issued.

5. Discussion

Many recent recommendations on solar forecast verification encourage the simultaneous use of several complementary performance metrics, as to obtain a complete overview of the different aspects of the forecast quality (Yang et al., 2020). This trend is clearly an improvement over the former practice, which relied only on a few arbitrarily (or sometimes deliberately) selected metrics. Since no point forecast can perform best in terms of all relevant metrics (Gneiting, 2011), the evaluation based on a wide range of performance indicators mostly lead to contradictory conclusions (i.e., a forecast is better in one respect but is worse in another), which may make the overall benefits of a superior forecast opaque. Understanding the relationship and trade-offs between different performance metrics is essential for the correct interpretation of verification results.

The variance of forecasts significantly influences most of the commonly used performance metrics, which has several important implications for forecast verification. As MSE-optimized forecasts are always under dispersed, the common practice of using RMSE skill score for evaluation overrates the forecasts with lower dispersion. Such under dispersed forecasts are less discriminatory, have a higher type-2 conditional bias, and are suboptimal in terms of MAE. Hence, a high forecast skill does not necessarily indicate an overall good forecast performance. This has long been pointed out by Murphy and Winkler (1987), that forecast skill, just like accuracy, is but one aspect of forecast quality.

Forecasts with different variance ratios can be best compared using the correlation coefficient between forecast and observation, which does not depend on the variance and the unconditional bias, i.e., it is not affected by a linear calibration. Nevertheless, the correlation coefficient does not give any information about the quality of forecasts in their raw form, as even forecasts with orders of magnitudes away from the observation can have a high correlation coefficient. Rather, it indicates how the forecasts could perform after proper linear calibration. This

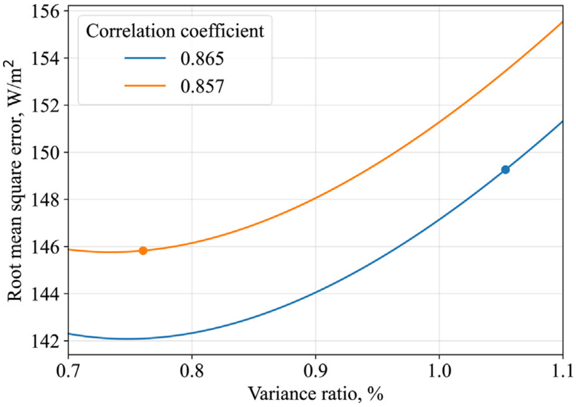

Fig. 4. Root mean square error (RMSE) of two forecasts with different correlation coefficients for the same location. Dots represent the actual variance-ratio and RMSE of the forecasts, while the lines represent all possible trade-offs that can be achieved by a linear calibration.

scale independence is mostly considered a drawback, but it is advantageous when the potential skills of different forecasting techniques are assessed.

The correlation coefficient is the best representation of the *potential* performance of the forecast of interest (Murphy & Epstein, 1989). In most research papers, the main aim is to develop forecasting methods with the biggest potential in practice. Therefore, the correlation coefficient should become a central element of the measure-oriented forecast verification. Fig. 4 demonstrates an example of how an MSE-based comparison could misidentify the overall better forecast. The two dots stand for two sets of forecasts for the same location. Whereas the orange dot has a lower MSE, the blue dot has a higher correlation coefficient. However, the solid lines represent all possible MSE values that can be achieved by a linear transformation of the given set of forecasts. Looking at these lines, it becomes clear that the lowest MSE of the orange forecasts is only due to their lower dispersion, while the blue forecasts can outperform the orange at any variance ratio after a simple linear transformation. Incidentally, the color blue stands for the bias-corrected NAM forecasts for GWN, while the orange is a modified version created by blending the raw forecasts with the 24-h persistence and applying a 3-h moving average. Even though such modification seems beneficial in terms of the RMSE skill score, it actually deteriorates the potential forecast performance.

The presented linear calibration formula is but one possible approach to bias–variance correction; but it can be further improved by, e.g., a regime-based linear calibration (Mejia, Giordano, & Wilcox, 2018). For example, the raw NAM forecasts are too "optimistic," in that they tend to issue exaggerate the amount of clear-sky situations, owing to the imperfect parameterization used in the model (there are more points in the top left quarter of the joint distribution scatterplot of Fig. 1 than in the bottom right quarter). This tendency results in a large positive bias in all seven stations and overdispersion in five out of seven stations. Metric-wise, the forecasts can be corrected via the linear calibration formula; however, a better solution is to deal with the uncertain sky states, specifically, and only correct the forecasts over those periods. Such an improvement would be directly reflected in the correlation coefficient. In general, any nonlinear or multivariate calibration method is beneficial if it improves the correlation coefficient.

Up to this point, the discussion is based on a quadratic error approach, because almost all aspects of quality can be obtained from the MSE factorizations. As an exception, the MAE does not fit into this context, and there is no strict relationship between the MAE and the other metrics. A higher correlation coefficient and lower bias are, in general, beneficial for a low MAE. Still, it does not mean that the forecasts with the highest correlation coefficient and zero bias always have the lowest MAE. An extensive comparison of physical model chains for forecast irradiance to power conversion revealed that the MSE is minimized by simpler model chains, while the MAE is by more complex ones (Mayer & Gróf, 2021). The MSE-optimized power forecasts were shown to be more under dispersed than MAE-optimized ones, which seems to be the general case in solar forecasting, as empirically demonstrated in Section 3.1 of the present paper.

It is now widely accepted that the verification should be based on a directive, which, in most cases, refers to an error metric under which a forecast is penalized. However, in practical applications, it is important to find a directive that corresponds to the highest economic value resulting from the decision-making based on the forecasts; the notion of "value" is another type of goodness of forecast, as argued by Murphy (1993). In the context of grid integration, the total imbalance caused by the inaccurate power forecasts is quasi-linear to the sum of the absolute errors, though with appropriate caveats. If the unit cost of imbalance is constant, the MAE-optimized forecasts can ensure the lowest imbalance costs, as has been identified by Antonanzas, Perpinan-Lamigueiro, Urraca, and Antonanzas-Torres (2020). On the other hand, as a general tendency, large imbalances may cause system stability issues and thus translate to more reserves (or flexible resources), of which the cost does not scale linearly with the amount allocated. In such cases, the MSE, which penalizes large errors heavily, can better indicate the economic benefits of forecasting. On this point, independent system operators (ISOs) around the world have diverse policies regarding how forecasts should be submitted and penalized. For instance, according to the Hungary market policies, individual PV plants that are subsidized by feed-in tariff have to pay penalties for the absolute deviation in their forecasts, but for China Southern Power Grid, the forecasts submitted by individual plants are penalized based on squared deviation (Yang et al., 2021). Indeed, the penalty scheme is often influenced by the market structure and the costs of the reserves.

As a general rule, evaluating a point forecast under a different metric other than the one that it is optimized for can lead to unfair comparisons (Kolassa, 2020). Consequently, if a forecaster uses MAE as the main performance indicator, they should avoid optimizing the forecasts, even indirectly, for any MSE-related terms—this is related to another important concept gauging the goodness of a forecast, namely, consistency (Murphy, 1993). In other words, training a machine-learning model for the minimal MSE, or applying a calibration fitted by the least-squares method, and then evaluating the forecasts based on their MAE is methodologically wrong. Moreover, the type-1 conditional bias term of the Murphy–Winkler factorization only refers to the calibration if it is understood in the context of the MSE. In MAE-optimized forecasts, which are calibrated for the conditional medians instead of the conditional means, this term has no meaningful interpretation.

6. Conclusion

For a long while, solar forecasters were obsessed with introducing fancy forecasting models. This is beneficial to some extent because any emerging field needs to broaden its view. So, the more models the field is aware of, the more rapid its development can be. Notwithstanding, it should be clear, now, that novelty in modeling is not the only thing that matters to solar forecasting research. Without carefully considering the verification and results thereof, anyone can game the system by opting for accuracy measures that favor a particular model, e.g., comparing MAE-optimized forecasts to MSE-optimized forecasts based on MAE.

This paper has gone to great length in demonstrating the profound impacts of different calibration strategies on forecast verification. MSE-optimized, MAE-optimized, and variance-corrected forecasts are compared in terms of the various aspects of forecast quality outlined in the Murphy–Winkler distribution-oriented verification framework. It is found that all three calibration schemes have strengths and weaknesses. Additionally, we took a closer look at the role of the correlation coefficient under different calibration schemes. The correlation coefficient is a good overall measure of the *potential* performance of forecasts; its usage is highly recommended in studies that assess the goodness of forecasts in general instead of focusing on a particular application or context, i.e. when a clear forecast directive is lacking.

Following the findings of the current study, in practical applications and operational forecasting, forecast verification ought to be based on a directive, as also suggested in several previous papers. At present, many grid operators have issued directives under which the submitted forecasts are evaluated and penalized (e.g., see Yang et al., 2021). It is therefore important to identify the metric that is consistent with the directive, and calibrate the forecast for that metric—even a simple linear calibration can result in a noticeable (5%–10%) improvement of the metric of interest—such that the penalty due to forecast inaccuracy can be minimized, and value of the forecast can be maximized.

Declaration of competing interest

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

Acknowledgments

The research reported in this paper and carried out at BME has been supported by the NRDI Fund (TKP2020 NC, Grant No. BME-NCS) based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology.

Appendix A. Verification results for all surfrad stations

Supplementary material related to this article, including the conditional distribution plots, Taylor diagrams and metrics for all seven SURFRAD stations, can be found online at https://doi.org/10.1016/j.ijforecast.2022.03.008.

References

Antonanzas, J., Perpinan-Lamigueiro, O., Urraca, R., & Antonanzas-Torres, F. (2020). Influence of electricity market structures on deterministic solar forecasting verification. *Solar Energy*, (2019), 1–3. http://dx.doi.org/10.1016/j.solener.2020.04.017.

Armstrong, J. S. (2001). Evaluating forecasting methods. *Princ. Forecast.*, 443–472.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. http://dx.doi.org/10.1038/nature14956.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762. http://dx.doi.org/10.1198/jasa.2011.r10138.

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*(3), 896–913. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., & Zareipour, H. (2020). Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy*, *7*, 376–388. http://dx.doi.org/10.1109/OAJPE.2020.3029979.

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*(6), 535–576. http://dx.doi.org/10.1016/j.pecs.2013.06.002.

Jimenez, P. A., Hacker, J. P., Dudhia, J., Haupt, S. E., Ruiz-Arias, J. A., Gueymard, C. A., et al. (2016). WRF-solar: Description and clear-sky assessment of an augmented NWP model for solar power prediction. *Bulletin of the American Meteorological Society*, *97*(7), 1249–1264. http://dx.doi.org/10.1175/BAMS-D-14-00279.1.

Jolliffe, I. T. (2008). The impenetrable hedge: a note on propriety, equitability and consistency. *Meteorological Applications*, *15*(1), 25–29. http://dx.doi.org/10.1002/met.60.

Kazantzidis, A., Tzoumanikas, P., Blanc, P., Massip, P., Wilbert, S., & Ramirez-Santigosa, L. (2017). Short-term forecasting based on all-sky cameras. In G. Kariniotakis (Ed.), *Renewable Energy Forecasting*. Woodhead Publishing.

Kolassa, S. (2020). Why the best point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, *36*(1), 208–211. http://dx.doi.org/10.1016/j.ijforecast.2019.02.017.

Lorenz, E., Hurka, J., Heinemann, D., & Beyer, H. G. (2009). Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*(1), 2–10. http://dx.doi.org/10.1109/JSTARS.2009.2020300.

Makarov, Y. V., Etingov, P. V., Ma, J., Huang, Z., & Subbarao, K. (2011). Incorporating uncertainty of wind power generation forecast into power system operation, dispatch, and unit commitment procedures. *IEEE Transactions on Sustainable Energy*, *2*(4), 433–442. http://dx.doi.org/10.1109/TSTE.2011.2159254.

Makridakis, S., Hyndman, R. J., & Petropoulos, F. (2020). Forecasting in social settings: The state of the art. *International Journal of Forecasting*, *36*(1), 15–28. http://dx.doi.org/10.1016/j.ijforecast.2019.05.011.

Mayer, M. J. (2021). Influence of design data availability on the accuracy of physical photovoltaic power forecasts. *Solar Energy*, *227*, 532–540. http://dx.doi.org/10.1016/j.solener.2021.09.044.

Mayer, M. J., & Gróf, G. (2021). Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, *283*, Article 116239. http://dx.doi.org/10.1016/j.apenergy.2020.116239.

Mejia, J. F., Giordano, M., & Wilcox, E. (2018). Conditional summertime day-ahead solar irradiance forecast. *Solar Energy*, *163*, 610–622. http://dx.doi.org/10.1016/j.solener.2018.01.094.

Miller, S. D., Rogers, M. A., Haynes, J. M., Sengupta, M., & Heidinger, A. K. (2018). Short-term solar irradiance forecasting via satellite/model coupling. *Solar Energy*, *168*, 102–117. http://dx.doi.org/10.1016/j.solener.2017.11.049.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293. http://dx.doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Murphy, A. H. (1995). The coefficients of correlation and determination as measures of performance in forecast verification. *Weather and Forecasting*, *10*(4), 681–688. http://dx.doi.org/10.1175/1520-0434(1995)010<0681:TCOCAD>2.0.CO;2.

Murphy, A. H., & Epstein, E. S. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, *117*(3), 572–582. http://dx.doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), 1330–1338. http://dx.doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Nielsen, A. H., Iosifidis, A., & Karstoft, H. (2021). IrradianceNet: Spatiotemporal deep learning model for satellite-derived solar irradiance short-term forecasting. *Solar Energy*, *228*, 659–669. http://dx.doi.org/10.1016/j.solener.2021.09.073.

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Van Knowe, G., Hemker, K., et al. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, *94*, 305–326. http://dx.doi.org/10.1016/j.solener.2013.05.005.

Sahu, D. K., Yang, H., & Kleissl, J. (2018). Assimilating observations to simulate marine layer stratocumulus for solar forecasting. *Solar Energy*, *162*, 454–471. http://dx.doi.org/10.1016/j.solener.2018.01.006.

Stephan, K., & Martin, R. (2011). Percentage errors can ruin your day (and rolling the dice shows how). *Foresight: The International Journal of Applied Forecasting*, *23*, 21–29.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. http://dx.doi.org/10.1029/2000JD900719.

Vannitsem, S., & Hagedorn, R. (2011). Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression methods. *Meteorological Applications*, *18*(1), 94–104. http://dx.doi.org/10.1002/met.217.

Wilks, D. S., & Murphy, A. H. (1986). A decision-analytic study of the joint value of seasonal precipitation and temperature forecasts in a choice-of-crop problem. *Atmosphere-Ocean*, *24*(4), 353–368. http://dx.doi.org/10.1080/07055900.1986.9649257.

Yang, D. (2019a). Making reference solar forecasts with climatology, persistence, and their optimal convex combination. *Solar Energy*, *193*, 981–985. http://dx.doi.org/10.1016/j.solener.2019.10.006.

Yang, D. (2019b). Post-processing of NWP forecasts using ground or satellite-derived data through kernel conditional density estimation. *Journal of Renewable and Sustainable Energy*, *11*(2), http://dx.doi.org/10.1063/1.5088721.

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H. G., et al. (2020). Verification of deterministic solar forecasts. *Solar Energy*, *210*, 20–37. http://dx.doi.org/10.1016/j.solener.2020.04.019.

Yang, D., & Bright, J. M. (2020). Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. *Solar Energy*, *210*, 3–19. http://dx.doi.org/10.1016/j.solener.2020.04.016.

Yang, D., & Gueymard, C. A. (2021). Probabilistic merging and verification of monthly gridded aerosol products. *Atmospheric Enviroment*, *247*, Article 118146. http://dx.doi.org/10.1016/j.atmosenv.2020.118146.

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. http://dx.doi.org/10.1016/j.solener.2017.11.023.

Yang, D., Li, W., Yagli, G. M., & Srinivasan, D. (2021). Operational solar forecasting for grid integration: Standards, challenges, and outlook. *Solar Energy*, *224*, 930–937. http://dx.doi.org/10.1016/j.solener.2021.04.002.

Yang, D., & Perez, R. (2019). Can we gauge forecasts using satellite-derived solar irradiance? *Journal of Renewable and Sustainable Energy*, *11*(2), Article 023704. http://dx.doi.org/10.1063/1.5087588.

Yang, D., & van der Meer, D. (2021). Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, *140*, Article 110735. http://dx.doi.org/10.1016/j.rser.2021.110735.

Yang, D., Wang, W., Gueymard, C. A., Hong, T., Kleissl, J., Huang, J., et al. (2022). A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews*, *161*, Article 112348. http://dx.doi.org/10.1016/j.rser.2022.112348.

Yang, D., Wang, W., & Xia, X. (2022). A concise overview on solar resource assessment and forecasting. *Advances in Atmospheric Sciences*, http://dx.doi.org/10.1007/s00376-021-1372-8, In press.