The Solar Forecast Arbiter: An Open Source Evaluation Framework for Solar Forecasting

Clifford W. Hansen[1], William F. Holmgren[2], Aidan Tuohy[3], Justin Sharp[4], Antonio T. Lorenzo[2], Leland J. Boeman[2], Anastasios Golnas[5]

[1]Sandia National Laboratories, Albuquerque, NM, 87185, USA
[2]University of Arizona, Tucson, AZ, 85721, USA
[3]Electric Power Research Institute, Chicago, IL, 60616, USA
[4]Sharply Focused, Portland, OR, 97213, USA
[5]U. S. Department of Energy, Washington, D.C., 20585, USA

***Abstract* — We describe an open source evaluation framework for solar forecasting to support the DOE Solar Forecasting 2 program and the broader solar forecast community. The framework enables evaluations of solar irradiance, solar power, and net-load forecasts that are impartial, repeatable and auditable. First, we define the use cases of the framework. The use cases, developed from the project's initial stakeholder engagement sessions, include comparisons to reference data sets, private forecast trials, evaluation of probabilistic forecast skill, and examinations of forecast errors during critical periods. We discuss the framework's data validation toolkit, reference data sources, and data privacy protocols. We describe the framework's benchmark forecast capabilities for intra-hour and day ahead forecast horizons. Finally, we summarize the reports and metrics that communicate the relative merits of the test and benchmark forecasts. The reports are created from standardized templates and include graphics for quantitatively evaluating deterministic and probabilistic forecasts and standard metrics for quantitatively evaluating forecasts.**

Index Terms* — *Forecasting, Solar energy, Performance evaluation

I. Introduction

The Department of Energy Solar Energy Technologies Office Solar Forecasting 2 funding opportunity supports eight teams working to improve solar power forecasts and their application to grid management [1]. Our team is creating a framework to fairly and transparently evaluate solar power forecasts. The framework will support the seven other DOE-funded teams and the broader solar forecast community. The framework includes the Solar Forecast Arbiter, an open source tool to support impartial, repeatable, and auditable evaluations of solar forecast performance [2]. Here we introduce the major components of the framework: stakeholder engagement, forecast definitions, use cases, reference data, benchmark forecasts, data quality assurance, metrics and reporting. We anticipate that the framework will become operational by the end of the summer of 2019.

II. Stakeholder Engagement

The framework development model emphasizes frequent stakeholder engagement to ensure that the framework meets the needs of the DOE SETO Solar Forecasting 2 program and the broader forecast community. The team held a two-hour workshop in June 2018 to collect user stories (i.e. statements of desired capability), discuss data protection concerns and forecast metrics, and to gather views on benchmark forecast capabilities. Following the workshop, we formed a *stakeholder committee* of researchers, forecast providers, forecast users and others with an interest in solar forecast evaluation – persons may join the stakeholder committee on the project website [2].

During the first year of the project, information gathered at the stakeholder meeting was used to formulate proposals for use cases, data sharing, metrics and benchmark forecasts. Proposals were circulated with the stakeholder committee for review and comment. For example, the user stories from the workshop were synthesized into proposed use cases which were iteratively refined by stakeholder review resulting in the procedural descriptions of the framework functionality (see Sect. IV); similar processes ensured that the structure and content of a data model for the framework, data sharing and protection policies, and the selection of benchmark forecasts and metrics all received stakeholder input and consensus. In addition, key stakeholders were engaged individually on the thorny issues of a common non-disclosure agreement and data sharing policies to ensure that proposals were likely to gain broad acceptance. The stakeholder committee also guided a glossary defining forecast terms.

A second Stakeholder Workshop which was open to committee members and other interested parties was held in June 2019. This workshop was used to confirm the decisions made to date, demonstrate the functionality implemented so far, and gather feedback on essential course corrections and work priorities.

978-1-7281-0494-2/19/$31.00 ©2019 IEEE

III. Forecast Definitions

The precise definition of a *forecast* is often a source of confusion in the forecasting industry. Consider these statements that use the term *forecast*:

A. "The forecast for the next 48 hours is 5 MW, 10 MW, 7 MW..."
B. "The hour ahead forecast this morning was 5 MW, 10 MW, 7 MW..."

In Statement A, the user refers to a series of expected values issued at a **single** point in time. In Statement B, the user refers to a series of expected values with the **same** lead time that are issued at **different** points in time.

An evaluation challenge arises if Statement A is extended to include a new 48-hour duration forecast every hour. In this scenario, many forecasts exist at each evaluation time. More information is needed to focus the analysis to a particular application.

Statement B suggests an alternative strategy: the forecast user or the forecast provider parses forecast data into a timeseries for evaluation according to their **application**. For example, from many 48-hour length forecasts, a timeseries of forecast values with lead times of 1 hour ahead could be compiled to evaluate forecasts that support participation in a particular market. Or, a timeseries of forecast values for each hour of the day ahead could be compiled from forecasts issued at the same time each day to evaluate forecasts input to a production cost model. In each case, the forecast under evaluation is a continuous, non-overlapping timeseries that can be compared to observations.

Standard and precise definitions of forecast-related terms are essential to conducting a forecast evaluation and communicating the results. Our framework proposes a series of definitions to support evaluations of solar forecasts [2]. Among others, the definitions include:

- *forecast data point* – a single (*Time*, *Value*) pair, where *Time* labels a moment in time or an interval of time. Metadata associated with the forecast data point describes the interval labelling convention (e.g., period-beginning) and unit of *Value*.
- *forecast run* – a sequence of one or more *forecast data points* issued at the same time.
- *forecast evaluation time series* – a complete series of *forecast data points* spanning the time interval for evaluation. A *forecast evaluation time series* can result by concatenating a series of forecast runs with sequential issue times, as shown in Fig. 1.

A series of attributes (e.g. *issue time, interval label, forecast length*) fully describes a forecast. The same attributes may be used to describe intra-hour and day-ahead forecasts.

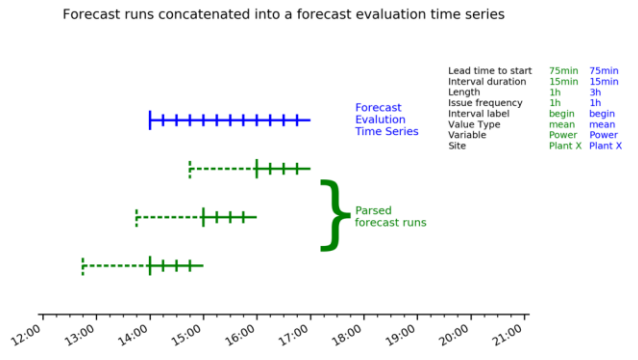


Fig. 1. Sample figure from definitions sections of Use Cases. The figure illustrates how three 75-minute lead time, hour length, 15-minute interval *forecast runs* (green) may be parsed to create a *forecast evaluation time series* (blue). The table summarizes the attributes of the forecasts.

IV. Use Cases

Our team surveyed stakeholders to determine the use cases for the framework. We determined the primary use cases for evaluation of forecasts to be:

1. **Compare a forecast to measurements**
2. Compare a probabilistic forecast to measurements
3. **Compare multiple forecasts to a common set of measurements**
4. Compare forecasts to measurements for sites and aggregates
5. Evaluate an event forecast
6. **Conduct a forecast trial**

Each use case is fully described on the project website [2]. Use cases in **bold type** are prioritized for initial capability of the Solar Forecast Arbiter.

Two additional use cases were identified as stretch goals for development. These use cases leverage the functionality of the primary use cases but are more involved:

7. Compare multiple forecast runs to measurements (stretch goal)
8. Establish long-term performance baseline of state-of-the-art operational forecasts (stretch goal)

Three use cases support the goal of analyzing forecast performance:

9. **Select subsets of forecasts and data**
10. Identify events
11. Find forecast errors with large impacts (stretch goal).

From the use cases, we derived a list of functional capabilities for the framework, including for example: calculate error metrics, communicate probabilistic forecasts, manage and protect forecasts and data, and provide reference forecasts.

V. Reference Data

The framework includes reference data to facilitate fair and consistent comparisons of forecast performance and to establish performance baselines in multiple climatic regions. Fig. 2 illustrates a selection of the reference data set. As of June 2019, the reference data currently comprises weather data from publicly available sources: the NOAA SURFRAD and SOLRAD networks [3, 4, 5]; the U.S. Climate Reference Network [6, 7], the University of Oregon's Solar Resource Measurement Laboratory (SRML) network [8]; NREL's Measurement Instrumentation Data Center (MIDC) [9]; the U.S. Department of Energy Atmospheric Radiation Measurement (ARM) network [10], Sandia National Laboratories (SNL); and the DOE Regional Test Centers (RTC) [11]. Power data for several small PV systems is available from the RTC. Parsers for the public data sets were added to the pvlib python library [12] to benefit the broader solar community.

Private parties can contribute weather and PV power data to the Solar Forecast Arbiter and can specify data protection and sharing permissions. The framework provides a data management system to maintain data security and data owners retain full control of contributed data with the ability to restrict access or remove data.

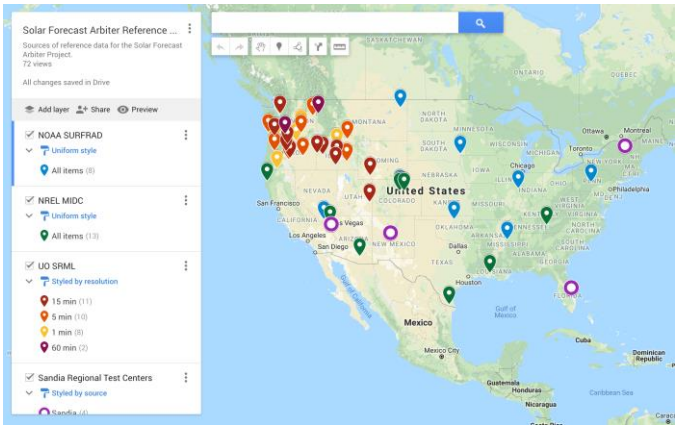


Fig. 2. Illustration of reference data from SURFRAD (blue), NREL MIDC (green), UO SRML (yellow, orange, brown), and DOE Regional Test Centers (purple circles). The RTC data includes PV power. CRN and SOLRAD data are available (not shown).

VI. Benchmark Forecasts

Benchmark irradiance and solar power forecasts provide a common reference for measuring the relative accuracy and value of other forecasts, and a consistent reference for quantifying forecast improvements. The Solar Forecast Arbiter includes several benchmark irradiance forecasts, and a common technique to translate from irradiance to power for PV systems. Benchmark forecasts were selected with stakeholder committee concurrence that meet three criteria: available throughout the United States; freely accessible or easily implemented; and provide forecast quantities of broad interest, e.g., global horizontal irradiance (GHI).

Benchmark irradiance forecasts for one hour to days-ahead horizons are selected from NOAA operational models and include:

- HRRR irradiance forecast (3 km grid, 18 to 36 hours ahead forecast length).
- RAP irradiance forecast (13 km, 21 to 39 hours)
- NAM cloud cover forecast (12 km, 72 hours)
- GFS cloud cover forecast (0.25 deg, 120 hours)

Cloud cover forecasts are translated to irradiance forecasts using a linear relationship [13]. To obtain benchmark forecasts derived from NOAA model data, a framework user supplies a location, model selection, and forecast time period. The Solar Forecast Arbiter selects the most appropriate NOAA model initialization time, accounting for typical data latency. To support accurate benchmark of hourly average quantities, the Solar Forecast Arbiter interpolates cloud cover and weather forecast data to 5 minutes, converts cloud cover to irradiance (and power if appropriate metadata is supplied), and then resamples the output to 60 minutes.

For intra-hour horizons the benchmark forecasts use persistence and "smart" persistence, i.e., persistence of the clear sky index. To obtain a persistence forecast, a framework user specifies a location and uploads recent data, either irradiance or power data with metadata about the PV system.

Translation of weather (irradiance, air temperature, wind speed) to power is performed in four steps:

1. Direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) are estimated from GHI using the Erbs model [14].
2. Plane of array irradiance is estimated from GHI, DNI and DHI using the Hay and Davies model [15] to transpose irradiance to the specified plane, and accounting for reflection of direct irradiance at the module surface as in [16].
3. Cell temperature is estimated using the PVsyst model [17]
4. DC and AC power are calculated using the PVWatts model [18]

All functions are implemented in pvlib python [12]. Future development of the Solar Forecast Arbiter may allow framework users to specify alternative models at each step.

VII. Data Quality Assurance

User-uploaded data is processed by a quality assurance module. The quality assurance component first checks for consistency of data time steps. The following quality tests are applied to weather data to flag each time point as meeting or not meeting the criteria for each test:

- Physical limits and consistency among GHI, DNI and DHI are checked using QCRad method [19].
- Air temperature is flagged if outside set limits.

- Wind speed is flagged if negative or greater than a set limit.
- GHI values are flagged where greater than 110% of the clear-sky value calculated by the Ineichen model using the SoDa climatological turbidity [20].
- Plane-of-array irradiance is flagged if greater than 110% of the transposed clear-sky model value.

The quality assurance component also checks daily time series of GHI, DC and AC power for stale or interpolated data, which can result when data communications are interrupted and the data acquisition system fills in missing values by repeating the last value or by interpolating between values. Stale or interpolated data are identified by detecting periods with nearly linear change in the data; linear change with zero slope is labeled as stale, and with non-zero slope the data are labeled as interpolated.

For a use case that involves assessing multiple forecasts against a common set of observations, a missing forecast is replaced with the last valid forecast from the same forecast provider. For assessment of a single forecast, periods with anomalous observation data are excluded. We envision that future development of the framework may provide the user with options to specify treatment of periods with missing forecasts or data.

VIII. Metrics

The framework offers a selection of metrics for measuring forecast performance for both deterministic and probabilistic forecasts. 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. The list of metrics leverages the findings [21, 22] of the first DOE Solar Forecasting research project. By default, the framework provides summary statistics and performance metrics with a number of more specialized metrics available as options. Metrics are defined at project website [2]. Implementation in open source code provides transparency in metric calculations. Specific periods can be selected (e.g. time of day or month of year) or excluded (e.g. nighttime values) from the calculation.

For deterministic forecasts, default metrics include mean absolute error, mean absolute percentage error, mean bias error, root mean square error, normalized root mean square error and forecast skill score. These are calculated for all analyses and normalized by ac rating for power forecasts. Other deterministic metrics will also be available, including the Pearson correlation coefficient (strength and direction of linear relationship between forecast and actual), coefficient of determination (extent to which variability in errors is explained by variation in observed values), centered root mean square error (variation in errors around the mean), and other related metrics. Many of these additional metrics do not directly measure how good a forecast is but allow for users to gain further insight that helps understand forecast performance.

Deterministic event forecasts are also assessed. An event is defined by values that fall below or exceed a certain threshold. A ramp (e.g. change in power over time) is a good example of an event, where the user specifies criteria that define ramps and the ability of forecasts to predict ramps is assessed. A contingency table quantifies event forecast accuracy (i.e. a 2x2 matrix counting true positive forecasts, false positive forecasts, etc.) From the contingency table event metrics are calculated such as probability of detection, false alarm ratio, probability of false detection, critical success index (how well a forecast predicts events) and event accuracy (fraction of events forecasted correctly).

For probabilistic forecasts the default metrics include the Brier score, Brier Skill score and continuous ranked probability score. The Brier Score is decomposed into Reliability, Resolution and Uncertainty, all three of which are important factors in understanding probabilistic forecast performance.

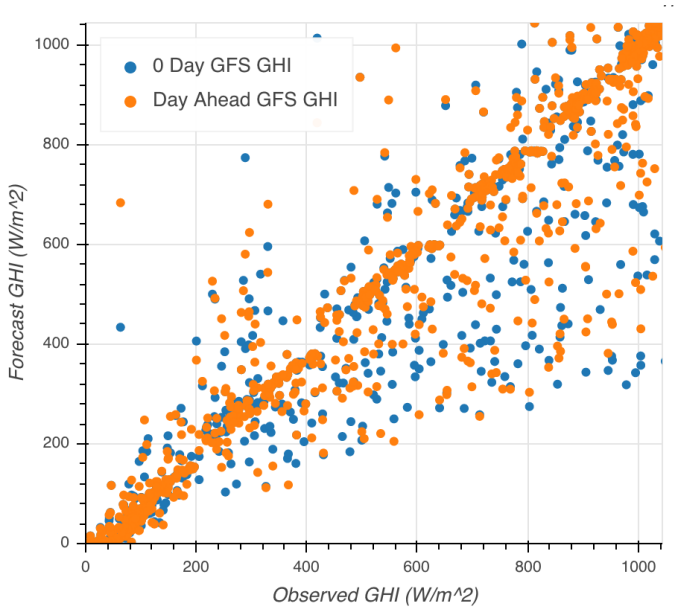
We plan to provide capability to estimate the cost of errors in a power forecast in a simplified manner. The user inputs the cost of power forecast error in $/MW as a constant, a time series or a value that depends on error magnitude (e.g. a lower cost at low MW error but higher cost for a greater MW error, to represent how the errors affect operations).

IX. Reporting

The Solar Forecast Arbiter produce reports including graphics and tables that assist users in determining the relative merits of forecasts and that aid in understanding forecast performance. Two broad categories of outputs comprise the reports. First, the forecasts themselves are displayed using scatter plots of forecast versus actuals, time series of forecasts, density plots and marginal distributions, allowing users to see how the forecasts compare to observations. Second, the metrics described in Sect. VIII are presented in various tables and plots, including bar charts, scatter plots of forecast versus error, and box and whisker plots. Values used in metric calculations can be filtered by time of day, month of year and for particular weather conditions. Reports include a summary of anomalous data, missing data and missing forecasts. Fig. 3 illustrates figures from a sample report.

X. Conclusions

We summarize the components of an open source forecast evaluation framework that supports the DOE SETO Solar Forecasting 2 program and the broader solar forecast community. Interested parties are encouraged to join the stakeholder committee to guide the project, participate in development and to contribute data to the reference data set, at https://solarforecastarbiter.org/stakeholdercommittee



Forecast	MAE	MBE	RMSE
0 Day GFS GHI	77.351	-50.019	105.213
Day Ahead GFS GHI	74.334	-44.402	97.759

Fig. 3. Sample figures from report of hourly average GHI forecast performance. Forecasts were evaluated against GHI measurements from the NREL MIDC OASIS station in Tucson, AZ [23, 24]. The evaluation period was April 1 through May 31, 2019. Two forecasts were created from daily 0Z GFS model data. The "0 day" (blue) forecast was issued at midnight local time for the following 24 hours. The "Day ahead" (orange) forecast was issued at midnight local time for the period 24-48 hours ahead. Top: scatter plot of forecast vs. observed values. Middle: average errors for each hour of the day for one forecast. Bottom: table of metrics for the total analysis period. Not shown: graphics of errors for each day and each month of the analysis period.

ACKNOWLEDGEMENTS

This work is funded in part or whole by the U.S. Department of Energy Solar Energy Technologies Office, under Award Number DE-EE0008214. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

REFERENCES

[1] U.S. Department of Energy Solar Energy Technologies Office, "Solar Forecasting 2," https://www.energy.gov/eere/solar/solar-forecasting-2 Accessed January 20, 2019.

[2] Solar Forecast Arbiter Contributors, "Solar Forecast Arbiter," https://solarforecastarbiter.org/ Accessed June 13, 2019.

[3] National Oceanic and Atmospheric Administration (NOAA). https://www.esrl.noaa.gov/gmd/grad/surfrad/ Accessed June 14, 2019.

[4] National Oceanic and Atmospheric Administration (NOAA). https://www.esrl.noaa.gov/gmd/grad/solrad/index.html Accessed June 14, 2019.

[5] B. Hicks et. al. (1996). "The NOAA Integrated Surface Irradiance Study (ISIS). A New Surface Radiation Monitoring Program." *Bull. Amer. Meteor. Soc.* 77, 2857-2864.

[6] National Oceanic and Atmospheric Administration (NOAA). https://www.ncdc.noaa.gov/crn/qcdatasets.html Accessed June 14, 2019.

[7] Diamond, H. J. et. al. (2013). "U.S. Climate Reference Network after one decade of operations: status and assessment". *Bull. Amer. Meteor. Soc.*, 94, 489-498.

[8] The University of Oregon Solar Radiation Monitoring Laboratory. http://solardat.uoregon.edu/ Accessed June 14, 2019.

[9] National Renewable Energy Laboratory Measurement and Instrumentation Data Center. https://midcdmz.nrel.gov/ Accessed June 14, 2019.

[10] U.S. Department of Energy Atmospheric Radiation Measurement Southern Great Plains Observatory. https://www.arm.gov/capabilities/observatories/sgp. Accessed June 14, 2019.

[11] U.S. Department of Energy Regional Test Centers. https://rtc.sandia.gov/. Accessed June 14, 2019.

[12] W. F. Holmgren, C. W. Hansen, and M. A. Mikofski (2018). "pvlib python: a python package for modeling solar energy systems," *The Journal of Open Source Software*. vol. 3, pp. 884.

[13] Larson et. al. (2016). "Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest" *Renewable Energy* 91, pp. 11-20.

[14] D. G. Erbs, S. A. Klein and J. A. Duffie, (1982). "Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation", *Solar Energy* 28(4), pp. 293-302.

[15] Hay, J.E., Davies, J.A. (1980). "Calculations of the solar radiation incident on an inclined surface". In: Hay, J.E., Won, T.K. (Eds.), *Proc. of First Canadian Solar Radiation Data Workshop*, 59. Ministry of Supply and Services, Canada.

[16] W. De Soto et al. (2006). "Improvement and validation of a model for photovoltaic array performance", *Solar Energy* 80, pp. 78-88.

[17] PVsyst 6 Help. https://www.pvsyst.com/help/. Accessed June 14, 2019.

[18] A. P. Dobos (2014). "PVWatts Version 5 Manual", *NREL Technical Report*, NREL/TP-6A20-62641, September 2014.

[19] C. N. Long and Y. Shi (2008). "An Automated Quality Assessment and Control Algorithm for Surface Radiation Measurements", *The Open Atmos. Science Journal* 2: 23-37, 2008.

[20] P. Ineichen and R. Perez (2002). "A New airmass independent formulation for the Linke turbidity coefficient", *Solar Energy* 73, pp. 151-157.

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

[22] T. L. Jensen, T. L. Fowler, B. G. Brown, J. K. Lazo and S. E. Haupt (2016). "Metrics for evaluation of solar energy forecasts," *NCAR Technical Notes*, NCAR/TN-527+STR.

[23] National Renewable Energy Laboratory Measurement and Instrumentation Data Center, data for Tucson, AZ. https://midcdmz.nrel.gov/apps/sitehome.pl?site=UAT Accessed June 14, 2019.

[24] Andreas, A.; Wilcox, S. (2010). Observed Atmospheric and Solar Information System (OASIS); Tucson, Arizona (Data); *NREL Technical Report* NREL Report No. DA-5500-56494.