Can we gauge forecasts using satellite-derived solar irradiance?

Cite as: J. Renewable Sustainable Energy **11**, 023704 (2019); https://doi.org/10.1063/1.5087588
Submitted: 02 January 2019 . Accepted: 18 February 2019 . Published Online: 04 April 2019

Dazhi Yang, and Richard Perez

Journal of Renewable and Sustainable Energy ARTICLE scitation.org/journal/rse

Can we gauge forecasts using satellite-derived solar irradiance?

Cite as: J. Renewable Sustainable Energy **11**, 023704 (2019); doi: 10.1063/1.5087588
Submitted: 2 January 2019 · Accepted: 18 February 2019 ·
Published Online: 4 April 2019

Dazhi Yang[1,a)] and Richard Perez[2]

AFFILIATIONS

[1]Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore
[2]Atmospheric Sciences Research Center, University at Albany, State University of New York, Albany, New York, USA

Note: This paper is part of the Special Collection on Best Practices in Renewable Energy Resourcing and Integration.
[a)]**Author to whom correspondence should be addressed:** yangdazhi.nus@gmail.com. **Tel.:** +65 9159 0888.

ABSTRACT

Satellite-derived irradiance data, as an alternative to ground-based measurements, offer a unique opportunity to verify gridded solar forecasts generated by a numerical weather prediction model. Previously, it has been shown that the mean square errors (MSE) evaluated against ground-based measurements and satellite-derived solar irradiance are comparable, which might warrant the use of satellite-based products for regional forecast verification. In this paper, the 24-h-ahead hourly forecasts issued by the North American Mesoscale forecast system are verified against both ground-based (Surface Radiation Budget Network, or SURFRAD) and satellite-based (National Solar Radiation Data Base, or NSRDB) measurements, at all 7 SURFRAD stations over 2015–2016. Three different MSE decomposition methods are used to characterize—e.g., through association, calibration, refinement, resolution, or likelihood—how well the two types of measurements can gauge the forecasts. However, despite their comparable MSEs, NSRDB is found suboptimal in its ability to verify forecasts as compared to SURFRAD. Nonetheless, if a new forecasting model produces significantly better forecasts than the benchmarking model, satellite-derived data are able to detect such improvements and make conclusions. This article comes with supplementary material (data and code) for reproducibility.

I. INTRODUCTION

The accuracy of a forecasting model is verified by comparing its forecasts to observations. Observations usually come from ground-based irradiance instruments (pyranometers, pyrheliometers), or sometimes, from the power output of a photovoltaic (PV) system, as they are deemed to have the least uncertainty. Besides ground-based measurements, other sources of irradiance measurements include remote-sensed observations and modeled reanalysis data. It is generally believed that the accuracy of reanalysis data is lower than that of satellite-derived irradiance, and the satellite-derived irradiance is less accurate than ground-based measurements.

That said, a distinct advantage of satellite-derived irradiance and reanalysis data is their extensive geographical coverage. On the one hand, because ground-based radiometry stations are scarce, in general, when a forecasting model generates gridded forecasts, one can only verify the forecasts at the few gauged locations. On the other hand, if those areal forecasts are verified against satellite or reanalysis data, they would provide more information, and thus a better understanding, on the spatial distribution of forecast error. In particular, such spatial distribution of forecast error is useful in correlating the model's behavior with climatic conditions. Thus, in addition to conventional forecast verification, where different forecasting models are being contrasted, whether gridded irradiance products can be used to evaluate forecasts is also a valuable topic worth investigating.

With the advent of remote-sensing technology and reanalysis modeling, there have been several recent works documenting the accuracies of these gridded irradiance products. For instance, Urraca *et al.*[1] compared the ERA5 and COSMO-REA6 reanalyses with two satellite-derived products, namely, the National Solar Radiation Data Base (NSRDB) and SARAH. Despite ERA5 and COSMO-REA6 being comparable to satellite-derived irradiance in terms of bias, the absolute error of reanalyses is still higher due to deficient cloud prediction and over-estimation of aerosols.[1] Hence, to benchmark gridded irradiance forecasts, satellite-derived data are currently more appropriate.

Among various satellite-based irradiance products, SolarAnywhere[2] is claimed to be the highest quality product.[3] The development of SolarAnywhere started since at least 1996, and its present popularity can be partly attributed to the seminal paper by Perez *et al.*[4] Today, SolarAnywhere data has evolved to a stage where it does not require site adaptation.[5] Furthermore, in a recent paper by Perez *et al.*,[6] it

was shown that the forecast root mean square errors (RMSE) evaluated against ground-based measurements and SolarAnywhere irradiance are comparable. To that end, Perez *et al.*[6] stated: "The similarity of these statistics warrants the use of satellite-data for regional validations." Whereas it is encouraging to see such improvement in the accuracy of satellite-derived irradiance, whether the forecast-verification capability can be concluded with a single error metric needs to be further studied.

A. A case study on verifying forecasts using satellite-derived data

In the solar forecasting literature, mean bias error (MBE) and RMSE are two most used error metrics.[7,8] While MBE examines the bias in the forecasts, the information contained in RMSE is more intricate (see below). Since readers often favor the use of a single percentage metric to gauge the forecasts, normalized RMSE is also popular.[9] Despite there being other popular metrics, such as the forecast skill score or mean absolute error, this paper focuses on the discussion related to MBE and RMSE. It is noted that the present goal is not to determine the accuracy of a particular forecasting model, instead, how MBE and RMSE respond to different types of benchmarks.

Following Perez's strategy, two years (2015–2016) of 24-h-ahead hourly global horizontal irradiance (GHI) forecasts generated by the North American Mesoscale (NAM) forecast system[10] are evaluated against measurements from both ground-based station and satellite-derived irradiance databases. The ground-based data come from the Surface Radiation Budget Network (SURFRAD), whereas the satellite-derived GHI is obtained from the National Solar Radiation Data Base (NSRDB) produced via the Physical Solar Model version 3. It is noted that both SURFRAD and NSRDB data have been used extensively in the literature,[11,12] their description and quality-control sequence are not iterated here; the R package "SolarData"[13] contains all necessary code to manipulate these datasets. Table I shows the MBE, RMSE, and nRMSE of NAM forecasts, denoted as f (forecast), verified against SURFRAD and NSRDB, denoted using x_g (ground) and x_s (satellite), respectively; for example, MBE(f, x_g) denotes the MBE between NAM forecast GHI and SURFRAD measured GHI. In addition, the accuracy of NSRDB data are also verified against SURFRAD using the three metrics.

By examining Table I, Perez's observation can be confirmed, namely, RMSE(f, x_g) and RMSE(f, x_s) are mostly comparable, except at the TBL (Table Mountain, Boulder) station, where the satellite-data-gauged RMSE is higher. However, Table I also reveals that MBE(f, x_s) and MBE(x_s, x_g) add up exactly to MBE(f, x_g). This is because

$$\begin{aligned}\mathrm{MBE}(x_s, x_g) + \mathrm{MBE}(f, x_s) &= \mathbb{E}(x_s) - \mathbb{E}(x_g) + \mathbb{E}(f) - \mathbb{E}(x_s)\\ &= \mathbb{E}(f) - \mathbb{E}(x_g)\\ &= \mathrm{MBE}(f, x_g),\end{aligned} \tag{1}$$

where symbol $\mathbb{E}$ denotes the expectation. Such linear propagation of MBE implies that when forecasts are gauged against satellite-derived irradiance, the bias in the satellite-derived irradiance is not being captured by the MBE metric. Hence, if the bias in satellite product is not negligible, or exhibits some level of spatial inhomogeneity across the lattice, it is risky to directly verify forecasts against the satellite-derived irradiance and draw conclusions.

Although MBE (using NSRDB data) is not in favor of using satellite-based irradiance to gauge forecast, one should not draw conclusions immediately. For instance, products such as SolarAnywhere[2] may have a higher accuracy than NSRDB,[14] and they might not have model-led bias in their irradiance estimates. In any case, additional analysis is still necessary to arrive at the conclusion that satellite-derived irradiance can be used to gauge forecasts.

B. Extended approaches for forecast verification

It is well known that no single metric could indicate the quality of a forecasting model. Hence, there are several approaches proposed to resolve this issue. Firstly, the Taylor diagram is an intuitive way of capturing the correlation, RMSE and the ratio of the variances of forecasts and observations.[15] However, it does not describe the bias in the forecasts. Alternatively, a suite of metrics can be used to jointly assess the forecast performance.[16] Notwithstanding, the suite therein proposed contains many metrics that have duplicated efforts in indicating forecast quality. Furthermore, it has been exemplified that two sets of forecasts with the exact same errors[17] can be dramatically different in distribution.[18] In any case, it appears that Zhang *et al.*[16] was essentially trying to describe the joint distribution of observations and forecasts, i.e., by using metrics that are functions of moments of the joint or marginal distribution.

The joint distribution contains all time-independent information relevant to forecast verification.[19] To that end, the utilization of the joint distribution has been a popular basis on which the quality of forecasts is studied, especially in meteorology.[18,20,21] Among the various approaches to characterize the joint distribution, the forecast-verification framework proposed by Murphy and Winkler[22] is perhaps

TABLE I. Forecast errors of 24-h-ahead hourly NAM (f) against ground-based GHI (x_g) and satellite-derived GHI (x_s), at 7 SURFRAD stations over 2015–2016. The unit for MBE and RMSE is W/m^2, whereas the unit for nRMSE is %.

Station	MBE(x_s, x_g)	MBE(f, x_g)	MBE(f, x_s)	RMSE(x_s, x_g)	RMSE(f, x_g)	RMSE(f, x_s)	nRMSE(x_s, x_g)	nRMSE(f, x_g)	nRMSE(f, x_s)
BON	5.96	38.34	32.38	84.48	142.51	145.67	18.18	30.67	31.23
DRA	−4.59	22.65	27.24	77.05	108.22	105.69	13.12	18.43	18.27
FPK	9.70	52.80	43.10	83.23	131.17	136.25	19.24	30.33	30.89
GWN	15.21	49.92	34.71	89.24	157.39	154.66	18.12	31.96	30.89
PSU	8.76	41.67	32.91	100.05	154.99	155.14	22.67	35.11	34.94
SXF	7.99	46.62	38.63	79.41	145.08	148.13	17.70	32.33	32.48
TBL	−21.50	63.37	84.87	121.20	163.67	184.06	24.11	32.56	38.40

most general. More interestingly, the entire framework is described by decomposing the mean square error (MSE), which is just the squared RMSE. Murphy's MSE decomposition is an established approach in meteorology, which can be extended to decomposing the skill score computed with climatological and persistence references.[23] In the remaining part of the paper, Murphy's forecast verification based on the joint distribution of observation and forecast is used to study the composition of the RMSEs seen in Table I. A total of three decomposition methods, namely, the bias–variance decomposition, calibration–refinement factorization, and likelihood–base rate factorization, are considered. The decomposed metrics examine the observation–forecast pairs from different aspects of what constitute good forecasts. In Sec. III, some discussions regarding the implication of the findings of Sec. II are presented. Conclusions follow at the end.

II. FORECAST VERIFICATION AGAINST DIFFERENT REFERENCE DATA

The joint distribution of observation, x, and forecast, f, is denoted as $p(f, x)$. Using the experimental data described earlier, $p(f, x_g)$ and $p(f, x_s)$ for all 7 locations are visualized in Fig. 1 in the form of a two-dimensional density plot. Based on visual inspection, the joint distributions constructed using ground-based data and those constructed using satellite-derived data are highly similar, except for GWN (Goodwin Creek, Mississippi). To characterize these joint distributions, they can be factorized in two ways:[22]

$$p(f,x) = p(x|f)p(f), \tag{2}$$

$$p(f,x) = p(f|x)p(x). \tag{3}$$

In meteorology, the conditional distribution $p(x|f)$ is related to the *calibration* of the forecasts. More specifically, the forecasts are perfectly calibrated if $\mathbb{E}(x|f) = \int xp(x|f)dx = f$. On the other hand, the marginal distribution $p(f)$ describes the "pool" of possible forecast values and their occurring probabilities. In an extreme case, suppose a forecaster uses only one value every time a forecast is needed, it is said that the forecasts are not *refined*.[22] Hence, the factorization that involves $p(x|f)$ and $p(f)$, i.e., Eq. (2), is called the *calibration–refinement* factorization.

The conditional distribution $p(f|x)$ indicates that given a particular weather condition, in this case, the irradiance value, how likely different forecasts are being issued. In other words, $p(f|x)$ is the *likelihood*. The marginal distribution $p(x)$ is irrelevant to forecasts, but describes the characteristics of the nature itself. These characteristics are known as *base rate* or sample climatology.[22] To that end, the factorization that involves $p(f|x)$ and $p(x)$, i.e., Eq. (3), is called the *likelihood–base rate* factorization.

A. Verification through bias-variance decomposition

With the above interpretation of the joint distribution, one can examine the composition of MSE. For instance, the well-known bias–variance decomposition of MSE

$$\begin{aligned}\mathrm{MSE} &= \int\int (f-x)^2 p(f,x)dfdx \\ &= \frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2 \\ &= \mathbb{V}(f-x) + [\mathbb{E}(f) - \mathbb{E}(x)]^2,\end{aligned} \tag{4}$$

can be further decomposed into

$$\mathrm{MSE} = \mathbb{V}(f) + \mathbb{V}(x) - 2\,\mathrm{cov}(f,x) + [\mathbb{E}(f) - \mathbb{E}(x)]^2. \tag{5}$$

It is clear that MSE in fact accounts for the means and variances of the marginal distributions, $p(x)$ and $p(f)$, as well as the covariance of the joint distribution, $p(f, x)$. It is noted that the covariance can be equivalent written as $\sqrt{\mathbb{V}(f)\mathbb{V}(x)}\rho(f,x)$, where $\rho(f, x)$ is the correlation between f and x. Correlation measures the linear relationship between the forecasts and observations, i.e., the *association* of observations with forecasts.[22] Following the decomposition shown in Eq. (5), the value of each term is listed in Table II with respect to both ground-based and satellite-derived references.

It is noticeable that the contribution to MSE from the bias term, $[\mathbb{E}(f) - \mathbb{E}(x)]^2$, is relatively small, as compared to the contributions from the variance terms. Interestingly, the bias between x_s and f

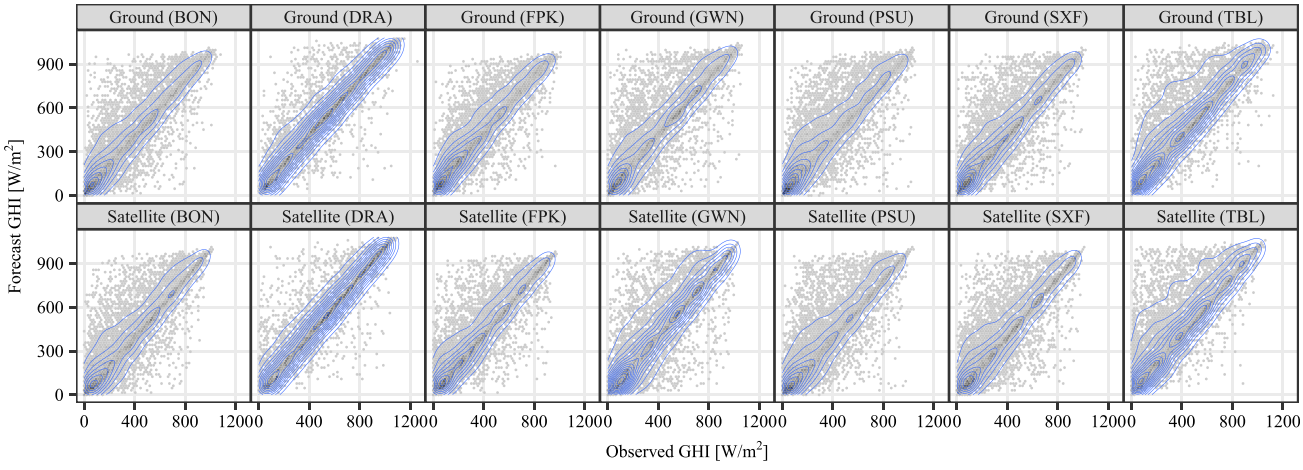

FIG. 1. Two-dimensional density plot for observation–forecast pairs. 24-h-ahead hourly NAM forecasts collocated with the 7 SURFRAD stations are plotted against both ground-based measurements and NSRDB satellite-derived irradiance, over 2015–2016.

TABLE II. Bias–variance decomposition [see Eq. (5)] of 24-h-ahead NAM (f) against ground-based GHI (x_g) and satellite-derived GHI (x_s), at 7 SURFRAD stations over 2015–2016. All metrics have the unit of W^2/m^4, except for correlation ρ, which is dimensionless.

Station	MSE(f, x_g)	MSE(f, x_s)	$\mathbb{V}(x_g)$	$\mathbb{V}(x_s)$	$\mathbb{V}(f)$	$\rho(f, x_g)$	$\rho(f, x_s)$	$[\mathbb{E}(f) - \mathbb{E}(x_g)]^2$	$[\mathbb{E}(f) - \mathbb{E}(x_s)]^2$
BON	20 309.30	21 219.20	74 132.00	71 202.70	76 288.57	0.87	0.86	1469.95	1048.47
DRA	11 712.00	11 170.48	88 784.79	83 187.89	83 377.44	0.94	0.94	513.06	742.13
FPK	17 206.79	18 564.97	65 169.80	65 793.04	68 704.31	0.89	0.88	2788.21	1857.92
GWN	24 770.99	23 919.23	80 042.52	75 729.94	84 326.75	0.86	0.86	2492.34	1204.96
PSU	24 023.14	24 069.73	72 843.19	68 942.66	75 845.66	0.85	0.84	1736.14	1082.75
SXF	21 049.64	21 942.88	68 861.07	69 568.19	72 935.21	0.87	0.86	2173.36	1491.99
TBL	26 787.33	33 876.71	78 478.84	73 060.04	78 287.15	0.85	0.82	4016.02	7203.58

appears to be smaller than that between x_g and f, except for DRA (Desert Rock, Nevada) and TBL. This suggests that the NSRDB data and the NAM forecasts are biased towards the same direction (over-prediction in this case); this is also reflected by the MBE metric in Table I.[24] In terms of correlation, $\rho(f, x_s)$ is lower than or equal to $\rho(f, x_g)$ for all stations, indicating a slightly weaker association between NSRDB GHI and NAM forecasts.

The three variances, namely, $\mathbb{V}(x_g)$, $\mathbb{V}(x_s)$, and $\mathbb{V}(f)$, reveal much information regarding the spread of the data at different locations. Firstly, at BON (Bondville, Illinois), DRA, GWN, PSU (Penn. State Univ., Pennsylvania), and TBL, $\mathbb{V}(x_s)$ is evidently smaller than $\mathbb{V}(x_g)$. This means that the satellite-derived data are under-dispersed, i.e., the variability in GHI is not fully captured by the satellite-to-irradiance model. Furthermore, the variance of the NAM forecasts, $\mathbb{V}(f)$, is higher than $\mathbb{V}(x_g)$ in general, except for DRA, where its cold-desert climate might have led NAM to generate more clear-sky forecasts (thus a lower variance).

Based on the above discussions, although the MSEs of NAM gauged against NSRDB and SURFRAD might be similar, the ground-based observations are preferred. Through the analysis of the decomposed MSE, it is found that: (1) NSRDB contains non-negligible bias, (2) has a weaker association with the NAM forecasts, and (3) is under-dispersed. These findings are usually not observed if only the numerical values of MSE are presented.

B. Verification through calibration–refinement factorization

The bias–variance decomposition demonstrated above has shown additional insights on the interpretation of MSE. Notwithstanding, only the moments and correlation are used to characterize the joint distribution $p(f, x)$. On the other hand, Eqs. (2) and (3) offer the opportunity to use conditional distribution to characterize the joint distribution. In this regard, Murphy and Winkler[22] showed another approach for MSE decomposition, namely, using the calibration–refinement factorization:

$$\mathrm{MSE} = \mathbb{V}(x) + \mathbb{E}_f[f - \mathbb{E}(x|f)]^2 - \mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2, \quad (6)$$

where $\mathbb{E}_f$ denotes the expectation with respect to the marginal distribution $p(f)$, and $\mathbb{E}(x|f)$ is the conditional expectation of x on f.

In Eq. (6), the first term is the variance of the observations, describing the base rate. For the second term, recall that the perfectly calibrated forecasts satisfy $\mathbb{E}(x|f) = f$. Hence, $\mathbb{E}_f[f - \mathbb{E}(x|f)]^2$ is a measure of calibration. Since the goal is to have calibrated forecasts, a small $\mathbb{E}_f[f - \mathbb{E}(x|f)]^2$ is preferred. The third term is known as *resolution*, which accounts for the difference between the conditional and marginal distribution of the observation. Since the sign in front of the third term is negative, it indicates that this term should be maximized.

In order to evaluate the latter two terms in Eq. (6), $\mathbb{E}(x|f)$ needs to be estimated. Since $\mathbb{E}(x|f)$ is the conditional mean, kernel conditional density estimation (KCDE) is employed to estimate $p(x|f)$. Once $p(x|f)$ is obtained, the estimation for $\mathbb{E}(x|f)$ becomes trivial. Given n observation–forecast pairs $(X_1, F_1), \ldots, (X_n, F_n)$, the kernel conditional density estimator of x on f is given by

$$\hat{p}(x|f) = \sum_{i=1}^{n} w_i(f) \mathscr{K}_{h_x}(x - X_i), \quad (7)$$

where

$$w_i(f) = \frac{\mathscr{K}_{h_f}(f - F_i)}{\sum_{i=1}^{n} \mathscr{K}_{h_f}(f - F_i)}, \quad (8)$$

where $\mathscr{K}$ is the scaled kernel, and h_f and h_x are bandwidth parameters controlling the smoothness of the density estimates in the f and x directions, respectively. Since in KCDE, the particular choice of kernel is not crucial,[25] the Gaussian kernel is used. Mathematically

$$\mathscr{K}_{h_x}(x - X_i) = \frac{1}{h_x} \mathscr{K}\left(\frac{x - X_i}{h_x}\right) = \frac{1}{h_x\sqrt{2\pi}} e^{-[(x - X_i)/h_x]^2/2}, \quad (9)$$

and $\mathscr{K}_{h_f}(f)$ is expressed by replacing x, h_x, and X_i in Eq. (9) with f, h_f, and F_i, respectively. With the computed $w_i(f)$, Hyndman *et al.*[26] showed that the conditional mean estimator is

$$\hat{\mathbb{E}}(x|f) = \sum_{i=1}^{n} w_i(f) x_i. \quad (10)$$

Figure 2 plots $\hat{\mathbb{E}}(x_g|f)$ and $\hat{\mathbb{E}}(x_s|f)$ for all 7 stations using SURFRAD, NSRDB, and NAM data described earlier. In this example, the bandwidth parameters are $h_f = h_x = 10\ \mathrm{W/m^2}$, obtained via trial-and-error based on the reconstructed MSE values.[27]

With $\hat{\mathbb{E}}(x_g|f)$ and $\hat{\mathbb{E}}(x_s|f)$, all terms in Eq. (6) can be calculated. The results are shown in Table III. It is found that at 4 out of 7 stations, the ground-based data offer a better calibration as compared to satellite-derived data. However, as compared to the other factors, namely, $\mathbb{V}(x)$ and $\mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2$, the amplitude of $\mathbb{E}_f[f - \mathbb{E}(x|f)]^2$ is small.

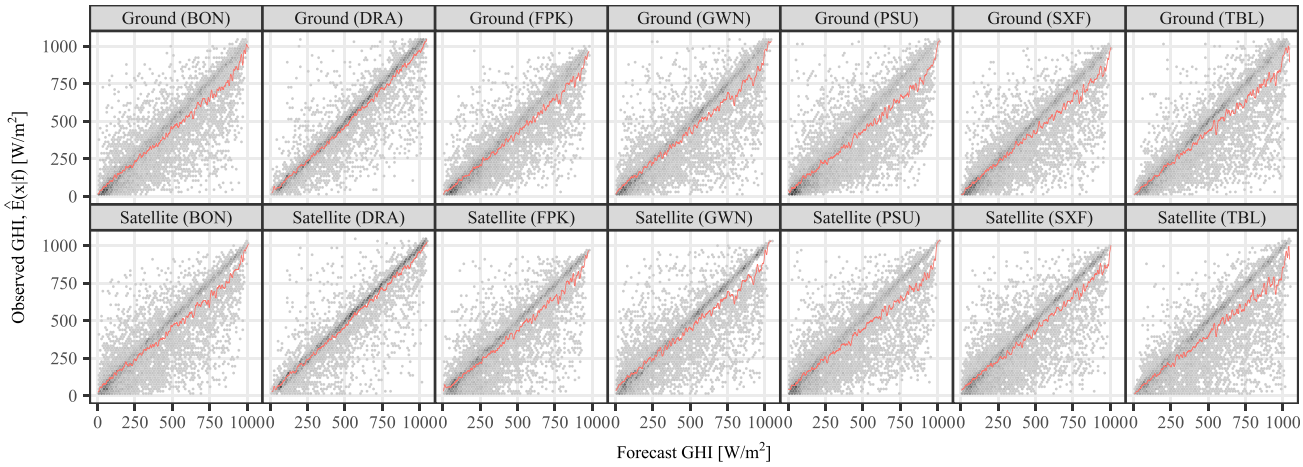

FIG. 2. Scatter plot of observed (SURFRAD and NSRDB) versus forecast (NAM) GHI. The conditional expectation $\mathbb{E}(x|f)$—estimated via kernel conditional density estimation—is overlaid for each case.

Whereas calibration might not be a good indicator—in this particular case—for forecast quality, resolution of the SURFRAD-gauged forecasts is strictly higher than that of the NSRDB-gauged forecasts. Hence, by using calibration–refinement factorization, the deficiency of using satellite-based observations to verify forecast is again shown.

C. Verification through likelihood–base rate factorization

In the last verification exercise, the likelihood–base rate factorization is considered. Following Murphy and Winkler,[22] the MSE is decomposed as

$$\mathrm{MSE} = \mathbb{V}(f) + \mathbb{E}_x[x - \mathbb{E}(f|x)]^2 - \mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2, \quad (11)$$

where $\mathbb{E}_x$ denotes the expectation with respect to the marginal distribution $p(x)$, and $\mathbb{E}(f|x)$ is the conditional expectation of f on x. Similar to Eq. (6), this factorization scheme offers additional insights on forecast verification.

The variance of forecast $\mathbb{V}(f)$ provides a summary of $p(f)$, i.e., the refinement (the metric is irrelevant for this case study, since only NAM is used). The second term in Eq. (11) examines the squared difference between each observation and the average forecast generated conditional on that observation. It can be viewed as the weighted average of the forecast error, thus the smaller the better. Lastly, $\mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2$ measures the difference between the average conditional forecast and the overall average forecast. Since $\mathbb{E}(f)$ is fixed, for different values of x, it is desired to have diversified conditional forecasts. In other words, this term should be maximized, also in accordance with the minus sign in front of it.

The exact procedure for estimating $\mathbb{E}(x|f)$ can be used for the estimation of $\mathbb{E}(f|x)$, with a change of variable, i.e., the conditional density of the forecasts is now being estimated. Figure 3 depicts a visualization of $\mathbb{E}(f|x_g)$ and $\mathbb{E}(f|x_s)$. Subsequently, the metrics for this decomposition are listed in Table IV.

It can be seen that the size of $\mathbb{E}_x[x - \hat{\mathbb{E}}(f|x)]^2$ is the smallest among the three factors under likelihood–base rate factorization. Nevertheless, the NSRDB data seem to produce better, i.e., smaller, values for $\mathbb{E}_x[x - \hat{\mathbb{E}}(f|x)]^2$, except for TBL. This may be partly attributed to the fact that NSRDB and NAM are biased towards the same direction, thus a smaller difference between the conditional and unconditional expectations of the forecasts. In terms of the diversity of the forecasts, $\mathbb{E}_{x_g}[\hat{\mathbb{E}}(f|x_g) - \mathbb{E}(f)]^2$ is marginally bigger than $\mathbb{E}_{x_s}[\hat{\mathbb{E}}(f|x_s) - \mathbb{E}(f)]^2$. This suggests that when the NAM forecasts are gauged using NSRDB, they appear to be less responsive to different

TABLE III. Calibration–refinement factorization [see Eq. (5)] of 24-h-ahead NAM (f) against ground-based GHI (x_g) and satellite-derived GHI (x_s), at 7 SURFRAD stations over 2015–2016. All metrics have the unit of $\mathrm{W}^2/\mathrm{m}^4$.

Station	$\mathrm{MSE}(f, x_g)$	$\mathrm{MSE}(f, x_s)$	$\mathbb{V}(x_g)$	$\mathbb{V}(x_s)$	$\mathbb{E}_f[f - \hat{\mathbb{E}}(x_g\|f)]^2$	$\mathbb{E}_f[f - \hat{\mathbb{E}}(x_s\|f)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x_g\|f) - \mathbb{E}(x_g)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x_s\|f) - \mathbb{E}(x_s)]^2$
BON	20 309.30	21 219.20	74 132.00	71 202.70	3229.08	3480.51	56 945.78	53 404.49
DRA	11 712.00	11 170.48	88 784.79	83 187.89	783.99	1235.24	77 840.66	73 219.72
FPK	17 206.79	18 564.97	65 169.80	65 793.04	4231.70	3525.86	52 160.20	50 695.72
GWN	24 770.99	23 919.23	80 042.52	75 729.94	4987.72	4469.98	60 206.60	56 253.86
PSU	24 023.14	24 069.73	72 843.19	68 942.66	4141.66	4275.36	52 904.06	49 085.43
SXF	21 049.64	21 942.88	68 861.07	69 568.19	4227.51	3727.86	51 997.02	51 336.34
TBL	26 787.33	33 876.71	78 478.84	73 060.04	6061.14	10 940.53	57 690.34	50 059.80

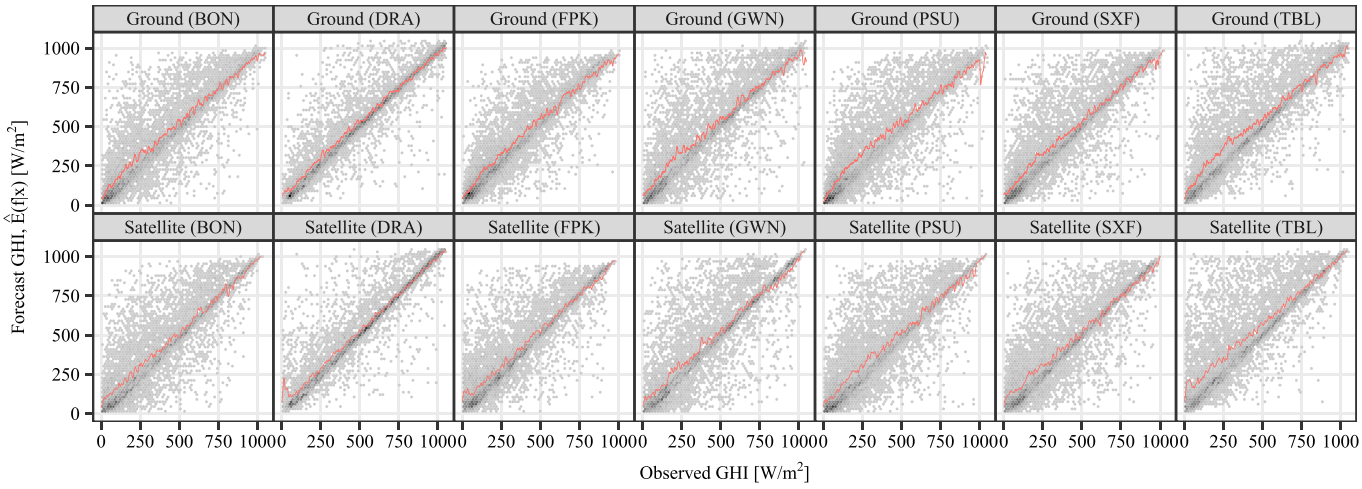

FIG. 3. Scatter plot of forecast (NAM) versus observed (SURFRAD and NSRDB) GHI. The conditional expectation $\mathbb{E}(f|x)$—estimated via kernel conditional density estimation—is overlaid for each case.

irradiance conditions, but in fact the diversity of the forecasts is higher.

III. DISCUSSION

The three decomposition methods shown in the Sec. II use one set of forecasts (NAM) and two sets of measurements (SURFRAD and NSRDB). In this section, some discussions related to the previous observations are presented.

A. Impacts of using satellite-derived data on decision making

When multiple sets of forecasts are present, the decision makers would either choose the best-performing model or use an ensemble to average those forecasts. Since the ensemble approach deals with the uncertainties in the forecasts, not measurements, it is not directly related to the present discussion.

Now consider the case where only satellite-derivate data are available, which could be biased and under-dispersed; this information is unknown to the forecasters. Hence, if the forecasts are biased toward the same direction as the satellite references, the *apparent* error appears to be smaller than the *true* error. Subsequently, this affects the effectiveness of the bias correction put up by the forecaster. More specifically, for positively (negatively) biased satellite reference, which over-predicts (under-predicts) GHI, the bias-corrected forecasts would still be positively (negatively) biased, causing the an under-estimated (over-estimated) conventional generation or over-estimated (under-estimated) curtailment; the decision maker would think there is more (less) energy from solar plants, thus schedule fewer (more) generators for the incoming time period, or curtail more (less) energy from solar. The exact effect will reverse if the forecasts and satellite references are biased towards opposite directions. On the other hand, if the satellite reference is under-dispersed, such as the cases for BON, DRA, GWN, PSU, and TBL, the best-performing model selected according to this criterion will also be under-dispersed. The additional irradiance variability not captured by the forecasting model implies more reserves will be involved during the operation, and thus a higher cost.

B. The potential of using satellite-derived GHI for regional forecast verification

The above discussion assumes that the bias and under-dispersion in satellite estimates are consistent in direction across the entire region. Nevertheless, owing to the well-known geographical smoothing effect,

TABLE IV. Likelihood–base rate factorization [see Eq. (5)] of 24-h-ahead NAM (f) against ground-based GHI (x_g) and satellite-derived GHI (x_s), at 7 SURFRAD stations over 2015–2016. All metrics have the unit of W^2/m^4.

Station	$\mathrm{MSE}(f, x_g)$	$\mathrm{MSE}(f, x_s)$	$\mathbb{V}(f)$	$\mathbb{E}_{x_g}[x_g - \hat{\mathbb{E}}(f\|x_g)]^2$	$\mathbb{E}_{x_s}[x_s - \hat{\mathbb{E}}(f\|x_s)]^2$	$\mathbb{E}_{x_g}[\hat{\mathbb{E}}(f\|x_g) - \mathbb{E}(f)]^2$	$\mathbb{E}_{x_s}[\hat{\mathbb{E}}(f\|x_s) - \mathbb{E}(f)]^2$
BON	20 309.30	21 219.20	76 288.57	2758.82	2064.27	58 713.17	57 108.73
DRA	11 712.00	11 170.48	83 377.44	1507.30	1324.40	73 150.92	73 459.85
FPK	17 206.79	18 564.97	68 704.31	3631.49	2841.84	55 072.45	52 930.24
GWN	24 770.99	23 919.23	84 326.75	4054.71	2219.93	63 595.20	62 631.21
PSU	24 023.14	24 069.73	75 845.66	3741.62	2365.37	55 516.66	54 069.39
SXF	21 049.64	21 942.88	72 935.21	3295.54	2822.93	55 168.92	53 743.09
TBL	26 787.33	33 876.71	78 287.15	6354.39	9085.74	57 785.68	53 467.90

the reality may be different from the hypothetical discussions above. For dispersed PV fleets, as the forecast regional footprint increases, the local irradiance variability decreases significantly. In this case, notwithstanding possible biases, the short-term differences between regional average GHI values from satellite and ground stations are thought to be considerably attenuated. However, as ground-based stations are scarce, it is difficult to validate this hypothesis across different geographical scales and climate types. This could be a valuable subject for future studies, e.g., using data and methods described in Yang.[28]

C. Uncertainties of the ground-based measurements

Up to this stage, the ground measurements have been assumed to have no error. Nevertheless, measurement uncertainty of the ground instruments also plays a part in deciding whether satellite-derived irradiance can be used to gauge forecasts. SURFRAD, as part of the Baseline Surface Radiation Database, is one of the most trusted, if not the most trusted, source of radiation data. As per the current best practice,[29] the expanded uncertainty in the reconstructed GHI (based on separate diffuse and direct irradiance measurements) is around 5.4%.[30] In those cases where suboptimal ground-based measurements are used, the uncertainties in the data are higher. Instrument calibration issues and lack of maintenance can cause the inaccurate data to be left undetected for a long time.[3] If that is the case, satellite data become the only acceptable decision making benchmark at hand.

D. Difference in MSE (caused by choice of reference data) relative to forecast improvements

In most solar forecasting studies, researchers propose new forecasting models that have *significant* improvements over a reference model. Usually, the improvements made by the new models should exceed the measurement uncertainty. That said, it would be useful to analyze the difference in MSE caused by the choices of reference data, with respect to the accuracy gain of the new models.

Without loss of generality, the model output statistics (MOS), as per Lorenz *et al.*,[31] is used to post-process the NAM forecasts. Here, MOS is conducted two times, using the ground-based measurements and satellite-derived GHI, respectively. The corrected NAM forecasts are denoted as $f_{g,mos}$ and $f_{s,mos}$. The MSE of the corrected NAM forecasts are shown in Table V, together with the MSEs of raw NAM forecasts. Two important observations are made: (1) the improvements brought by MOS are much bigger than the differences in MSE caused by the choice of reference data, and (2) the relative performance between MSE(f, x_g) and MSE(f, x_s) does not change. This implies: (1) if the improvement of a new model is substantial enough, satellite-based reference data are able to capture the change in apparent error; and (2) all analyses and discussions regarding the source of reference can still be applied.

IV. CONCLUSION

Two sources of data can be used to gauge solar forecasts, namely, ground-based measurements and satellite-derived irradiance products. It has been observed that the forecast MSEs gauged against these two data sources can be very similar, which might warrant the use of satellite-derived data for forecast verification. This paper employs three different methods to study the composition of MSE. These decomposition methods are useful in detecting the differences in MSE caused by the choice of reference data, which is otherwise unobservable if only a summary statistic is used. More specifically, given the SURFRAD ground-based measurement, x_g, and NSRDB satellite-derived GHI, x_s, the performance of NAM forecast, f, can be summarized as follows:

1. $[\mathbb{E}(f) - \mathbb{E}(x_g)]^2 > [\mathbb{E}(f) - \mathbb{E}(x_s)]^2$ indicates a non-negligible bias in NSRDB data;
2. $\rho(f, x_g) > \rho(f, x_s)$ suggests that the linear association between NSRDB and NAM is weaker;
3. $\mathbb{V}(x_g) > \mathbb{V}(x_s)$ indicates that the NSRDB is under-dispersed, and could not fully capture the variability in GHI;
4. $\mathbb{E}_f[f - \hat{\mathbb{E}}(x_g|f)]^2 < \mathbb{E}_f[f - \hat{\mathbb{E}}(x_s|f)]^2$ implies that the NAM forecasts may appear to be less calibrated when gauged using NSRDB reference; and
5. $\mathbb{E}_f[\hat{\mathbb{E}}(x_g|f) - \mathbb{E}(x_g)]^2 > \mathbb{E}_f[\hat{\mathbb{E}}(x_s|f) - \mathbb{E}(x_s)]^2$ reveals that one tends to under-estimate the resolution of NAM forecasts with NSRDB references.

Based on the above-mentioned differences, it can be concluded that NSRDB is not optimal to gauge forecast, if the "true" error of a forecasting model is of interest. However, if one is only interested in the "apparent" errors among different forecasters, satellite-based observations can be used to gauge those forecasts, provided that the improvements made by the new forecaster are substantial enough. This is because the difference between a good and a bad model is likely to be larger than the difference caused by the choice of reference data.

Lastly, the present study is conducted using NSRDB data. There are, however, other satellite-based products that may have a higher accuracy than NSRDB, for instance, SolarAnywhere data. Whether these high-quality satellite data can lead to a different conclusion is unknown. In any case, the framework herein used can be considered for future studies of this kind.

TABLE V. MSEs of the MOS-corrected NAM forecasts. The corrections done using SURFRAD and NSRDB data are denoted as $f_{g,mos}$ and $f_{s,mos}$, respectively.

Station	MSE(f, x_g)	MSE(f, x_s)	MSE($f_{g,mos}$, x_g)	MSE($f_{s,mos}$, x_s)	MSE(f, x_g) − MSE($f_{g,mos}$, x_g)	MSE(f, x_s) − MSE($f_{s,mos}$, x_s)
BON	20 309.30	21 219.20	16 479.78	16 983.34	3829.53	4235.86
DRA	11 712.00	11 170.48	10 541.87	9495.65	1170.13	1674.83
FPK	17 206.79	18 564.97	12 810.60	14 708.52	4396.19	3856.45
GWN	24 770.99	23 919.23	18 956.08	18 607.38	5814.92	5311.85
PSU	24 023.14	24 069.73	19 257.11	19 526.64	4766.03	4543.09
SXF	21 049.64	21 942.88	16 500.71	17 265.39	4548.93	4677.49
TBL	26 787.33	33 876.71	20 394.83	23 190.32	6392.50	10 686.39

SUPPLEMENTARY MATERIAL

See supplementary material for the instructions for reproducing all results reported in the article, i.e., all tables and figures. The code is written in R, and several packages (and, of course, their dependencies, see http://stat.ethz.ch/R-manual/R-patched/library/tools/html/package.dependencies.html) need to be installed before the scripts can be executed.

- Package dplyr (https://cran.r-project.org/web/packages/dplyr/index.html) is a fast, consistent tool for working with data frame like objects, both in memory and out of memory.
- Package lubridate (https://cran.r-project.org/web/packages/lubridate/index.html) makes it easier to work with dates and times.
- Package ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html) offers a powerful graphics language for creating elegant and complex plots.
- Package xtable (https://cran.r-project.org/web/packages/xtable/index.html) exports tables to LATEX.

Data: The folder data contains the arranged data files for 7 SURFRAD stations over a two-year period, 2015–2016. The file names start with the stations' abbreviations. In each file, there are 7 columns, namely, date and time, zenith angle, GHI, persistence forecast, NSRDB GHI, and the raw NAM GHI.

Code: A total of 5 R scripts, Tables I.R–V.R, reproduce the results shown in Tables I–V, respectively. In addition, the code snippets for Figs. 1–3 are attached at the end of Tables I.R, III.R, and IV.R, respectively. To execute these scripts, the user only needs to change the working directory.

ACKNOWLEDGMENTS

D.Y. would like to thank Elynn Wu and Jan Kleissl from UCSD for providing the NAM forecasts used in this work. R.P. would like to acknowledge California Energy Commission EPC-17-003 Forecast Accuracy Improvement Project for insights relevant to this paper's discussion.

REFERENCES

[1]R. Urraca, T. Huld, A. Gracia-Amillo, F. J. M. de Pison, F. Kaspar, and A. Sanz-Garcia, "Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data," Sol. Energy **164**, 339–354 (2018).

[2]https://www.solaranywhere.com/

[3]R. Perez, J. Schlemmer, A. Kankiewicz, J. Dise, A. Tadese, and T. Hoff, "Detecting calibration drift at ground truth stations: A demonstration of satellite irradiance models' accuracy," in *Proceedings of the 2017 IEEE 44th Photovoltaic Specialist Conference* (2017), pp. 1104–1109.

[4]R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, and F. Vignola, "A new operational model for satellite derived irradiances: Description and validation," Sol. Energy **73**, 307–317 (2002).

[5]M. André, R. Perez, T. Soubdhan, J. Schlemmer, R. Calif, and S. Monjoly, "Preliminary assessment of two spatio-temporal forecasting technics for hourly satellite-derived irradiance in a complex meteorological context," Sol. Energy **177**, 703–712 (2019).

[6]R. Perez, J. Schlemmer, K. Hemker, S. Kivalov, A. Kankiewicz, and J. Dise, "Solar energy forecast validation for extended areas and economic impact of forecast accuracy," in *Proceedings of the 2016 IEEE 43rd Photovoltaic Specialists Conference* (2016), pp. 1119–1124.

[7]D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining," Sol. Energy **168**, 60–101 (2018).

[8]MBE $= \frac{1}{N}\sum_{n=1}^{N}(\text{forecast}_n - \text{reference}_n)$, RMSE $= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\text{forecast}_n - \text{reference}_n)^2}$, where N is the number of forecasts being verified.

[9]nRMSE $= \sqrt{\sum_{n=1}^{N}(\text{forecast}_n - \text{reference}_n)^2}/\sqrt{\sum_{n=1}^{N}\text{reference}_n^2}$.

[10]The NAM forecast is run four times a day at 0:00, 6:00, 12:00, and 18:00 UTC. The forecasts issued by the 0:00 run are used herein. In addition, all validations are performed for daylight hours only.

[11]D. Yang, "A correct validation of the National Solar Radiation Data Base (NSRDB)," Renewable Sustainable Energy Rev. **97**, 152–155 (2018).

[12]D. Yang, "Kriging for NSRDB PSM version 3 satellite-derived solar irradiance," Sol. Energy **171**, 876–883 (2018).

[13]D. Yang, "SolarData: An R package for easy access of publicly available solar datasets," Sol. Energy **171**, A3–A12 (2018).

[14]SolarAnywhere is a paid product. Its adoption in research might be limited as compared to other publicly available alternatives, such as NSRDB. Currently, site adaptation is still considered to be an important preprocessing step when dealing with satellite-based irradiance data; J. Polo, S. Wilbert, J. Ruiz-Arias, R. Meyer, C. Gueymard, M. Súri, L. Martín, T. Mieslinger, P. Blanc, I. Grant, J. Boland, P. Ineichen, J. Remund, R. Escobar, A. Troccoli, M. Sengupta, K. Nielsen, D. Renne, N. Geuder, and T. Cebecauer, "Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets," Sol. Energy **132**, 25–37 (2016).

[15]K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," J. Geophys. Res.: Atmos. **106**, 7183–7192, https://doi.org/10.1029/2000JD900719 (2001).

[16]J. Zhang, A. Florita, B.-M. Hodge, S. Lu, H. F. Hamann, V. Banunarayanan, and A. M. Brockway, "A suite of metrics for assessing the performance of solar power forecasting," Sol. Energy **111**, 157–175 (2015).

[17]A set of 5 error metrics were considered, namely, bias, MSE, correlation coefficient, coefficient of determination, and MSE skill score.

[18]Y. Tian, G. S. Nearing, C. D. Peters-Lidard, K. W. Harrison, and L. Tang, "Performance metrics, error modeling, and uncertainty quantification," Mon. Weather Rev. **144**, 607–613 (2016).

[19]A. H. Murphy, B. G. Brown, and Y.-S. Chen, "Diagnostic verification of temperature forecasts," Weather Forecasting **4**, 485–501 (1989).

[20]H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," J. Hydrol. **377**, 80–91 (2009).

[21]T. R. Stewart, "A decomposition of the correlation coefficient and its use in analyzing forecasting skill," Weather Forecasting **5**, 661–666 (1990).

[22]A. H. Murphy and R. L. Winkler, "A general framework for forecast verification," Mon. Weather Rev. **115**, 1330–1338 (1987).

[23]A. H. Murphy, "General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality," Mon. Weather Rev. **124**, 2353– 2369 (1996).

[24]By using the bias–variance decomposition, MBE becomes redundant.

[25]L. Wasserman, *All of Nonparametric Statistics* (Springer Science and Business Media, New York, 2006).

[26]R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald, "Estimating and visualizing conditional densities," J. Comput. Graph. Stat. **5**, 315–336 (1996).

[27]The automatic bandwidth-selection algorithm [D. Ruppert, S. J. Sheather, and M. P. Wand, "An effective bandwidth selector for local least squares regression," J. Am. Stat. Assoc. **90**, 1257–1270 (1995)] produced an overly smoothed fit, which makes the reconstructed MSE bigger than it should be.

[28]D. Yang, "Ultra-fast preselection in lasso-type spatio-temporal solar forecasting problems," Sol. Energy **176**, 788–796 (2018).

[29]C. A. Gueymard and D. R. Myers, "Evaluation of conventional and high-performance routine solar radiation measurements for improved solar resource, climatological trends, and radiative modeling," Sol. Energy **83**, 171–185 (2009).

[30]C. A. Gueymard, "Direct and indirect uncertainties in the prediction of tilted irradiance for solar engineering applications," Sol. Energy **83**, 432–444 (2009).

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