Solar Power Forecast Evaluation Metrics

This report presents a suite of metrics for evaluating deterministic and probabilistic solar power forecasts within the Solar Forecast Arbiter evaluation framework. These metrics will be used for different purposes, e.g. comparing the forecast and the measurement, comparing the performance of multiple forecasts, and evaluating an event forecast. Related topics, such as anomalous and missing data issues, the periods during which forecasts are evaluated, forecast visualization, and cost metrics, will be the focus of upcoming stakeholder engagement efforts.

In the metrics below, n is the number of samples, F is the forecasted value, and O is the observed (actual) value.

Metrics for Deterministic Forecasts

1) *Mean Absolute Error (MAE)*: The absolute error is the absolute value of the difference between the forecast value and the observed value. The MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|F_i - O_i\right| \tag{1}$$

MAE is one of the most common metrics to evaluate the performance of solar power forecasting. One shortcoming of MAE is that many relatively small errors may disguise a few large errors.

2) *Mean Absolute Percentage Error (MAPE)*: The absolute percentage error is the absolute value of the difference between the forecast value and the observed value expressed as a percentage of a normalizing factor.

$$MAPE = 100\%\frac{MAE}{Norm} = \frac{100\%}{n \times Norm}\sum_{i=1}^{n}\left|F_i - O_i\right| \tag{2}$$

where *Norm* is the normalizing factor. The normalizing factor must have the same units as the forecast and observed values. For example, for power forecasts it is common to use AC capacity as the normalizing factor. For net load forecasts, one may use the average or peak system load as the normalizing factor.

3) *Mean Bias Error (MBE)*: The bias is the difference between the forecast and the observed value. The MBE is:

$$MBE = \frac{1}{n}\sum_{i=1}^{n}\left(F_i - O_i\right) \tag{3}$$

MBE should be interpreted cautiously because positive and negative errors will cancel out.

4) *Root Mean Square Error (RMSE)*: The RMSE is the square root of the average of squared differences between forecast and observed values. The RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)^2} \tag{4}$$

RMSE is a frequently used measure for evaluating forecast accuracy. It can be interpreted as a kind of average error one can expect. Since the errors are squared before they are averaged, the RMSE gives higher weight to large errors. The RMSE is most useful when evaluating forecasts in situations where large errors are undesirable.

5) *Normalized Root Mean Squared Error (NRMSE)*: The NRMSE is the normalized form of the RMSE. The NRMSE is:

$$NRMSE = 100\% \times \frac{RMSE}{Norm} = \frac{100\%}{Norm}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)^2} \tag{5}$$

Similar to the MAPE metric, the normalizing factor of NRMSE must have the same units as the forecast and observed values.

6) *Forecast Skill Score (FS)*: The forecast skill score measures the change in a metric for a forecast relative to a reference forecast. The Solar Forecast Arbiter will calculate forecast skill based on RMSE:

$$FS = 1 - \frac{RMSE_{forecast}}{RMSE_{ref}} \tag{6}$$

where $RMSE_{forecast}$ is calculated based on the forecast of interest, and $RMSE_{rfe}$ is calculated for the reference forecast. The Solar Forecast Arbiter provides several benchmark forecasts that can serve as a reference [3].

7) *Pearson Correlation Coefficient (r)*: Correlation indicates the strength and direction of a linear relationship between two variables (for example model output and observed values). The Pearson correlation coefficient (also called the sample correlation coefficient) measures the linear dependency between the prediction (F) and the observation (O) [1] and is obtained by dividing the covariance of two variables by the product of their standard deviations:

$$r = \frac{\sum_{i=1}^{n}\left(F_i - \bar{F}\right)\left(O_i - \bar{O}\right)}{\sqrt{\sum_{i=1}^{n}\left(F_i - \bar{F}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2}} \tag{7}$$

where $\bar{F}$ and $\bar{O}$ are the averages of forecast and observation respectively. The correlation r is +1 in the case of a perfect increasing linear relationship, and -1 in case of a decreasing linear relationship, with values in between indicating the degree of linear relationship between forecast

and observations. A correlation coefficient of 0 means there is no linear relationship between the variables.

8) *Coefficient of Determination* (R^2): The coefficient of determination measures the extent that variability in the forecast errors is explained by variability in the observed values. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(O_{\mathrm{i}} - F_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(O_{\mathrm{i}} - \bar{O}\right)^2} \tag{8}$$

If a perfect forecast is made, the R^2 is 1.

9) *Centered (unbiased) Root Mean Squared Error (CRMSE)*: The CRMSE describes the variation in errors around the mean. CRMSE is given by:

$$CRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(F_i - \bar{F}) - (O_i - \bar{O})\right]^2} \tag{9}$$

The CRMSE is related to RMSE and MBE through $\mathrm{RMSE}^2 = \mathrm{CRMSE}^2 + \mathrm{MBE}^2$. In [14], it was shown that the CRMSE could also be decomposed into components related to the standard deviation and the correlation coefficient:

$$CRMSE^2 = \sigma_F^2 + \sigma_O^2 - 2\sigma_F\sigma_O r \tag{10}$$

where σ_F and σ_O are the standard deviations of forecast and observation, respectively, and r is the correlation coefficient defined in (7).

10) *Kolmogorov-Smirnov test Integral (KSI)*: The KSI quantifies the level of agreement between the CDFs of forecast and observed values. KSI is calculated as [19]:

$$KSI = \int_{p_{\min}}^{p_{\max}} D_n(p)\,dp \tag{11}$$

where p_{max} and p_{min} are the maximum and minimum values of the observations and $D_n(p)$ is the absolute difference between the two empirical cumulative distribution functions, defined as

$$\begin{aligned} &D_n(p) = \max\left|CDF_O(p) - CDF_F(p)\right| \text{ for } p \in \left[p_k, p_{k+1}\right] \\ &p_k = p_{\min} + kd,\ k = 0,1,\mathrm{K}\ ,K \text{ and } d = \frac{p_{\max} - p_{\min}}{K} \end{aligned} \tag{12}$$

In practice, $K = 100$ is typical. A KSI value of zero implies that the CDFs of forecast and observed values are equal. KSI can be normalized as

$$KSI\,(\%) = \frac{100}{a_{critical}} KSI \text{ where } a_{critical} = V_c\left(p_{\max} - p_{\min}\right) \text{ and } V_c = \frac{1.63}{\sqrt{n}} \tag{13}$$

When $n \geq 35$ the normalized KSI can be interpreted as a statistic that tests the hypothesis that the two empirical CDFs represent samples drawn from the same population [19].

11) *The OVER Metric*: In concept the OVER metric [19] modifies the KSI to quantify the difference between the two CDFs but only where the CDFs differ by more than the critical limit V_c (13). The OVER is calculated as:

$$OVER = \int_{p_{min}}^{p_{max}} D_{\mathrm{n}}^{*} dp \tag{14}$$

where

$$D_{\mathrm{n}}^{*} = \begin{cases} D_n - V & \text{if } D_n > V_c \\ 0 & \text{if } D_n \leq V_c \end{cases} \tag{15}$$

The OVER metric can be normalized using the same approach as for KSI.

12) *Combined Performance Index (CPI)*: The CPI "combines in a single statistic the descriptive power of KSI and OVER (for CDF agreement) and of RMSD (for overall dispersion)" [6]. RMSD is the root mean square difference. In this work, the observations are taken to be known exactly and thus RMSD can be replaced with RMSE:

$$CPI = \frac{1}{4}\left(KSI + OVER + 2 \times RMSE\right) \tag{16}$$

where the KSI, OVER and RMSE are defined above.

Metrics for Deterministic Event Forecasts

An event is defined by values that exceed or falling below a threshold [12]. A typical event is the ramp in power of solar generation, which is determined by:

$$\left|P(t + \Delta t) - P(t)\right| > \text{Ramp Forecasting Threshold} \tag{17}$$

where *P(t)* is the solar power output at time *t* and Δt is the duration of the ramp event.

Based on the predefined threshold, all observations or forecasts can be evaluated by placing them in either the "event occurred" or "event did not occur" categories. The 2x2 contingency table, as shown in Figure 1, can be formed to categorize individual pairs of forecasts and observations into different groups. The numbers *a, b, c, d* count the number of times the event forecast agrees (or disagrees) with events in the observed values.

		Observed: Yes	Observed: No
Forecast	Yes	a (Forecast = Yes, Observed = Yes)	b (Forecast = Yes, Observed = No)
	No	c (Forecast = No, Observed = Yes)	d (Forecast = No, Observed = No)

Figure 1. The 2×2 contingency table—relationship between counts of forecast/observation pairs

Based on the contingency table in Figure 1, the metrics to evaluate deterministic events can be classified as follows [12]:

13) *Probability of Detection (POD)*: The POD is the fraction of observed "Yes" correctly forecast as "Yes":

$$POD = \frac{a}{a+c} \tag{18}$$

14) *False Alarm Ratio (FAR)*: The FAR is the fraction of forecast "Yes" events that did not occur:

$$FAR = \frac{b}{a+b} \tag{19}$$

15) *Probability of False Detection (POFD)*: The POFD is the fraction of observed "No" that were forecast to be "Yes":

$$POFD = \frac{b}{b+d} \tag{20}$$

16) *Critical Success Index (CSI)*: The CSI evaluates how well an event forecast predicts observed events, e.g. ramps in irradiance or power. The CSI is the relative frequency of hits, i.e. how well predicted "yes" events correspond to observed "yes" event, formulated in (21):

$$CSI = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}+\mathrm{c}} \tag{21}$$

17) *Event Bias (EBIAS)*: The EBIAS is the ratio of counts of forecast and observed events:

$$EBIAS = \frac{a+b}{a+c} \tag{22}$$

18) *Event Accuracy (EA)*: The EA is the fraction of events that were forecast correctly, i.e. Forecast = "Yes" and Observed = "Yes", or Forecast = "No" and Observed = "No", formulated in (23)

$$EA = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \tag{23}$$

Metrics for Probabilistic Forecasts

Probabilistic forecasts represent uncertainty in the forecast quantity by providing a probability distribution or a prediction interval rather than a single value.

1) *Brier Score (BS):* The BS measures the accuracy of forecast probability for one or more events. The BS is:

$$BS = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2 \tag{24}$$

where n is the number of forecast events, f_i is the forecast probability of event i, and o_i is the actual outcome of event i ($o_i = 1$ if the event occurs, $o_i = 0$ otherwise). Smaller values of BS indicate better agreement between forecasts and observations. The shortcoming of the BS is that it becomes inadequate for very rare events, because it does not sufficiently discriminate between small changes in forecasts that are significant for rare events [9].

2) *Brier Skill Score (BSS)*: The BSS is based on the BS and measures the performance of a probability forecast relative to a reference forecast:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \tag{25}$$

where BS_{ref} is the Brier score achieved by the reference forecast. BSS equal to zero indicates the forecast is no better (or worse) than the reference. BSS less than zero indicates the forecast is worse than the reference.

When the probability forecast takes on a finite number of values (e.g., 0.0, 0.1, …, 0.9, 1.0), the BS can be decomposed into a sum of three metrics that give additional insight into a probability forecast [13]:

$$\begin{aligned} \mathrm{BS} &= \text{Reliability} + \text{Resolution} + \text{Uncertainty} \\ &= \mathrm{REL} + \mathrm{RES} + \mathrm{UNC} \end{aligned} \tag{26}$$

Table 1. Terms used in definition of REL, RES and UNC

$F\left(t_k\right), k = 1,\mathrm{K}\,,n$	Probability forecast for an event o at each time t_k
$f_i, i = 1,\mathrm{K}\,,I$	Discrete values that appear in the probability forecast F

$o(t_k)$	Indicator for event o: $o(t_k)=1$ if event occurs at time t_k, $o(t_k)=0$ otherwise
N_i:	The number of times each forecast value f_i appears in the forecast F. By definition, $n=\sum_{i=1}^{I} N_i$
$p(f_i)=\frac{N_i}{n}$	The relative frequency of each forecast value f_i in the forecast F.
$\bar{o}_i = p(o_1 \mid f_i) = \frac{1}{N_i}\sum_{k\in N_i} o_k$	$\bar{o}_i$ is the average of $o(t_k)$ at the N_i times t_k when $F(t_k)=f_i$
$\bar{o}=\frac{1}{n}\sum_{k=1}^{n} o(t_k)=\frac{1}{n}\sum_{i=1}^{I} N_i\bar{o}_i$	$\bar{o}$ is average of $o(t_k)$ for all times t_k.

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i\left(f_i-\bar{o}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i\left(\bar{o}_i-\bar{o}\right)^2 + \bar{o}\left(1-\bar{o}\right) = REL + RES + UNC \tag{27}$$

3) *Reliability (REL)*: The REL is given by

$$REL = \frac{1}{n}\sum_{i=1}^{I} N_i\left(f_i-\bar{o}_i\right)^2 \tag{28}$$

Reliability is the weighted average of the squared differences between the forecast probabilities f_i and the relative frequencies of the observed event in the forecast subsample of times where $F(t_k)=f_i$. A forecast is perfectly reliable if $REL=0$. This occurs when the relative event frequency in each subsample is equal to the forecast probability for the subsample.

4) *Resolution (RES)*: The RES is given by:

$$RES = \frac{1}{n}\sum_{i=1}^{I} N_i\left(\bar{o}_i-\bar{o}\right)^2 \tag{29}$$

Resolution is the weighted average of the squared differences between the relative event frequency for each forecast subsample, and the overall event frequency. Resolution measures the forecast's ability to produce subsample forecast periods where the event frequency is different. Higher values of RES are desirable.

5) *Uncertainty (UNC)*: The UNC is:

$$UNC = \bar{o}\left(1-\bar{o}\right) \tag{30}$$

Uncertainty is the variance of the event indicator $o(t)$. Low values of UNC indicate that the event being forecasted occurs only rarely.

6) *Sharpness (SH)*: The SH [7] represents the degree of "concentration" of a forecast comprising a prediction interval of the form $[f_l, f_u]$ within which the forecast quantity is expected to fall with probability $1-\beta$. A good forecast should have a low sharpness value. The prediction interval endpoints are associated with quantiles α_l and α_u where $\alpha_u - \alpha_l = 1-\beta$. For a single prediction interval, the SH is:

$$SH = f_u - f_l \tag{31}$$

and for a timeseries of prediction intervals (arising from, e.g., a forecast for a sequence of times, or from a series of forecasts) SH is given by the average:

$$SH = \frac{1}{n}\sum_{i=1}^{n} f_{u,i} - f_{l,i} \tag{32}$$

7) *Continuous Ranked Probability Score (CRPS)*: The CRPS [13] is a score that is designed to measure both reliability and sharpness of a probabilistic forecast. For a timeseries of forecasts comprising a CDF at each time point, the *CRPS* is:

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int \left|F_i(x) - O_i(x)\right| dx \tag{33}$$

where $F_i(x)$ is the CDF of the forecast quantity x at time point i, and $O_i(x)$ is the CDF associated with the observed value x_i

$$O_i(x) = \begin{cases} 0 & x < x_i \\ 1 & x \geq x_i \end{cases} \tag{34}$$

The *CRPS* reduces to the mean absolute error (MAE) if the forecast is deterministic.

Reference:

[1] Carlos Ruberto Fragoso Júnior, Universidade Federal de Alagoas (UFAL) . http://www.ctec.ufal.br/professor/crfj/Graduacao/MSH/Model%20evaluation%20methods.doc

[2] J. Zhang, et al., A suite of metrics for assessing the performance of solar power forecasting, Solar Energy 111 (2015) 157–175.

[3] https://solarforecastarbiter.org/benchmarks/

[4] W. Lieberman-Cribbin, C. Draxl and A. Clifton, A Guide to Using the WIND Toolkit Validation Code, NREL/TP-5000-62595, December 2014.

[5] D.W. van der Meer and J. Widén, J. Munkhammar, Review on probabilistic forecasting of photovoltaic power production and electricity consumption, Renewable and Sustainable Energy Reviews 81 (2018) 1484–1512.

[6] Christian A. Gueymard, Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models, Solar Energy 86 (2012) 2145–2169.

[7] Pinson P, Nielsen HA, Mller JK, H. Madsen H, Kariniotakis GN. Nonparametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 2007;10(6):497–516.

[8] Alexis Bocquet, et al., Assessment of probabilistic PV production forecasts performances in an operational context, 6th Solar Integration Workshop - International Workshop on Integration of Solar Power into Power Systems, Nov 2016, Vienna, Austria.

[9] Riccardo Benedetti, "Scoring Rules for Forecast Verification". Monthly Weather Review. 138 (1): 203–211, 2010.

[10] Quan H, Srinivasan D, Khosravi A. Short-term load and wind power forecasting using neural network-based prediction intervals. IEEE Trans on Neural Network and Learning System 2014;25(2):303–15.

[11] Yinghao Chu, Carlos F.M. Coimbra, Short-term probabilistic forecasts for Direct Normal Irradiance, Renewable Energy 101 (2017) 526-536.

[12] T. Jensen, et al., Metrics for evaluation of solar energy forecast. NCAR Technical notes. NCAR/TN-527+STR, June 2016.

[13] Daniel Wilks, “Statistical Methods in the Atmospheric Sciences, 3rd Edition”, Academic Press, 2011.

[14] Karl E. Taylor, Summarizing multiple aspects of model performance in a single diagram, JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 106, NO. D7, PAGES 7183-7192, APRIL 16, 2001.

[15] https://solarforecastarbiter.org/usecases/#definitions

[16] A. H. Murphy and R. L. Winkler, A general framework for forecast verification. Monthly Weather Review, vol. 115, pp. 1330-1338. 1987.

[17] E. Gilleland, Forecast verification for solar power forecasts. NCAR. July 19, 2018. http://opensky.ucar.edu/islandora/object/conference%3A3331/datastream/PDF/download/citation.pdf

[18] Coimbra CF, Kleissl J, Marquez R, Chapter 8 - Overview of Solar-Forecasting Methods and a Metric for Accuracy Evaluation. In: Kleissl J, editor. Solar energy forecasting and resource assessment. Boston: Academic Press; 2013. p. 171–194.

[19] B. Espinar, Let al., Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations, Solar Energy, vol. 83, pp. 118-125, 2009. https://doi.org/10.1016/j.solener.2008.07.009.