Available online at www.sciencedirect.com

ScienceDirect

Solar Energy 111 (2015) 157–175

SOLAR ENERGY

www.elsevier.com/locate/solener

A suite of metrics for assessing the performance of solar power forecasting

Jie Zhang [a], Anthony Florita [a], Bri-Mathias Hodge [a,*], Siyuan Lu [b], Hendrik F. Hamann [b], Venkat Banunarayanan [c], Anna M. Brockway [c]

[a] National Renewable Energy Laboratory, Golden, CO 80401, USA
[b] IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA
[c] U.S. Department of Energy, Washington, D.C. 20585, USA

Received 8 August 2014; received in revised form 2 October 2014; accepted 12 October 2014

Communicated by: Associate Editor Jan Kleissl

Abstract

Forecasting solar energy generation is a challenging task because of the variety of solar power systems and weather regimes encountered. Inaccurate forecasts can result in substantial economic losses and power system reliability issues. One of the key challenges is the unavailability of a consistent and robust set of metrics to measure the accuracy of a solar forecast. This paper presents a suite of generally applicable and value-based metrics for solar forecasting for a comprehensive set of scenarios (i.e., different time horizons, geographic locations, and applications) that were developed as part of the U.S. Department of Energy SunShot Initiative's efforts to improve the accuracy of solar forecasting. In addition, a comprehensive framework is developed to analyze the sensitivity of the proposed metrics to three types of solar forecasting improvements using a design-of-experiments methodology in conjunction with response surface, sensitivity analysis, and nonparametric statistical testing methods. The three types of forecasting improvements are (i) uniform forecasting improvements when there is not a ramp, (ii) ramp forecasting magnitude improvements, and (iii) ramp forecasting threshold changes. Day-ahead and 1-hour-ahead forecasts for both simulated and actual solar power plants are analyzed. The results show that the proposed metrics can efficiently evaluate the quality of solar forecasts and assess the economic and reliability impacts of improved solar forecasting. Sensitivity analysis results show that (i) all proposed metrics are suitable to show the changes in the accuracy of solar forecasts with uniform forecasting improvements, and (ii) the metrics of skewness, kurtosis, and Rényi entropy are specifically suitable to show the changes in the accuracy of solar forecasts with ramp forecasting improvements and a ramp forecasting threshold.
Published by Elsevier Ltd.

Keywords: Grid integration; Nonparametric statistical testing; Solar power forecasting; Solar power ramps; Sensitivity analysis

1. Introduction

Solar power penetration in the United States is growing rapidly, and the SunShot Vision Study reported that solar power could provide as much as 14% of U.S. electricity demand by 2030 and 27% by 2050 (Margolis et al., 2012). At these high levels of solar energy penetration, solar power forecasting will become very important for electricity system operations. Solar forecasting is a challenging task, and solar power generation presents different challenges for transmission and distribution networks. On the transmission side, solar power takes the form of centralized solar power plants, a non-dispatchable component of the

* Corresponding author. Tel.: +1 303 384 6981.
E-mail address: bri.mathias.hodge@nrel.gov (B.-M. Hodge).

http://dx.doi.org/10.1016/j.solener.2014.10.016
0038-092X/Published by Elsevier Ltd.

generation pool. On the distribution side, solar power is generated by a large number of distributed arrays installed on building rooftops and other sites. These arrays can alter traditional load patterns by offsetting electricity use behind the meter. Integrating large amounts of solar power into the grid can magnify the impact of steep ramps in solar power output, which poses challenges to system operators' ability to account for solar variability. Forecast inaccuracies of solar power generation can result in substantial economic losses and power system reliability issues because electric grid operators must continuously balance supply and demand.

1.1. Overview of solar forecasting

Solar power output is directly proportional to the magnitude of solar irradiance incident on the panels. To integrate high penetrations of solar energy generation, accurate solar forecasting is required in multiple spatial and temporal scales. Solar irradiance variations are caused primarily by cloud movement, cloud formation, and cloud dissipation. In the literature, researchers have developed a variety of methods for solar power forecasting, such as statistical approaches using historical data (Hammer et al., 1999; Sfetsos and Coonick, 2000; Paoli et al., 2010), the use of numerical weather prediction (NWP) models (Marquez and Coimbra, 2011; Mathiesen and Kleissl, 2011; Chen et al., 2011), tracking cloud movements from satellite images (Perez et al., 2007), and tracking cloud movements from direct ground observations using sky cameras (Perez et al., 2007; Chow et al., 2011; Marquez and Coimbra, 2013a). NWP models are the most popular method for forecasting solar irradiance several hours or days in advance. Mathiesen and Kleissl (2011) analyzed the global horizontal irradiance in the continental United States forecasted by three popular NWP models: the North American Model, the Global Forecast System, and the European Centre for Medium-Range Weather Forecasts. Chen et al. (2011) developed a statistical method for solar power forecasting based on artificial intelligence techniques. Crispim et al. (2008) used total sky imagers (TSI) to extract cloud features using a radial basis function neural network model for time horizons from 1 min to 60 min. Chow et al. (2011) also used TSI to forecast short-term global horizontal irradiance. The results suggested that TSI is useful for forecasting time horizons up to 15 min to 25 min-ahead. Marquez and Coimbra (2013a) presented a method using TSI images to forecast 1-min averaged direct normal irradiance at the ground level for time horizons between 3 min and 15 min. Lorenz et al. (2007) showed that cloud movement-based forecasts likely provide better results than NWP forecasts for forecast timescales of 3 h to 4 h or less; beyond that, NWP models tend to perform better. In summary, forecasting methods can be broadly characterized as physical or statistical. The physical approach uses NWP and PV models to generate solar power forecasts; whereas the statistical approach relies primarily on historical data to train models (Pelland et al., 2013). Recent solar forecasting studies (Chu et al., 2014; Quesada-Ruiz et al., 2014) integrated these two approaches by using both physical and historical data as inputs to train statistical models. A brief description of these solar forecasting methods is summarized in Table 1.

As solar penetration increases, considerable research is underway to improve the accuracy of solar forecasting models. In the United States, the Department of Energy's SunShot Initiative has created the solar forecasting accuracy improvement program to significantly improve the state of the art in solar forecasting.

1.2. Research motivation and objectives

A key gap in developing solar forecasting models is the unavailability of a consistent and robust set of metrics to measure and assess the improvement in forecasting accuracy, because different researchers use improvements described by different metrics as their own evaluation criteria. In addition, it is not clear that the traditional statistical metrics used to evaluate forecasts best represent the needs of power system operators. Because weather patterns and locational atmospheric conditions vary considerably both spatially and temporally, solar forecasting accuracy is dependent on geographic location and timescale of the data. Conventional measures of solar forecasting accuracy include root mean square error (RMSE), mean bias error (MBE), and mean absolute error (MAE). Marquez and Coimbra (2013b) proposed a metric for using the ratio of solar uncertainty to solar variability to compare different solar forecasting models. Espinar et al. (2009) proposed several metrics based on the Kolmogorov–Smirnov test

Table 1
Solar forecasting methodologies (Pelland et al., 2013).

	Methods	Description/comment	Forecast horizons
Physical approach	NWP models	NWP models are the most popular method for forecasting solar irradiance more than 6 h or days in advance	6 h to days ahead
	Total sky imagers (TSI)	TSIs are used to extract cloud features or to forecast short-term global horizontal irradiance	0–30 min ahead
Statistical approach	Statistical methods	Statistical methods were developed based on autoregressive or artificial intelligence techniques for short-term forecasts	0–6 h ahead
	Persistence forecasts	Persistence of cloudiness performs well for very-short-term forecasts	0–4 h ahead

integral (KSI) to quantify the differences between the cumulative distribution functions (CDFs) of actual and forecast solar irradiation data. However, many of the developed forecast metrics do not take into account the types of errors that have the most impact on power system operations. Extreme forecasting errors can have disproportionate economic and reliability impacts on operations; therefore, a set of metrics that emphasizes these errors is needed to capture the true impact of the forecasts on power system operations.

The objective of this paper is to develop a suite of generally applicable, value-based, and custom-designed metrics for solar forecasting for a comprehensive set of scenarios (different time horizons, geographic locations, and applications) that can assess the economic impacts of improved solar forecasting. The sensitivity of the proposed metrics to improved solar forecasts is also analyzed. Section 2 presents the developed metrics for different types of forecasts and applications. Section 3 summarizes the solar power data used in the paper. The methodologies for sensitivity analysis and nonparametric statistical testing of different metrics are developed in Sections 4 and 5, respectively. The results and discussion of the case study are presented in Section 6. Concluding remarks and ideas on areas for future exploration are given in the final section.

2. Metrics development

One of the objectives of the SunShot Initiative's solar forecasting accuracy improvement program is to establish a standard set of metrics for assessing solar forecast accuracy with stakeholder guidance. The metrics proposed in this paper are intended to be responsive to feedback gathered from stakeholders during three workshops (U.S. Department of Energy SunShot Initiative, 2013a,b, 2014).

Two key factors that impact the accuracy of solar forecasting are geographic locations and forecast timescales. Therefore, in this paper, solar power plants at multiple geographic regions were analyzed at multiple timescales to quantify the effects of geographic location and forecast horizon on the forecasting accuracy. The proposed solar forecasting metrics in this paper can be broadly divided into four categories: (i) statistical metrics for different time and geographic scales, (ii) uncertainty quantification and propagation metrics, (iii) ramp characterization metrics, and (iv) economic metrics. A brief description of the metrics in each of the four categories is given in Table 2, and these metrics are described more fully in Sections 2.1–2.4. A detailed formulation of each statistical metric can be found in Appendix A.

2.1. Statistical metrics

Distributions of forecast errors at multiple temporal and spatial scales were analyzed to investigate the variability of solar forecasts. The distribution of forecast errors is a graphical representation of the raw forecasting error data, which provides a good overview of the performance of the forecasts for longer time periods. In addition, interval forecasts of solar power can help determine the reserve requirements needed to compensate for forecast errors, which is an important consideration in the commitment and dispatching of generating units. Multiple distribution types have been analyzed in the literature to quantify the distribution of solar (or wind) power forecast errors, including the hyperbolic distribution, kernel density estimation (KDE), the normal distribution, and Weibull and beta distributions. In this paper, the distribution of solar power forecast errors is estimated using the KDE method, which has been widely used in the renewable energy community (Zhang et al., 2013a,e; Juban et al., 2007).

In conjunction with the distribution of forecast errors, statistical moments (mean, variance, skewness, and kurtosis) can provide additional information to evaluate forecasts. Assuming that forecast errors are equal to forecast power minus actual power, a positive skewness of the forecast errors leads to an over-forecasting tail, and a negative skewness leads to an under-forecasting tail. A distribution with a large kurtosis value indicates a peaked (narrow) distribution; whereas a small kurtosis indicates a flat (wide) data distribution.

The KSI and OVER metrics were proposed by Espinar et al. (2009). The KSI test is a nonparametric test to determine if two data sets are significantly different. The KSI parameter is defined as the integrated difference between the two CDFs. Instead of comparing forecast errors directly, the KSI metric evaluates the similarities between the forecasts and the actual values. In addition, the KSI metric contains information about the distribution of the forecast and actual data sets, which are not captured by metrics such as RMSE, MAE, MaxAE, and MBE. A smaller value of KSI shows that the forecasts and actual values behave statistically similarly, which thereby indicates a better performance of the solar power forecast. A zero KSI index means that the CDFs of two sets are equal. The OVER metric characterizes the integrated differences between the CDFs of the actual and forecast solar power. In contrast to the KSI metric, the OVER metric evaluates only large forecast errors beyond a specified value, because large forecast errors are more important for power system reliability. KSIPer and OVERPer are used to represent the KSI and OVER in the form of percentages, respectively.

2.2. Metrics for uncertainty quantification and propagation

Two metrics are proposed to quantify the uncertainty in solar forecasting: (i) the standard deviation of solar power forecast errors and (ii) the Rényi entropy of solar power forecast errors. Forecasting metrics such as RMSE and MAE are unbiased only if the error distribution is Gaussian; therefore, new metrics are proposed based on the use of concepts from information theory, which can

Table 2
Proposed metrics for solar power forecasting.

Type	Metric	Description/comment
Statistical metrics (See Appendix A for detailed information)	Distribution of forecast errors	Provides a visualization of the full range of forecast errors and variability of solar forecasts at multiple temporal and spatial scales
	Pearson's correlation coefficient	Linear correlation between forecasted and actual solar power
	Root mean square error (RMSE) and normalized root mean square error (NRMSE)	Suitable for evaluating the overall accuracy of the forecasts while penalizing large forecast errors in a square order
	Root mean quartic error (RMQE) and normalized root mean quartic error (NRMQE)	Suitable for evaluating the overall accuracy of the forecasts while penalizing large forecast errors in a quartic order
	Maximum absolute error (MaxAE)	Suitable for evaluating the largest forecast error
	Mean absolute error (MAE) and mean absolute percentage error (MAPE)	Suitable for evaluating uniform forecast errors
	Mean bias error (MBE)	Suitable for assessing forecast bias
	Kolmogorov–Smirnov test integral (KSI) or KSIPer	Evaluates the statistical similarity between the forecasted and actual solar power
	OVER or OVERPer	Characterizes the statistical similarity between the forecasted and actual solar power on large forecast errors
	Skewness	Measures the asymmetry of the distribution of forecast errors; a positive (or negative) skewness leads to an over-forecasting (or under-forecasting) tail
	Excess kurtosis	Measures the magnitude of the peak of the distribution of forecast errors; a positive (or negative) kurtosis value indicates a peaked (or flat) distribution, greater or less than that of the normal distribution
Uncertainty quantification metrics	Rényi entropy	Quantifies the uncertainty of a forecast; it can utilize all of the information present in the forecast error distributions
	Standard deviation	Quantifies the uncertainty of a forecast
Ramp characterization metrics	Swinging door algorithm	Extracts ramps in solar power output by identifying the start and end points of each ramp
Economic metrics	95th percentile of forecast errors	Represents the amount of non-spinning reserves service held to compensate for solar power forecast errors

utilize all of the information present in the forecast error distributions. An information entropy approach was proposed in the literature (Hodge et al., 2012; Bessa et al., 2011) for assessing wind forecasting methods. This information entropy approach based on Rényi entropy is adopted here to quantify the uncertainty in solar forecasting. The Rényi entropy is defined as:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log_2\sum_{i=1}^{n} p_i^\alpha \quad (1)$$

where α (where $\alpha > 0$ and $\alpha \neq 1$) is the order of the Rényi entropy, which allows the creation of a spectrum of Rényi entropies; and p_i is the probability density of the ith discrete section of the distribution. Large values of α favor higher probability events; whereas smaller values of α weight all of the instances more evenly (Hodge et al., 2012). Generally, a larger value of Rényi entropy indicates a higher uncertainty in the forecasting.

2.3. Metrics for ramps characterization: swinging door algorithm

One of the biggest concerns associated with integrating a large amount of solar power into the grid is the ability to handle large ramps in solar power output, which are often caused by cloud events and extreme weather events (Mills and Wiser, 2010). Different time and geographic scales influence the severity of up- or down-ramps in solar power output. Forecasting solar power can help reduce the uncertainty involved with the power supply. In this paper, the swinging door algorithm is used to identify ramps over varying time frames because of its flexibility and simplicity (Florita et al., 2013).

The swinging door algorithm extracts ramp periods in a series of power signals by identifying the start and end points of each ramp. The user sets a threshold parameter that influences the algorithm's sensitivity to ramp variations. This threshold parameter, the only tunable parameter in the algorithm, is the width of a "door," represented by ε in Fig. 1. The parameter ε directly characterizes the threshold sensitivity to noise and/or insignificant fluctuations to be specified. With a smaller ε value, many small ramps will be identified; with a larger ε value, only a few large ramps will be identified. It is important to note that the scale in Fig. 1 is arbitrary for the purpose of explanation, and in general the signal magnitude is much larger than the scale of the threshold bounds. A detailed description of the swinging door algorithm can be found in (Bristol, 1990; Florita et al., 2013).

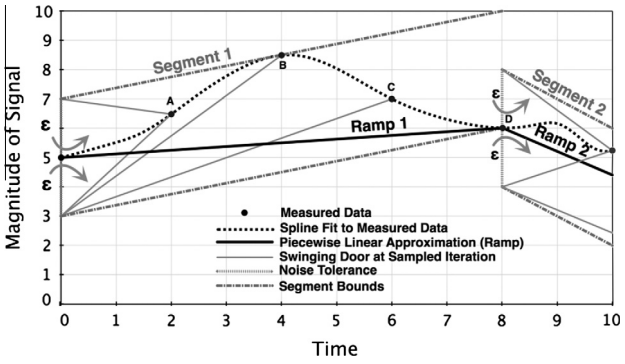


Fig. 1. The swinging door algorithm for the extraction of ramps in power from the time series (Florita et al., 2013).

2.4. Economic metrics

Power system operators typically rely on reserves to manage the anticipated and unanticipated variability in generation and load. These reserves are usually referred to as "operating reserves" and are used to manage variability in the timescale of minutes to multiple hours, which is also the time frame of solar variability. High solar penetration can necessitate additional operating reserves that need to be procured to manage the inherent variability of solar generation. Improving solar forecasting accuracy is expected to decrease the amount of these additional operating reserves: the greater the predictability and hence the certainty of power output from solar, the less variability from solar that needs to be managed with additional operating reserves. Therefore, reduction in the cost of additional operating reserves that need to be procured for managing solar variability is a good metric to assess the economic impact of accuracy improvements in solar forecasting. Using the 95th percentile of forecast errors is a generally accepted method in the power industry for load and other variability forecasts to determine the amount of operating reserves needed; therefore, this paper uses the 95th percentile of solar power forecast errors as an approximation of the amount of reserves that need to be procured to accommodate solar generation.

3. Data summary

The data used in this work is obtained from the Western Wind and Solar Integration Study Phase 2 (WWSIS-2), which is one of the world's largest regional renewable integration studies to date (Lew et al., 2013). The Western Interconnection subregions are shown in Fig. 2. The WWSIS-2 study examined the impacts of up to 25% solar energy penetration in the western United States. Day-ahead and 1-hour-ahead solar forecasts were investigated in this study. Although the cases presented in this study use WWSIS-2 data, the insights gained are universally applicable regardless of the data set.

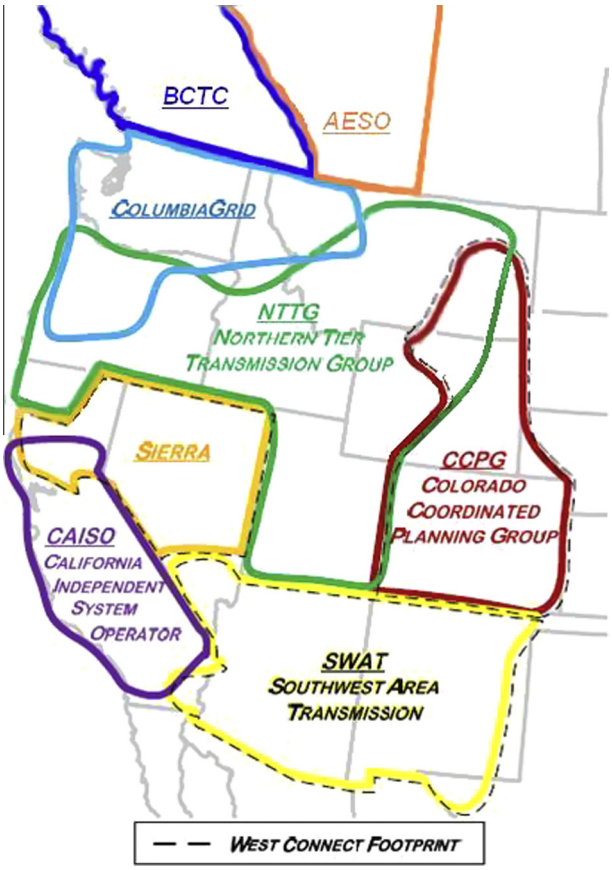


Fig. 2. Map of the Western Interconnection subregions (Lew et al., 2013).

The WWSIS-2 solar data is synthesized based on a 1-min interval using satellite-derived, 10-km × 10-km gridded, hourly irradiance data. In this paper, the 60-min solar power plant output for 2006 is used as the actual data. The solar power output data includes distributed generation rooftop photovoltaic, utility-scale photovoltaic, and concentrating solar power with thermal storage. Day-ahead solar forecasts were produced by 3TIER based on NWP simulations. The 1-hour-ahead forecasts were synthesized using a 1-hour-ahead persistence-of-cloudiness approach. With this method, the solar power index—which represents the ratio of actual power (P) to clear-sky power (P_{CS})—is first calculated. Next, the solar forecast power is estimated by modifying the current power output by the expected change in clear-sky output. For the 1-hour-ahead persistence-of-cloudiness approach, the forecast solar power at time $t+1$ can be calculated as follows.

$$P(t+1) = P(t) + SPI(t) \times [P_{CS}(t+1) - P_{CS}(t)] \quad (2)$$

where $P_{CS}(t+1)$ and $P_{CS}(t)$ represent the clear-sky solar power at time $t+1$ and t, respectively; $P(t)$ is the actual solar power output at time t; and $SPI(t)$ is the solar power index at time t.

3.1. Different geographic locations

Four scenarios are analyzed based on latitude and longitude locations of solar power plants. The first scenario analyzes the forecast of a single solar power plant with a 100-MW capacity. The second scenario analyzes the forecast of 46 solar power plants near Denver, Colorado, with an aggregated 3463-MW capacity. The third scenario investigates 90 solar power plants in the state of Colorado with an aggregated 6088-MW capacity. The fourth scenario analyzes solar power plants in the entire Western Interconnection in the United States, including 1007 solar power plants with an aggregated 64495-MW capacity. Fig. 3 shows the locations of solar power plants for different scenarios.

3.2. Improved solar power forecasts

To adequately study the value of improved solar power forecasts, we devised a number of scenarios that enable the analysis of different types of forecast improvements. The improvements are categorized by the appearance of large solar ramps, which are one of the biggest concerns of high-penetration solar power scenarios. First, the start and end points of all significant ramps are extracted using the swinging door algorithm. The definition of significant ramps is based on the magnitude of solar power change. In this paper, a significant ramp is defined as the change in solar power output that is greater than a ramp forecasting threshold (θ), expressed as:

$$|P(t+\Delta_t) - P(t)| > \theta \tag{3}$$

where $P(t)$ is the solar power output at time t; and Δ_t is the duration of the ramp. Three types of forecasting improvements are performed on the day-ahead solar power forecasts of the entire Western Interconnection scenario. These improvements are generated through the following procedures:

i. Uniform forecasting improvements of the time series excluding ramping periods: The forecast errors of the time series when there is not a significant ramp are uniformly decreased by a percentage (I_u).
ii. Ramp forecasting magnitude improvements: Only significant ramps that are identified as a change greater than or equal to a threshold value (e.g., 10% of the capacity) are modified in the improved forecasts. The forecast errors of the time series with ramps are decreased by a percentage (I_r).
iii. Ramp forecasting threshold changes: The ramp forecasting threshold (θ) is changed from 10% to 20% of the solar power capacity.

Based on the three improvements, the improved day-ahead forecasts (P_{nf}) are expressed as:

$$P_{nf}(t) = \begin{cases} P(t) + E_f(t) * (1 - I_u), & \text{when there is not a ramp} \\ P(t) + E_f(t) * (1 - I_r), & \text{when there is a ramp} \end{cases} \tag{4}$$

where P is the actual solar power generation and E_f is the forecast error of the original solar power forecast.

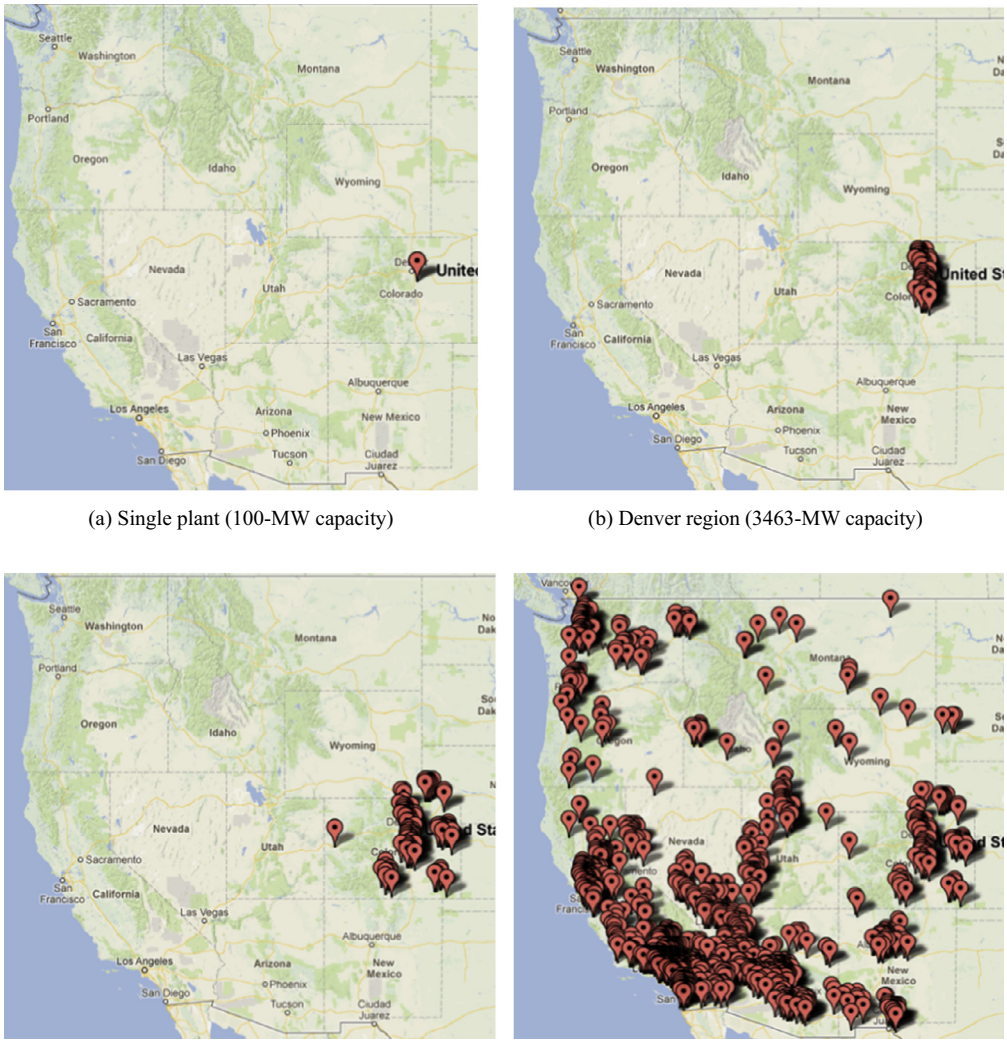
Sensitivity analysis is generally performed by running the model a very large number of times. Since the ramp extraction using the swinging door algorithm is relatively computational expensive, a surrogate approach allows for the examination of the entire parameter space while maintaining computational tractability. To analyze the sensitivity of the proposed metrics to the three types of solar forecasting improvements, a design-of-experiments (DoE) methodology is proposed, in conjunction with response surface development and sensitivity analysis. The inputs of the response surface are the three improvement parameters: I_u, I_r, and θ; the output of the response surface is the metric value (e.g., RMSE, MAE, etc.). The ranges of the three parameters are defined as: $0\% < I_u < 100\%$, $0\% < I_r < 100\%$, and $10\% < \theta < 20\%$. The Sobol's quasi-random sequence generator, a widely used method in response surface (Forrester et al., 2008; Zhang et al., 2013c), was adopted to generate training points. Sobol's sequences (Sobol, 1976) use a base of two to form successively finer uniform partitions of the unit interval and reorder the coordinates in each dimension. The Sobol sequence generator produces highly uniform samples of the unit hypercube. In addition, nonparametric statistical testing is proposed to compare the distributions of each metric and discern whether their differences are statistically significant. Fig. 4 shows the overall structure of evaluating the sensitivity of the proposed metrics to different types of solar forecasting improvements. The methodologies for sensitivity analysis and nonparametric statistical testing of different metrics are described in Sections 4 and 5, respectively.

4. Response surface and sensitivity analysis

The response surface methodology is concerned with the construction of approximation models to estimate system performance and to develop relationships between specific system inputs and outputs (Wang and Shan, 2007; Zhang et al., 2012). In this paper, multiple response surfaces are constructed to represent the metric values as functions of the parameters of the three types of solar forecasting improvements. Support vector regression (SVR) was adopted for this purpose. The extended Fourier Amplitude Sensitivity Test (eFAST) was adopted for sensitivity analysis.

4.1. Support vector regression (SVR)

SVR has gained popularity within the statistical learning community, engineering optimization community, and renewable energy community (Che and Wang, 2010;



(a) Single plant (100-MW capacity)

(b) Denver region (3463-MW capacity)

(c) State of Colorado region (6088-MW capacity)

(d) Western Interconnection (64495-MW capacity)

Fig. 3. Solar power plants at different geographic locations.

Zhang et al., 2013b) because it provides a unique way to construct smooth, nonlinear regression approximations by formulating the response surface construction problem as a quadratic programming problem. SVR can be expressed as (Chang and Lin, 2011)

$$\tilde{f}(x) = \langle w, \Phi(x) \rangle + b \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product; w is a set of coefficients to be determined; and $\Phi(x)$ is a map from the input space to the feature space. To solve the coefficients, we can allow a predefined maximum tolerated error ξ (with respect to the actual function value) at each data point, given by (Chang and Lin, 2011)

$$\left| \tilde{f}(x_i) - f(x_i) \right| \leq \xi \tag{6}$$

where $f(x)$ is the actual function to be approximated. The flatness of the approximated function can be characterized by w. By including slack variables x_i to the constraints and a cost function, the coefficient w can be obtained by solving a quadratic programming problem. In this paper, we use an efficient SVR package, a library for support vector machines developed by Chang and Lin (2011). The epsilon-SVR and the radial basis function kernel are adopted in this paper. The parameters are selected based on cross-validation to achieve a desirable training accuracy.

4.2. Extended Fourier Amplitude Sensitivity Test (eFAST)

The eFAST algorithm is a variance-based sensitivity analysis method (Saltelli and Bolado, 1998). The sensitivity value is defined based on conditional variances that indicate the individual or joint effects of the uncertain inputs on the output. Two indexes are calculated: (i) the main effect index, which measures the contribution of one type of solar forecasting improvement alone (e.g., uniform improvements) to the uncertainty (variance) in a metric value (e.g., Rényi entropy); and (ii) the total effect index, which gives the total variance in the metric value caused

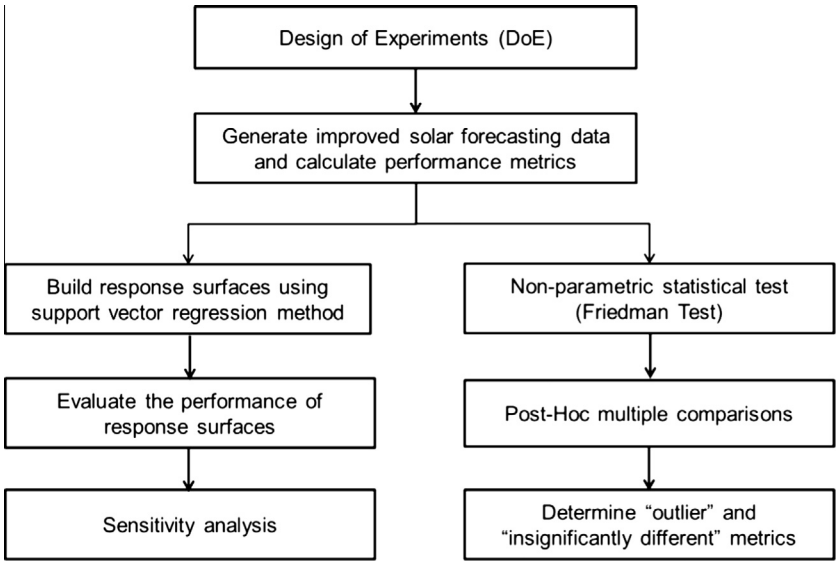


Fig. 4. Overall structure of evaluating the sensitivity of metrics to solar forecasting improvements.

by the type of solar forecasting improvements (e.g., uniform improvements) and its interactions with any of the other types of forecasting improvements. The eFAST method was implemented in the R statistical environment (R, 2008) using the *sensitivity* package (Pujol et al., 2014)

5. Nonparametric statistical testing

It was unclear which iteration of the DoE provided the "best forecast" because of the nonlinearity introduced by the ramp threshold (forecasting improvement) parameter. That is, the Sobol sequence used within the forecasting improvement study precluded the ranking of forecasts from best to worst, or vice versa, and metrics could not be directly or relatively compared to an ordered array of "forecasting accuracies." Nevertheless, the situation presented itself as an opportunity to compare the distributions of each metric to discern whether their differences were statistically significant. As such, the metrics output from the DoE was analyzed as a randomized complete block design, which allowed the relative assessment of metrics for the range of forecasting performances expected or desired. Examining each metric's distributions revealed the violation of normality as well as the assumption of equal variance; thus, nonparametric statistical testing approaches were required.

Because the metrics can be viewed as a proxy for a time series of forecast errors—i.e., for inference on the severity of future events—it was deemed of value to understand (i) which metrics provided indistinguishable information and (ii) which metrics did not, so that recommendations could be made on their use within the appropriate context. Testing the former case was accomplished with a nonparametric Friedman test for statistical significance coupled to post hoc analysis. Results from the Friedman analysis guided an extended metric study for known forecasting accuracies to deduce relative metric performance and make progress toward the latter case. Each metric represents a mapping of errors to a unique metric continuum, in which each metric summarizes a time series of errors with a point measure. The units of the metrics have little correspondence, and their comparison is a nontrivial task.

To ensure the validity of multiple comparisons of metrics, the continuum of metrics measured on the DoE space was first normalized and then inverted where applicable to maintain a consistent metric setting with equivalent scale. The smallest value (0) was equated with the smallest metric error, or low anticipated severity of forecast deviation from actuality; and the largest value (1) was equated with the largest error, or high anticipated severity of forecast deviation from actuality. The argument for this linear normalization is threefold: (i) the DoE spans the range of anticipated forecasts; (ii) the nonparametric approach relies on relative ranking, and the continuous information of the metrics is lost; and (iii) a consistent metric scale, or some form of normalization, is required and unavoidable. Further, it should be noted that each iteration of the DoE could be considered a nuisance factor effect—i.e., a background effect that needs to be considered but is not the primary interest—because greater interest is placed on detecting statistical differences among the distributions of each treatment, i.e., the 16 metrics.

5.1. The Friedman test

The Friedman test, as detailed in Hollander and Wolfe (1999), is a nonparametric randomized block design alternative to the one-way repeated measures analysis of variance (ANOVA) used when either the assumption of normality and/or equal variances are violated. The Friedman test was used to assess whether statistically significant differences exist among the 16 metrics (treatments) while

simultaneously blocking the nuisance effect from the DoE iteration variance. Although the Friedman test is less powerful than the ANOVA, it is of no concern that the metrics' residuals violate ANOVA assumptions. The Friedman test relies on a rank sum procedure and the comparison of parameter location, i.e., the median. The only requirements of the test are that the data comes from a continuous population and that all observations are mutually independent. The former is satisfied because each metric is distance based or probabilistically derived, and the latter is satisfied by scripting independence. Although it considers minor nuisance block effects, the null hypothesis, H_0, states that the parameter location is equal for each treatment—i.e., there is no difference in the distributions of each metric. The alternative hypothesis, H_a, is that a significant difference exists among the metrics (treatments). The Friedman statistic is given by:

$$S = \left[\frac{12}{nk(k+1)}\sum_{j=1}^{k} R_j^2\right] - 3n(k+1) \tag{7}$$

where within n independent blocks (DoE iterations) the k metrics (treatments) are ranked according to the best performing treatment, which is assigned rank 1, and the worst performing, rank n. The rank, $K_{i,j}$, within the i-th ($i \in 1, 2, \ldots, k$) block is the j-th ($j \in 1, 2, \ldots, k$) treatment. There are $k!$ possible ranking arrangements within a given block, which equates to $n(k!)$ ranking possibilities when considering n blocks. An adjustment for ties leads to a modification of the above equation but was not a concern of the data set. An equally likely mean ranking of a metric (treatment) is evidence to support the null hypothesis. The Friedman test statistic, S, approaches a chi-square distribution with $k-1$ degrees of freedom. When deciding to accept or reject the null hypothesis, the S statistic is compared to the respective χ^2 at an α level of significance, also known as an upper-tailed critical value test; the prevalent value of $\alpha = 0.05$ was utilized in this analysis.

5.2. Post-hoc multiple comparisons

If the p-value from the Friedman test is significant, a post hoc multiple comparison can be performed. The post-hoc comparison assesses which metric (treatment) differences led to the rejection of the null hypothesis. The task requires performing all pairwise comparisons to determine the specific pairwise differences responsible, and it entails the control of the familywise error rate (FWER). FWER is the probability of making Type I errors among all hypotheses when considering multiple hypotheses tests, and it seeks to reduce the probability of any false discoveries. FWER is controlled by considering the permutations of each variable of interest to calculate pointwise empirical p-values as well as by accounting for its test statistic relative maximum of permuted statistics (maxT) over all variables. A permutation-based approach to Friedman's test was implemented in the R statistical environment (R, 2008) using the *coin* package (Hothorn et al., 2013), and multiple comparisons using the *multcomp* R package (Hothorn et al., 2014) as described in Hollander and Wolfe (1999) as the Wilcoxon–Nemenyi–McDonald–Thompson multiple comparison test.

The post-hoc analysis gives the p-value of metric (treatment) differences and allows the assessment of whether a significant difference ($p < 0.05$) exists. Because it is possible that multiple statistically significant differences exist, a strategy was devised to iteratively reduce the metrics considered (within the Friedman test with post-hoc analysis code) to crudely classify the metrics and provide a basis for extended metric performance evaluation. The strategy was to (i) evaluate all metrics with a Friedman test, as blocked by the DoE iterations, (ii) perform the post-hoc multiple comparison if the Friedman test is statistically significant, (iii) record which metric is most frequently a part of the set of statistically significant differences, (iv) remove the recorded metric from the original metric set and retain the subset, and (v) repeat the process. The strategy allowed for the formation of metrics that could be considered "outliers" and a remaining set that could be considered "insignificantly different."

6. Results and discussion

6.1. Metrics evaluation for the WWSIS scenario

6.1.1. Distributions of solar power forecast errors

Distributions of day-ahead and 1-hour-ahead solar power forecast errors at the four analyzed regions are shown in Fig. 5. To compare the results from different spatial and temporal scales of forecast errors, we normalized the forecast error using the capacity of the analyzed solar power plants. As indicated by the narrower distribution curves for the 1-hour-ahead forecasting, it is observed that the 1-hour-ahead forecasting performs better than the day-ahead forecasting. The 1-hour-ahead forecast has a larger probability density value than the day-ahead forecast when the forecast error is smaller; the day-ahead forecast has a larger probability when the forecast error is larger. In addition, the distribution of errors at a larger geographic area has a more pronounced peak, slimmer shoulders, and longer tails. This observation indicates that relative forecast errors are smaller for a large geographic area, which shows the smoothing effect from geographic diversity for solar (Mills and Wiser, 2010).

6.1.2. Metrics evaluation for day-ahead and 1-hour-ahead solar forecasting

The values of different metrics to evaluate solar power forecasts at multiple spatial and temporal scales are shown in Table 3. As expected and inferred from the correlation coefficients—normalized root mean square error (NRMSE), normalized root mean quartic error (NRMQE), MaxAE, MAE, mean absolute percentage error (MAPE), KSIPer, OVERPer, and 95th percentile—1-hour-ahead forecasting

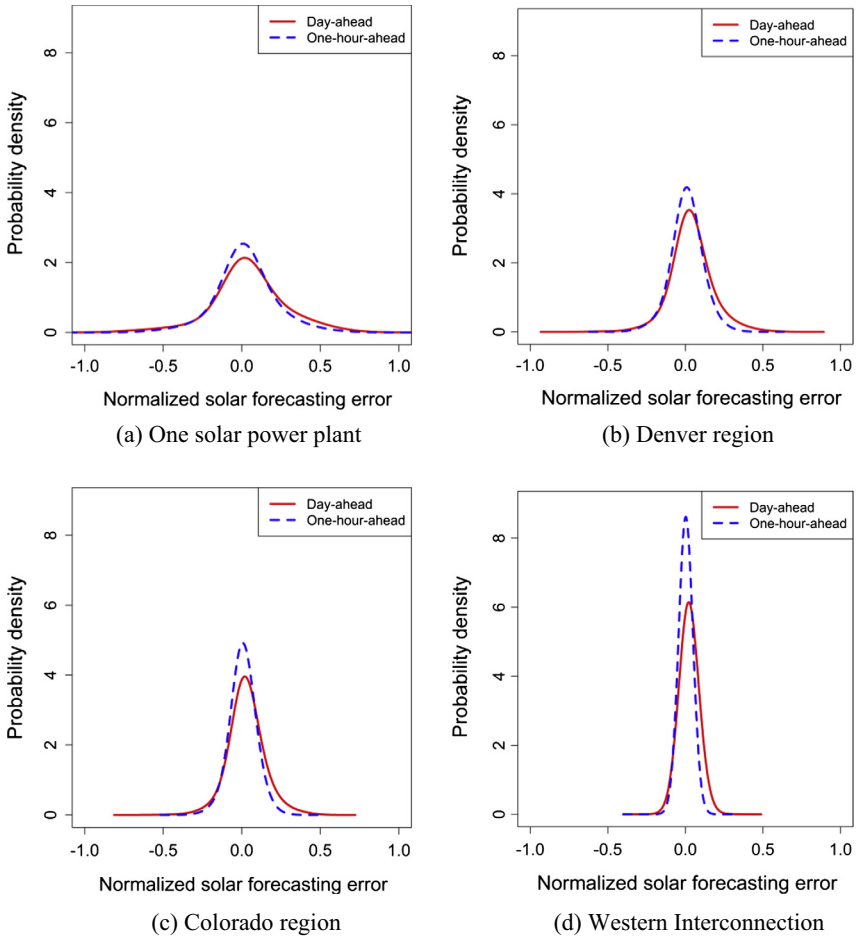


Fig. 5. Distributions of forecast errors at different geographic locations. (a) One solar power plant. (b) Denver region. (c) Colorado region. (d) Western Interconnection.

Table 3
Metrics values estimated by using an entire year of data.

Metrics	One plant		Denver		Colorado		Western Interconnection	
	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead
Correlation coefficient	0.65	0.76	0.87	0.94	0.91	0.96	0.990	0.995
RMSE (MW)	22.07	17.12	438.25	284.36	624.19	378.65	2711.31	1488.28
NRMSE	0.22	0.17	0.13	0.08	0.10	0.06	0.04	0.02
RMQE (MW)	32.58	26.05	695.25	432.95	978.04	575.01	4136.96	2476.55
NRMQE	0.33	0.26	0.20	0.13	0.16	0.09	0.06	0.04
MaxAE (MW)	84.10	74.33	2260.94	1304.73	3380.28	1735.24	17,977.53	16,127.32
MAE (MW)	14.81	11.34	286.65	191.17	413.11	256.69	1973.90	1064.52
MAPE	0.15	0.11	0.08	0.06	0.07	0.04	0.03	0.02
MBE (MW)	4.27	2.19	131.82	31.64	172.54	43.32	1497.29	132.13
KSIPer (%)	216.73	104.42	184.30	52.84	143.38	48.28	132.92	47.76
OVERPer (%)	136.36	28.16	94.43	0.77	54.65	0.37	41.43	0.00
Standard dev. (MW)	21.65	39.57	418.00	282.62	599.94	376.20	2260.09	1482.44
Skewness	−0.19	0.08	0.20	−0.20	0.18	−0.21	0.62	−0.23
Kurtosis	2.04	2.40	3.79	2.52	3.35	2.47	3.76	4.82
95th percentile (MW)	50.59	39.57	990.66	637.45	1394.85	838.27	5652.60	3079.32
Capacity (MW)	100.00	100.00	3463.00	3463.00	6088.00	6088.00	64495.00	64495.00

performs better than day-ahead forecasting. This matches the observation from the forecast error distributions shown in Fig. 5.

The NRMSE values become smaller with increasing geographic area, which shows that the solar forecast performs relatively better for larger regions. Because of the weighting property of the NRMSE metric, the difference between day-ahead and 1-hour-ahead forecasts is more significant than that observed by the Pearson's correlation coefficient metric, especially for large geographic

areas. Root mean quartic error (RMQE) and NRMQE have larger metrics values than RMSE and NRMSE, respectively, because of the quartic weight of large forecast errors. The MaxAE metric shows that (i) the maximum forecast error is greater than 70% of the capacity for the single plant scenario, and (ii) for the Western Interconnection scenario, the maximum errors of day-ahead and 1-hour forecasts are 27.9% and 25% of the capacity, respectively. It is very important for power system operators to be aware of such large solar power forecast errors that might cause significant economic losses for the power system. The MAE values of day-ahead forecasts for the single plant scenario, the Denver region scenario, the Colorado region scenario, and the Western Interconnection scenario are, respectively, 31%, 50%, 61%, and 85% larger than those of the 1-hour-ahead forecasts. These results show that the accuracy difference between forecasts at different timescales is increasing with the area of aggregation. MAE could be used to characterize the difference in solar forecasting performance attributed to spatial aggregation. The positive MBE metrics indicate an over-forecast characteristic for both day-ahead and 1-hour-ahead forecasting. For the KSI metric, we observe that (i) the value of KSIPer decreases with increasing geographic area, indicating that the solar forecast performs relatively better for larger spatial aggregations; and (ii) the difference between day-ahead and 1-hour-ahead forecasts is more significant for scenarios 2 through 4 (Denver, Colorado, and Western Interconnection) than for the single plant scenario. The zero OVERPer value of the 1-hour-ahead forecasts for the Western Interconnection scenario shows that the forecasts and actual values are statistically the same beyond the critical value V_c. The large skewness value 0.62 in the day-ahead time frame for the Western Interconnection scenario indicates a significant tendency to over-forecast. The larger the kurtosis values, the narrower the distribution of forecast errors. As shown in Table 3, the 1-hour-ahead forecast for the Western Interconnection scenario has the largest kurtosis value, which indicates the best forecasting performance.

Table 4 shows the forecast uncertainty as evaluated by Rényi entropy for both day-ahead and 1-hour-ahead forecasting at the four geographic regions. Five cases are analyzed for each scenario based on forecasting time periods: (i) forecasting throughout a whole year, (ii) forecasting in January, (iii) forecasting in July, (iv) forecasting at the time of 14:00 each day throughout a whole year, and (v) forecasting at the solar peak time of 10:00 to 16:00 each day throughout a whole year. We observe that the length of the forecasting period affects the uncertainty in the forecasting. In general, the uncertainty in the forecasting of using the whole year's data is relatively less than that in any of the other cases (January, July, 14:00, and 10:00 to 16:00), which can be partly attributed to the nature of the forecasting continuity.

6.1.3. Ramp extraction results

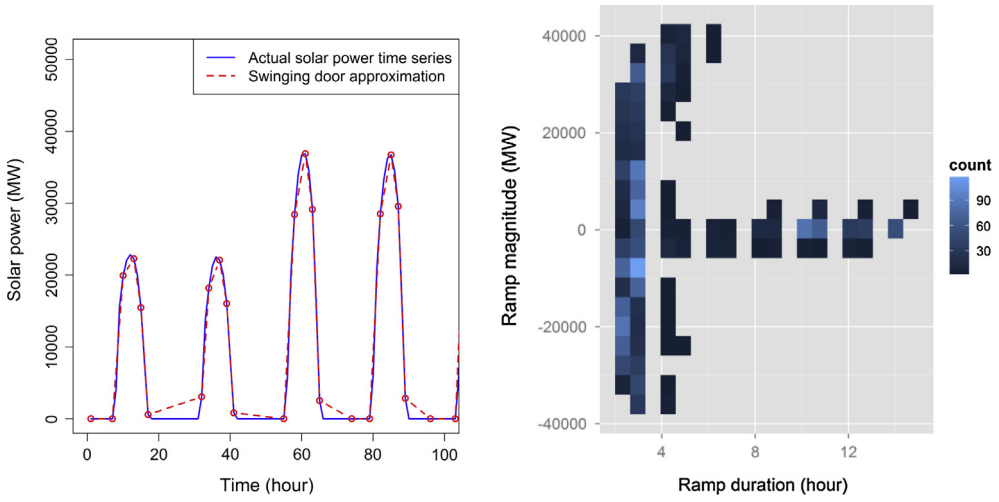
Fig. 6 shows a typical example in the extraction of ramps from actual solar power generation of the Western Interconnection scenario during a 100-h period. The tolerance value, ε, is set at 2.5% of solar capacity. In the figure, the solid and dashed lines represent the actual solar power and the piecewise linear approximation (generated by the swinging door algorithm), respectively. An accurate piecewise linear approximation to the actual solar power profile is obtained as shown in Fig. 6(a). The figure presents the nature of up and down ramps with large, medium, and insignificant changes in power. Fig. 6(b) shows the bivariate distribution of all solar power ramps. The extracted ramps in the actual solar power generation are visualized in terms of ramp duration and ramp magnitude. It is observed that the distribution spreads within the more immediate ramp region (the left side of Fig. 6(b)). Most ramps occur within a time span of 5 h or less.

6.2. Response surfaces of metrics

The response surface development, sensitivity analysis (in Section 6.3), and nonparametric statistical testing (in Section 6.4) are based on the day-ahead forecasts of the entire Western Interconnection scenario (with a 64495-MW capacity). A response surface is constructed to represent the metric value as a function of the parameters of the three types of solar forecasting improvements. The assumed forecasting improvement capability is defined as (i) 0–100% for uniform forecasting improvements, (ii) 0–100% for ramp forecasting improvements, and (iii) 10–20% of solar power capacity for the ramp forecasting threshold. The number of training points is defined as 10 times the given number of problem dimensions. Thus, 30 training points are used for this 3-dimensional problem. The accuracy of the constructed response surface is evaluated by cross-validation technique. Response surfaces of six typical metrics are shown in Fig. 7. A constant value

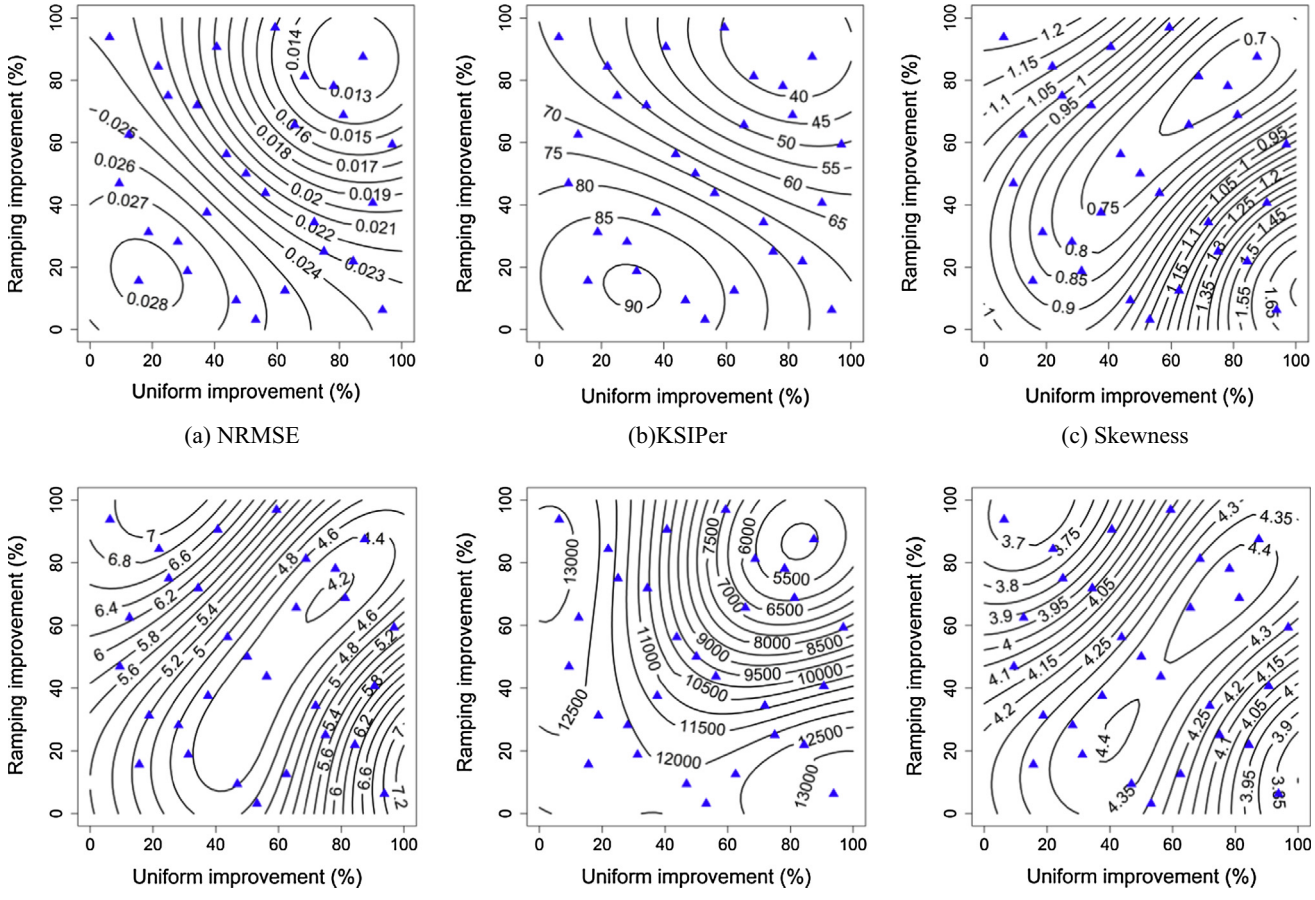
Table 4
The uncertainty metric of Rényi entropy at multiple spatial and temporal scales.

Cases	One plant		Denver		Colorado		Western Interconnection	
	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead	Day-ahead	1-Hour-ahead
Year	4.83	4.64	4.24	4.63	4.33	4.73	4.47	4.01
January	4.71	5.06	5.18	5.06	5.46	4.79	5.24	5.11
July	4.64	4.74	4.25	4.87	5.02	5.09	4.75	4.86
14:00	5.07	5.00	4.83	4.99	5.13	5.27	4.97	5.72
10:00–16:00	4.95	4.73	4.60	4.79	4.82	4.94	4.90	4.45



(a) Swinging door approximation (b) Bivariate distribution of ramps

Fig. 6. Ramp extraction from the actual solar power generation of the Western Interconnection scenario.



(a) NRMSE (b)KSIPer (c) Skewness

(d) Kurtosis (e) MaxAE (f) Rényi entropy

Fig. 7. Response surfaces of metrics constructed by SVR.

of the ramp forecasting threshold is used in the plots, which is 10% of the solar power capacity. The triangle points in the figures represent the 30 training points for the development of the response surfaces.

Fig. 7(a) shows that the NRMSE value decreases with both uniform forecasting improvements and ramp forecasting improvements; a similar trend is also observed from the KSIPer metric shown in Fig. 7(b). Fig. 7(c) through (f)

show that the response surfaces of skewness, kurtosis, MaxAE, and Rényi entropy are highly nonlinear. For the skewness metric shown in Fig. 7(c), it is observed that (i) a consistent positive skewness is obtained through uniform and ramp improvements, leading to an over-forecasting tail; and (ii) the minimum skewness is observed with approximately 80% uniform forecasting improvements and 80% ramp forecasting improvements. In Fig. 7(f), high uncertainty is observed in two regions; whereas low uncertainty is obtained at the top left and bottom right corners, resulting primarily from one type of improvement (ramp or uniform forecasting improvements).

6.3. Sensitivity analysis results

The main effects and total effects of the proposed metrics to the three types of forecasting improvements are listed in Table 5. The larger the value of the main effect (or total effect) index, the more sensitive the metrics are to the type of forecasting improvement. Most metrics are highly sensitive to the uniform improvements (compared to ramp forecasting improvements and ramp threshold changes), indicating that these metrics can consistently and effectively show the difference in the accuracy of solar forecasts with uniform improvements. In addition, the skewness, kurtosis, and Rényi entropy metrics are observed to be sensitive to all three types of forecasting improvements. These three metrics (skewness, kurtosis, and Rényi entropy) could be adopted to evaluate the improvements in the accuracy of solar forecasts with ramp forecasting improvements and ramp threshold changes that are important to the economics and reliability of power system operations.

6.4. Results from nonparametric statistical testing

Table 6 provides the statistically significant results from the first iteration of the Friedman test with post-hoc multiple comparisons. The significant differences are in absolute terms, so the ordering of the differences is irrelevant. Any p-value less than 0.05 means that a significant difference exists between the two metrics (treatments) according to the nonparametric test, and the OVERPer is identified as significantly different from other metric distributions in half (i.e., two or four) of the cases. Although the $\alpha = 0.05$ threshold is somewhat arbitrary, the OVERPer is also the most frequent metric identified near the threshold according to a list of p-values sorted in ascending order.

Table 6
The first iteration of the Friedman Test with post-hoc multiple comparisons.

Significant differences	p-Value
OVERPer – MAPE	0.0186
Rényi entropy – OVERPer	0.0411

The first iteration of the Friedman test with post-hoc multiple comparisons found that the OVERPer was most frequently significantly different. According to the devised strategy of Section 5.2, the OVERPer metric was removed from consideration and the process was again applied. The OVERPer was considered to be an "outlier metric," because it gave considerably different information than the other metrics. This did not imply that the information provided by the metric was "wrong," especially because it was not known *a priori* which iteration within the DoE forecast improvement study was the "best," but that it could be considered as its own class of information.

Table 7 provides the first five differences of the statistically significant results from the second iteration of the Friedman test with post-hoc multiple comparisons. A total of 26 significant differences were found in the second iteration, but to save space all results are not provided here. However, the table illustrates the general trend: the RMSE was most frequently a significantly different comparison. The RMSE was considered an "outlier metric" because of its considerably different distribution, and it could be considered to contain its own class of information.

Table 5
Sensitivity analysis of metrics to the three types of forecasting improvements.

Metrics	Uniform improvement		Ramp improvement		Ramp threshold	
	Main effect	Total effect	Main effect	Total effect	Main effect	Total effect
Correlation coefficient	0.836	0.905	0.070	0.119	0.004	0.069
RMSE	0.783	0.862	0.114	0.169	0.001	0.072
NRMSE	0.783	0.862	0.114	0.169	0.001	0.072
RMQE	0.771	0.883	0.099	0.187	0.001	0.061
NRMQE	0.771	0.883	0.099	0.187	0.001	0.061
MaxAE	0.753	0.900	0.065	0.196	0.008	0.093
MAE	0.788	0.849	0.112	0.164	0.004	0.085
MAPE	0.788	0.849	0.112	0.164	0.004	0.085
MBE	0.659	0.734	0.211	0.282	0.085	0.113
KSIPer	0.657	0.731	0.211	0.285	0.113	0.114
OVERPer	0.803	0.889	0.067	0.143	0.010	0.094
Standard deviation	0.815	0.899	0.083	0.143	0.001	0.060
Skewness	0.436	0.876	0.113	0.528	0.004	0.058
Kurtosis	0.313	0.887	0.061	0.546	0.031	0.218
95th percentile	0.788	0.891	0.088	0.162	0.001	0.071
Rényi entropy	0.207	0.716	0.221	0.682	0.052	0.197

Table 7
The second iteration of the Friedman test with post-hoc multiple comparisons.

Significant Differences	p-value
RMSE – Rényi entropy	4.27e−6
RMSE – KSIPer	1.33e−5
RMSE – R^2	3.34e−5
Skewness – Rényi entropy	1.13e−4
RMSE – RMQE	2.15e−4

It is likely that 26 significant differences were identified in the second iteration of the strategy, compared to only two in the first iteration, because of an artifact of the FWER control: the OVERPer was vastly different than the other metrics and thus contained the majority of errors relative to the errors of all pairwise differences. This makes intuitive sense because the OVERPer only contains information when forecast errors are above a threshold. In any case, the RMSE metric was deleted from consideration, and the process was again applied. In the third iteration in the strategy, no significant differences among the remaining metrics were found. This meant the null hypothesis could not be rejected, and the remaining metrics provided very similar information in terms of their nonparametric distributions over the blocked DoE space.

6.5. Discussion on sensitivity analysis and nonparametric statistical testing

The response surface-based sensitivity analysis found that the metrics of skewness, kurtosis, and Rényi entropy are sensitive to uniform improvements, ramp improvements, and ramp threshold. The nonparametric statistical testing found that the proposed metrics could be broadly divided into three classes: (i) the OVERPer metric, (ii) the RMSE metric, and (iii) the remaining metrics. One should be aware of a few important considerations when selecting the appropriate metrics for evaluating the performance of solar power forecasting. First, because skewness and kurtosis are not stand-alone metrics, it is recommended that the metrics of MBE, standard deviation, skewness, kurtosis, and the distribution of forecast errors should be used as a group. In addition, it is important to select at least one metric from each class determined through the nonparametric statistical testing. Thus, for a comprehensive, consistent, and robust assessment of the performance of solar power forecasts, a suite of metrics consisting of MBE, standard deviation, skewness, kurtosis, distribution of forecast errors, Rényi entropy, RMSE, and OVERPer is recommended. In addition, the four statistical moments (MBE, standard deviation, skewness, kurtosis) can be extracted from the distribution of forecast errors in many cases. When the four statistical moments can be easily extracted from the distribution of forecast errors, the final set of metrics can be further reduced to the following four metrics: distribution of forecast errors, Rényi entropy, RMSE, and OVERPer.

6.6. Metrics evaluation for a 1-MW solar power plant at Smyrna Airport, Tennessee

To further evaluate the effectiveness of the solar forecasting metrics, an additional case study was performed for a 1-MW solar power plant installed at Smyrna Airport, Tennessee (as shown in Fig. 8). Hourly measured solar power outputs at the plant between May 1, 2013, and October 31, 2013, were used. Day-ahead forecasts for the solar power plant were performed using the North American Mesoscale Forecast System (NAM) simulations. NAM is a regional weather forecast model that covers North America with a horizontal resolution of 12 km (Rogers et al., 2009). A two-stream radiative transfer model and the PV_LIB TOOLBOX developed at Sandia National Laboratories (Stein, 2012) were used for solar power estimation. One-hour-ahead forecasts were obtained using a 1-hour-ahead persistence-of-cloudiness approach.

Table 8 lists the values of different metrics for day-ahead and 1-hour-ahead forecasts. According to the final set of metrics (distribution of forecast errors, Rényi entropy, RMSE, and OVERPer), we observe that: (i) the

Fig. 8. The solar power plant at Smyrna Airport, Tennessee (Soltas Energy, 2014).

Table 8
Solar forecasting metrics values for the solar plant at Smyrna Airport, Tennessee.

Metrics	Day-ahead	1-Hour-ahead
Correlation coefficient	0.78	0.91
RMSE (MW)	0.18	0.10
NRMSE	0.18	0.10
RMQE (MW)	0.27	0.15
NRMQE	0.27	0.15
MaxAE (MW)	0.73	0.44
MAE (MW)	0.12	0.07
MAPE	0.12	0.07
MBE (MW)	0.07	−0.03
KSIPer (%)	195.96	87.89
OVERPer (%)	105.10	16.65
Standard deviation (MW)	0.17	0.09
Skewness	0.43	0.33
Kurtosis	2.22	2.62
95th percentile (MW)	0.42	0.22
Rényi entropy	4.52	4.81

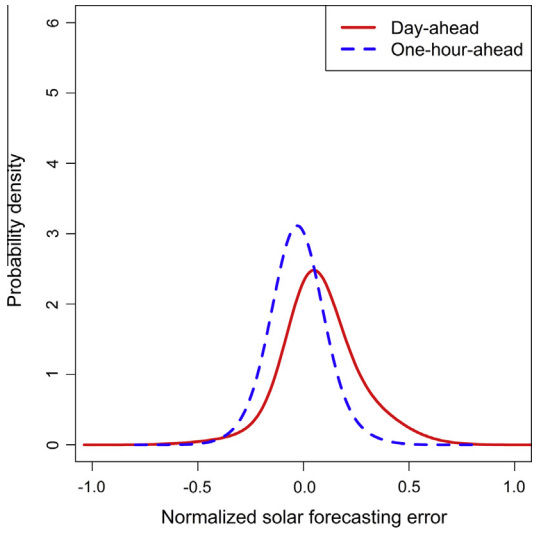


Fig. 9. Distributions of day-ahead and 1-hour-ahead solar power forecast errors for the plant at Smyrna Airport, Tennessee.

1-hour-ahead forecasting generally performs better than the day-ahead forecasting; and (ii) there is relatively more uncertainty in the 1-hour-ahead forecasting than that in the day-ahead forecasting. All the metrics can successfully evaluate the expected performance of solar forecasting. Fig. 9 shows the distribution of solar power forecast errors. Both the forecast error distributions and the MBE metric in Table 8 indicate an over-forecast characteristic for day-ahead forecasts and an under-forecast characteristic for 1-hour-ahead forecasts.

7. Conclusion

This paper proposed a suite of metrics for evaluating the performance of solar power forecasting. The performance of the proposed metrics was evaluated using the actual and forecast solar power data from the Western Wind and Solar Integration Study Phase 2. The distribution of forecast errors indicates that relative forecast errors are smaller for a large geographic area. The results showed that the all proposed metrics can successfully evaluate the quality of a solar forecast.

To analyze the sensitivity of the proposed metrics to improved solar forecasts, a sensitivity analysis methodology was developed based on a DoE and response surfaces. Nonparametric statistical testing was performed to compare the distributions of each metric to discern whether their differences were statistically significant. The results showed that (i) all proposed metrics were sensitive to solar forecasts with uniform forecasting improvements; (ii) the metrics of skewness, kurtosis, and Rényi entropy were also sensitive to solar forecasts with ramp forecasting improvements and ramp forecasting threshold; and (iii) the differences among the metrics of OVERPer, RMSE, and the remaining metrics were statistically significant. In addition, a small suite of metrics were recommended based on the sensitivity analysis and nonparametric statistical testing results, including MBE, standard deviation, skewness, kurtosis, distribution of forecast errors, Rényi entropy, RMSE, and OVERPer.

Currently, there are two main customers for solar forecasting technologies: the utility company and the independent system operator of a power market. As solar penetration increases, solar forecasting will become more important to solar energy producers and solar power plant developers. A suite of metrics such as those presented in this paper are expected to assist all stakeholders in assessing the performance of solar power forecasts and using them for various applications. To know whether the values obtained from applying these metrics do indeed represent a significant improvement in forecasting accuracy, it is very important to establish baselines and target values for these metrics that are relevant to each stakeholder. Future work will determine these baselines and target values for the metrics proposed in this paper on the basis of specific customer types and geographic regions.

Acknowledgements

This work was supported by the U.S. Department of Energy under Contract No. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory as part of the project work performed under the SunShot Initiative's Improving the Accuracy of Solar Forecasting program. Comments and suggestions from the following researchers are also gratefully acknowledged: Dr. Brad Lehman from Northeastern University, Dr. Joseph Simmons from University of Arizona, Dr. Edwin Campos from Argonne National Laboratory, Dr. Melinda Marquis from National Oceanic and Atmospheric Administration, Tara Jensen and Tressa Fowler from National Center for Atmospheric Research, and Jari Miettinen from VTT Technical Research Centre of Finland.

Appendix A. Statistical metrics

A.1. Kernel density estimation (KDE)

KDE is a nonparametric approach to estimate the probability density function of a random variable. KDE has been widely used in the renewable energy community for wind speed distribution characterization (Zhang et al., 2011, 2013a) and wind and solar power forecasting (Zhang et al., 2013e; Juban et al., 2007). For an independent and identically distributed sample, $x_1, x_2, \ldots, x_n$, drawn from some distribution with an unknown density f, KDE is defined as (Jones et al., 1996)

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad \text{(A.1)}$$

In the equation, $K(\cdot) = (1/h)\,K(\cdot/h)$ has a kernel function K (often taken to be a symmetric probability density) and a bandwidth h (a smoothing parameter).

A.2. Pearson's correlation coefficient

Pearson's correlation coefficient is a measure of the correlation between two variables (or sets of data). In this paper, the Pearson's correlation coefficient, ρ, is defined as the covariance of actual and forecast solar power variables divided by the product of their standard deviations, which is mathematically expressed as:

$$\rho = \frac{\mathrm{cov}(p, \hat{p})}{\sigma_p \sigma_{\hat{p}}} \tag{A.2}$$

where p and $\hat{p}$ represent the actual and forecast solar power output, respectively. Pearson's correlation coefficient is a global error measure metric; a larger value of Pearson's correlation coefficient indicates an improved solar forecasting skill. The Pearson's correlation coefficient shows the similarity between the overall trend of the forecasts and actual values, though it does not necessarily account for relative magnitudes. Because of geographic smoothing, this metric may be better used to evaluate forecast accuracy at individual plants or in small groupings of plants instead of for large balancing authority areas or interconnections. This is because the geographic smoothing will decrease the magnitude of the difference between a good and a bad forecast.

A.3. Root mean square error (RMSE), normalized root mean square error (NRMSE), root mean quartic error (RMQE) and normalized root mean quartic error (NRMQE)

The RMSE metric also provides a global error measure throughout the entire forecasting period, which is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2} \tag{A.3}$$

where p_i represents the actual solar power generation at the ith time step, $\hat{p}_i$ is the corresponding solar power generation estimated by a forecasting model, and N is the number of points estimated in the forecasting period. To compare the results from different spatial and temporal scales of forecast errors, we normalized the RMSE using the capacity of the analyzed solar power plants. The RMSE (or NRMSE) metric tends to penalize large forecast errors because of the squaring of each error term, which effectively weights large errors more heavily than small errors. The metric is useful for evaluating the overall performance of the forecasts, especially when extreme events are a concern.

The RMQE metric also provides a global error measure throughout the entire forecasting period, which is given by

$$\mathrm{RMQE} = \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^4\right]^{1/4} \tag{A.4}$$

The NRMQE metric is calculated by normalizing the RMQE using the total capacity of the analyzed solar power plants. The RMQE (or NRMQE) metric penalizes large forecast errors more than the RMSE (or NRMSE) metric.

A.4. Maximum absolute error (MaxAE), mean absolute error (MAE), mean absolute percentage error (MAPE), and mean bias error (MBE)

The MaxAE metric is indicative of the largest forecast errors and is given by:

$$\mathrm{MaxAE} = \max_{i=1,2,\ldots,N}|\hat{p}_i - p_i| \tag{A.5}$$

The MaxAE metric is useful to evaluate the forecasting of short-term extreme events in the power system. A smaller MaxAE indicates a better forecast. The MaxAE metric can capture the largest forecast error in the forecast period, which is very important for a power system. However, this metric likely gives too much weight to the extreme event, so it is more useful when evaluated during very short time periods.

The MAE metric has been widely used in regression problems and by the renewable energy industry to evaluate forecast performance and is given by:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{p}_i - p_i| \tag{A.6}$$

The MAE metric is also a global error measure metric, but it does not punish extreme forecast events as much compared to the RMSE metric. Smaller values of MAE indicate better forecasts. One concern about the MAE metric is that a large number of very small errors can easily overwhelm a small number of large errors. This can be problematic in systems for which extreme events are a concern.

The MAPE and MBE metrics are expressed as:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{p}_i - p_i}{p_0}\right| \tag{A.7}$$

$$\mathrm{MBE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i) \tag{A.8}$$

where p_0 is the capacity of analyzed solar power plants. The MAPE metric can be used to compare the results from different spatial and temporal scales of forecast errors. The MBE metric intends to indicate an average forecast bias. A larger MBE shows more forecast bias. Assuming the forecast error is equal to the forecast minus the actual power generation, a positive MBE indicates over-forecasting; whereas a negative MBE indicates under-forecasting. Understanding the overall forecast bias (over- or under-forecasting) would allow power system operators to better allocate resources to compensate for forecast errors in the dispatch process. The MBE metric provides important information that can be used to improve the forecasts, but it does not give a good indication of the total range

of forecast errors. For example, the same MBE value could represent many significantly different error distributions, some of which may be more favorable than others.

A.5. Kolmogorov–Smirnov test integral (KSI) and OVER metrics

The KSI metric is more appropriate for comparing forecasts during longer time periods and for measuring how similar the distributions of forecasts and actual values are for the time period under consideration. The KS statistic D is defined as the maximum value of the absolute difference between two CDFs, expressed as (Espinar et al., 2009)

$$D = \max\left|F(p_i) - \hat{F}(p_i)\right| \tag{A.9}$$

where F and $\hat{F}$ represent the CDFs of actual and forecast solar power generation data sets, respectively. The associated null hypothesis is elaborated as follows: if the D statistic characterizing the difference between one distribution and the reference distribution is lower than the threshold value V_c, the two data sets have a very similar distribution and could statistically be the same. The critical value V_c depends on the number of points in the forecast time series, which is calculated for a 99% level of confidence (Espinar et al., 2009).

$$V_c = \frac{1.63}{\sqrt{N}}, \quad N \geq 35 \tag{A.10}$$

The difference between the CDFs of actual and forecast power is defined for each interval as (Espinar et al., 2009)

$$\begin{aligned} &D_j = \max\left|F(p_i) - \hat{F}(p_i)\right|, \quad j = 1, 2..., m \\ &\text{where } p_i \in [p_{\min} + (j-1)d, p_{\min} + jd] \end{aligned} \tag{A.11}$$

Here the value of m is chosen as 100, and the interval distance d is defined as (Espinar et al., 2009)

$$d = \frac{p_{\max} - p_{\min}}{m} \tag{A.12}$$

where $p_{\max}$ and $p_{\min}$ are the maximum and minimum values of the solar power generation, respectively. The KSI parameter is defined as the integrated difference between the two CDFs, expressed as (Espinar et al., 2009)

$$\text{KSI} = \int_{p_{\min}}^{p_{\max}} D_n dp \tag{A.13}$$

To compare the results from different spatial and temporal scales of forecast errors, a relative value of KSI (KSIPer) is calculated by normalizing the KSI value by $a_c = V_c \times (p_{\max} - p_{\min})$ (Espinar et al., 2009).

$$\text{KSIPer}(\%) = \frac{\text{KSI}}{a_c} \times 100 \tag{A.14}$$

The OVER metric characterizes the integrated difference between the CDFs of the actual and forecast solar power. In contrast to the KSI metric, the OVER metric evaluates only large forecast errors beyond a specified value, because large forecast errors are more important for a power system. In this paper, the OVER metric considers only the points at which the critical value V_c (given in Eq. (A.9)) is exceeded. The OVER metric and its relative value are given by (Espinar et al., 2009)

$$\text{OVER} = \int_{p_{\min}}^{p_{\max}} t dp \tag{A.15}$$

$$\text{OVERPer}(\%) = \frac{\text{OVER}}{a_c} \times 100 \tag{A.16}$$

The parameter t is defined by (Espinar et al., 2009)

$$t = \begin{cases} D_j - V_c & \text{if} \quad D_j > V_c \\ 0 & \text{if} \quad D_j \leq V_c \end{cases} \tag{A.17}$$

As with the KSIPer metric, a smaller value of OVERPer indicates a better performance of the solar power forecasting.

A.6. Skewness and kurtosis

Because the MAE and RMSE metrics cannot distinguish between two distributions with the same mean and variance but different skewness and kurtosis values, they ignore additional information about the forecast errors that could potentially have a significant impact on system operations. Skewness is a measure of the asymmetry of the probability distribution and is the third standardized moment. Assuming that forecast errors are equal to forecast power minus actual power, a positive skewness of the forecast errors leads to an over-forecasting tail, and a negative skewness leads to an under-forecasting tail. The tendency to over-forecast (or under-forecast) is important in that the system actions taken to correct for under-forecasting and over-forecasting events are not equal. An over-forecasting tendency could lead to the commitment of a less-than-optimal number of large thermal units, which would need to be corrected through the use of more dispatchable, and therefore more expensive, generation units (Zhang et al., 2013d,e). Skewness is not a stand-alone metric, but it can provide additional information to the system operator about the tendencies of the forecasting system that can be utilized to prepare appropriate counteractions.

Kurtosis is a measure of the magnitude of the peak of the distribution of forecast errors, or conversely the width of the distribution, and is the fourth standardized moment. The difference between the kurtosis of a sample distribution and that of the normal distribution is known as the excess kurtosis. In this work, the term *kurtosis* is treated synonymously with *excess kurtosis*. A distribution with a positive kurtosis value is known as leptokurtic, which indicates a peaked (narrow) distribution; whereas a negative kurtosis indicates a flat (wide) data distribution, known as platykurtic. The pronounced peak of the leptokurtic distribution indicates a large number of very small forecast errors, which shows a better performance of the forecasting system (Hodge and Milligan, 2011). In general, a narrow distribution of forecast errors indicates a better, more

accurate, performance of the forecasting method. Like skewness, kurtosis is not an appropriate stand-alone metric. However, it can provide information to the system operator about the relative frequency of extreme events.

References

Bessa, R.J., Miranda, V., Botterud, A., Wang, J., 2011. 'Good' or 'bad' wind power forecasts: a relative concept. Wind Energy 14, 625–636.

Bristol, E., 1990. Swinging door trending: adaptive trend recording? ISA Natl. Conf. Proc., 749–753.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27:1–27:27.

Che, J., Wang, J., 2010. Short-term electricity prices forecasting based on support vector regression and auto-regressive integrated moving average modeling. Energy Convers. Manage. 51 (10), 1911–1917.

Chen, C., Duan, S., Cai, T., Liu, B., 2011. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol. Energy 85 (11), 2856–2870.

Chow, W.C., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy test bed. Sol. energy 85 (11), 2881–2893.

Chu, Y., Nonnenmacher, L., Inman, R.H., Liao, Z., Pedro, H.T.C., Coimbra, C.F.M., 2014. A smart image-based cloud detection system for intra-hour solar irradiance forecasts. J. Atmos. Ocean. Technol. 31 (9), 1995–2007.

Crispim, E.M., Ferreira, P.M., Ruano, A.E., 2008. Prediction of the solar radiation evolution using computational intelligence techniques and cloudiness indices. Int. J. Innovative Comput. Inform. Control 4 (5), 1121–1133.

Espinar, B., Ramírez, L., Drews, A., Beyer, H.G., Zarzalejo, L.F., Polo, J., Martín, L., 2009. Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations. Sol. Energy 83 (1), 118–125.

Florita, A., Hodge, B.-M., Orwig, K., 2013. Identifying wind and solar ramping events. In: IEEE 5th Green Technologies Conf., Denver, Colorado.

Forrester, A., Sobester, A., Keane, A., 2008. Engineering Design via Surrogate Modelling: A Practical Guide. Wiley, New York.

Hammer, A., Heinemann, D., Lorenz, E., Lückehe, B., 1999. Short-term forecasting of solar radiation: a statistical approach using satellite data. Sol. Energy 67 (1–3), 139–150.

Hodge, B.-M., Milligan, M., 2011. Wind power forecasting error distributions over multiple timescales. In: IEEE PES GM Proc., San Diego, CA.

Hodge, B.-M., Orwig, K., Milligan, M., 2012. Examining information entropy approaches as wind power forecasting performance metrics. In: 12th Int. Conf. on Probabilistic Methods Applied to Power Systems, Istanbul, Turkey.

Hollander, M., Wolfe, D.A., 1999. Nonparametric Statistical Methods. John Wiley & Sons, Inc., Kawashima, New York.

Hothorn, T., Hornik, K., Wiel, M., Zeileis, A., 2013. Coin: A Computational Framework for Conditional Inference. http://cran.r-project.org/web/packages/coin/index.html.

Hothorn, T., Bretz, F., Westfall, P., 2014. Simultaneous Inference in General Parametric Models. http://cran.r-project.org/web/packages/multcomp/index.html.

Jones, M., Marron, J., Sheather, S., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. 91 (433), 401–407.

Juban, J., Siebert, N., Kariniotakis, G. N., 2007. Probabilistic short-term wind power forecasting for the optimal management of wind generation. In: IEEE Power Eng. Society, Lausanne Power Tech Conf. Proc., pp. 683–688.

Lew, D. et al., 2013. The Western Wind and Solar Integration Study Phase 2. NREL/TP-5500-55888. National Renewable Energy Laboratory, Golden, CO.

Lorenz, E., Heinemann, D., Wickramarathne, H., Beyer, H.G., Bofinger, S., 2007. Forecast of ensemble power production by grid-connected PV systems. In: Proc. 20th European PV Conference, Milano, Italy.

Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol. Energy 85 (5), 967–977.

Margolis, R., Coggeshall, C., Zuboy, J., 2012. SunShot Vision Study. U.S. Department of Energy, Washington, D.C..

Marquez, R., Coimbra, C.F.M., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Sol. Energy 85 (5), 746–756.

Marquez, R., Coimbra, C.F.M., 2013a. Intra-hour DNI forecasting based on cloud tracking image analysis. Sol. Energy 91, 327–336.

Marquez, R., Coimbra, C.F.M., 2013b. Proposed metric for evaluation of solar forecasting models. J. Sol. Energy Eng. 135, 011016-1.

Mills, A., Wiser, R., 2010. Implications of Wide-Area Geographic Diversity for Short-Term Variability of Solar Power. LBNL-3884E. Lawrence Berkeley National Laboratory, Environmental Energy Technologies Division, Berkeley, CA.

Paoli, C., Voyant, C., Muselli, M., Nivet, M., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol. Energy 84 (12), 2146–2160.

Pelland, S., Remund, J., Kleissl, J., Oozeki, T., Brabandere, K.D., 2013. Photovoltaic and Solar Forecasting: State of the Art. Technical report, IEA PVPS T14-01:2013.

Perez, R., Moore, K., Wilcox, S., Renné, D., Zelenka, A., 2007. Forecasting solar radiation—preliminary evaluation of an approach based upon the national forecast database. Sol. Energy 81 (6), 809–812.

Pujol, G., Iooss, B., Janon, A., Gilquin, L., Le Gratiet, L., Lemaitre, P., 2014. Sensitivity: Sensitivity Analysis. http://cran.r-project.org/web/packages/sensitivity/index.html.

Quesada-Ruiz, S., Chu, Y., Tovar-Pescador, J., Pedro, H.T.C., Coimbra, C.F.M., 2014. Cloud-tracking methodology for intra-hour DNI forecasting. Sol. Energy 102, 267–275.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

Rogers, E., DiMego, G., Black, T., Ek, M., Ferrier, B., Gayno, G., Janjic, Z., Lin, Y., Pyle, M., Wong, V., Wu, W., Carley, J., 2009. The NCEP North American mesoscale modeling system: recent changes and future plans. In: 23rd Conference on Weather Analysis and Forecasting, Boston, Massachusetts.

Saltelli, A., Bolado, R., 1998. An alternative way to compute fourier amplitude sensitivity test (FAST). Comput. Stat. Data Anal. 26 (4), 445–460.

Sfetsos, A., Coonick, A.H., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol. Energy 68 (2), 169–178.

Sobol, I.M., 1976. Uniformly distributed sequences with an additional uniform property. USSR Comput. Math. Math. Phys. 16 (5), 236–242.

Soltas Energy, 2014. http://www.soltasenergy.com/.

Stein, J.S., 2012. The photovoltaic performance modeling collaborative (PVPMC). In: 38th IEEE Photovoltaic Specialists Conference (PVSC), Austin, Texas.

U.S. Department of Energy SunShot Initiative, 2013a. In: 93rd American Meteorological Society Annual Meeting: Solar Forecasting Metrics Workshop, Austin, Texas.

U.S. Department of Energy SunShot Initiative, 2013b. In: UVIG Workshop on Variable Generation Forecasting Applications to Power System Planning and Operations: Solar Forecasting Metrics Workshop, Salt Lake City, Utah.

U.S. Department of Energy SunShot Initiative, 2014. In: UVIG Workshop on Variable Generation Forecasting Applications to Power System Planning and Operations: Solar Forecasting Metrics Workshop, Tucson, AZ.

Wang, G., Shan, S., 2007. Review of metamodeling techniques in support of engineering design optimization. J. Mech. Des. 129 (4), 370–380.

Zhang, J., Chowdhury, S., Messac, A., 2012. An adaptive hybrid surrogate model. Struct. Multidisciplinary Optimiz. 46 (2), 223–238.

Zhang, J., Chowdhury, S., Messac, A., Castillo, L., 2011. Multivariate and multimodal wind distribution model based on kernel density estimation. In: ASME 5th Int. Conf. on Energy Sustainability, Washington, DC.

Zhang, J., Chowdhury, S., Messac, A., Castillo, L., 2013a. A multivariate and multimodal wind distribution model. Renew. Energy 51, 436–447.

Zhang, J., Chowdhury, S., Messac, A., Hodge, B.-M., 2013b. Assessing long-term wind conditions by combining different measure-correlate-predict algorithms. In: ASME International Design Engineering Technical Conferences, Portland, Oregon.

Zhang, J., Chowdhury, S., Zhang, J., Messac, A., Castillo, L., 2013c. Adaptive hybrid surrogate modeling for complex systems. AIAA J. 51 (3), 643–656.

Zhang, J., Hodge, B.-M., Florita, A., 2013d. Investigating the correlation between wind and solar power forecast errors in the Western Interconnection. In: ASME 7th Int. Conf. on Energy Sustainability, Minneapolis, MN.

Zhang, J., Hodge, B.-M., Florita, A., 2013e. Joint probability distribution and correlation analysis of wind and solar power forecast errors in the Western Interconnection. J. Energy Eng.. http://dx.doi.org/10.1061/(ASCE)EY.1943-7897.0000189.